Analysis and Visualization of Large-Scale Gene Expression Microarray Compendia

Matthew Aaron Hibbs

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

Over the past decade, gene expression microarray data has become one of the most important tools available for biologists to understand molecular processes and mechanisms on the whole-genome scale. Microarray data provides a window into the inner workings of the transcriptional process that is vital for cellular maintenance, development, biological regulation, and disease progression. While an exponentially increasing amount of microarray data is being generated for a wide variety of organisms, there is a severe lack of methods designed to utilize the vast amount of data currently available. In my work, I explore several techniques to meaningfully harness large-scale collections of microarray data both to provide biologists with a greater ability to explore data repositories, and to computationally utilize these repositories to discover novel biology.

First, effective search and analysis techniques are required to guide researchers and enable their effective use of large-scale compendia. I will present a user-driven similarity search algorithm designed to both quickly locate relevant datasets in a collection and to then identify novel players related to the user's query. Second, I will discuss techniques for visualization-based analysis of microarray data that incorporate statistical measures into visualization schemes and utilize alternative views of data to reveal previously obscure patterns. Third, I will focus on novel methods that allow users to simultaneously view multiple datasets with the goal of providing a larger biological context within which to understand these data. Finally, I will discuss how we have successfully

used these approaches to discover novel biology, including successfully directing a large-scale experimental investigation of *S. cerevisiae* mitochondrial organization and biogenesis.

The combination of visualization-based analysis methods and exploratory algorithms such as those presented here are vital to future systems biology research. As data collections continue to grow and as new forms of data are generated, it will become increasingly important to develop methods and techniques that will allow experts to intelligently sift through the available information to make new discoveries.

Acknowledgements

First, I would like to thank my friends and family for their love and support throughout graduate school and the process of creating this dissertation. I am especially grateful to my parents, Jerry and Cathy; my sister Debbie; my brother Steve; and the rest of my family. I also thank my friends and collaborators who have made this work possible: Chad Myers, David Hess, Nathaniel Dirksen, Paul Chang, Tom Briggs, Sasha Myers, Shirley Gaw, Grant Wallace, Curtis Huttenhower, Rachel Sealfon, Erin Smith, Kara Dolinski, Maitreya Dunham, and Amy Caudy.

I must especially thank Olga Troyanskaya and Kai Li, who have been unparalleled mentors throughout my graduate career. Importantly, I thank Olga for sparking my interest in the field of computational biology and for fostering a collaborative laboratory environment. The members of the Troyanskaya lab have created an exciting and fun environment to conduct research, and I thank them all for their advice, input, and friendship: Chad Myers, Curtis Huttenhower, Daniel Barrett, David Hess, Edo Airoldi, Florian Markowetz, Maria Chikina, Rachel Sealfon, Patrick Bradley, Yuanfang Guan, Zafer Barutcuoglu, and Camelia Chiriac.

I would also like to thank the members of my dissertation committee, who have provided invaluable feedback and advice: David Botstein, Tom Funkhouser, and Leonid Kruglyak. Additionally, the members of the Botstein and Kruglyak laboratories have given me excellent advice and taught me a great deal

about the study of biology. I also thank our excellent technical support staff, especially John Wiggins.

Chapter 2 is joint work with David Hess, Chad Myers, Curtis Huttenhower, Kai Li, and Olga Troyanskaya. It has been published in Bioinformatics [32]. In particular, I must thank David Hess for his laboratory efforts and literature searches to verify the biological conclusions presented in this chapter. Chapter 3 is joint work with Nathaniel Dirksen, Kai Li, and Olga Troyanskaya; it has been published in BMC Bioinformatics [31]. Chapter 4 is joint work with Grant Wallace, Maitreya Dunham, Kai Li, and Olga Troyanskaya; it was presented at the 11th International Conference on Information Visualization (IV'07) [33]. Chapter 5 is joint work with Grant Wallace, Maitreya Dunham, Kai Li, and Olga Troyanskaya; a version of this work will soon be submitted for publication. Chapter 6 is joint work with Chad Myers, Curtis Huttenhower, David Hess, Amy Caudy, Kai Li, and Olga Troyanskaya; a version of this work will soon be submitted for publication. I must pay special thanks to David Hess and Amy Caudy, who conducted the experimental efforts described in this chapter.

This work was financially supported by many sources, including by NSF grants CNS-0406415, DBI-0546275, IIS-0513552, and EIA-0101247; by NIH grants R01 GM071966 and T32 HG003284; by the Program in Integrative Information, Computer, and Application Sciences (PICASso) program which is funded by NSF grant DGE-9972930; by a Google Research Award; and by the NIGMS Center of Excellence grant P50 GM071508.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biology is currently experiencing a period of transition from a study of small-scale, specific phenomenon to the understanding of entire systems and genomes. This transition is enabled by the development of many new technologies, including whole genome sequencing, gene expression microarrays, physical protein interaction assays, genetic interaction assays, and tandem mass spectrometry. These experimental techniques can be utilized to generate data on an immense scale, which holds the promise of elucidating the functions of genes and proteins, their regulatory mechanisms, and their modes of interaction [36, 47].

However, this wealth of data remains largely underutilized. Over the past 10 years we have experienced an exponential increase in the amount of functional genomics data generated, but the rate at which novel gene functions are discovered has remained fairly constant [67] (Figure 1.1). There are several difficult challenges for bridging this gap between data and knowledge. In particular, the sheer scale of newly generated datasets prevents traditional biological analysis from performing comprehensive evaluations of individual studies. Moreover, it is increasingly clear that these data are best understood within the greater context of other available data. While a single study may shed light on a particular specific question, the conjunction of many studies can be much more powerful with the ability to address more general biological concerns.

Computational methods have the ability to address several of these challenges to data analysis and to provide insight into many biological questions.

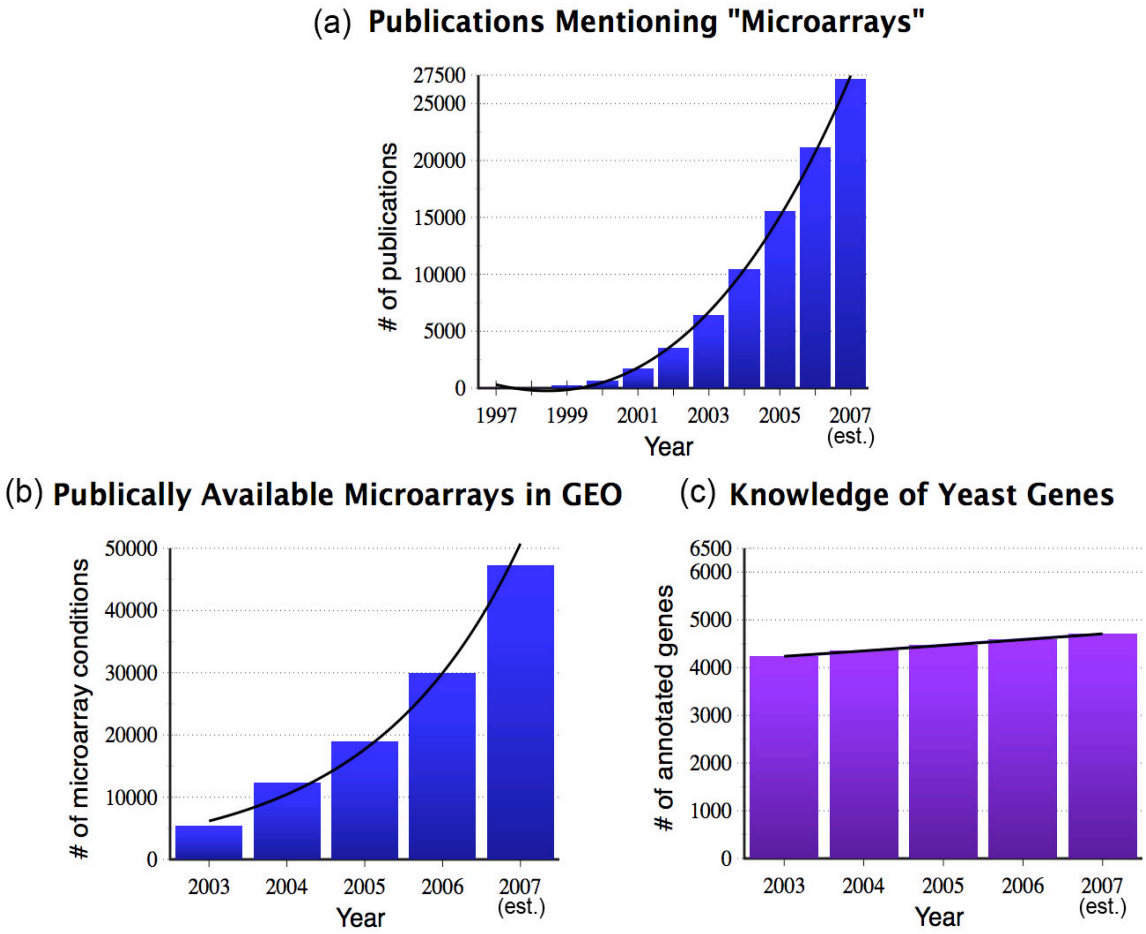


Figure 1.1: The growing data-knowledge gap. While the amount of data generated to investigate gene and protein function grows at increasing rates each year, the rate at which we gain knowledge of specific functions has remained constant. Here (a) shows the number of publications each year that mention gene expression microarrays and (b) shows the number of publicly available microarray conditions added each year to the National Center for Biotechnology Information (NCBI) microarray repository, the Gene Expression Omnibus (GEO) [22]. While both of these measures of data generation exhibit increasing rates of growth from year to year, (c) shows the number of genes in *S. cerevisiae* with a known function as determined by the *Saccharomyces* Genome Database (SGD) [18], which demonstrates a constant growth rate of biological knowledge.

1.1 The promise of computational biology

For several years, bioinformatics and computational biology have broadly promised to increase our biological understanding of gene function through the application of machine learning, data mining, visualization, and statistical analysis methods. Despite this effort, the majority of new discoveries of gene and protein function continue to be generated by laboratories without the aid of computational prediction methods. This failure of computational biology is the result of several factors, the largest of which is a disconnection between computational predictions of gene function and large-scale laboratory studies of the roles of genes in biological processes.

In general, computational function prediction methods are based on the premise of "guilt by association," meaning that given partial knowledge of gene functions, they infer the function of other genes. Typically these methods utilize collections of biological data to identify characteristic patterns associated with known functions, to discriminate classes of gene function, or to search for similar profiles of data. Thus, given collections of data and a "gold standard" of prior knowledge of gene function, these methods computationally predict the novel involvement of genes in biological processes. Several machine learning and data mining approaches have been utilized for this task in computational biology, including Bayesian networks [38, 42, 44, 52, 61, 92], support vector machines [9, 50, 66], feature selection methods [65], and others [59, 63].

While these methods often produce biologically interesting and sensible results, the predictions occur with varying rates of accuracy, false positives (FPs,

i.e. genes incorrectly predicted to a function), and false negatives (FNs, i.e. genes incorrectly omitted from a function). As such, follow-up laboratory work is required to confirm or deny these predictions. However, while individual predictions of these methods have been verified through further experiments, the vast majority of these predictions remain unconfirmed. This lack of follow-up is problematic both for computationalists, whose work remains unverified, and for traditional biologists, whose work could be guided to promising experimental targets to greatly accelerate discoveries.

Ideally, computational prediction methods integrated with laboratory investigations would complete a cycle of prediction, experimentation, and verification, where newly generated data and confirmed gene functions would become additional inputs to further iterations of the cycle (Figure 1.2). However, as few large-scale validation studies are performed, this cycle is often broken shortly after the generation of new predictions.

A further problem encountered by several computational techniques is the reliance on a “gold standard” or a set of curated, known assignments of genes to specific biological functions or pathways, such as those provided by the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and the Munich Information Center for Protein Sequences (MIPS). While these repositories contain a great deal of our collective knowledge of biology, individual experts still retain superior knowledge, particularly for their domain of study. However, the majority of computational function prediction methods do not leverage this expert knowledge.

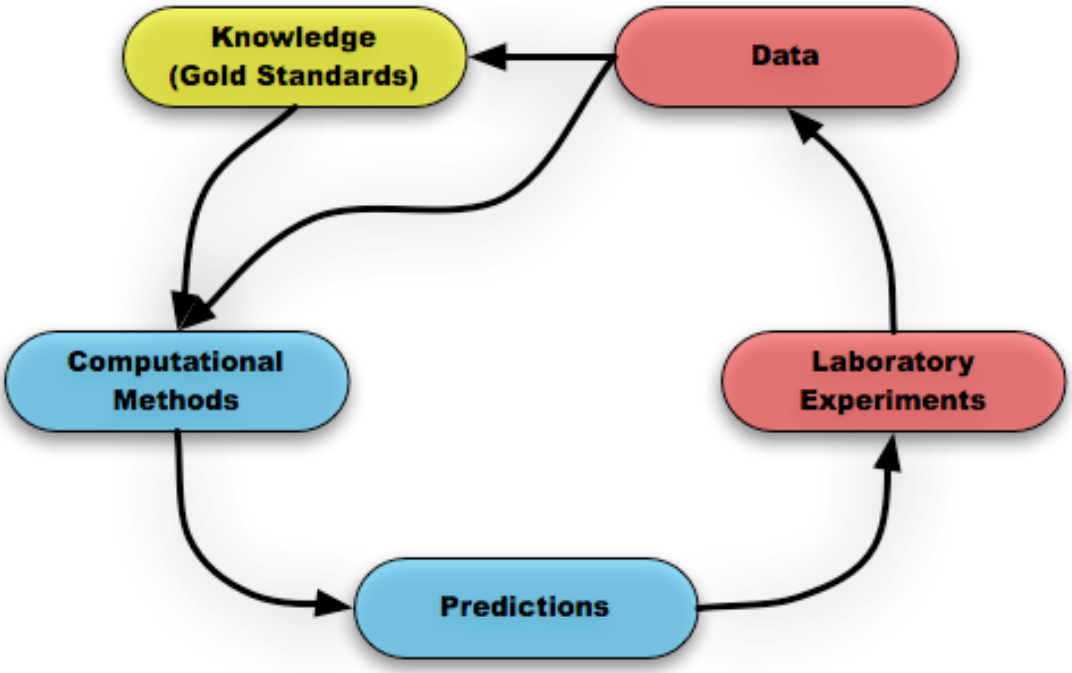


Figure 1.2: Iterative cycle of computational biology. Computational techniques can be used to augment traditional biology and speed the rate of novel discoveries. Given existing data and knowledge, computational approaches can formulate new predictions of gene function. These predictions can drive laboratory experiments, which generate data either confirming or denying the predictions. Armed with this new data and knowledge, the process can iteratively progress leading to further discovery.

The work presented here begins to address many of these shortcomings of modern computational biology. This work focuses on user-driven search methods and exploratory visualization techniques designed for gene expression microarray data that incorporate expert biologists into the earliest stages of computational analysis. Further, we validated these approaches with large-scale laboratory experiments that demonstrate the utility of these methods to accelerate real-world biological discoveries.

1.2 Gene expression microarrays

1.2.1 Biological motivation

While the genetic code, in the form of DNA, contains all of the instructions necessary to build a working cell or organism, proteins are the workhorses that comprise the building blocks of cells and perform the functions necessary for life. Portions of DNA, called genes, encode the amino acid sequences needed to create all of the proteins in an organism. In the simplest case, proteins are created by a two-step process of transcription and translation. During transcription, a gene's sequence is copied from the nuclear DNA to an mRNA, which is free to leave the nucleus. This mRNA is later translated by ribosomes in the cytoplasm to create the specific sequence of amino acids that form the coded protein.

Specific proteins are not produced at all times. The amount of each protein required by a cell can change dramatically from virtually none to a very large number of copies depending on cellular and environmental factors. As such, transcription and translation both play a regulatory role in determining which proteins are produced in what quantities at which times. Thus, understanding the timings and quantities of protein production can provide indications about the potential roles of proteins within cells. For example, if a protein is produced only when a cell is exposed to conditions of unusually high heat, that protein may be involved in a cell's response to a hotter environment, perhaps by maintaining the proper fold of other proteins or catalyzing a heat-specific metabolic reaction. Further, if two proteins are regulated and produced

with similar timings and/or quantities, they may perform the same or related functional roles. For these reasons, understanding protein production and regulation is vital for our understanding of gene/protein function. While both transcription and translation are important for the regulation of protein production, the transcriptional response of many genes is both strong and easily measurable using modern technology.

1.2.2 Technology

Gene expression microarrays provide a quantitative measure of the transcription levels of thousands of genes in a genome simultaneously [12]. In general, microarray technology relies on the chemical nature of mRNA to hybridize to its complementary nucleotide sequence. Microarrays typically place known complementary sequences for large numbers of genes in a genome at specific positions on slides. These slides are then exposed to pools of mRNA (or cDNA) isolated from samples and labeled with a measurable dye. The mRNA binds to the slide location containing its complementary sequence, and the amount of dye present at each position on the slide is used to quantify the amount of hybridization that occurred, which is used as a measure for the level of transcription for each gene. A schematic overview of a microarray experiment is shown in Figure 1.3.

While all gene expression microarrays share the goal of quantifying the transcription level of genes, there are several variations in specific technologies employed. Microarray techniques such as that shown in Figure 1.3 are often referred to as “two-color” microarray experiments because they utilize two

differently labeled samples – a reference sample and a test sample. Typically the test sample is drawn from a population of interest (e.g. cells exposed to a chemical, growing in an altered environmental state, or containing a genetic mutation), while the reference sample is often drawn from a population of wild type or normal cells. Thus, the transcription level of a gene under a condition of interest can be quantified by its change from a normal condition. Many of the original microarray experiments, such as those performed at Stanford, as well as some modern commercial platforms, such as Agilent, employ a two-color approach.

Several other commercial platforms are currently available that employ variations of this basic approach. For example, Affymetrix produces a "one-color" microarray platform utilizing highly calibrated "perfect match" (PM) and "mis-match" (MM) probes to quantify expression. In this case a PM probe contains the proper complementary sequence for an mRNA, while a MM probe contains an incorrect nucleotide in its sequence. Thus if a PM probe and paired MM probe achieve a similar level of hybridization, the probe can be discarded since the mRNA binding was not specific for the PM probe's sequence. While this does not allow a gene's transcriptional levels to be directly measured as changes from a baseline in the manner of a two-color array, it does ensure that the reported hybridization levels are specific. Often, additional one-color microarrays are used to establish such a baseline, and the resulting data can be treated in a similar manner to two-color approaches.

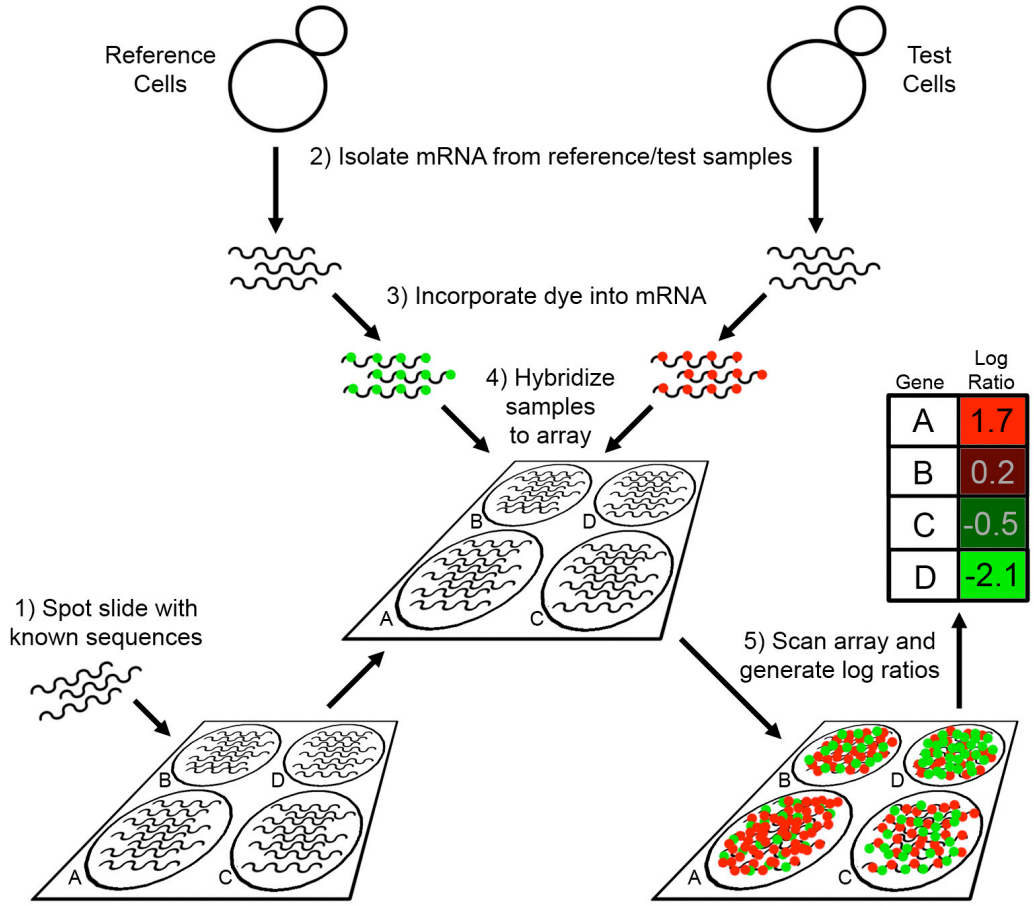


Figure 1.3: Schematic of microarray methodology. Gene expression microarrays simultaneously quantify the transcriptional activity of many genes. This schematic shows the basic methodology of a "two-color" microarray platform. First, slides are spotted with known, complementary sequences. Then mRNA is harvested from two samples, a test sample of interest and a reference sample. The mRNA is differentially dyed, and then the samples are hybridized to the slide. The amounts of dye at each spot on the resulting slide are measured to determine transcription levels for each gene.

While the particulars of downstream microarray analysis must consider the source platform used to generate the data, there are several common important properties of microarray data for analysis. Foremost, microarray data contains high levels of noise, stemming from biological, chemical, and experimental sources. While some noise sources can be minimized through careful usage of

protocols and normalizing the environment where experiments are conducted, transcription is an inherently noisy proxy of protein abundance. Post-transcriptional modifications of mRNA as well as post-translational regulation of proteins limit the amount of information that can be represented by measures of transcriptional levels of genes. Thus, any microarray analysis method must take into account the inherent large levels of noise.

1.2.3 Analysis challenges

The end results of most microarray experiments are very large matrices of numbers representing the expression level of many genes under a variety of experimental conditions. Most studies perform between 5 and several hundred hybridizations to measure transcriptional responses in a variety of related conditions. The resulting datasets are often represented with rows corresponding to genes and columns corresponding to the conditions examined.

Typical analyses of microarray datasets are based on the “guilt by association” principle of observing common patterns between genes and groups of genes. As such, clustering techniques are particularly popular for the initial phases of analysis [23]. Most clustering approaches attempt to reorder rows and/or columns of a data matrix to place similar genes and/or conditions closer to one another. An example of a simple hierarchically clustered dataset is shown in Figure 1.4. Additional analysis methods, as well as their benefits and shortcomings, are discussed in the next several chapters.

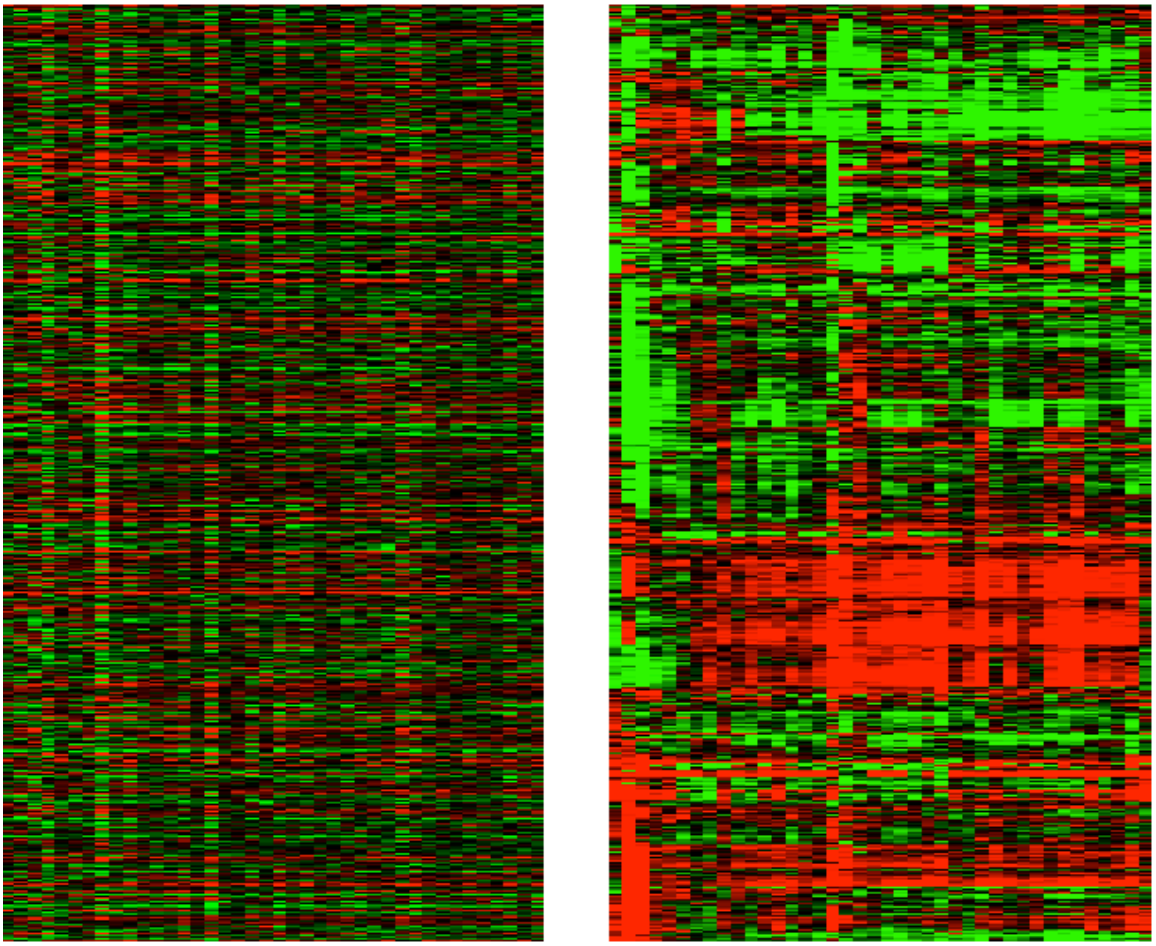


Figure 1.4: Example of microarray clustering. The same data is shown un-clustered and clustered. On the left the data is not clustered (rows in genome order, columns in random order), and on the right both rows and columns have been hierarchically clustered. Several major "guilt by association" patterns are already evident using this simple clustering technique.

1.3 Contributions and Overview

This work describes algorithms and methods for the analysis, exploration, and visualization of microarray data with the goal of elucidating patterns and structure that are important to characterize novel functions of genes and proteins. Our approach differs from previous methods in several key manners in order to address many of the biological and computational concerns described above. First, these methods incorporate user input into the earliest phases of analysis,

which allows researchers to discover patterns and information related to their specific area of expertise. Second, these methods are executable at interactive rates, which allows users to further refine and direct their research in an exploratory fashion. Third, these methods utilize novel visualization interfaces to present the actual underlying data rather than simply reporting summary statistics or gene lists. Finally, these methods have been extensively validated by using them to predict and experimentally confirm novel biology, which strongly affirms their biological utility.

The remainder of this dissertation describes these methods and approaches in detail. Chapter 2 describes a query-driven similarity search method for utilizing large collections of microarray data to robustly locate meaningful “guilt by association” patterns. In this chapter the method is validated on a small scale using specific examples, but a more comprehensive evaluation appears later. Chapter 3 focuses on visualization-based analysis methods of clustering techniques that incorporate statistical measures used during the clustering process into the visualization scheme. Chapter 4 introduces a visualization approach for simultaneously exploring multiple gene expression microarray datasets. Chapter 5 incorporates aspects of the previous three chapters into a unified platform for more comprehensive analysis and visualization of large microarray compendia. Chapter 6 describes a large-scale approach integrating the computational function prediction methodology described in chapter 2 with rigorous experimental laboratory methods to

iteratively explore the role of genes in mitochondrial function in *S. cerevisiae*.

Finally, chapter 7 summarizes and concludes this work.

Chapter 2

Exploring the functional landscape of gene expression: directed search of large microarray compendia

2.1 Introduction

The recent, rapid expansion in the amount of functional genomics data created by the biology community promises to provide broad understanding of protein function and regulation on a systems level. In particular, the increased accessibility and lower cost of gene expression microarrays has led to the publication of hundreds of studies in a variety of organisms. However, these data have thus far remained vastly underutilized. While much work has been done investigating individual datasets, advancement of knowledge in the field requires intuitive methods for biology researchers to quickly and easily explore the totality of existing data, to identify the datasets and publications relevant to their area of interest, and to locate the important information within those datasets. For example, a biologist interested in DNA damage repair should not be limited to analysis of a single dataset concerned with exposure to DNA damaging agents, but rather should be able to quickly determine which published microarray experiments elicit a DNA damage response, find the relevant portions of those datasets, and then be able to examine that data to draw conclusions and form hypotheses.

No existing approach for microarray analysis allows for fast, intuitive exploration of the large, diverse collection of published gene expression data. The utility and necessity of exploration-based techniques has been demonstrated for microarray data on the much smaller scale of one or a few datasets. General clustering techniques and bi-clustering methods have been successfully used to allow biologists to find relevant information in this small-scale setting. However, these methods are not appropriate for application to very large-scale microarray compendia due to sensitivity to noise that is compounded when aggregating data, an inability to work with data generated under diverse conditions, and/or prohibitively slow running times.

Typical clustering approaches group genes together to minimize a distance function between genes. While these distances can be quickly calculated across the concatenation of many datasets, their biological accuracy greatly decreases when taken over heterogeneous conditions. This approach is sometimes referred to as “mega-clustering” in the literature [8, 27, 74] and while appropriate in limited experimental settings involving small numbers of biologically related datasets, it is not appropriate for analysis of large-scale, heterogeneous collections of gene expression data [55]. Signals present in only a few of the datasets in a compendium are lost when the total data collection is large, causing clustering techniques to capture only the global signals in the compendium and miss more specific signals. Thus, clustering is best limited to initial exploratory analysis of single datasets.

Bi-clustering methods seek gene similarity in only a subset of available conditions, which is more appropriate for functionally heterogeneous data [17, 55]. However, the most basic formulations of bi-clustering allow for the selection of any subset of conditions, which is often not biologically meaningful when the selected conditions bear no relationship to each other. As data compendia increase in size, it becomes more conceivable for these bi-clustering formulations to find patterns in the noise, as finding arbitrary subsets of conditions where genes exhibit similar levels of expression becomes easier by pure chance as the number of conditions increases. Further, the general bi-clustering problem is NP-complete [55], meaning that these methods can require unreasonable running times to find complete solutions, particularly on large data collections.

As the general bi-clustering problem is often intractable, a variety of heuristics and normalization steps are utilized in practice. For example, some approaches obtain faster running times by limiting the types of bi-clusters they can identify [89], or by focusing on specific types of data, such as time courses [54]. Other bi-clustering methods achieve tractable complexity by starting with a query set of related seed genes and iteratively growing out maximal bi-clusters around the seed [39].

Another approach for microarray data exploration is a query-driven search process, such as the feature selection-based Gene Recommender algorithm [65]. This approach has proven very useful on the scale of smaller data compendia, however it is not as effective when applied to very large-scale collections. As with some formulations of bi-clustering, feature selection

techniques may find noisy patterns among unrelated conditions, and can require lengthy computation times for complete analysis.

To address all of these shortcomings, we propose a more scalable, context-specific search methodology that enables biology researchers to explore the entirety of very large microarray compendia in a biologically meaningful manner. Our approach offers many fold higher biological accuracy and running speeds many times faster than current techniques. We have also categorized the functional coverage and biases of this collection to assess which biological areas are well characterized in the current microarray compendium and which areas are open to further study. Based on this compendium of data we demonstrate the effectiveness and usefulness of our approach for information exploration and hypothesis formulation. We have implemented our algorithm in an interactive, web-based search engine available at http://function.princeton.edu/SPELL.

2.2 Methods

In this section we briefly discuss our collection of microarray data and our functional coverage analysis of this compendium. We then discuss in detail our fast, context-sensitive search procedure, called SPELL.

2.2.1 Creation of the *S. cerevisiae* gene expression data compendium

We collected 117 microarray datasets from 81 publications totaling 2394 array hybridizations from a variety of sources [15, 18, 22, 51, 81]. Missing values were imputed using the KNN impute algorithm with K=10 using Euclidean distance [91] and technical replicates (i.e. spot repeats and dye swaps) were

averaged together, resulting in data files of complete matrices with one entry per gene appearing in the dataset (see Appendix A for further details and the complete list of datasets).

Gene similarities are calculated within a dataset containing *n* conditions using the Pearson correlation coefficient, ρ, as defined by:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x\sigma_y},$$

where *x* and *y* are expression level data vectors for two genes, μ_x and μ_y are means, and σ_x and σ_y are standard deviations. However, the distribution of all pair-wise Pearson correlations varies greatly from one dataset to the next. This is a function of several factors, including the number of experimental conditions in a dataset, the biological process targeted, and the microarray technology employed. In order to better compare correlations between datasets, we apply Fisher's z-transform to improve comparability [24]. The Fisher z-transformed correlations, *z*, are defined as:

$$z_{x,y} = \frac{1}{2}\log\left(\frac{1+\rho_{x,y}}{1-\rho_{x,y}}\right),$$

where ρ is defined as above. As a final step, we standardize these quantities by subtracting the mean correlation within each dataset and dividing by the corresponding standard deviation which results in approximately normal distributions [$\sim\mathcal{N}(0,1)$] of correlations within each dataset under the assumption,

based on empirical observation, that the true underlying distribution of the data is approximately normal (Figure 2.1).

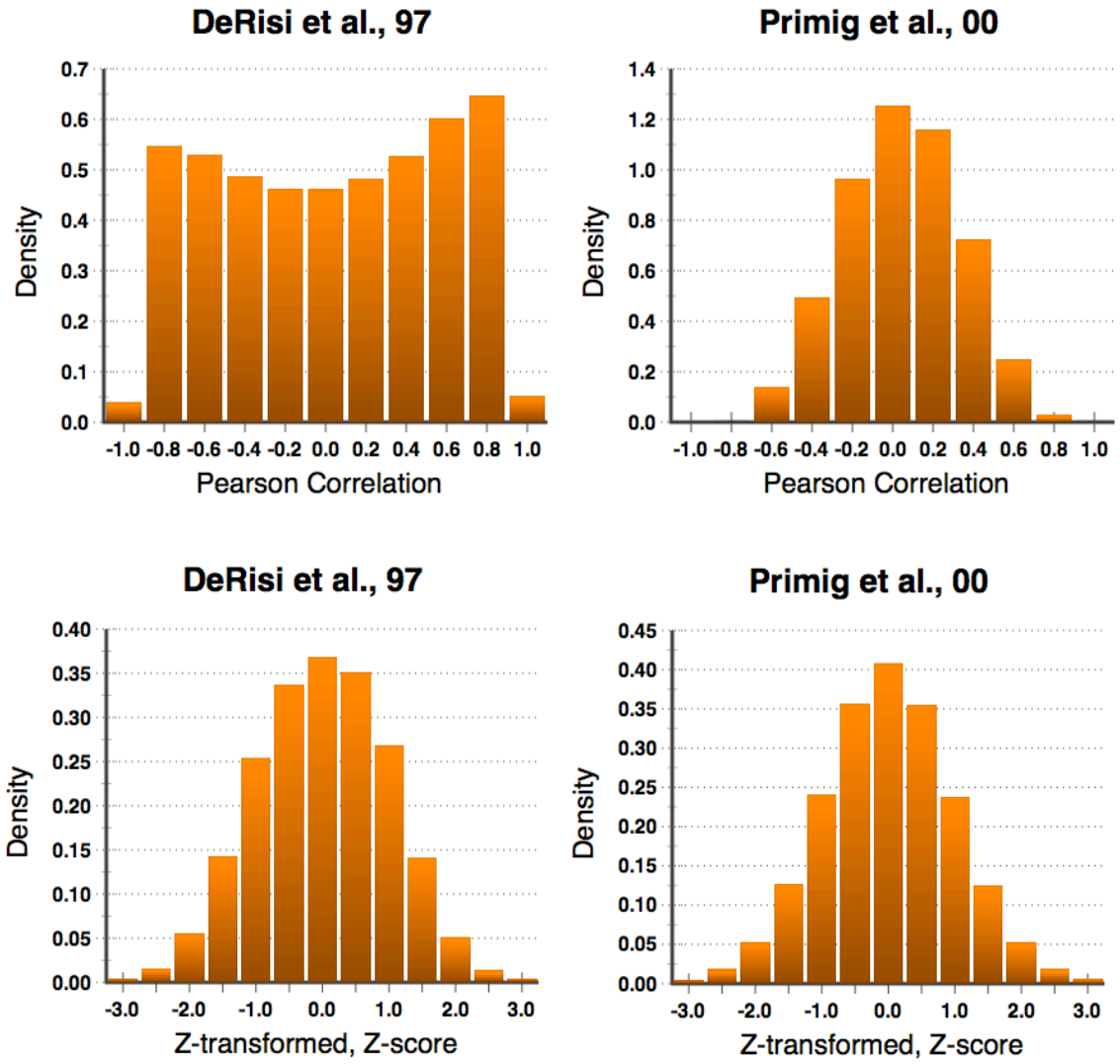


Figure 2.1: Example results of Fisher Z-transformation. The distribution of Pearson correlations between all pairs of genes within each dataset varies greatly depending on the number of conditions in each dataset, the process targeted, and the array platform used. In order to ensure comparable measures of correlation from one dataset to the next, we employ the Fisher z-transformation. Here we see the distribution of all pair-wise correlations for two datasets before and after applying the z-transformation. After the transform, the distributions are approximately normal for all datasets.

2.2.2 Functional coverage analysis

As motivation for our search algorithm presented in the next section, and in order to characterize which biological processes are represented in the compendium, we analyzed the functional coverage of each dataset over a variety of Gene Ontology (GO) terms [4] using the z-test for significance. Given the background of all pair-wise z-scores within a dataset, *d*, for each GO term, *g*, we calculated all pair-wise correlations for the n_g genes annotated to the term and find the mean sample correlation, μ_g. The z-test statistic for each GO term/dataset pair, $\zeta_{g,d}$, was calculated as:

$$\zeta_{g,d} = \sqrt{\frac{n_g(n_g-1)}{2}} \quad \frac{\mu_g - \mu_b}{\sigma_b} ,$$

where μ_b is the mean of the background distribution and σ_b is the background standard deviation. Approximate significance of these z-statistics was computed based on an upper-tailed hypothesis test [58]. The calculated p-values are approximate due to the assumption of underlying normality in the data and because correlations among genes annotated to the same GO term are not necessarily independent. For display in Figure 2.6, the resulting matrix of pseudo p-values was hierarchically clustered in both dimensions. In addition to the z-test presented here, we have calculated significance using the non-parametric Kolmogorov-Smirnov test (see Appendix B for details).

2.2.3 Search algorithm details

Motivated by our characterization of the functional coverage of the compendium, we have devised a search procedure to leverage the

compendium's diversity. Our search algorithm is based on two components: a signal balancing technique that enhances biological information; and dataset relevance weighting to identify functional patterns within datasets that are meaningful given a set of user-provided query genes. (Note that this algorithm is independent of the functional coverage analysis presented in section 2.2.2.) We refer to this algorithm as SPELL (Serial Patterns of Expression Levels Locator). A schematic overview of this method is shown in Figure 2.2.

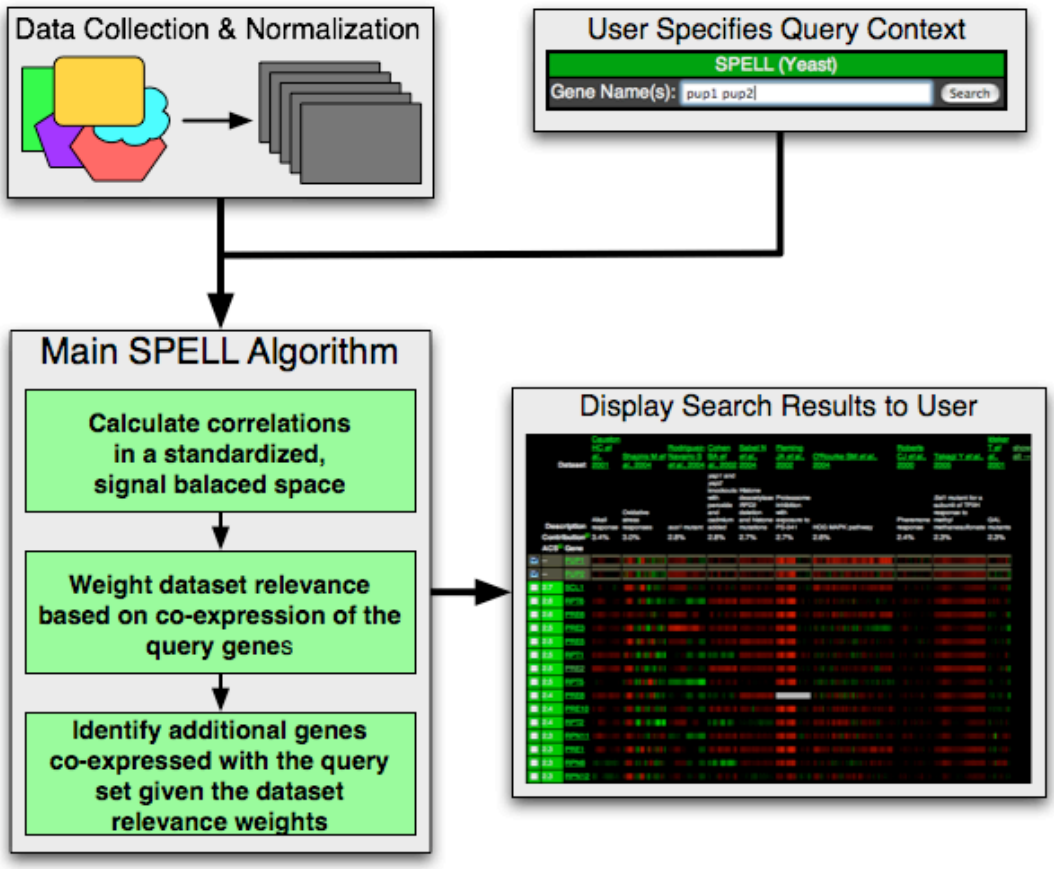


Figure 2.2: Schematic view of the SPELL search engine framework. Our system consists of several key components and phases shown here. Input to the main algorithm consists of a collection of normalized gene expression datasets and a set of researcher-provided query genes of interest. Our algorithm relies on signal balancing coupled with a method to select datasets relevant to the specified query. The algorithm identifies additional genes highly co-expressed with the query set and returns that list to the researcher.

2.2.3.1 Identification of functional patterns through signal balancing

While correlations between the original data vectors in microarray datasets are biologically meaningful, the high levels of noise in these datasets can lead to spurious results, particularly in the context of very large compendia. Singular value decomposition (SVD) has been applied to several other problems in microarray analysis, and it has been shown that this process can lead to substantial noise reduction [1, 95]. We apply SVD in a novel way to re-balance the signals present in datasets.

Briefly, SVD factors an original $m \times n$ data matrix, X, into 3 component matrices of the form:

$$X_{m \times n} = U_{m \times n} \Sigma_{n \times n} V^T_{n \times n} ,$$

such that Σ contains the singular values of X along its diagonal in decreasing order and U and V^T contain the left- and right- singular vectors, respectively. In practice, V^T defines an orthonormal basis for the columns of X in decreasing order of corresponding singular values, while U defines the projection of each original data vector in this new basis.

In contrast to typical applications of SVD for microarray analysis, we calculate correlations between genes' coefficients in U rather than re-project to an approximation of X. In this case, U can be interpreted as the "balanced" projection of X onto its right singular basis, where the balancing weights are inversely proportional to the singular values defined by Σ, i.e. $U=XV\Sigma^{-1}$. Correlations between genes in U equally weight each dimension of the orthonormal basis and balance their contributions such that the least prominent

patterns are amplified and more dominant patterns are dampened. This process helps reveal biological signals, as some of the dominant patterns in many microarray datasets are not biologically meaningful.

We have quantitatively found that our application of SVD to microarray data for the purpose of signal balancing performs much more accurately than the traditional use of SVD for noise reduction. In the traditional use of SVD, low singular values and their corresponding singular vectors are removed from the decomposed matrices ($U\Sigma V^T$), then the matrices are multiplied back together to reconstruct a version of the original data matrix (X). Often, enough singular values are retained to account for some percentage of the variation of the original data. However, in our analysis we find that performance generally degrades when using this traditional application of SVD. Rather, by calculating correlations within the left singular vectors (U) we perform our analysis in a space where the more dominant patterns are dampened and the less dominant patterns are magnified, which produces better results (Figure 2.3). Note that this process is related to some applications of SVD to microarray data, such as the work by Alter *et al.* [1], which found that dominant eigengenes are sometimes highly correlated with systematic effects in the data.

We apply this signal balancing approach to each dataset in our compendium separately. All correlations calculated during our search procedure in the next section are calculated in the resulting signal balanced U matrices rather than the original data matrices.

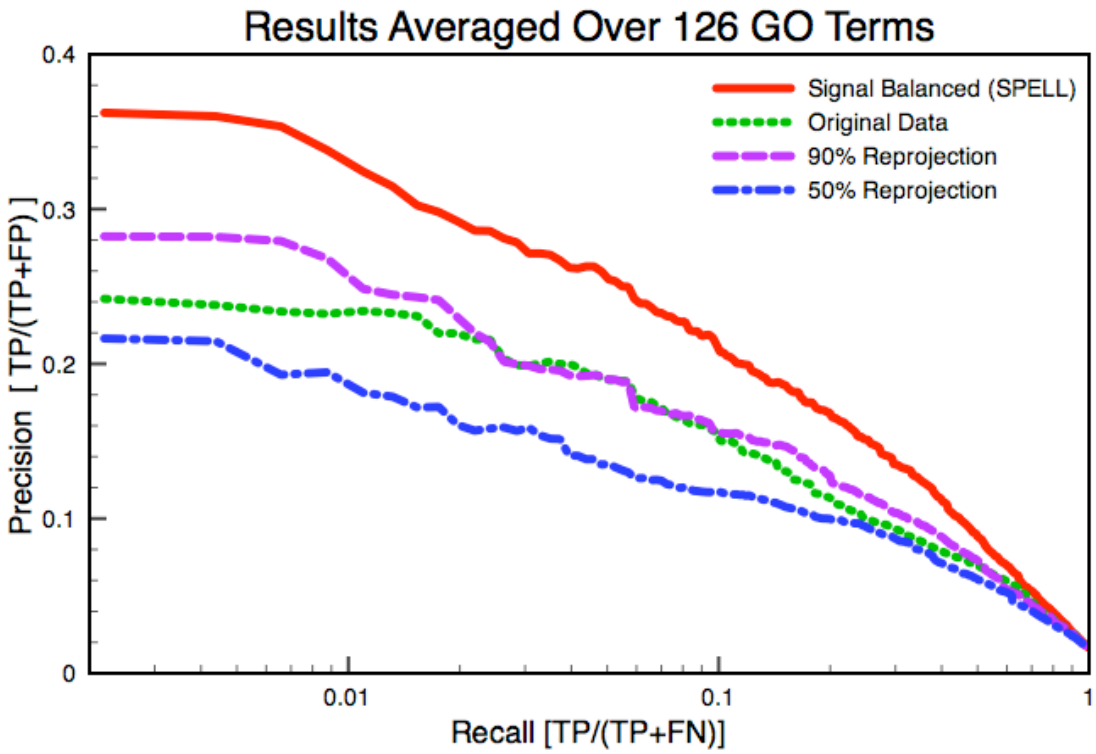


Figure 2.3: Results of SVD-based signal balancing. This graph evaluates our method in the manner described in section 2.2.4. In all four cases the same evaluation is applied, however different input matrices are used corresponding to our use of SVD for signal balancing, the original data matrix, and retaining 90% and 50% of data variance and reconstructing the original data matrix. Our use of SVD-based signal balancing outperforms both the original data and traditional applications of SVD.

2.2.3.2 Query-based search

Given a compendium of signal balanced microarray datasets, *D*, and a query set of genes of interest, *Q*, our approach assigns a relevance weight to every dataset in the compendium. We then identify additional genes closely related to the query set within the weighted datasets. Given a set of query genes, $q_i \in Q$, we determine a relevance weight, *w*, for each dataset, *d*, in our compendium as the mean of all pair-wise z-transformed correlations, *z*, among the query genes:

$$w_d = \left(\frac{2}{|Q|(|Q|-1)} \right) \sum_{i=1}^{|Q|-1} \sum_{j=i+1}^{|Q|} f(z_{q_i,q_j}) ,$$

where the function *f* is used to control the contribution of the correlations to the dataset relevance weights. Empirically, we have found that a quadratic function of the z-transformed correlations produces more accurate results (as compared to linear, cubic, or exponential functions) by giving relatively more weight to higher correlations. Also, we find that negative correlations are generally less biologically meaningful than positive correlations (Figure 2.4). Therefore we also limit the influence of negative correlations by disregarding z-transformed correlations less than one standard deviation away from the mean, resulting in the following:

$$f(z) = \begin{cases} z^2 & if \quad z \geq 1 \\ 0 & otherwise \end{cases}$$

Given these weights for each dataset, we calculate a per-gene score, *s*, as the mean of weighted correlations to the query set for each gene *x*, across all *D* datasets in the compendium as:

$$s_x = \frac{1}{|Q| \sum_{d \in D} w_d} \sum_{d \in D} \sum_{q \in Q} w_d f(z_{x,q})$$

Once scores are calculated for all genes, the results are sorted, and the top results are returned. The effect of this process is to select those datasets most relevant to the biological context defined by the query and identify additional genes related in these datasets.

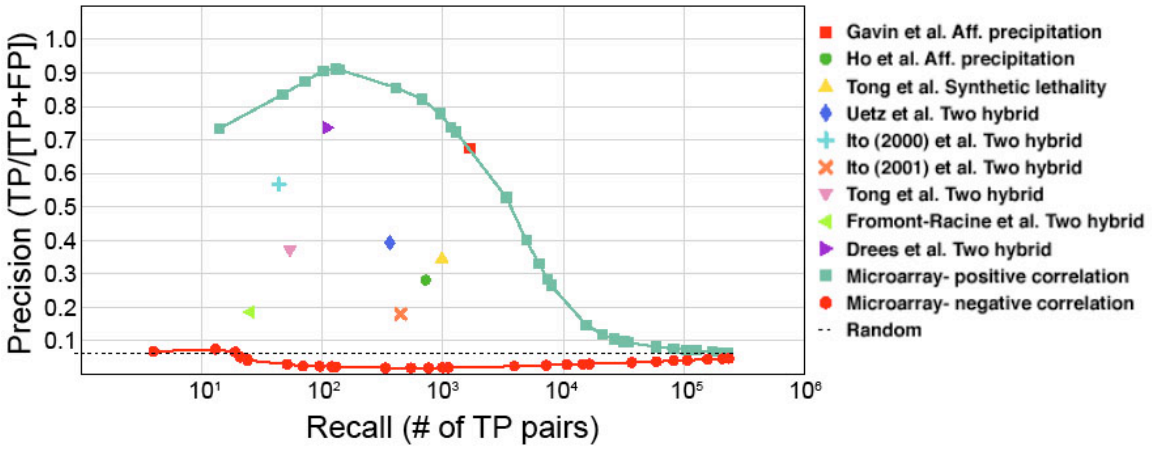


Figure 2.4: Positive vs. negative correlation performance. We have found that negative correlations tend to not be functionally informative in many cases. As an example of this effect we have examined the precision-recall plot of positive correlations across all microarray data and negative correlations across all data. The following graph was created using the GRIFn system [60]. Several reference datasets are included for comparison.

2.2.4 Performance evaluation methodology

In order to evaluate our method's performance, we assessed the ability of our approach to recapitulate known biology by examining a set of 126 functionally distinct GO terms selected by an expert curation of the hierarchy performed by [60]. These GO terms were identified as both specific enough such that predicted annotations could be validated through laboratory testing, but also general enough to reasonably expect high-throughput data to be informative. We excluded very small terms (less than 10 annotated genes), as results can be misleading with such small numbers of positive examples.

We estimated precision-recall characteristics of our method through extensive cross-validation. For each GO term examined, we executed a separate search with each possible pair of annotated genes as the query set (i.e. "leave-two-in" cross-validation). Each of these queries resulted in an ordered list

of all genes in the genome as ranked by the algorithm tested. We combined these lists by calculating the average rank of each gene across all lists (excluding the query genes) and producing an ordered master list for each GO term from best average rank to worst. Precision-recall curves were generated based on the master list's performance over the GO term examined, and average precision was used as a summary statistic for comparisons. To create precision-recall graphs averaged across GO terms, mean precisions were calculated at the scale of the smallest recall step examined (i.e. the inverse of the number of genes annotated to the largest GO term tested). The average precision, *AP*, for each GO term, *G*, is calculated as:

$$AP_G = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{i}{rank_i} ,$$

where $rank_i$ is the rank placement of the i^{th} gene annotated to the term in the ordered list of results. Note that this metric is a quantized form of the area under the precision-recall curve.

In addition to testing the performance of our SPELL algorithm, we compare our results with commonly used mega-clustering techniques based on both raw Pearson correlation and Fisher z-transformed, standardized z-scores. For Pearson correlation, results were calculated across the concatenation of all data into a single large matrix. For z-scores, results were calculated in individual datasets and the z-scores were averaged together. We also compared SPELL with another unsupervised, query-driven search technique, the Gene Recommender algorithm [65]. However, as this algorithm was not designed for

analysis on this scale over such a large collection of data, the running time limited this comparison to the 82 smallest of the 126 GO terms used in other comparisons. In all cases, the same cross-validation and bootstrapping procedure was used. Several results of these comparisons are shown in Figures 2.7 and 2.8 (see Appendix C for further results).

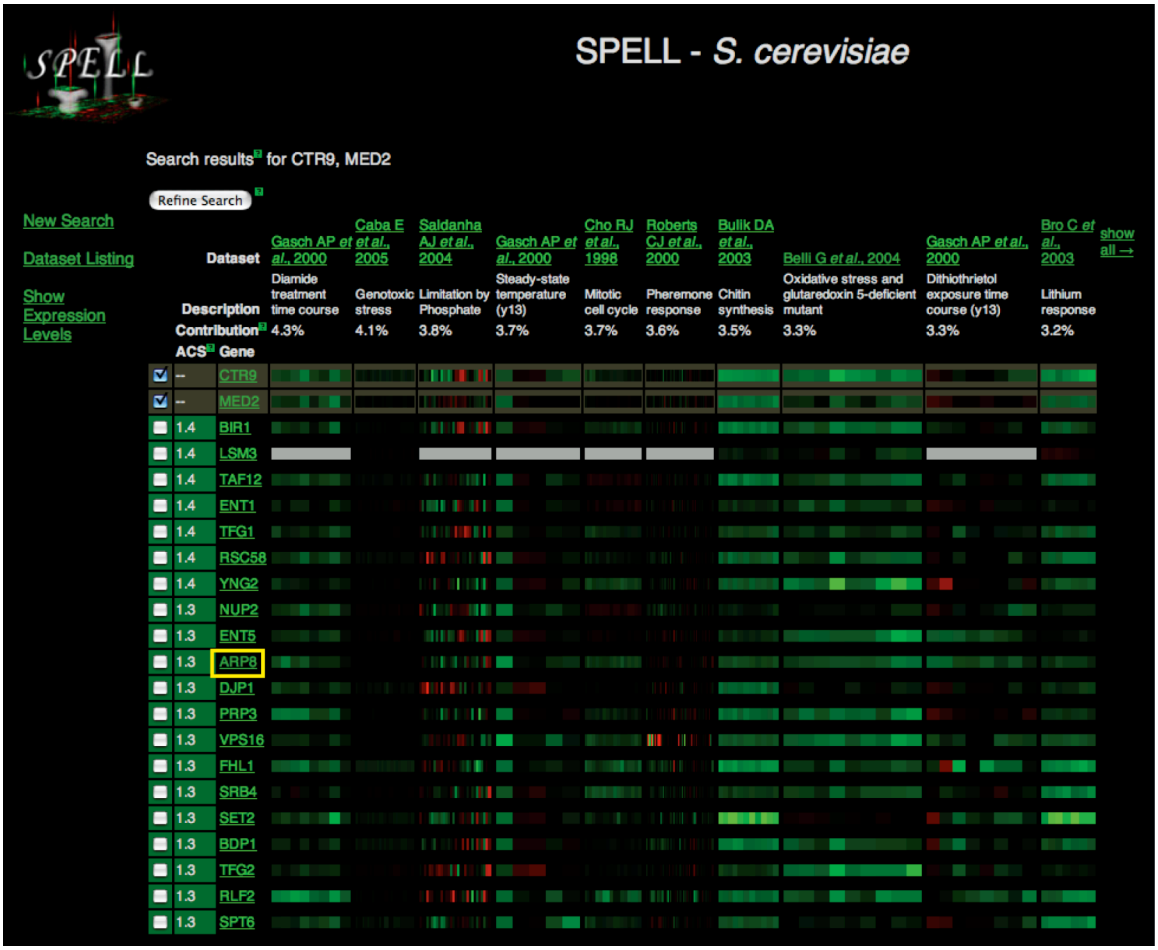


Figure 2.5: Example result page from the SPELL search engine. This is a screenshot of the results page from a query performed using the web-accessible search engine of our SPELL algorithm. In this example, the user specified a query of 2 genes related to transcription, *CTR9* and *MED2*. The resulting list of related genes is significantly enriched for the GO biological process "transcription from RNA polymerase II promoter" as expected. The un-annotated gene *ARP8* is also in this list (highlighted), and subsequent investigation confirms that this gene likely plays a role in this process.

2.3 Implementation

Our SPELL methodology is implemented in a web-accessible search engine at http://function.princeton.edu/SPELL. Our interface allows a researcher to provide a list of query genes, then the search engine reports which datasets are most relevant to that query, lists additional genes related to the query within the relevant conditions, and displays the expression levels of these genes. Links to extra information about each dataset, the original publications, and gene information are also provided. Queries are processed in seconds, which allows researchers to quickly locate and observe the relevant portions of the data compendium.

In addition to processing initial searches, users can refine and direct their search in a serial fashion, which allows researchers to more fully explore the data compendium by observing which biological conditions induce stronger or weaker correlations among varying sets of query genes. Thus a user can target the query to particular biological processes, which is especially valuable when investigating genes that are involved in multiple functions. A screenshot of this search engine is shown in Figure 2.5.

2.4 Results and Discussion

2.4.1 Functional coverage analysis of the microarray compendium

To map out the functional landscape of existing gene expression microarray data in *S. cerevisiae*, we have collected a large data compendium and examined it for coverage of known pathways and biological processes. Our

collection contains 117 distinct datasets spanning 2394 array hybridizations. To our knowledge, this is the largest single microarray data compendium for *S. cerevisiae*.

In general, we expect different datasets to activate different pathways depending on the experimental condition studied. For example, stress response datasets should show a strong signal for ribosomal processes, but not necessarily meiosis, for which a sporulation time course may be better suited. We quantified this effect for our *S. cerevisiae* microarray compendium over a broad selection of biological processes as defined by GO and the Saccharomyces Genome Database (SGD) annotations [18]. For each GO term and dataset combination, we examined the statistical difference between the expression correlation among annotated genes and the background correlation among all genes within the dataset (see Methods for details). The results of this evaluation are summarized in Figure 2.6 (see Appendix B for further information).

This analysis illustrates both which datasets are informative of each biological area and which biological areas are represented in the compendium at large. Some subsets of GO terms are significant in nearly all datasets, such as ribosomal processes (Figure 2.6b). In contrast, many biological processes are active in only a few datasets, generally those where experimental conditions were specifically targeting the process in question. An example of this is GO terms that relate to the process of meiosis (Figure 2.6c), which are significant in only a few, targeted datasets.

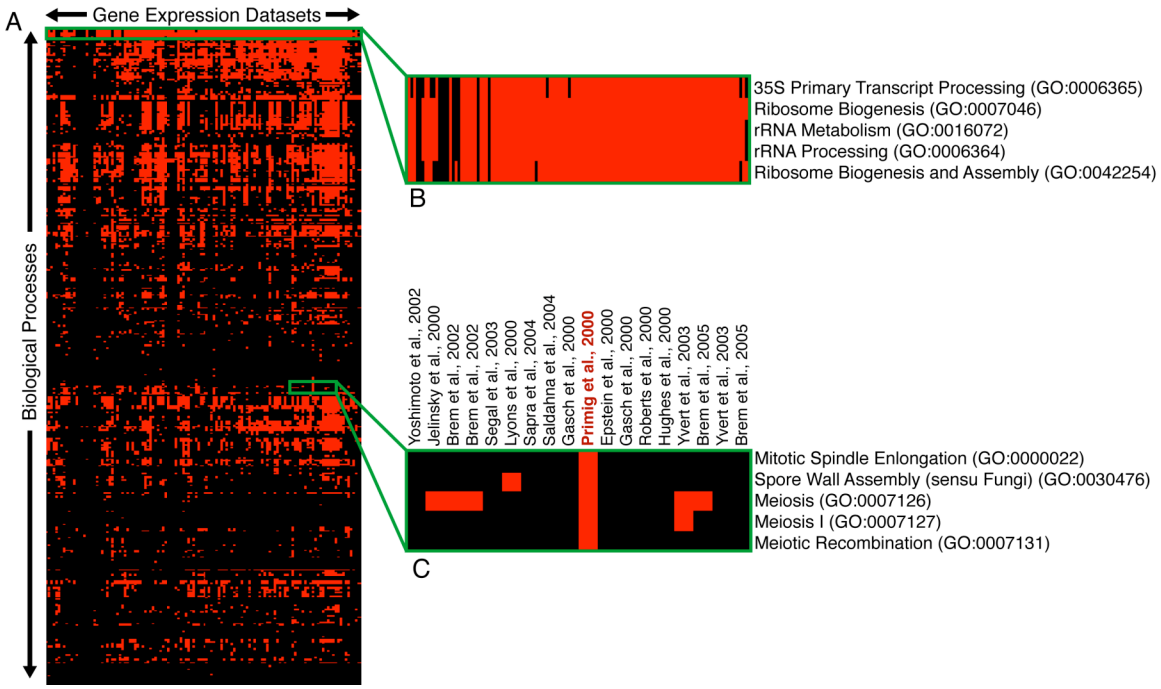


Figure 2.6: Functional coverage within the *S. cerevisiae* microarray compendium. We examined the functional coverage of the datasets from our yeast microarray collection in a very broad selection of 403 biological pathways and processes defined by GO. We measured the approximate significance of the differences in distributions of pair-wise correlations between genes annotated to a GO term and the background distribution of all genes within each dataset. A) shows the full result plotting every dataset in columns versus GO terms in rows. Dataset/GO term pairs with significant signal enrichment are colored red (p-value < 10^{-4}, Bonferroni corrected). B) shows a detail of a group of ribosome related processes that are significantly enriched in almost all datasets. C) shows detailed results for a group of meiosis related processes that are enriched in only a subset of datasets, including the highlighted sporulation time course [68]. This analysis demonstrates both which functional areas are represented in each dataset as well as which areas remain to be studied through gene expression assays (see Appendix B for additional results).

Finally, our analysis identifies several functional groups not significantly represented in our compendium, and thus likely not covered by currently available microarray data. These fall into several categories: pathways not believed to be transcriptionally regulated, functions that do not occur in many lab strains, and finally, functional areas which may not have been targeted by a specific assay to induce co-regulation (see Appendix B for details).

2.4.2 Query-driven search

Our approach to analysis relies on signal balancing coupled with context-sensitive search to provide fast, accurate performance. Given a set of query genes from a user, we weight the relevance of each dataset based on the query genes' correlation within that dataset. We then calculate the context-weighted correlation of every other gene back to the query set to identify the genes most related to the query set to report as results. Note that this approach is unsupervised in that the search process is independent of the functional coverage analysis discussed above.

By considering correlations only in entire logical datasets (e.g. a heat shock time course), we harness the biological diversity in the collection in a meaningful way. As we know that different datasets contain signals from different biological processes, it is vital to examine signals in those subsets of the compendium that are relevant to a particular area. By determining dataset relevance based on the query sets' correlation, our method uses the data itself to determine which datasets are important for a specific query, rather than relying on a literature search or curation. This approach allows specific signals that may be present in only a few datasets in the compendium to be found without explicit prior knowledge of what the compendium contains. Another important benefit of examining correlations only in functionally coherent units is that this approach is able to compare and combine information from datasets generated using diverse technologies. Regardless of inter-dataset differences in signal or noise, our method is able to isolate and identify the most important information.

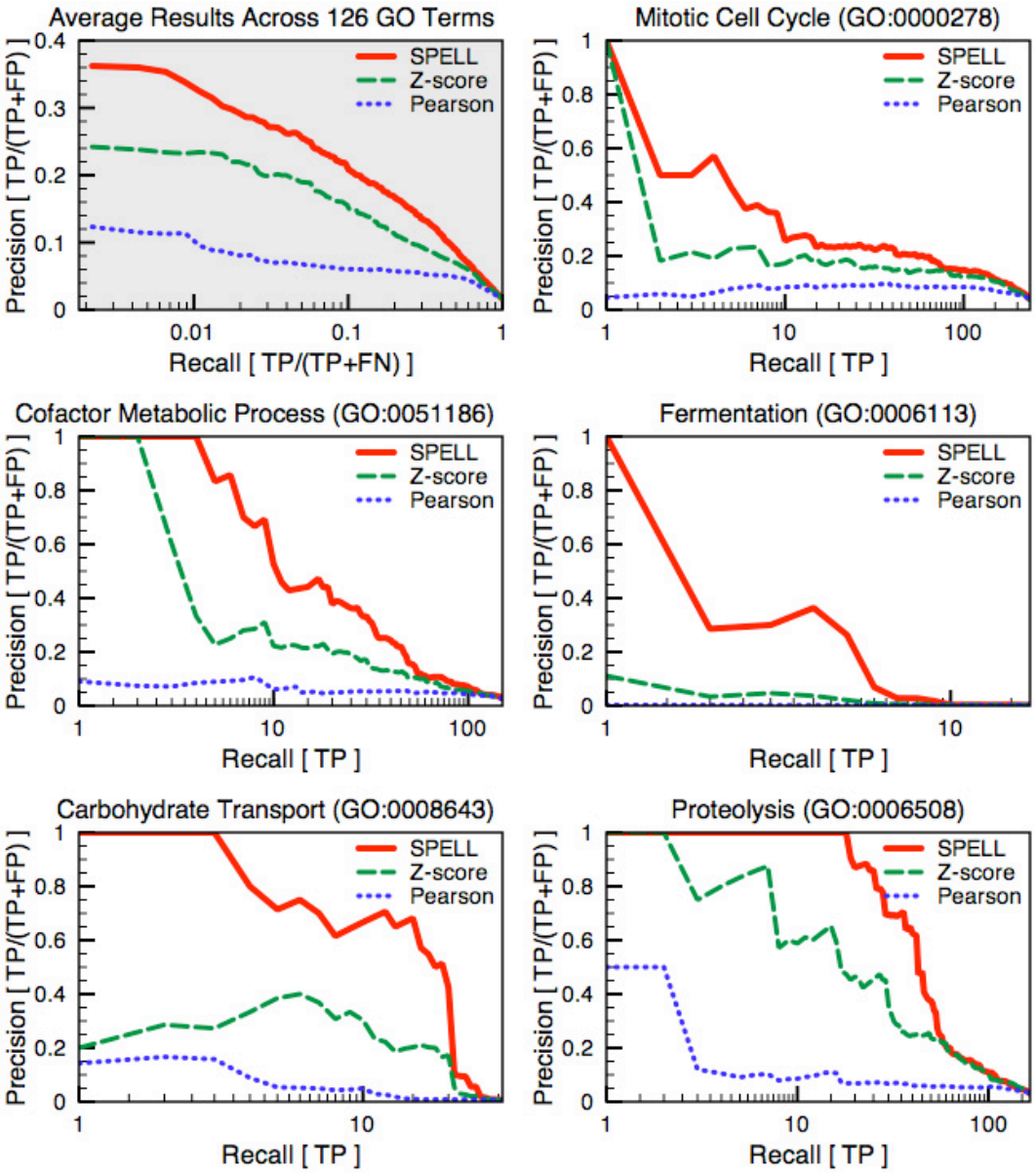


Figure 2.7: SPELL vs. mega-clustering performance. This figure compares the biological performance between the SPELL search engine and mega-clustering approaches. These graphs show the tradeoff between precision (the fraction of genes correctly identified) versus recall (the number of genes found). Results are shown for our methodology (SPELL), Pearson correlation calculated over all data concatenated together (Pearson), and average z-scores across all datasets (Z-score). The top left graph displays results averaged over all 126 GO terms examined. The remaining five graphs are a sample of the terms examined. On average, our method shows a more than 250% improvement in performance over Pearson correlation on concatenated data (see Appendix C for further details).

2.4.3 Performance evaluation in 126 biological areas

We have evaluated the ability of SPELL and other methods to reconstruct a known pathway given only a subset of genes in that pathway as input (see Methods for details). We find that SPELL recovers known process proteins with substantially higher accuracy than other commonly used approaches (see Figures 2.7 and 2.8). For instance, measured in average precision, SPELL improves by a mean of 273% over the typical Pearson correlation concatenation approach. In 35 of the 126 GO terms examined, performance increases by more than 200%, in 71 cases performance increases by more than 100%, and in a total of 101 cases performance increases by more than 50%. We find a performance decrease in only 5 GO terms, each of which has no biological signal in our gene expression compendium. Specifically, 4 of these 5 GO terms were identified as underrepresented in the collection during our functional coverage analysis, meaning no datasets in the compendium can be confidently deemed relevant to these processes. The remaining GO term where performance decreased is “DNA recombination” which contains many genes with very high sequence similarity (transposons), causing cross-hybridization effects that make dataset co-expression not biologically meaningful. Thus, for all GO terms examined where a biologically meaningful signal is present in the microarray compendium, our approach leads to an increase in biological accuracy over mega-clustering.

We also compared the performance of SPELL with another unsupervised search approach, Gene Recommender [65]. On average, SPELL exhibits a 67%

performance increase over this approach and is dramatically faster (Figure 2.8). In this analysis using a very large data collection, SPELL demonstrates a substantial improvement in biological accuracy over both simple mega-clustering techniques and the sophisticated feature selection-based Gene Recommender algorithm.

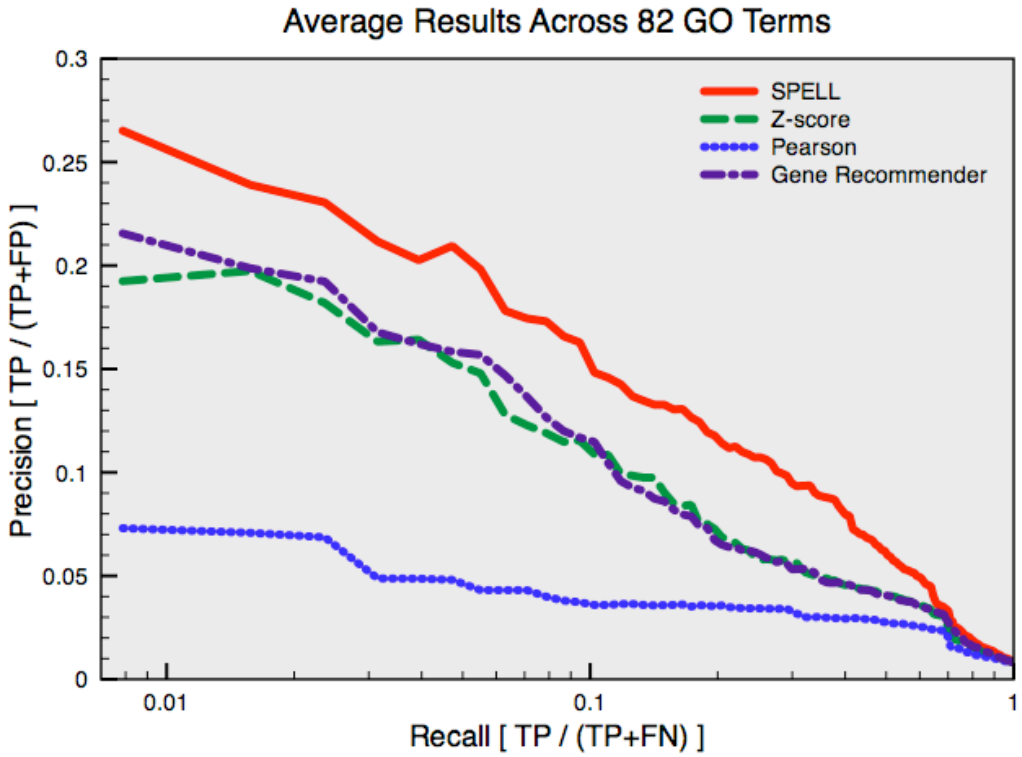


Figure 2.8: SPELL vs. Gene Recommender performance. This graph compares the biological performance of SPELL and the feature selection based search method, Gene Recommender [65]. This analysis is similar to that of Fig. 2.7, except that due to run time limitations of the Gene Recommender algorithm, this comparison was conducted on a subset of 82 GO terms. SPELL exhibits an average performance increase of 67% over Gene Recommender. (See Appendix C for further details.)

2.4.4 Novel biological predictions and confirmation

The results of our cross-validation and bootstrapping analysis can also be used to make novel gene function predictions. We examined the high-precision,

low-recall area of the SPELL results to identify potential functions for genes currently lacking any annotations to the GO biological process branch. In many cases we have found supporting evidence for these predictions in the literature, and/or conducted laboratory experiments that support the hypotheses.

2.4.4.1 Multiple functions of un-annotated gene ARP8 are predicted by SPELL

SPELL makes 13 novel functional predictions for the gene, *ARP8*, which fall into 3 categories: processes related to the cell cycle, processes related to transcription by RNA polymerase II, and processes related to cellular morphogenesis and structure (see Appendix D for complete list). Although this gene is not annotated to the GO biological process branch, several studies have been conducted that support these predictions.

Arp8 is a component of the 12 protein complex INO80. INO80 is a chromatin remodeling complex that is involved in regulation of transcription and in DNA damage response [80]. The role of ATP-dependent chromatin remodeling complexes in transcriptional regulation is well documented [16], and thus it comes as no surprise that an important component of the INO80 complex was predicted to the GO terms involved in transcriptional regulation. Perhaps more interesting, SPELL also predicted a recently characterized function of INO80—its role in both repairing double stranded DNA breaks and homologous recombination [94]. Mutants that cripple INO80 function have been shown to be sensitive to DNA damaging agents, and temperature sensitive alleles of INO80 arrest at G2/M [80]. Thus the series of GO terms related to progress through the cell cycle are extremely relevant to the function of *Arp8* in the INO80 complex.

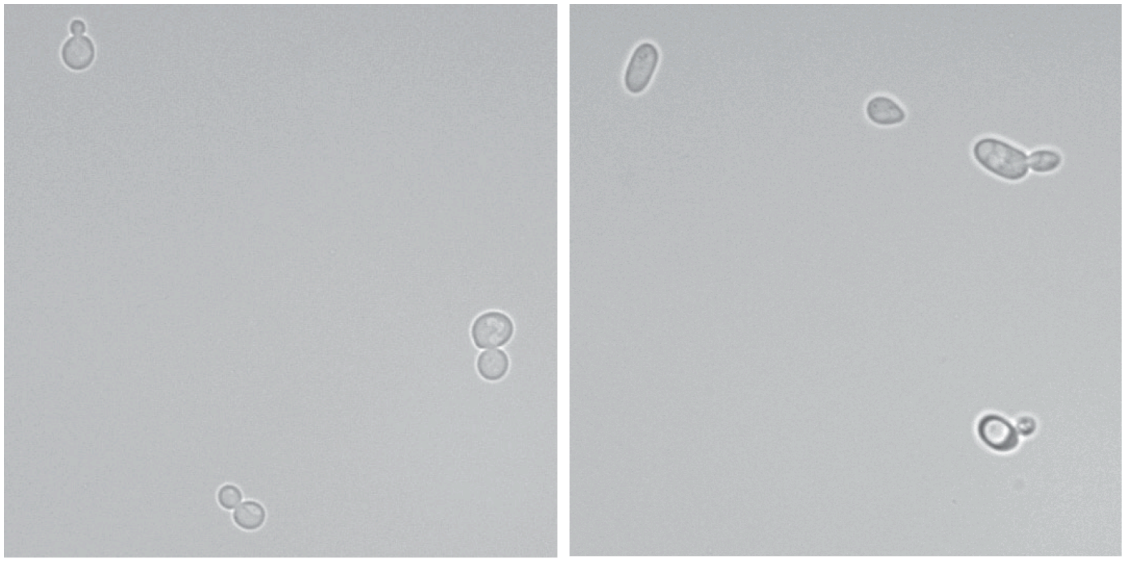


Figure 2.9: Cell morphology defect of *arp8Δ*. Our system, SPELL, predicted that the gene *ARP8* is involved in cellular morphology. Subsequent laboratory testing shows that an *arp8*Δ strain exhibits an abnormal growth phenotype. Wild-type cells (left) have a cell volume much less than the *arp8*Δ strain (right). Further, the *arp8*Δ cells have an irregular, elongated morphology when compared to the wild-type cells. This is strong confirmation of our system's prediction that *ARP8* is related to cell morphology.

A novel predicted function for the *ARP8* gene was a role in cellular morphogenesis and cytoskeleton organization. Using a complete deletion of the *ARP8* gene from the yeast deletion set [29], we grew four independent colonies of both wild-type yeast and an *arp8*Δ in rich media. We measured the cell volume for these cultures and found a dramatic increase in cell volume to 66.7±2.1 fl for *arp8*Δ, up from 36.9±0.7 fl for wild-type. Furthermore, by observing these cultures with microscopy we discovered that *arp8*Δ cells had an abnormal, enlarged ellipsoid shape compared to the rounded shape of wild-type yeast as shown in Figure 2.9. These data verify that the *ARP8* gene plays a

critical role in maintaining normal cellular shape and size, which supports these predictions of our system.

The ability of SPELL to identify several distinct functions of *ARP8* demonstrates the effectiveness of our methodology. By searching through the available data in a context-sensitive manner, our approach has the ability to identify signals in biologically diverse subsets of the compendium in a meaningful way.

2.4.4.2 SPELL predicts YDL089W is involved in sporulation

Another biological prediction made by our system is that the previously uncharacterized ORF YDL089W is involved in sporulation. Several lines of evidence strongly support this prediction. First, overexpression of YDL089W suppresses the sporulation defect of a *csm1*Δ strain [99]. *Csm1* is involved in chromosome segregation during meiosis and *Csm1* was demonstrated to have a physical interaction with YDL089W. Furthermore, a protein chip screen for targets of the *Cdc28* kinase (an important regulator of chromosome segregation at G2/M) found YDL089W as a target [93]. These results experimentally support our prediction that YDL089W plays a role in sporulation.

2.4.4.3 Support for other novel GO biological process annotation predictions by SPELL

SPELL predicts that the un-annotated protein *SET7* is involved with protein amino acid alkylation. The most common alkylation event in cells is the transfer of a methyl group to an amino acid. The SET domain has been shown to catalyze the methylation of lysine residues [100]. The assignment of the

process amino acid alkylation to *SET7* is consistent with the lysine methylation function of the *Set7* protein.

Another novel annotation prediction that is consistent with recently published data is the assignment of *TVP38* to glycoprotein metabolism. The *Tvp38* protein was recently identified as one of nine novel components in the Golgi apparatus where much of protein glycosylation occurs [40]. Furthermore, the copurification with glycosylation proteins found in this study strongly supports this functional prediction.

2.4.4.4 Effectiveness of SPELL for novel biological process annotations

The biological diversity of these verified predictions of our system demonstrate the effectiveness of our approach. Novel functions for genes as diverse as double stranded break repair, sporulation, glycosylation, and transcriptional regulation have been correctly predicted by our approach using only publicly available gene expression microarray data. We believe systems such as SPELL that can enable fast generation of meaningful hypotheses given existing data will play a key role in directing future laboratory work.

2.5 Conclusions

As the biology community is producing a very large amount of gene expression data, it is critical to develop fast, biologically relevant search methods to enable researchers to leverage all of the available data in their own analyses. To this end, we have gathered the largest single collection of *S. cerevisiae* microarray data and studied the representation of various pathways and functions within the datasets contained in this collection. Our study exhibits the

biological diversity of publicly available data and also points to several biological areas that are not yet covered by the gene expression collection.

We propose a general, effective search method for harnessing very large gene expression data compendia. We have implemented this method, called SPELL, in a web-based, context-sensitive search engine for the large scale *S. cerevisiae* data collection. The accuracy of our approach is on average more than 250% improved over existing mega-clustering techniques when recapitulating known biology. Further, our system makes several novel biological predictions that we have verified through recent publications in the literature and additional laboratory tests. While we believe that our system will be very useful for biologists, there is still room for the development of additional methods for query-driven data exploration. For example, modifications to bi-clustering algorithms or the further development of feature selection techniques may also be useful paths for future research. These types of approaches will prove invaluable for the research community by providing an easy, direct link to biologically relevant information that exists within published gene expression data.

Chapter 3

Visualization methods for statistical analysis of microarray clusters

3.1 Introduction

Recent high-throughput and whole-genome experimental methods create new challenges in data analysis and visualization. Gene expression and protein microarrays output hundreds of thousands of data points that can be used for prediction of gene function over the entire genome. However, there are serious and fundamental challenges in the analysis of these data. Microarray data contain substantial experimental noise and as our knowledge of biology is incomplete, no perfect gold standard exists for verification of microarray analysis methods.

In order to determine gene/protein relationships and functions from microarray data, methods must be robust to noise and must identify groups of genes that may be functionally related. Statistical methods, such as clustering, attempt to identify data patterns and group genes together based on various distance metrics and algorithms. The lack of a true gold standard makes it impossible to verify the absolute accuracy of any clustering method. Several statistical approaches have been presented for assessing cluster quality [20, 46, 57, 101], but these are all either internal validation methods or methods that rely on incomplete external standards such as MIPS [56] or Gene Ontology [4] functional protein classifications. Further, these methods do not address the

issue of identifying specific problems within clusters of microarray profiles or assessing the relationships between clusters of genes. Well designed visualization methods are capable of aiding in these tasks by helping to bridge the gap between raw data and the analysis of that data [2]. To perform more comprehensive cluster analysis, statistically integrative, dynamic, noise-robust data visualizations are required to complement purely analytical evaluation methods.

Existing visualization tools do not include methods to statistically and dynamically evaluate clusterings of genes. Several tools can display expression data in various static ways suitable for publication [79] or provide useful dynamic views of tabular data [45], but are not specifically intended for cluster analysis. JavaTreeView [75] and the HierarchicalClusteringExplorer [78] dynamically display hierarchically clustered data for analysis and VxInsight [98] displays the result of a built-in clustering algorithm in an interactive 3D topology, but none are able to display results of other clustering methods for analysis. TreeMap [6] provides an innovative way to visualize hierarchically clustered data as well as data organized in the context of the GO hierarchy, but is not intended for cluster analysis. New tools such as GeneXplorer [72] provide an interactive method for visualization and analysis of microarray data on websites, but do not focus on the task of cluster analysis. Several tools, including the MultiExperimentViewer [73] and Genesis [88], provide multiple methods of performing clustering as well as some visualization methods to analyze the resulting clusters. Commercial tools, such as GeneSpring [28] and SpotFire [86], offer various statistical and

visualization tools for general analysis, but neither offer visual methods specific to analyzing the results of clustering algorithms. Therefore, there is a need for visualization-based methodologies designed specifically to statistically and dynamically evaluate clusters produced by the variety of available algorithms and software tools.

Here we present a suite of interactive microarray analysis methods that integrate relevant statistical information into visualizations for the purpose of assessing the quality and relationships of clusters in a noise-robust fashion. Our methodology is general and can be used to analyze the results of most clustering algorithms performed on either protein or gene expression microarray datasets.

3.2 Results and discussion

3.2.1 Noise robust visualization

Microarray data contain a substantial amount of noise; therefore, visualizations must facilitate tasks like pattern identification and outlier detection in a noise-robust fashion. Microarray data span a rather large and noisy numerical range, so traditional microarray visualizations use a cutoff value that specifies where maximum saturation occurs. While this is necessary in order to see variation around zero, it obscures variation in highly over or under expressed areas (Figure 3.1a-c). At a minimum this cutoff value should be dynamically controlled by the user so that they have the ability to see both types of variation. Several currently available tools include this ability, as does our method, but while the ability to change the cutoff value helps to increase dynamic range and decrease the effects of noise in visualizations, it fails to address the entire

problem. Traditional visualization methods essentially display the Euclidean distance between gene expression profiles, a measure that is not robust to outliers. Distance metrics more robust to noise, such as a rank-based Spearman correlation coefficient, can be used for numerical analysis of microarray data. We propose a rank-based visualization method to serve as the complement to these noise robust distance metrics (Figure 3.1d).

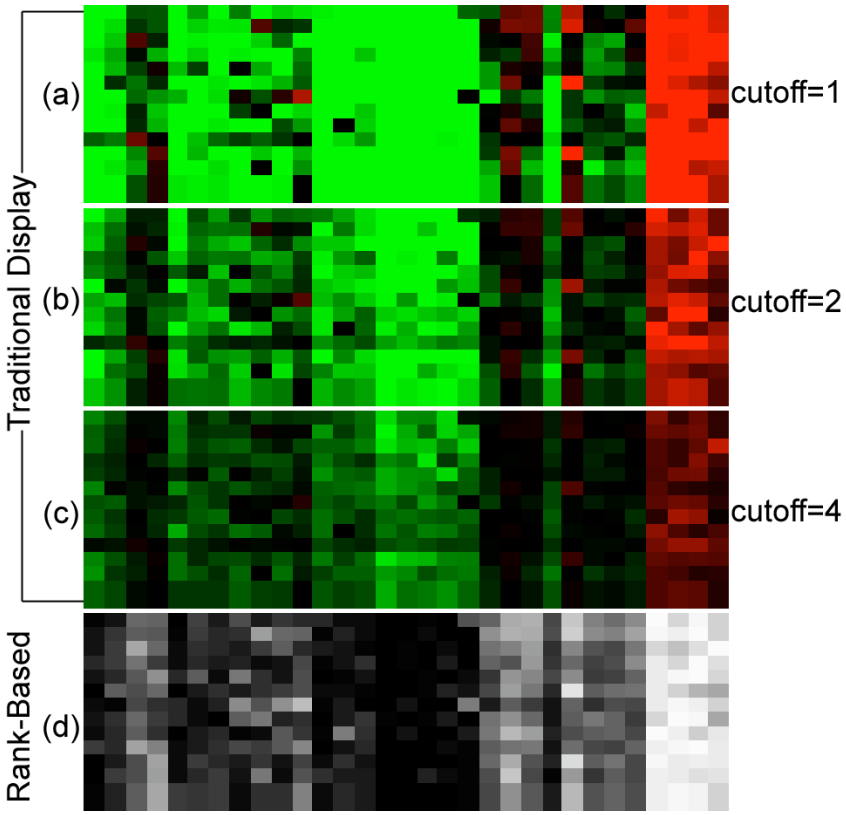


Figure 3.1: Example of noise in microarray visualization. Four views of the same data displayed in different ways. (a-c) show a traditional display using different cutoff values. Note that in (a) variation in the highly over and under expressed regions cannot be seen due to saturation, while in (c) variation in the highly expressed regions can be seen, but variation near zero cannot. (d) uses our rank-based visualization method. In this rank-based view (d), the experiment with the lowest expression for each gene is colored black, the experiment with the highest expression is colored white, and the other experiments interpolate between in grayscale. Using this method, users can see the overall pattern of variation in the data, which makes it clear that heterogeneity in the traditional view is mostly the result of noise. (Data from [26])

Our method performs a rank transform on each gene by sorting the gene's expression levels, then ranking the experiment for each gene with the lowest expression 0, the next lowest 1, and so on to the highest expression which is ranked *N-1*, where *N* is the number of experiments. Each experiment is then displayed as a grayscale percentage of rank/(*N-1)*. In this display, the experiment with lowest expression for each gene is colored black, the experiment with the highest expression is colored white, and the intermediate experiments gradate between them in shades of gray.

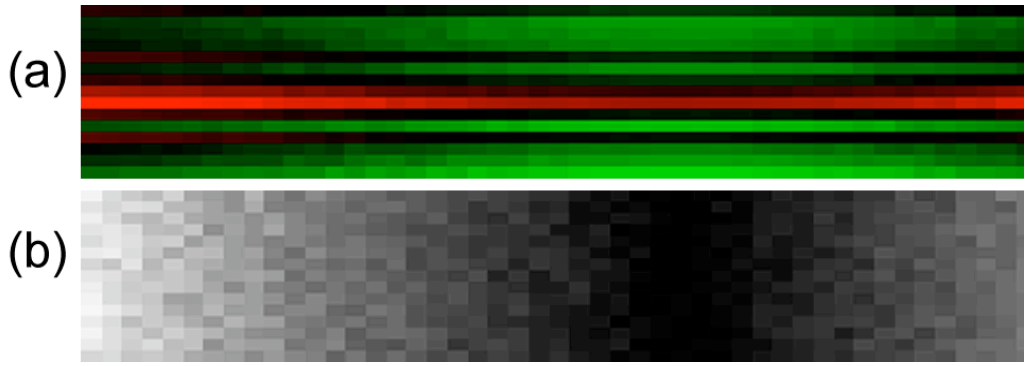


Figure 3.2: Rank-based visualization of synthetic data. Synthetic data displayed (a) traditionally and (b) using our rank-based method. This data was generated by creating a single sinusoidal expression profile and for each gene (row) randomly shifting that profile up or down and introducing small amounts of Gaussian random noise throughout. The result is that the genes generally follow the same shape/trend over experiments, but the shapes are shifted up/down from one another. Traditional view (a) masks the similarity between genes, but their relationship is clear in the rank-based view (b).

In addition to being more robust to noise, this rank-based visualization allows users to easily see patterns of shape/trend that are not apparent in traditional visualizations. Clustering algorithms that use a rank-based distance metric will group together genes based on their pattern of expression, which can result in clusters that look very non-uniform when traditionally displayed (Figure 3.2). However, in our rank-based visualization it is clear that these genes do

belong together because they share expression profiles with the same shape/trend.

While the example in Figure 3.2 is an extreme case, this rank-based visualization approach is useful in a variety of biological settings. For example, in many time series data sets it is useful to observe changes in expression over time in response to some process such as environmental changes, drug introduction, or cell cycle phase. In particular, a group of genes which all rise in expression over a period of samples in a cell cycle experiment, but whose absolute expression levels are not the same will appear heterogeneous when displayed traditionally. However, when displayed using our rank-based method, the pattern of expression is much clearer, which can aid users to identify biologically meaningful trends of expression (Figure 3.3). Genes exhibiting a coherent progression of shape/trend over time may be co-regulated. Thus, it is important to identify trends and not just examine similarities of absolute expression level.

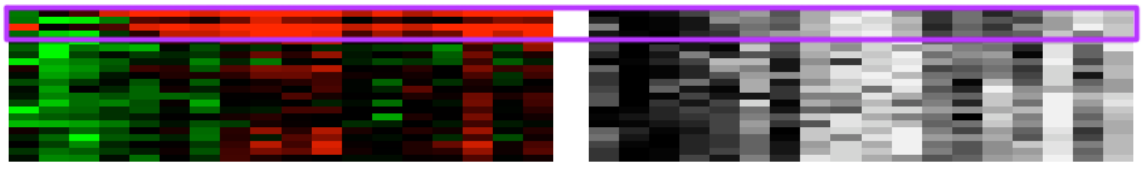


Figure 3.3: Rank-based visualization of time series data. Yeast cell cycle data displayed (a) traditionally and (b) using our rank-based method. In the traditional visualization the top 4 genes (within the purple box) appear to be very different from the rest of the genes in this cluster. However, using the rank-based method it becomes clear that these genes follow the same general pattern of the entire cluster, with initially low expression building up to highest expression in the central time points and then falling to roughly middle values. (Data from [85])

3.2.2 Assessing cluster quality

While multiple statistical methods have been developed for assessing the quality of clusters produced by different algorithms [20, 46, 57] the most appropriate clustering algorithm choice depends on the dataset, distance metric, and goal of the analysis [101]. Due to the limitations of these methods, it is important to effectively display clustered data in a manner that allows researchers to examine the variation and consistency of the results of different clustering algorithms. We propose two new visualization techniques that can be used to assess overall cluster quality, and also identify individual outliers and other anomalies in the data quickly and efficiently.

First, to analyze the overall cohesion of each cluster, we developed a "difference display" method. For each cluster, we display the cluster average bar to show the general expression of the cluster as a whole. We calculate the vector of the cluster average $\vec{\bar{g}}$ from the vectors of expression profiles of each gene, $\vec{g}_i$, for each cluster containing *M* genes with expressions measured over *N* experiments using the standard formula:

$$\vec{\bar{g}} = \frac{\sum_{i=1}^{M} \vec{g}_i}{M}$$

Each gene's expression is displayed as a difference, $\vec{d}_i$, from the cluster average, $\vec{\bar{g}}$:

$$\vec{d}_i = \vec{\bar{g}} - \vec{g}_i$$

Thus if a gene is shaded green in an experiment, it is expressed lower than the cluster average for this experiment, and if shaded red it is expressed more in an experiment than the cluster average for that experiment. In this visualization a cluster that is relatively dark is more uniform since the genes are generally close to the average (Figure 3.4a). Individual genes that differ from the average more than others will stand out as brighter than their neighbors, which allows for easy visual detection of outliers (Figure 3.4b). Thus, this visualization allows researcher to easily identify genes that do not fit well with the cluster's expression profile, and thus may be functionally distinct from the rest of the cluster.

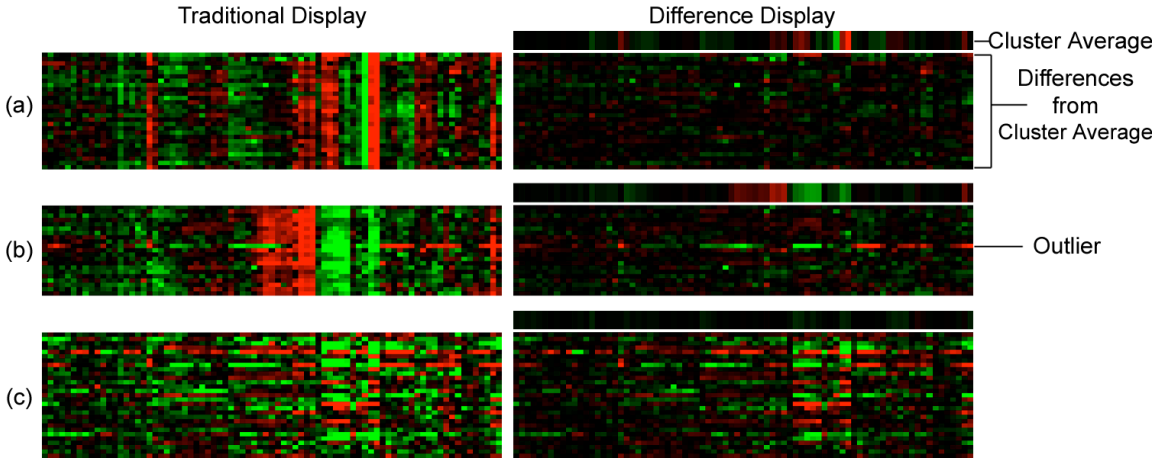


Figure 3.4: Difference display visualization. Three clusters displayed traditionally on the left and in our difference image visualization on the right. In the difference display, the large top bar on each cluster shows the cluster average, each gene is displayed as its difference from that average (green indicates expressed less than the cluster average, red shows more expressed, and black means equally expressed with the cluster average). Cluster (a) is a coherent cluster of genes and appears very dark because of its homogeneity. Cluster (b) is another dark, uniform cluster, but it also contains one randomly inserted gene, which can be easily identified in our difference display. Cluster (c) contains a random selection of genes, and its randomness is clear from the brightness of the difference display. This difference display allows for quick assessment of overall cluster homogeneity and facilitates quick outlier detection. (Data and clusters a & b from [23])

Second, in addition to assessing overall cluster quality and identifying gene outliers, it is important to look at variation of individual experiments within each cluster. We calculate the standard deviation, *s*, of each experiment, *j*, within a cluster in the normal manner:

$$s_j = \sqrt{\frac{\sum_{i=1}^{M} (\overline{g_j} - g_{i,j})^2}{M}}$$

Where *M* is the number of genes in the cluster, $\overline{g_j}$ is the cluster average for experiment *j*, and $g_{i,j}$ is the expression level of gene *i* in experiment *j*. We display the standard deviation of each experiment within the cluster below the cluster average bar. Here black indicates a standard deviation of zero and white indicates higher standard deviations, saturating at a user defined cutoff value. This allows a user to quickly identify high and low variation experiments on a per-cluster basis (Figure 3.5). High variation experiments may imply that the genes in this cluster were less related under those particular experimental conditions.

Visualizing clusters in this difference display method allows users to see variations in expression level that may be biologically significant that are not visible in traditional visualization methods. For example, the data shown in Figure 3.5 is the glycolysis cluster (2E) from [23]. When viewed traditionally this cluster appears very homogenous and consistent. However, when viewed as a difference from the cluster average, we can observe that in the region of highly under-expressed experiments some genes are more expressed than the average while others are less expressed than average (red and green boxes are shown in

this area). This suggests that the cluster could be split into two smaller clusters that would be even more homogenous.

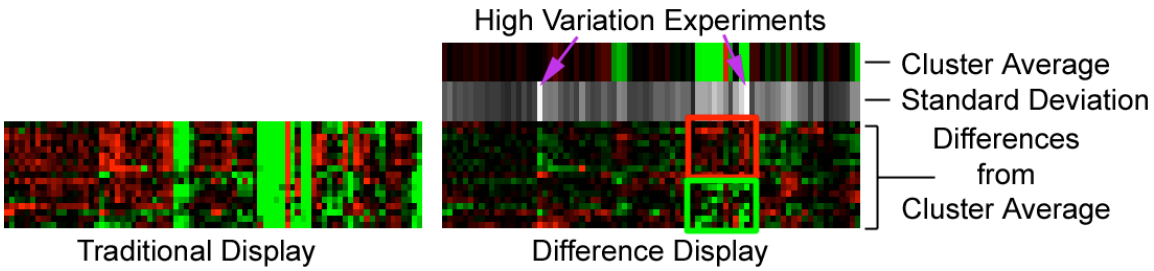


Figure 3.5: Experiment variation display. A cluster displayed traditionally on the left and in our difference image visualization on the right also showing the standard deviation within the cluster for each experiment. Black on the standard deviation bar indicates a standard deviation of zero, while white indicates a higher value. Purple arrows point to several experiments in this cluster that show high variance. In general, the high variance among some experiments may indicate that this cluster is unregulated under those conditions. In this example, we can inspect the differences from the cluster average in the high variance experiments and see that for these conditions the upper group of genes (indicated by a red box) is less under expressed than the lower group of genes (indicated by a green box) which suggests that the cluster could be split into two sub-clusters to reduce this variation as shown in Figure 3.6.

In this example 8 of the 9 genes indicated by the red box in Figure 3.6 are annotated to glycolysis (TPI1, GPM1, PGK1, TDH3, TDH2, ENO2, TDH1, and FBA1), but only 3 of the 8 genes indicated by the green box have this annotation (CDC19, TYE7, PFK1). The red grouping of genes is significantly enriched for the glycolysis biological process (p-value = 9×10^{-20}). However, the genes in the green box are significantly enriched for the more general process of alcohol metabolism (p-value = 1.7×10^{-11}) as 7 of these 8 genes are involved in this process (PDC5, PDC6, PDC1, CDC19, HXK2, TYE7, PFK1). Thus, there is a sound biological basis to draw a distinction between these two groups, but traditional visualization is unable to show this type of biologically meaningful variation in highly over or under expressed regions.

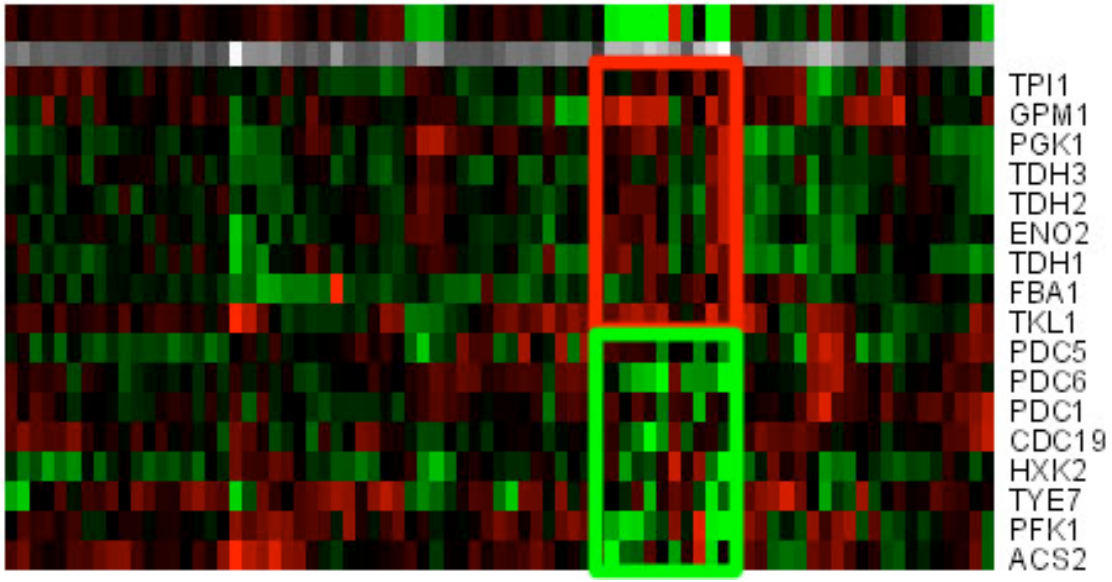


Figure 3.6: Experiment variation example detail. This figure shows further detail of the cluster seen in Figure 3.5. The genes in the red box are less under-expressed than the genes in the green box, which is evident in the difference display visualization. These groups are biologically different, as the red genes are best characterized as specifically related to glycolysis, while the genes in the green box are better characterized as more generally related to alcohol metabolism.

3.2.3 Assessing cluster relationships

In addition to assessing the quality of clusters produced by an algorithm, it is also important to understand how the clusters and genes in different clusters relate to each other. Clusters with similar overall expression profiles may functionally interact with one another. One method to show high level cluster-to-cluster relationships is to calculate a hierarchical clustering using only the averages of each cluster. We can then hierarchically arrange the cluster averages and display the dendrogram relating the averages to each other (Figure 3.7). As this method only creates a hierarchy for the cluster averages, rather than for individual genes as in the case of hierarchical clustering of the entire

dataset, it allows us to show cluster relationships for arbitrary clustering algorithms.

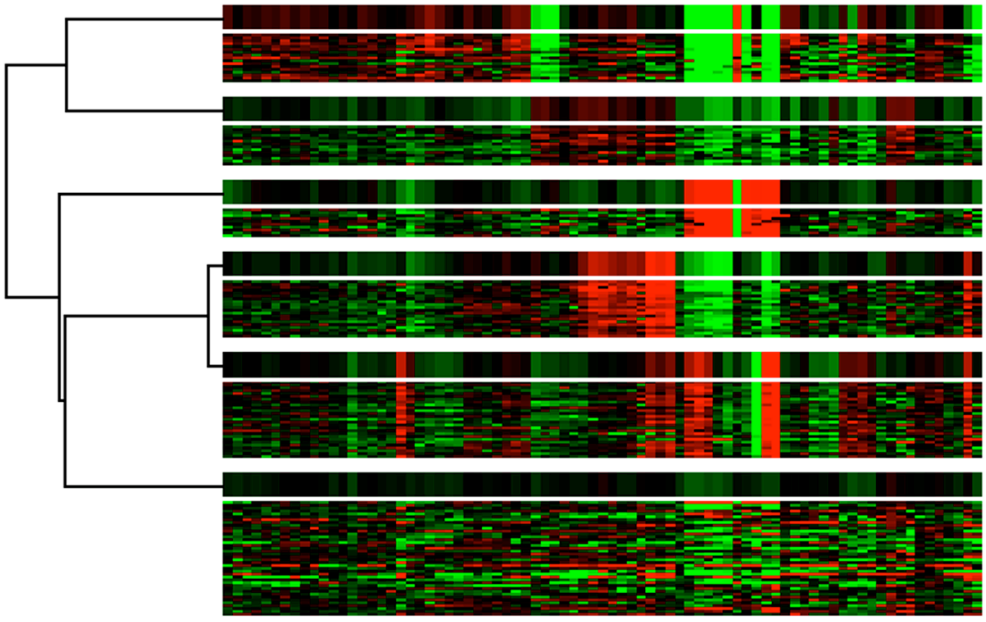


Figure 3.7: Dendrogram of averages. A dendrogram created from cluster averages with the genes in a cluster displayed below each average. The length of each branch of the tree is proportional to the distance between the averages. We create the hierarchy from the cluster averages, which allows us to show high level relationships between clusters generated by arbitrary clustering algorithms. (Data and clusters from [23])

However, this dendrogram of averages fails to show the relationships between genes in different clusters. It is important to examine gene-to-gene and gene-to-cluster relationships to assess whether or not genes are included in the most appropriate cluster. In order to view the lower level relationships among genes in clusters we can project high dimensional microarray data into a lower dimensional space such that genes with similar expression profiles are spatially closer to each other than genes with different expression profiles. We use Principal Component Analysis (PCA) to define the axes of a three-dimensional space to project the genes and clusters onto. PCA has been used previously in

microarray data analysis for dimensionality reduction to facilitate easier analysis and comparisons [20, 71] and to identify patterns of noise [1]. Our method is interactive and navigable which allows users to examine individual genes and view relationships between clusters as they separate out spatially.

To perform PCA on the microarray datasets, we use Singular Value Decomposition (SVD). SVD decomposes an $m \times n$ matrix of the full microarray data, *X*, into three additional matrices:

$$X_{m \times n} = U_{m \times n} \Sigma_{n \times n} V^T_{n \times n}$$

Where *M* is the number of genes and corresponds to rows of the matrix, and *N* in the number of experimental conditions and corresponds to the columns of the matrix. We use the eigengenes, or Principal Components (PCs), defined in the rows of V^T as the axes for our PCA visualization. The position of each gene in that space is determined by the corresponding column of $U\Sigma$. The square of the singular values, contained on the diagonal of Σ, correspond to the variance included by each PC such that the percent of variation, *p*, captured by the k^{th} PC is determined by:

$$p_k = \frac{\sigma_k^{\,2}}{\sum_{i=1}^{M} \sigma_i^{\,2}}$$

In this formulation, the singular values are in decreasing order, meaning that the first PC includes more variation than the second, and so on. Thus, using the top 3 PCs includes the most variation possible in a three dimensional projection. We would expect that well-formed clusters would separate out the

most when using the top PCs as the axes of projection. However, in some data sets the top PCs are not the most appropriate space for projection. For example, in the Spellman *et al.* cell cycle data set [85] using our tool we can see that the first PC does not show the "banded" pattern typical of ordered cell cycle data, which the second, third, and fourth PCs do display (Figure 3.8a). Accordingly, a projection into the first two PCs does not separate out cell cycle regulated genes/clusters spatially (Figure 3.8b).

This is consistent with previous PCA analysis done by Alter *et al.* [1], which identified the first PC of this data as highly correlated to noise rather than meaningful information. Our method allows the user to dynamically specify which PCs define each axis, which allows exploration of which PCs are most appropriate for analysis and identification of potential noise-correlated patterns in the data. In the case of Spellman *et al.* cell cycle data, we can use the 2^{nd}, 3^{rd}, and 4^{th} PCs for projection, which leads to much better spatial separation (Figure 3.8c). In this projection, we can see that each phase of the cell cycle spatially separates in temporal order around the origin and that the G1 and M phases appear opposite each other, which is consistent with the underlying patterns of expression for cell cycle genes. Our projection of genes and clusters into a space defined by user selected PCs allows the user to view and analyze relationships on both a cluster-to-cluster basis and a gene-to-gene basis.

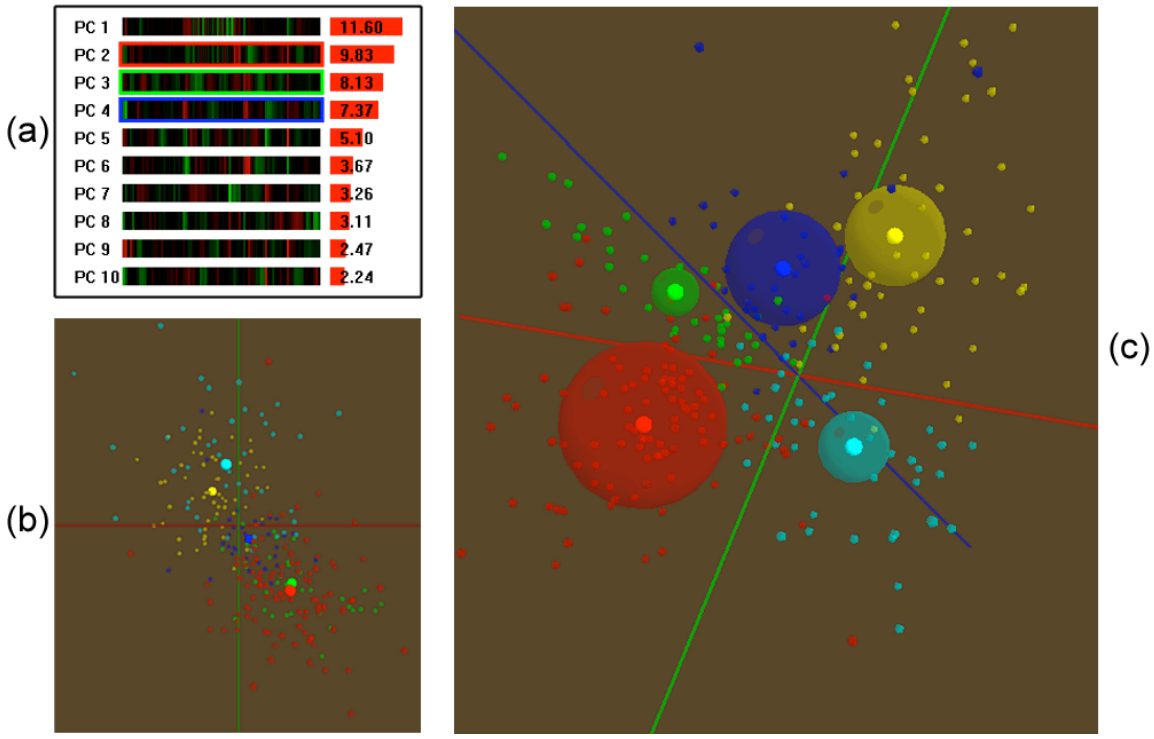


Figure 3.8: Principal component projection visualization. A projection of genes from a cell cycle data set into a 3D space defined by user selected Principal Components. Genes in each cluster are colored by phase (Red-G1, Green-S, Blue-G2, Yellow-M, and Cyan-M/G1). Cluster averages are displayed by larger solid spheres. The much larger transparent spheres show the region included by one standard deviation away from the average. (a) shows the top ten PCs of this data set and the percent of variance accounted for by each PC. (b) is a projection of cell cycle genes onto a space defined by the 1st and 2nd PCs. The separation is poor due to the first PC being highly correlated to noise in this data set. (c) shows the same data projected into a space defined by the 2nd, 3rd, and 4th PCs. These PCs are highlighted in (a) corresponding to the axis colors in (c). Notice that the cell cycle phases are separated in order around the origin, and that G1 and M phase genes are opposite each other, which is consistent with their opposing expression profiles. (Data and clusters from [85]).

3.2.4 Multiple simultaneous views and scaleable architecture

In our system each of the visualizations described above are dynamically linked to each other, so that selections, colorations, etc. are shared among views. This allows users to perform tasks in conjunction with one another. For example, using the difference image visualization and the PC projection, users

can assess the quality of a clustering as well as the relationship between clusters very easily (Figure 3.9).

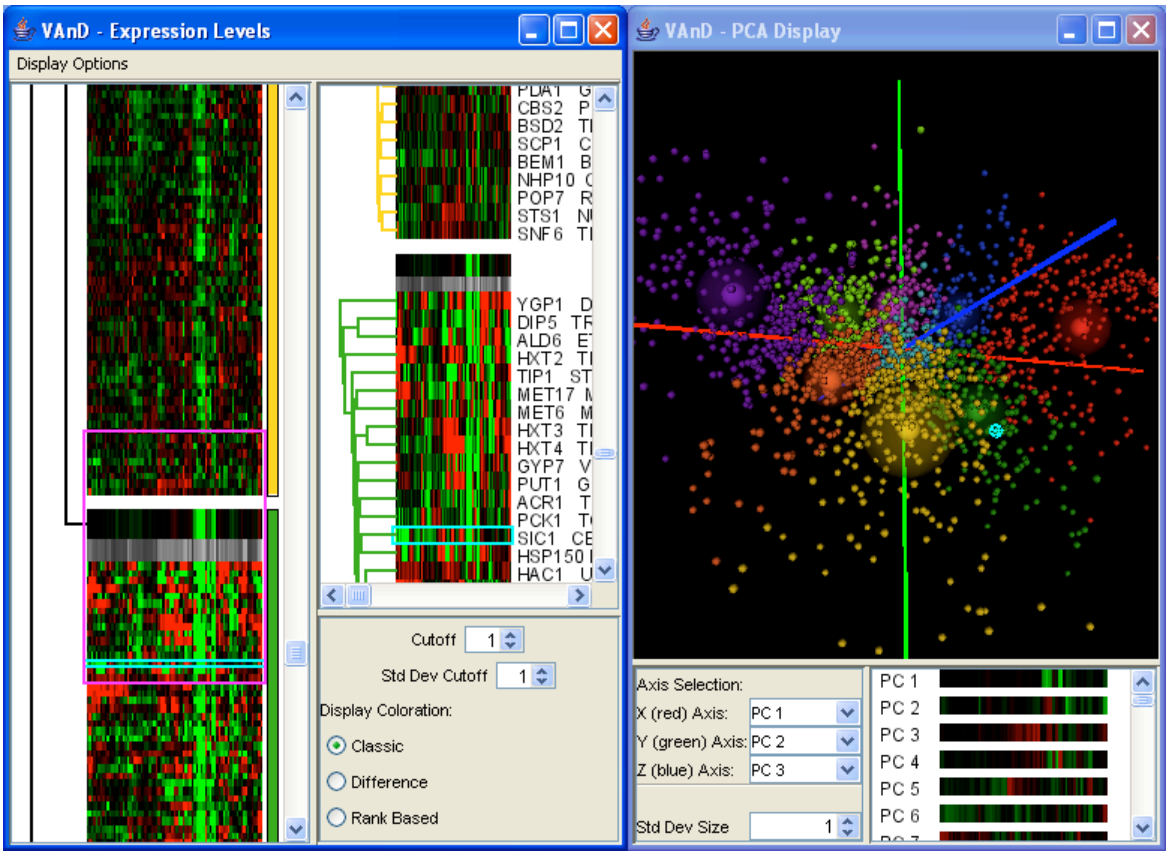


Figure 3.9: Multiple simultaneous views. A screenshot of GeneVAnD displaying clustered data. The panels shown are the expression level window on the left, which can toggle between traditional, difference, and rank-based displays, and the PC projection window on the right. A selected gene is highlighted in blue in all views.

Our implementation of these methods is both modular and scalable. Although all of the visualizations share a common data structure for dynamic linking, each visualization is displayed in its own panel, allowing for easy addition or removal of new visualization components. Each of the panels is fully scalable for use on both desktop/laptop size displays as well as large display walls. The ability to use these visualizations on large, high-resolution displays facilitates

collaboration among researchers and allows users to view greater portions of their datasets simultaneously (Figure 3.10).

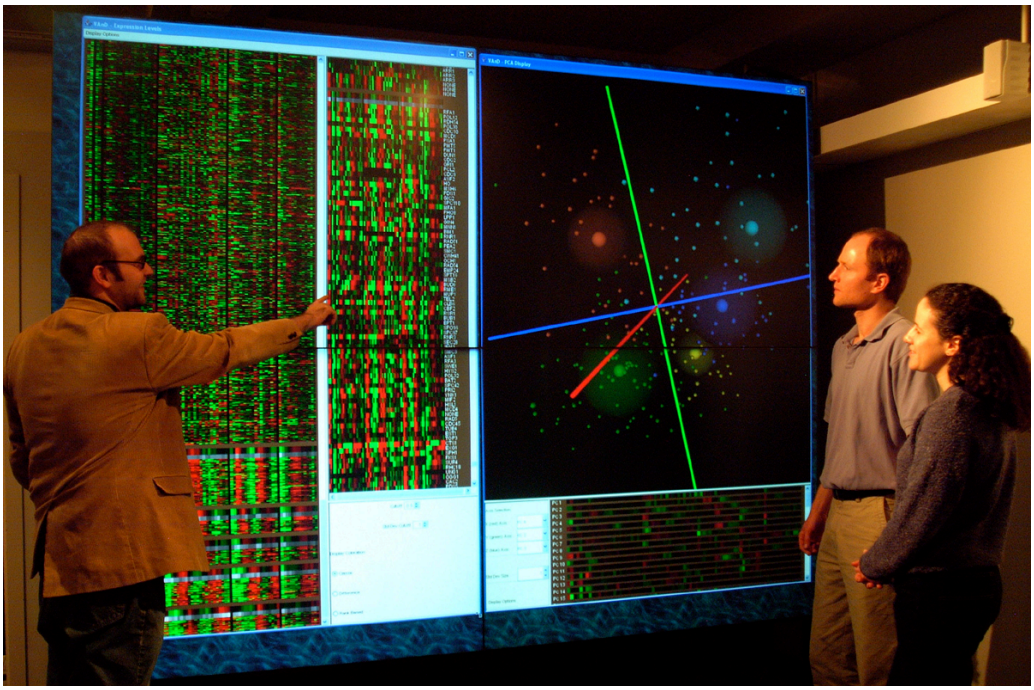


Figure 3.10: Large scale display. GeneVAnD in use on a large-scale display wall. The high resolution enables display of more information simultaneously and the large scale creates an environment conducive for collaboration between multiple researchers.

3.3 Implementation

Our methodology has been implemented in GeneVAnD (Genomic Visual Analysis of Datasets). GeneVAnD is written in Java and is cross platform for use on Windows, Linux/Unix, and Macintosh operating systems. We use Java3D [41] to display the PC projections and Piccolo [10] to display the expression profiles. The JAva MAtrix Library (JAMA) [43] is used to perform the SVD calculation. The package is designed in a modular way to allow future extensions and inclusion of additional information and visualizations. The executables and

source code of GeneVAnD can be found at http://function.princeton.edu /GeneVAnD.

3.4 Conclusions

Statistical clustering of microarray data is vital for identifying groups of genes that may be functionally related. However the high level of noise in microarray data and the lack of a gold-standard for comparison deeply complicate the evaluation of clustering algorithms. Here we have presented a set of visualization methods geared specifically toward evaluating clustering of microarray datasets. Our rank-based method allows for more noise-robust visualizations of expression levels, our difference display method facilitates visual assessments of general cluster quality as well as outlier detection, and our PC projection method allows for visual assessments of cluster relationships. Our methodology integrates meaningful statistics into an interactive and noise-robust data visualization package for use in analyzing the results of clustering algorithms. Through several examples we have demonstrated the effectiveness of these methods to aid researchers in the analysis of the results of clustering algorithms by facilitating noise-robust assessments of cluster quality and cluster relationships. We believe that more statistically integrative and targeted visualization methods can benefit not only cluster analysis, but also many other important data analysis problems in genomics.

Chapter 4

Viewing the Larger Context of Genomic Data through Horizontal Integration

4.1 Introduction

Scientifically meaningful data visualization is vital for the advancement of knowledge in many fields, particularly molecular biology. Genomics is one of the fastest growing modern scientific disciplines, as it promises a better understanding of the inner workings of cells, is vital to understand diseases, elaborates our understanding of evolution, moves towards the era of personalized medicine, and reveals the root causes of cancer. One of the most powerful new tools molecular biologists wield to solve these problems are gene expression microarrays, and the majority of microarray analysis is done through visualization techniques[23, 70].

Gene expression microarrays simultaneously measure the activation or suppression of every gene in a genome at a particular point in time. These studies result in data matrices containing hundreds of thousands to millions of observations, and the majority of researchers rely on visualization tools to mine these data to discover new biological information. Biologists face the challenges of understanding not only the data that they generate, but also of comprehending their results in the broader context of previous studies. As microarray technology matures, decreases in cost, and becomes more accessible, the number of

microarray studies produced is growing exponentially, which further complicates thorough analysis.

No existing method for microarray visualization enables researchers to directly understand and analyze their data within the greater context of previously published findings. This severely limits research capabilities by forcing users to focus on their own data during the initial analysis phase and to compare with other studies only at later stages to confirm or contradict their conclusions. Integrating the vast amount of available data into the analysis phase as early and seamlessly as possible will allow researchers to build upon previous results, observe inconsistencies, and form more powerful conclusions.

We propose a novel methodology for the analysis and exploration of multiple microarray datasets simultaneously. By leveraging visual paradigms that are commonly used for small-scale microarray analysis, our approach remains easily interpretable by researchers. Due to the sheer size of these datasets, we employ an “overview + detail” approach on a per-dataset basis to allow users to view specific genes as well as their context within the whole genome. However, we extend this paradigm to include the larger context of additional available datasets as well, which we call an “overview + detail + setting” paradigm.

We have implemented our approach into a system called HIDRA (Horizontally Integrated Dataset Relationship Analysis), and we have deployed this system to experimental genomics researchers both for individual use and for collaborative use on a large-format display device. In the next section we

discuss existing microarray visualization approaches in more detail. We then outline our specific visualization goals and what techniques we use to achieve those goals. And finally we show two case studies where meaningful biological observations have been made by researchers using our system.

4.2 Related Work

Existing microarray visualization tools focus on the analysis of single datasets, and many of these tools are used on a daily basis by the research community[76]. The majority of visual displays of microarray data fall into two major categories: heat maps [73, 75, 78] and parallel coordinates[34]. Other approaches are also used, such as scatterplots, histograms, and spreadsheets, but these are generally complementary techniques used in conjunction with a heat map and/or parallel coordinate display[28, 31, 73, 86].

Heat map displays traditionally show a clustered data matrix of values represented as colors interpolated from red to green. This type of display allows a user to quickly identify prevalent patterns among genes in a dataset by looking for bands of data with similar profiles. These displays are often accompanied by a dendrogram created from hierarchical clustering, which dictates the order in which genes are displayed and visually encodes a distance metric relationship between genes.

Heat maps have seen near universal adoption amongst biologists, and their results are the canonical representation of gene expression used in the majority of microarray publications. While these displays allow the full matrix of data values to be viewed, the patterns and labels of individual genes are only

visible by zooming into more detailed portions of the map. Many tools support this type of exploration by using an "overview + detail" paradigm[7], where users see the entire dataset, but can then select a smaller region to see in greater detail.

Parallel coordinate systems display genes as a collection of segmented lines overlaid on a measurement grid. These displays have the ability to show all of the available data in a relatively small area. This approach is also well suited for the identification of desired patterns, as users are able to select only those genes that pass through defined portions of the grid.

While parallel coordinate views show all of the available data, the results can be difficult to interpret. When viewing a large number of genes simultaneously, it is difficult to distinguish one expression profile from another. As with heat maps, this approach suffers from not being able to label individual genes within the total plot. The absence of a dendrogram created from hierarchical clustering presents both benefits and complications. The dendrogram visually indicates a quantitative distance metric between two genes in a dataset, but it also enforces an ordering and structure on the data that may be somewhat artificial. Parallel coordinate displays do not suffer from this imposition of ordering, but do not visually quantify arbitrary distances between profiles.

Many of the most successful microarray visualization approaches combine both heat map displays and parallel coordinates views, along with several other views of the same data[76]. We refer to these approaches as "vertically

integrated" as they allow researchers to see the same data from many complementary angles. These methods have been very successful and have gained wide use among the microarray analysis community.

We propose extending the power of multiple simultaneous views in an orthogonal direction. Rather than displaying multiple viewpoints of the same data, our approach displays the same type of viewpoint on multiple datasets at the same time -- we refer to this paradigm as a "horizontally integrated" approach. This expansion of the amount of visualized data enables researchers to view a broader setting of known biology and place their own results within this larger context.

4.3 Design & Implementation

We established several goals for the design of our microarray visualization methodology that incorporates broader context. The following goals are a combination of our initial aspirations and the desires of our research collaborators that used our system:

- *Ease of use*. A successful system must be usable and intuitive for the target audience, in this case biology researchers.
- *Dynamic, consistent interaction*. The approach must be adaptive to user input as their desire to explore and observe information changes over time, but these adaptations must feel natural to the user.
- *Scalability*. Our approach must scale both with the amount of data visualized, and with available screen space.

- *Biologically meaningful*. Perhaps the most important criterion is that a microarray visualization system must enable researchers to explore their data in a way that facilitates biological observations and insights.

4.3.1 Single dataset visualization

In order to maintain a baseline of usability and comfort with the microarray analysis community, we have chosen to adopt the use of heat maps accompanied by dendrograms as the basis for our methodology. This approach is by far the most common presentation format for microarray data in biology literature. For individual dataset display we leveraged the codebase of the commonly used, open-source tool, JavaTreeView[75], which we then modified for our purposes. This provides the immediate advantage of utilizing pre-existing abilities and biases of the microarray research community. On the level of a single dataset we also utilize the "overview + detail" paradigm to allow users to view both the entire dataset, as well as a more detailed view of a subset of that data. An example of this visualization for a single dataset is shown in Figure 4.1.

Users have several options for interacting with this display of information. Subsets of genes to view in the detail portion can be selected by dragging a box on the heat map, or by choosing branches of the dendrogram. These selections can be refined by traversing up or down the dendrogram using the keyboard. This allows users to isolate particular desired areas of the larger dataset to view with greater scrutiny.

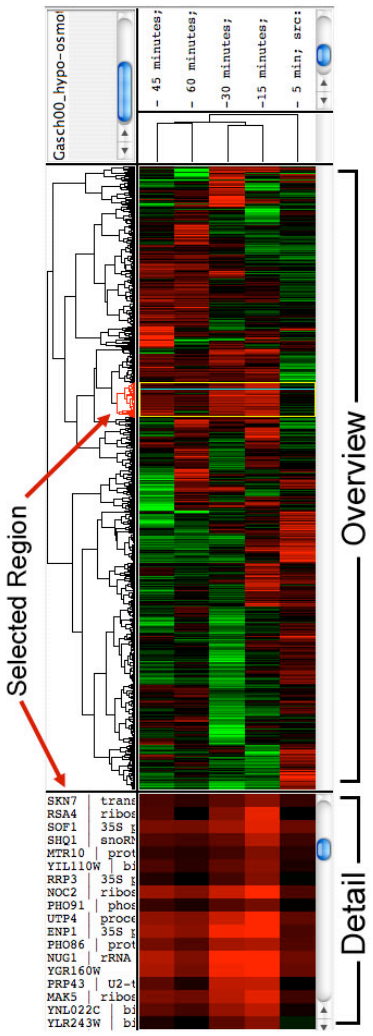


Figure 4.1: Individual dataset display. A single dataset displayed using a heat map and dendrogram in an overview + detail format. Rows correspond to genes and columns to experimental conditions. Each intersection is colored on a continuous scale from green through black to red. The data was hierarchically clustered in both dimensions. A region of the dataset was selected in the overview and the corresponding section is shown in greater detail below.

Due to differences in experimental technologies and personal preferences/abilities, it is also important for users to maintain control regarding the parameters of the heat map coloration. In general, microarray data lies in a broad, noisy range of values that depends on several laboratory factors. For this reason, values above/below a cutoff are saturated out to a maximum intensity, but this cutoff is not universal, and should thus default to a reasonable value, but

be in the control of the user. Additionally, the color scheme used for display must be adjustable by the user. The red/green gradient is commonly employed because it has a direct link to the chemical dyes used in microarray experiments, however such a scheme is clearly unacceptable for color-blind researchers.

4.3.2 Multiple dataset visualization

Several factors are important to consider when incorporating additional datasets into microarray visualization. The common features of microarray datasets are genes, while the experimental conditions vary between datasets, which indicates that between dataset comparisons should be visible on a per-gene basis. However, microarray datasets are often created using disparate technologies or experimental practices, and individual datasets are generally targeted to investigate a specific area or process, which indicates that information such as clustering and normalization are appropriate only on a per-dataset basis.

In order to address these biological requirements, we have developed an approach we refer to as “overview + detail + setting”. On the level of each dataset it is vital to observe both the entire dataset (overview) as well as more specific information (detail). However, for the larger goal of placing an individual researcher’s data in the greater context of available data, datasets must be linked together (setting). In particular, we applied this approach to microarray data with the goal of making comparisons between datasets as intuitive as possible, while maintaining important per-dataset information.

The most common paradigm in microarray literature is for the expression of genes to correspond to rows of a visualized data matrix. As genes are the common element of interest between datasets, we place the datasets next to each other horizontally to preserve gene-row orientation across all data. However, the ordering of genes is determined by clustering, and the clustering process is biologically meaningful on the level of individual datasets. To address these issues, we have synchronized the detail views across all datasets to facilitate comparisons, while preserving the cluster order of individual datasets in the overviews.

By synchronizing the detail views, we preserve the expectation that gene measurements are aligned along rows, even across multiple datasets. The order of the genes shown in the detail views corresponds to the order of those genes in the dataset where the selection was made. To provide information about the per-dataset context of the selected genes, a thin line is displayed in each dataset's overview to indicate where each selected gene falls within that dataset. An example of this multiple dataset visualization is shown in Figure 4.2. The gene-level synchronization of the detail views enables low-level comparisons of a gene's specific behavior in different datasets; while in the overviews, the selection highlights indicate a higher-level comparison of gene group relationships between datasets.

For example, a user can select a tight group of genes in one dataset, and immediately observe how those genes cluster together in every other dataset at a general level. A researcher can then examine the detail views to investigate

the specific expression levels that led to the observed global patterns. This type of exploratory analysis across a large amount of diverse datasets is impossible with existing tools, but is vital for experimental microarray analysis, as we demonstrate in our validation.

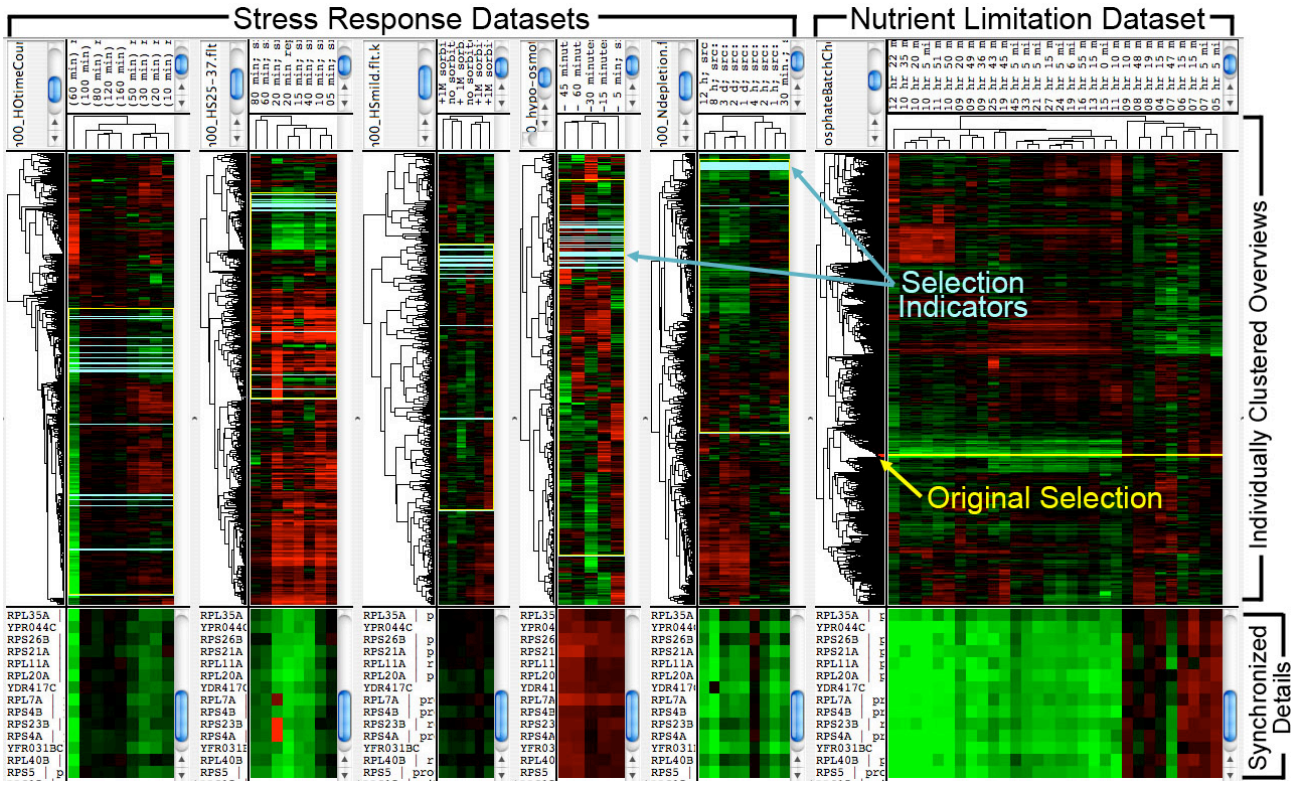


Figure 4.2: Multiple disparate datasets viewed in HIDRA. Six different datasets are shown here tiled horizontally. Each dataset was individually hierarchically clustered in both dimensions. A selection has been made in the rightmost dataset (from a nutrient limitation study [74]) and the thin light blue lines in the left five datasets (from a stress response study [27]) indicate where these genes are located in their overviews. A user can quickly observe that the selected genes are non-randomly grouped in the clustering of the other datasets. Further inspection of the aligned genes in the detail views shows cases where these genes are behaving similarly/differently.

4.3.3 Scalability, interactions, and interfaces

The inclusion of multiple datasets also requires addressing scalability, interaction and user interface concerns[64]. First, as more data is viewed

simultaneously, screen space quickly becomes an issue. While several datasets can be viewed at once on even the smallest desktop/laptop displays, users may be in situations where they still feel limited. By default, when enough datasets are loaded to overflow the available display space, a scrollbar becomes active to pan between datasets. We also provide the ability to dynamically re-order, remove, and/or add new datasets as the researcher explores their data. In this manner users can choose the most relevant datasets to occupy the visible area as their needs change over time.

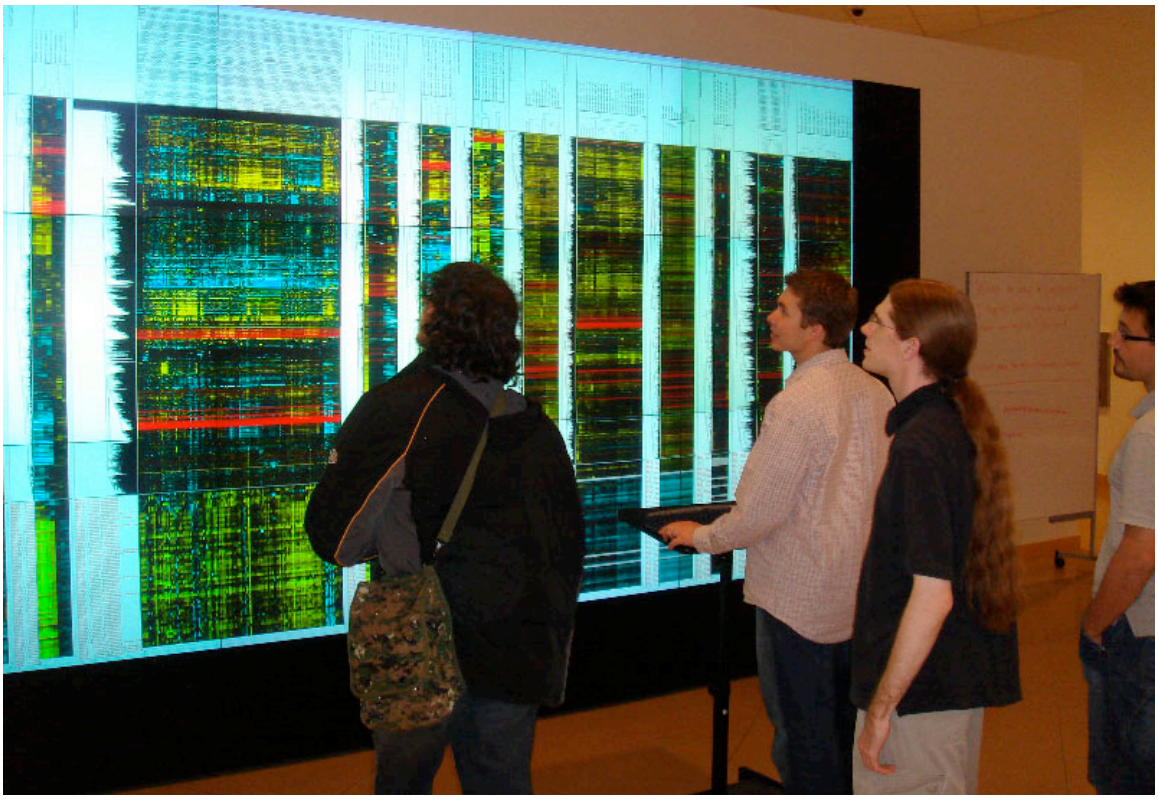


Figure 4.3: Scalability of HIDRA. A group of collaborators are using HIDRA on the large-scale display wall at the Lewis-Sigler Institute for Integrative Genomics. This display is capable of simultaneously showing an order of magnitude more data than traditional desktop/laptop displays, which is helpful when dealing with very large data repositories.

Another option to see more data is to move to large-scale display devices if they are available to the user[53, 96, 97]. Our approach scales very well to large-format devices by providing control over text size, column widths, row heights, etc (Figure 4.3). Using displays of this magnitude allows users to see as much as an order of magnitude more data at once. These very high-resolution displays are also helpful for collaboration, which is very common among microarray analysts.

Regarding the user interface, several visualization choices must be made on a per-dataset basis. In particular, the desired color scale, saturation cutoffs, dendrogram widths, etc. often vary greatly from one dataset to another, due to technological and experimental differences. We provide controls to alter all of these parameters for any selected subset of datasets, including the individual level. Further, we store these choices on a per-dataset basis, so that as users re-order, remove, and/or re-load data these per-dataset choices remain intact.

Further, some interactions should be consistent from one dataset to the next. In order to preserve the gene-row alignment of the detail views, the heights of each panel are slaved to any panel being resized, such that all detail views maintain the same height. Additionally, scrolling in any detail view causes synchronous scrolls in all detail views to maintain consistency.

4.3.4 Implementation

We have implemented these methods into a Java-based system called HIDRA. The use of Java as a development language allows us to more easily produce a cross-platform result, which is of particular importance to the biology

community, who use a variety of operating system platforms. Among our immediate collaborators, individuals use Windows, Macintosh, and Linux operating systems to perform their analysis. The Java language also easily permits future expansion of our approach to include additional features, which is vital as genomic research is rapidly evolving.

4.4 Validation

The ultimate validation of a scientific data visualization approach is its usefulness and adoption within the research community. In particular, a successful approach should aid in the discovery of novel biology. We are working with many collaborators spread across five laboratories to assess how our multiple dataset visualization approach aids their research as well as how to improve HIDRA. We have deployed our system for these users both on their own desktop/laptop machines and on the large-scale shared display wall at the Lewis-Sigler Institute for Integrative Genomics at Princeton. While we are still receiving feedback from these users, here we discuss two of the user experiences that led to biological insights made using our approach. These examples demonstrate the power of our technique as these observations could not be easily made using any previously existing methodology.

4.4.1 User experience #1 – Stress response effects in yeast

One scientist using our system is interested in studying stress response and growth rate effects in yeast. By utilizing our multi-dataset visualization capabilities applied to several existing datasets, she was able to draw several novel, biologically meaningful conclusions. She was able to simultaneously

examine the expression levels of genes in a set of standard stress response datasets[27] as well as results from a nutrient limitation study[74] and a collection of gene knockout experiments[35]. The biological question this user wished to examine is whether or not the traditional global stress response signal is present in other types of data.

Using our approach, she was able to easily find and select clusters of genes in the nutrient limitation and knockout studies that she suspected may be the result of a stress response effect, and then examine how those genes related to each other within the standard collection of stress datasets (see Figure 4.2). Performing this type of analysis is simple in our multiple dataset approach; however, using previously existing techniques we would need to launch over a dozen independent instances of a program and continually cut and paste selections between instances, rendering such analysis practically impossible.

Our collaborator identified several groups of genes in these datasets that exhibited a strong pattern of correlation within the stress response datasets as well. This suggests that the effect on gene expression of various nutrient limitations and gene knockouts may be superceded by the more general stress response effect. Our collaborators are currently performing further analysis, both in the lab and with our visualization system to better characterize this phenomenon. Thus, by observing the relationships between these very different datasets in HIDRA, this scientist quickly identified unexpected commonalities that may prove biologically interesting.

4.4.2 User experience #2 – Cell cycle synchronization effects

A second example of an important observation was made by another biologist using HIDRA to investigate disparities among related datasets. In this case the scientist was examining several datasets all purportedly studying the same phenomenon, the yeast cell cycle. In particular, two studies used a variety of means to synchronize cell populations to create time courses of gene expression throughout the phases of the cell cycle[19, 85].

A group of genes in one of these time courses were tightly clustered with very high over-expression at early points of the time course. However, using HIDRA we could quickly see that these genes were largely unrelated in the other time courses, and during the early time points they were not over-expressed in the datasets produced from other means of synchronization (see Figure 4.4).

Upon further inspection, a significant number of these genes are known to be involved in cell conjugation and mating. The time course where these genes are tightly clustered was synchronized by exposing the cell population to a pheromone that induces a mating response, which halts cell cycle progression. Our collaborator quickly realized that the expression response seen in these early time points was an artifact of the synchronization method, rather than a change caused by the cell cycle. In this case, observing differences between datasets studying the same phenomenon helped focus efforts on important portions of the datasets.

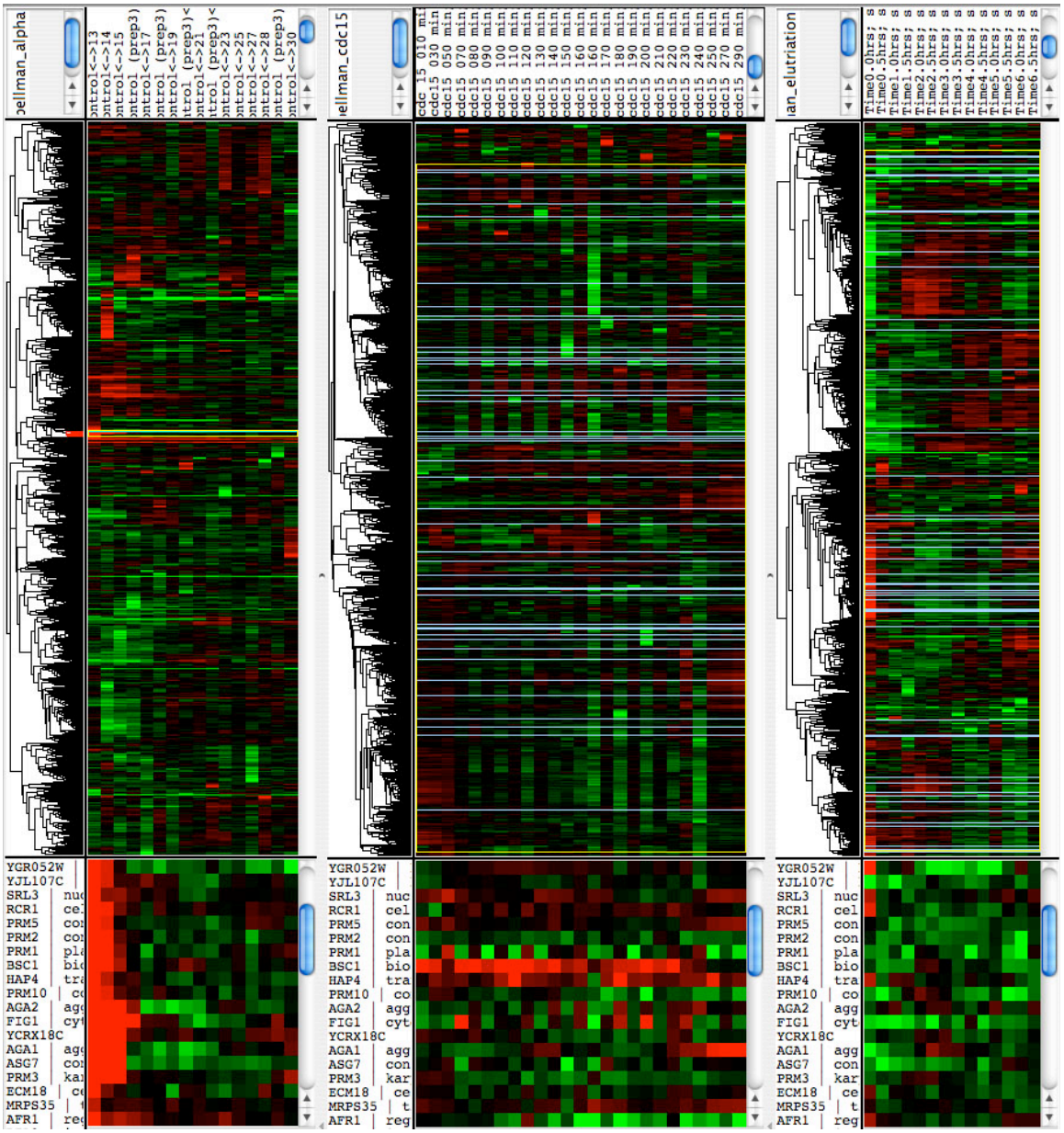


Figure 4.4: Exploration of differences between multiple similar datasets. In this case, three time courses studying the same phenomenon (the yeast cell cycle) from two studies [19, 85] are shown. A group of genes with very high over-expression at early time points is selected in the leftmost dataset, but these genes show little relationship to one another in the other two time courses. Further study revealed that the over-expression of these genes in the left dataset was an experimental artifact.

4.4.3 Discussion

The two examples described above are representative of the types of interactions users have had with HIDRA. By quickly observing commonalities among disparate datasets collaborators have been able to identify common trends that could indicate meaningful relationships between experimental conditions. Conversely, by finding key differences between related datasets users can explore phenomena unique to particular assays. This type of exploration allows microarray researchers to quickly make key insights and form hypotheses that would be difficult to make viewing the data independently.

4.5 Conclusions

We have presented a novel methodology for the concurrent analysis of multiple gene expression microarray datasets. Our approach allows researchers to understand how their own data relates to data previously published in the literature, which is vital for continued analysis. By exploring the larger context of available data, users can overcome the limitations of existing approaches for higher-level analysis of their own data. Observing a more global view of expression data allows biologists to make more insights and formulate novel hypotheses.

Our approach to the inclusion of greater data context is based on expanding common visualization practices to create an “overview + detail + setting” system. We include the concept of the greater information setting by horizontally integrating and linking separate overview and detail views for individual datasets. This type of data integration – inclusion of multiple parallel

views – is in contrast to the integration of a variety of viewpoints based on the same underlying data, which we call a vertically integrated approach.

Although we apply the concepts of including a broader setting of information through horizontal integration to a specific solution for microarray visualization, these principles are much more general. For example, a system similar to HIDRA for microarray analysis could be created based on parallel coordinates, rather than heat maps. Horizontally incorporating additional datasets into a system based on vertically integrated multiple views could potentially provide both the benefits of more complete understanding of single datasets and the benefits of understanding the greater information context.

The concept of visualizing the broader setting of available data is vital for future analysis and comparisons within the biology community. We have shown real-world examples of insights that can be made using our approach for microarray visualization that are difficult or impossible to discover using existing techniques. We believe integrating additional datasets into visualization systems is a powerful paradigm not only for genomics data, but potentially for many other disciplines as well.

Chapter 5

A Platform for Integrated, Scalable Analysis and Visualization of Gene Expression Microarray Data Compendia

5.1 Introduction

The previous three chapters of this dissertation describe algorithms and approaches that can be utilized by biologists to aid their own research by identifying the portions of available microarray data that are relevant to their area of interest (chapter 2), by incorporating statistical choices made during microarray analysis into visualization schemes (chapter 3), and by providing a simultaneous, interactive view of multiple datasets to explore the broader context of other available data. These methods are very complimentary, and combined together they create an excellent platform for meaningful biological analysis and visualization of large gene expression data compendia.

For example, a researcher interested in studying breast cancer treatment may have performed laboratory work to collect gene expression time points while exposing a cell line to a chemotherapy drug. They will likely cluster this dataset using a distance metric such as centered Pearson correlation to perform their initial analysis within a visualization system. Currently available tools, including integrated analysis and visualization packages[28, 73, 79, 86], end their applicability to this type of research at this point, as they are only designed to

work with one dataset at a time. However, the next natural questions for this study revolve around comparing the transcriptional response of this particular chemotherapy treatment time course with other studies to find important commonalities and differences.

To this end, a researcher could select a cluster of interest from their own data and perform a SPELL similarity search (as described in chapter 2) to identify additional datasets where these genes are behaving in a similar manner [32]. It would then be useful to simultaneously view these related datasets in a visualization system such HIDRA (described in chapter 4) to explore the relationships between these data [33]. Most likely, these related datasets will be independently clustered for visualization and the principles used in GeneVAnD (described in chapter 3) could be used to ensure that the data is shown in a statistically truthful fashion [31]. Further, additional common analysis capabilities, such as gene function enrichment analysis could be performed.

In order facilitate this type of comprehensive data analysis and visualization, we have integrated each of the techniques discussed in the previous chapters, as well as other approaches, into a single microarray analysis platform, called bioHIDRA (Biological Holistically Integrated Dataset Relationship Analysis). The remainder of this chapter describes this software platform, and outlines real-world examples where this system can be used to explore novel biology.

5.2 The integrated platform

Beginning with the visualization capabilities of the HIDRA system, we have incorporated additional analysis functionality to provide users with the tools needed to broadly explore large gene expression compendia. The HIDRA system incorporates hierarchically clustered views of multiple datasets aligned using an "overview + detail + context" paradigm. Each dataset is shown in its entirety within an overview panel, while a selection of genes can be viewed in more detail in a zoom panel. Datasets are tiled next to each other, and the gene selections and detail panels are synchronized between all datasets to display the greater context that can be seen by exploring multiple datasets simultaneously (Figure 5.1).

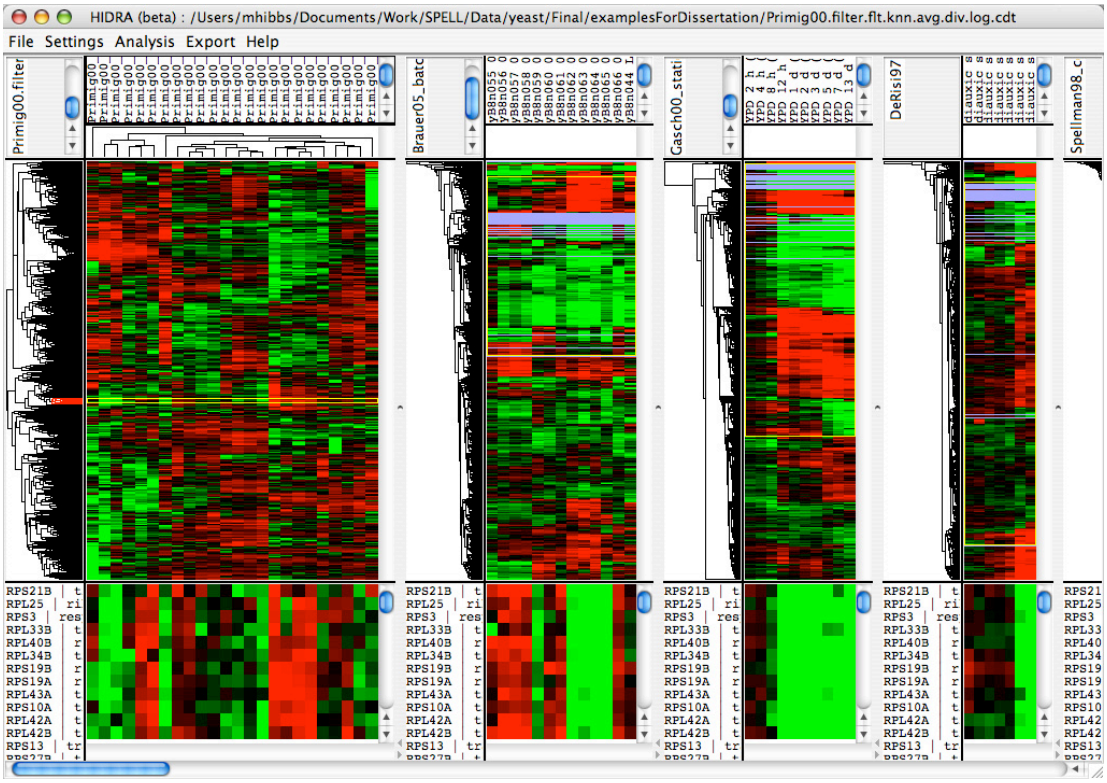


Figure 5.1: Screenshot of the bioHIDRA system. Based on the visualization paradigm described in chapter 4, multiple datasets are shown tiled next to each other. Selections and iterations are synchronized across all datasets.

5.2.1 Finding relevant datasets and genes with SPELL

While HIDRA has the ability to comfortably display between 6 and 10 average sized datasets simultaneously on a typical laptop or desktop monitor, many more datasets exist in the public domain. For *S. cerevisiae*, we have collected roughly 120 datasets spanning 2400 experimental conditions, while for human we have collected nearly 600 datasets spanning more than 14,000 conditions. Thus, given the limitations of available screen space, it is important to optimally choose which of the available datasets are worth exploring.

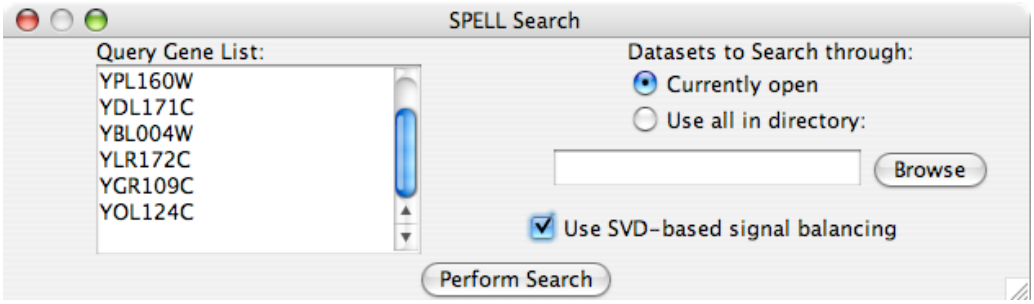


Figure 5.2: The SPELL search dialog. Users can specify a set of query genes either by manually entering the gene names, or by selecting a cluster from an open dataset. A search can be performed either in open datasets, or in all datasets in a directory. Also, SVD-based signal balancing can be employed if desired.

The SPELL similarity search algorithm has the ability to identify datasets relevant to a set of query genes very quickly and efficiently. We have incorporated this search algorithm into this software platform with two different modes: searching through an entire compendium, or searching through only selected datasets (Figure 5.2). In either case, a user provides a set of query genes, either by manually specifying the genes, or by selecting a cluster of genes within an open dataset. When searching through a large compendium, the software then loads each of the datasets of the compendia into memory,

calculates the SPELL dataset relevance weight for each dataset, and displays the results back to the user in a list. The user can then select which datasets they would like to explore, and then those datasets are displayed in decreasing order of the calculated relevance weight.

The process is similar when searching through only selected datasets, except that the relevance weight calculations are limited to those datasets that have already been opened. While the SPELL search algorithm runs very quickly, loading several hundred datasets into memory can take quite some time, so this option is more appropriate when a researcher knows beforehand which datasets are generally relevant to their area of interest. Further, the SPELL relevance weights can be used as a robust measure of cluster coherence. While a user can select a set of genes in one cluster and observe their expression patterns and visually assess their co-clustering in other datasets, using the SPELL search algorithm assigns a statistical measure of co-expression among the query genes for each dataset.

In addition to identifying and selecting relevant datasets from large compendia, the SPELL similarity search analysis option can be used to identify additional genes related to a query set. This can be useful in several ways, both in the context of individual datasets and large compendia. When performing a search through a single dataset, a SPELL search is essentially performing a targeted clustering seeded with a set of query genes. Whereas most clustering algorithms start from scratch to group co-expressed genes together, this process allows a user to effectively force a set of genes to co-cluster, and then expands

that cluster to include additional closely related expression profiles. When searching through a large compendium, users gain the benefit of the additional signal available in the collection at large to refine their search results and locate genes related to the query set.

In either case, after performing a SPELL search, the user is presented with a list of all genes in the genome, ranked by search score from most related to the query set to least related to the query set. The user can select which of these genes they would like to see in more detail, then those genes are selected in all of the open datasets. As the detail views shown for each dataset are synchronized by selection, the expression profiles for the selected genes are easily viewable for all open datasets at once.

5.2.2 Assessing functional enrichments among clusters

We have also included the ability to quickly perform Gene Ontology (GO) enrichment analysis from within this software platform. Given a set of co-clustered genes it is common practice to determine if the cluster is over-represented for genes involved in a particular process or molecular function, or localized to the same cellular compartment. To this end, GO term enrichment analysis is often performed using available tools[14, 18, 77]. However, none of these tools are integrated into a microarray visualization platform.

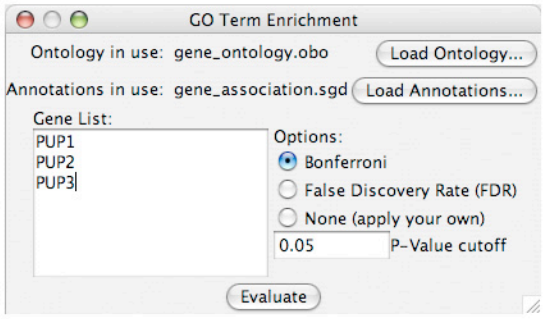
We have incorporated GO term enrichment analysis into this software platform based on the hypergeometric distribution [14]. Given a set of n query genes where k of these genes belong a GO term of interest, in an organism with

N genes where *M* of those genes are annotated a that GO term, a p-value for the significance of this overlap can be calculated using the summation

$$p-value = \sum_{x=k}^{n} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

We calculate this statistic for all GO terms, and then correct these p-values for multiple hypothesis testing using either a very conservative Bonferroni correction [11] or a less stringent false discovery rate (FDR) correction [11].

(a)



(b)

GO Term Enrichment Results

Branch	Term Name	Term ID	% Query	% Total	P-value	Genes
cellular_component	proteasome core complex (sensu Eukary...	GO:0005839	3 of 3	15 of 6474	5.23E-7	PUP3,PUP2,PUP1
cellular_component	proteasome complex (sensu Eukaryota)	GO:0000502	3 of 3	47 of 6474	1.87E-5	PUP3,PUP2,PUP1
molecular_function	endopeptidase activity	GO:0004175	3 of 3	57 of 6474	3.37E-5	PUP3,PUP2,PUP1
cellular_component	proteasome core complex, beta-subunit...	GO:0019774	2 of 3	7 of 6474	1.56E-4	PUP3,PUP1
biological_process	modification-dependent protein cataboli...	GO:0019941	3 of 3	145 of 6474	5.72E-4	PUP3,PUP2,PUP1
biological_process	ubiquitin-dependent protein catabolic p...	GO:0006511	3 of 3	145 of 6474	5.72E-4	PUP3,PUP2,PUP1
biological_process	proteolysis involved in cellular protein ca...	GO:0051603	3 of 3	147 of 6474	5.97E-4	PUP3,PUP2,PUP1
biological_process	modification-dependent macromolecule ...	GO:0043632	3 of 3	152 of 6474	6.60E-4	PUP3,PUP2,PUP1
molecular_function	peptidase activity	GO:0008233	3 of 3	156 of 6474	7.14E-4	PUP3,PUP2,PUP1
biological_process	cellular protein catabolic process	GO:0044257	3 of 3	156 of 6474	7.14E-4	PUP3,PUP2,PUP1
biological_process	protein catabolic process	GO:0030163	3 of 3	170 of 6474	9.25E-4	PUP3,PUP2,PUP1
biological_process	proteolysis	GO:0006508	3 of 3	174 of 6474	9.93E-4	PUP3,PUP2,PUP1
biological_process	biopolymer catabolic process	GO:0043285	3 of 3	269 of 6474	3.69E-3	PUP3,PUP2,PUP1
biological_process	cellular macromolecule catabolic process	GO:0044265	3 of 3	303 of 6474	5.28E-3	PUP3,PUP2,PUP1
biological_process	macromolecule catabolic process	GO:0009057	3 of 3	335 of 6474	7.14E-3	PUP3,PUP2,PUP1
biological_process	cellular catabolic process	GO:0044248	3 of 3	412 of 6474	1.33E-2	PUP3,PUP2,PUP1
biological_process	catabolic process	GO:0009056	3 of 3	426 of 6474	1.47E-2	PUP3,PUP2,PUP1

Figure 5.3: GO term enrichment dialogs. (a) Go term enrichment can be performed by either selected a cluster of genes from an open dataset, or by specifying genes manually. Several options are available for multiple hypothesis correction. (b) Results of the enrichment analysis are displayed in a table.

This software platform provides an option to perform GO term enrichment analysis either by selecting a cluster from an open dataset, or by providing a list of genes manually. The full table of statistically significant results is displayed to

the user (Figure 5.3). Performing this type of analysis can very quickly direct researchers to interesting phenomenon within datasets, as demonstrated in the next section.

5.3 Example usage scenarios

5.3.1 Sporulation specific expression effects

Beginning with the Primig *et al.* [68] sporulation time course dataset, we can use bioHIDRA to identify signals unique to this dataset as well as signals common between this time course and other available datasets. While the functional coverage analysis presented in chapter 2 indicated that this dataset contained a largely unique signal specific to the processes of meiosis and sporulation, using bioHIDRA allows us to directly identify this specific transcriptional pattern. We can select the most closely related cluster by choosing the section of the dendrogram with the shortest branches (Figure 5.4a). Performing GO term enrichment analysis on this selection reveals that this group of genes is strongly enriched for processes such as ‘spore wall assembly,’ ‘sporulation,’ and ‘reproductive process’ which is consistent with the goal of the study that produced this dataset (Figure 5.4b). We can then perform a SPELL search using this set of genes as the query to identify other datasets within our compendia where these genes are also co-expressed. The results of this search indicate that this cluster is largely unique to the Primig dataset, as no other dataset achieves a very high SPELL relevance score (Figure 5.4c).

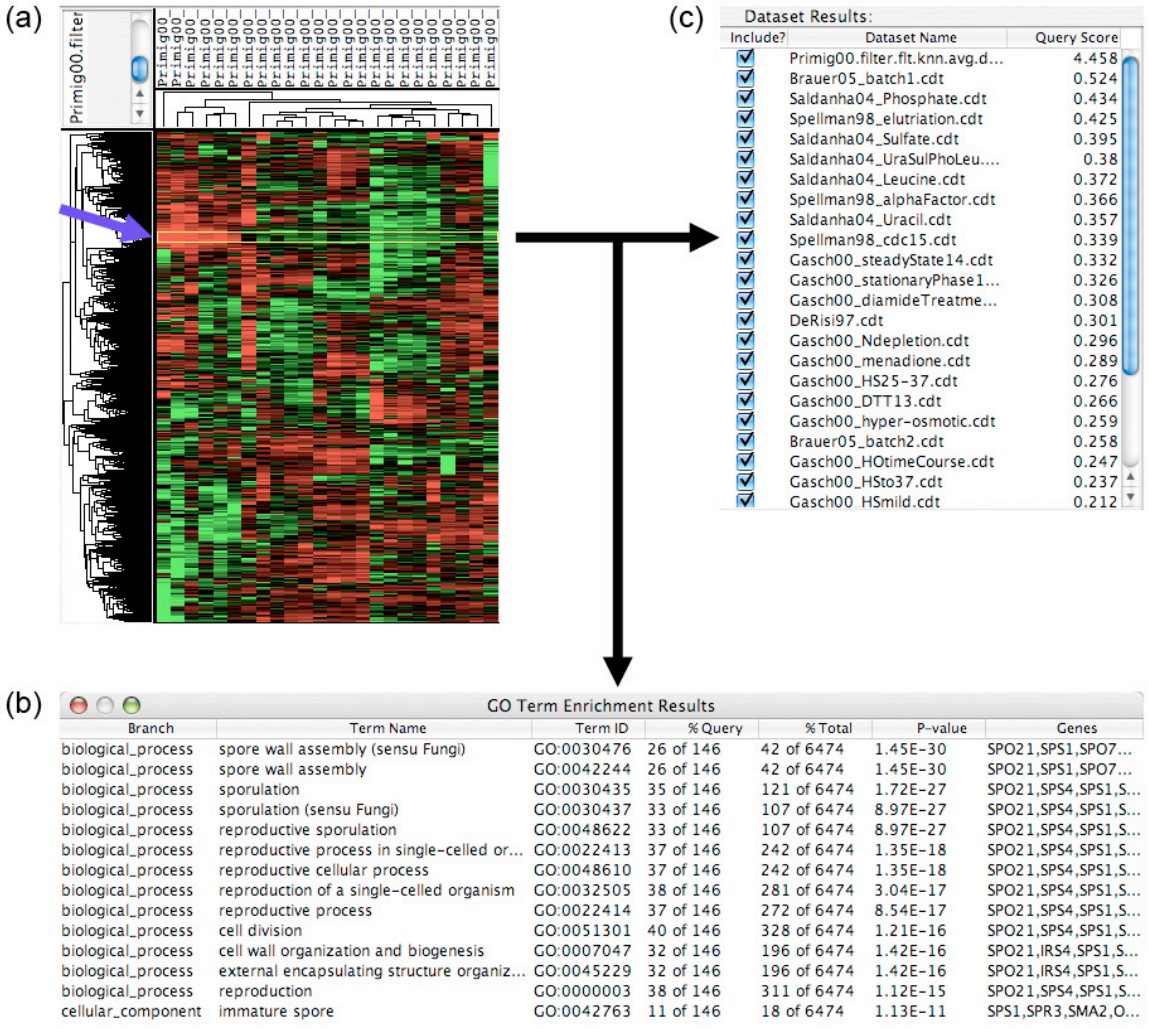


Figure 5.4: Unique sporulation signal example. (a) A coherent cluster was selected from a sporulation time course. (b) GO term enrichment performed on this cluster shows that these genes are significantly enriched for processes related to sporulation and meiosis. (c) A SPELL search using these genes reveals that no other dataset induces a high level of co-expression among these genes.

While this sporulation signal is specific to the Primig dataset, other signals are common between this dataset and many others. We can select the next most tightly clustered group of genes and perform a similar analysis process (Figure 5.5a). In this case these genes are strongly enriched for GO terms such as 'cytosolic ribosome,' 'ribonucleoprotein complex,' and 'translation.' As such, this cluster appears to be generally related to ribosomal processes in yeast

(Figure 5.5b). Performing a SPELL search, we can see that many other datasets within this selection also induce strong co-expression between this group of genes (Figure 5.5c), which is consistent with our knowledge of global ribosomal translation responses from the functional coverage analysis in chapter 2. Thus, beginning with just one dataset, we were able to quickly identify important specific and global signals and characterize their biological meanings.

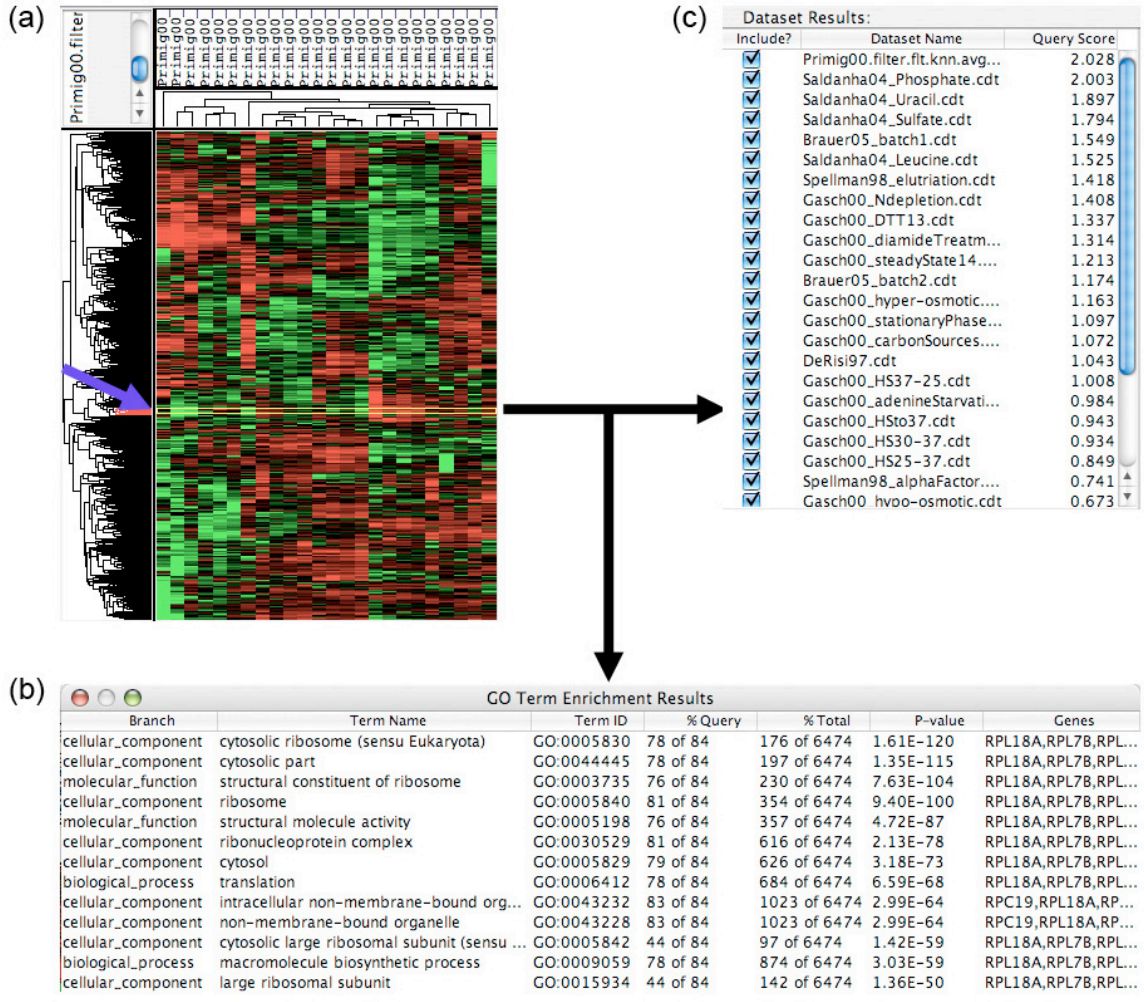


Figure 5.5: General ribosomal signal example. (a) Another coherent cluster was selected from the same dataset as in Figure 5.4. (b) GO term enrichment analysis shows that this cluster is mostly comprised of ribosomal genes. (c) A SPELL search using these genes shows that a large number of datasets induce co-expression of this group of genes.

5.3.2 Expression diversity among stress response studies

As a second demonstration of the bioHIDRA software platform, we examined several datasets produced by the Gasch *et al.* [27] stress response study. This investigation perturbed yeast cultures in a wide variety of ways, including various levels of heat shock, exposure to drugs, deprivation of nutrients, etc. In the original study, the authors observed a large class of genes with a common transcriptional response across nearly all of the perturbations examined. They refer to this collection of genes as the "generalized stress response" (GSR) group. By examining this data using bioHIDRA we can identify not only members of the GSR, but also signals that more specific to only some perturbations.

We started by loading 17 of the perturbations from this study and began to select groups of genes, assess their enrichment for GO terms, and identify other perturbations where the genes were also co-expressed. Many of the strongest clusters in these datasets belong to the GSR group, including many ribosomal proteins and general stress response proteins. However, we were able to quickly identify an interesting group of highly over-expressed genes within the hydrogen peroxide (H_2O_2) exposure time course (Figure 5.6a). Beginning with this small set of genes, we performed a SPELL search and found that the other perturbations where these genes were strongly co-expressed were the menadione and diamide exposure time courses (Figure 5.6c). All three of these compounds are known causes of oxidative stress [83], and the set of genes resulting from the SPELL search were significantly enriched for GO oxidation

related terms such as ‘response to oxidative stress,’ ‘antioxidant activity,’ and ‘cell redox homeostasis’ (Figure 5.6b).

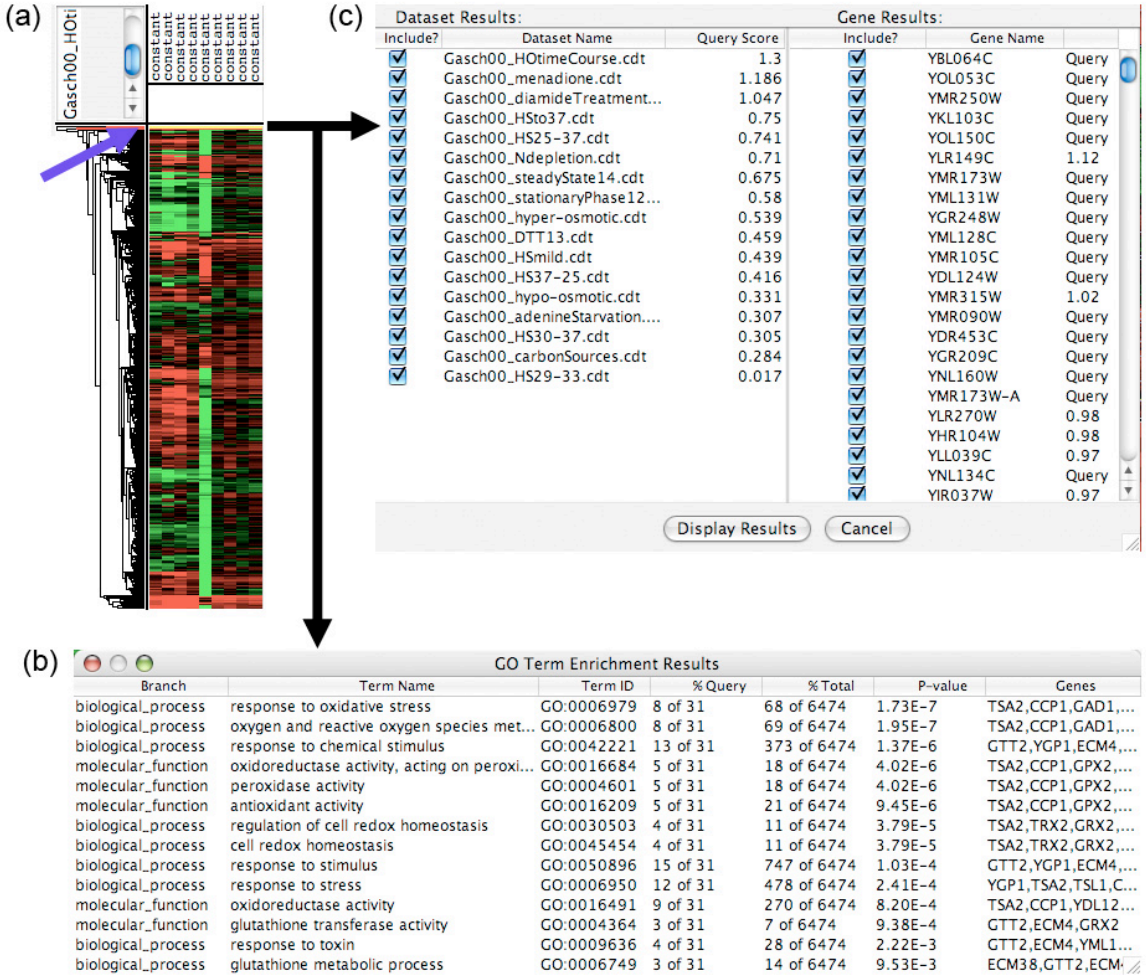


Figure 5.6: Oxidative stress effects among many perturbations. (a) A cluster of very highly expressed genes was selected in a hydrogen peroxide exposure time course. (c) A SPELL search using these genes as a query found that these genes were also co-expressed in other oxidative stress conditions. (b) GO term enrichment performed on the genes resulting from the SPELL search are enriched for areas related to cellular redox.

For each of these three perturbations the genes found by the search are highly over-expressed. While the remaining perturbations did not induce the same level of co-expression, we can still examine the expression patterns of these genes within the other datasets. Interestingly, between the two hyper- and

hypo- osmotic datasets we can observe very different patterns of expression for this set of genes. In the hyper-osmotic case the genes are also over-expressed (though not to the same degree as in the oxidative stress conditions), but in the hypo-osmotic case these genes are severely under-expressed.

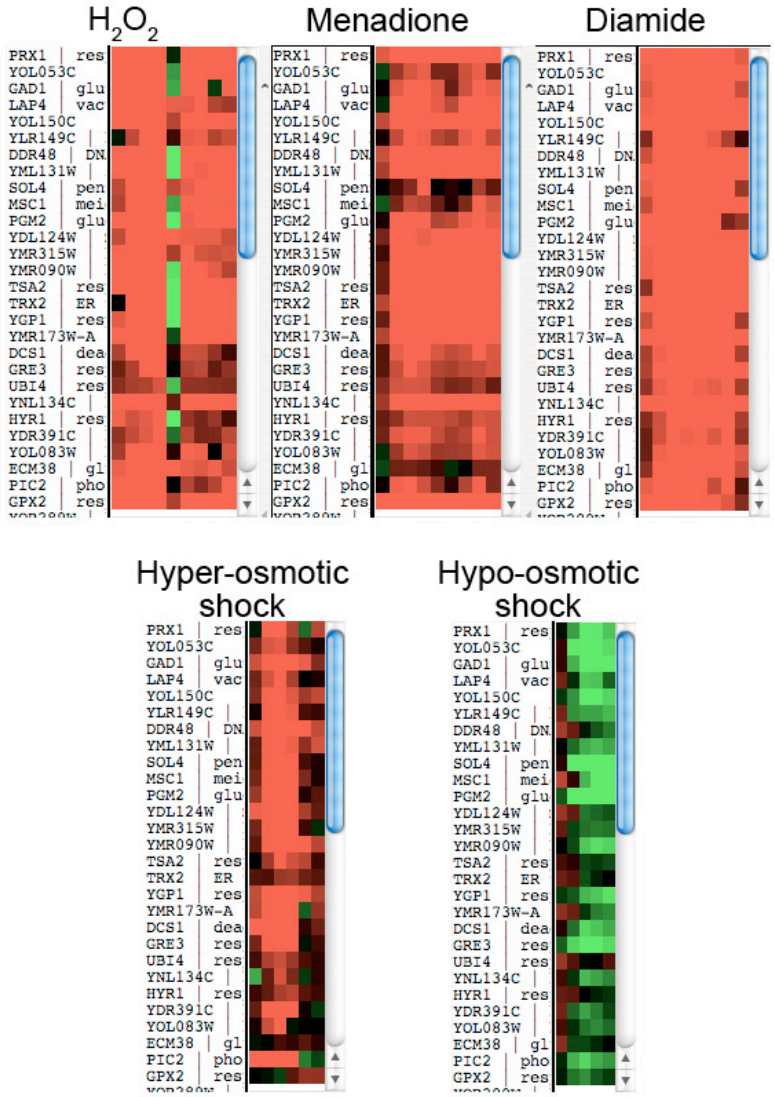


Figure 5.7: Oxidative and osmotic stresses. Based on a SPELL search performed as in Figure 5.6, the resulting genes are highly over-expressed in the three datasets receiving the largest relevance weight (shown on top). By examining other stress perturbations we can see that these genes are also over-expressed in hyper-osmotic stress conditions, and under-expressed in hypo-osmotic stress conditions. This observation is consistent with recent publications indicating that oxidative stress and osmotic stress share important response pathway components.

This observation is also consistent with recent studies that have determined that several genes and regulatory pathways are shared between the response to oxidative stress and the response to hyper-osmotic shock [49]. As such, under hyper-osmotic conditions these genes related to oxidative stress are induced, while in hypo-osmotic conditions the transcription of these genes is repressed (Figure 5.7).

5.4 Conclusion

The integration of analysis and visualization methods for the exploration of large-scale microarray compendia is very valuable for biology researchers. By providing a single, unified platform for gene expression analysis it is much easier for investigators to compare their own findings with other available data to draw novel conclusions and form new hypotheses. While the examples discussed above find previously known information, the clusters observed also contain additional genes not known to be involved in these functions. Thus, these genes may be excellent targets for future studies.

As the majority of microarray analysis performed by biologists is done within the context of visualization tools, it is vital for these tools to be able to answer the questions that are important to these biologists. Incorporating the ability to view additional relevant datasets, discover genes with a common signal across many experiments, and evaluate the functional enrichment of clusters are excellent advances that enable researchers to gain a better understanding of transcriptional responses.

Chapter 6

A Computationally Driven System for Iterative Experimental Discovery of Novel Biology

6.1 Introduction

Machine learning and data mining techniques have been applied to a wealth of genome-scale data to produce meaningful predictions of gene/protein involvement in biological processes or pathways. As we attempt to discover novel biology in a wide range of organisms with limited experimental resources, these approaches have promised to direct investigations toward the most promising targets first, with the hope of greatly accelerating the discovery process [36, 47]. However, computational prediction methods and novel laboratory investigations remain largely disconnected. Perhaps as a result, the rate at which gene functions are characterized has not kept pace with the rate at which data are generated [67].

Surprisingly few studies of gene function have been performed on the basis of computational predictions, despite their great potential to inform and guide such investigations. While individual predictions have been confirmed in the lab, no studies have been performed on a large-scale that truly integrate computational and laboratory aspects to fully explore novel gene functions. This lack of follow-up has led to several concerns and challenges within the computational biology community. Foremost, it remains unproven how

effectively computational methods can be utilized by biologists in the context of their own research goals. It also remains unclear which classes of computational prediction methods are appropriate for specific laboratory use to find novel biological discoveries. Lastly, it is generally unknown how computational methods and laboratory testing should interact in order to best advance our knowledge of biology.

We have performed a large systematic study of gene/protein function predictions made by computational methods in order to address these concerns, to clarify the potential role of computation in biological laboratories, and to discover novel biology in a particular area. We have used an ensemble of computation methods, including both supervised and unsupervised techniques based on diverse underlying data, to predict genes/proteins involved in the process of mitochondrial organization and biogenesis in *S. cerevisiae*. Mitochondria and their mechanisms of proliferation and inheritance are an excellent area for this type of study for several reasons. Mitochondrial defects are implicated in a variety of human diseases [25, 87], including neurodegenerative disorders [5, 48] and muscular diseases [21]. Also, the biological mechanisms of mitochondrial organization are largely conserved from yeast through humans (60% of mitochondrial yeast genes have a human ortholog), and as many as 1 in 5 mitochondrial proteins are known to be involved in human disease [3, 21]. Finally, this process is understood well enough to provide a reasonable number of training examples for prediction methods, but still, it is thought that at least a quarter of the genes/proteins involved have not

yet been identified [69, 82]. Thus, mitochondrial organization and biogenesis is an important, relevant, and tractable area where computational methods can demonstrate their utility.

Our approach combines these computational predictions with a rigorous set of experimental tests designed to confidently assess if each gene is involved in this process. These assays are specific and reliable enough to convincingly confirm a gene's involvement in mitochondrial maintenance. As such, these assays are much more time consuming than high-throughput assays, but we have been able to scale up these methods in order to test 150-200 genes in a 4-6 month time frame, which is significantly faster than traditional low-throughput approaches. We refer to this level of experimental testing as "medium-throughput" biology, and we demonstrate that these types of assays are excellent for integration with computational approaches. By maintaining a high level of reliability while also testing a moderately large number of genes this data is able to meaningfully identify gene/protein function as well as produce enough novel information to inform the computational prediction methods. By combining these assays and computational predictions we have begun to iterate the cycle of prediction and experimentation to further explore mitochondrial pathways.

The results of our study identify and verify 98 novel genes involved in mitochondrial organization, over half of which have a human ortholog, including several disease genes. Our predictions were experimentally confirmed at a rate of 59%, which demonstrates the utility of computational approaches to guide laboratory work. Further, we have examined the nature of these predictions and

their relationship to the ensemble of computational methods. From this analysis we draw several conclusions vital for computationalists to consider when designing and using computational prediction methods. Finally, we demonstrate that by deeply incorporating computation and biology in an iterative fashion that we can greatly expand our knowledge of gene functions.

6.2 Results

6.2.1 Overview of study and results

We employed an ensemble of three diverse computational methods[32, 37, 62] to predict novel genes/proteins involved in the process of mitochondrial organization and biogenesis. Each of these methods integrated high-throughput data sources and utilized existing biological knowledge from the Gene Ontology (GO) [4] and *Saccharomyces* Genome Database (SGD) [18] to identify candidates for involvement. Our computational predictions of gene function were validated using two rigorous laboratory assays, each of which can indicate involvement in mitochondrial proliferation. The first round of prediction and evaluation used only annotations to GO as a training set. We then performed a second iteration of this process after updating our training set to include genes confirmed in the first iteration. A schematic view of our system for prediction, verification, and iteration is shown in Figure 6.1.

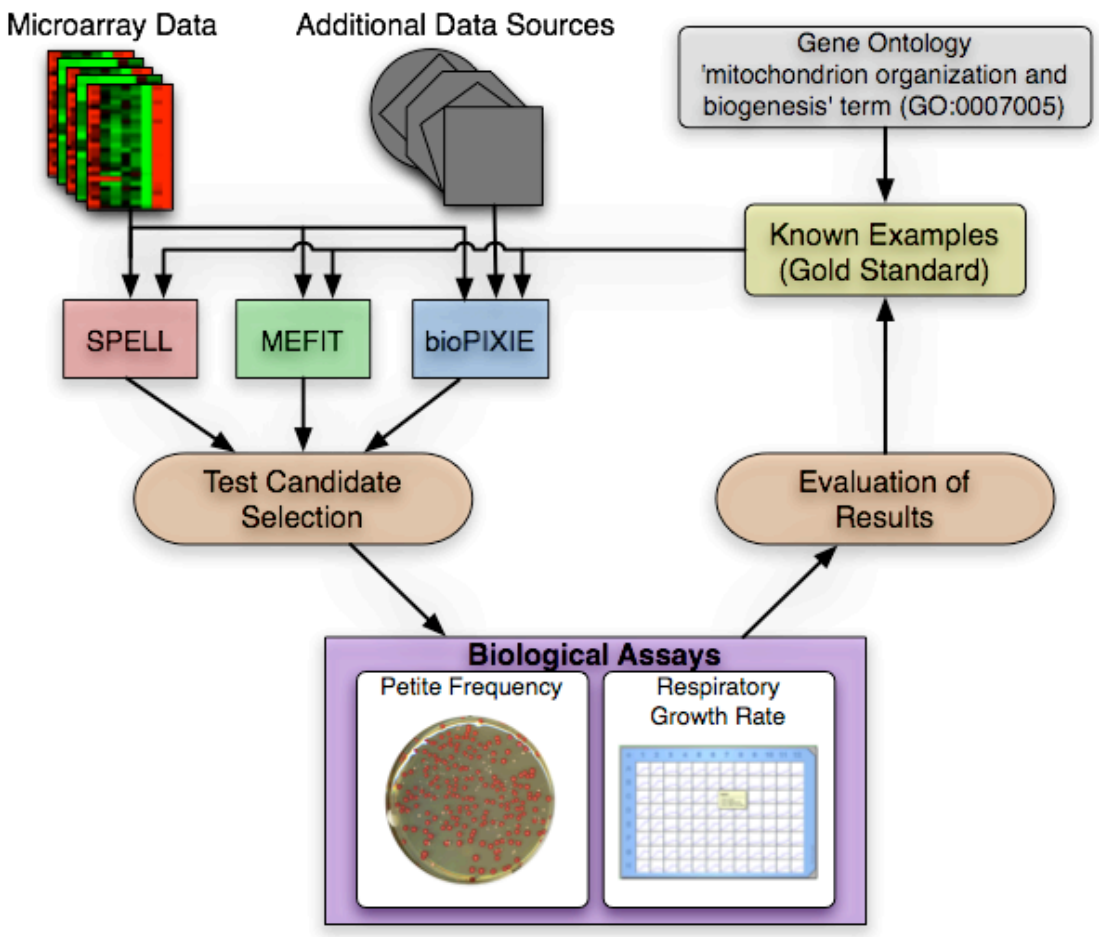


Figure 6.1: An overview of our iterative approach integrating computational and experimental methodologies. Our study uses an ensemble of computational gene function prediction methods (bioPIXIE, MEFIT, and SPELL) trained and/or evaluated on known biology to predict novel annotations to the GO term 'mitochondrial organization and biogenesis.' We selected test candidates based on these computational approaches, and then validated the predictions experimentally using two rigorous, quantitative biological assays. Upon evaluating the results of these tests, the verified predictions were added to existing knowledge, and the process was repeated to further explore this biological process.

When the study was undertaken, 106 genes were annotated by SGD to the "mitochondrial organization and biogenesis" GO term (GO:0007005, as of 4/15/2007); these 106 genes were used as input to the computational methods during the first iteration of testing. We initially evaluated our 186 most confident

computational predictions, and 122 (66%) were validated as exhibiting a significant phenotype indicative of involvement in organization. Upon further inspection of these confirmed predictions, we found existing literature evidence for 41 of these genes. By following this literature, we identified an additional 69 genes with strong evidence that have not yet been annotated as such by SGD. We have presented this list of "under-annotations" to SGD and they are in the process of updating the annotations to these genes.

Our second iteration of prediction and validation used a set of 297 genes as input to the computational methods (106 original annotations, 122 newly confirmed genes, and 69 "under-annotated" genes). We evaluated the 48 most confident predictions that were not previously tested, and 17 (35%) were validated. All together, our study identified 208 new annotations to the process of mitochondrial organization and biogenesis, which nearly triples the number of genes previously known to be involved in this area (Figure 6.2a). While this biological result is striking and important, it also has significant ramifications in the application of computational techniques as a whole and in their integration with experimental biology, which we discuss in detail below.

6.2.2 Guiding laboratory experiments with computation greatly increases discovery rates

Among our 234 laboratory evaluated computational predictions, 139 were confirmed, resulting in an overall true positive rate of 59%, which is excellent confirmation that computational predictions can successfully direct laboratory experiments. Further, our rate of discovery is much improved over the background rate of observing the same phenotypic classes. In addition to the

predicted genes tested, we also chose 48 genes at random to establish baseline rates of phenotypes. Of these 48 genes, 12 (25%) exhibited a phenotype consistent with involvement in mitochondrial organization. Based on these results, the use of computational methods to guide our investigation increased our discovery rate by 236%.

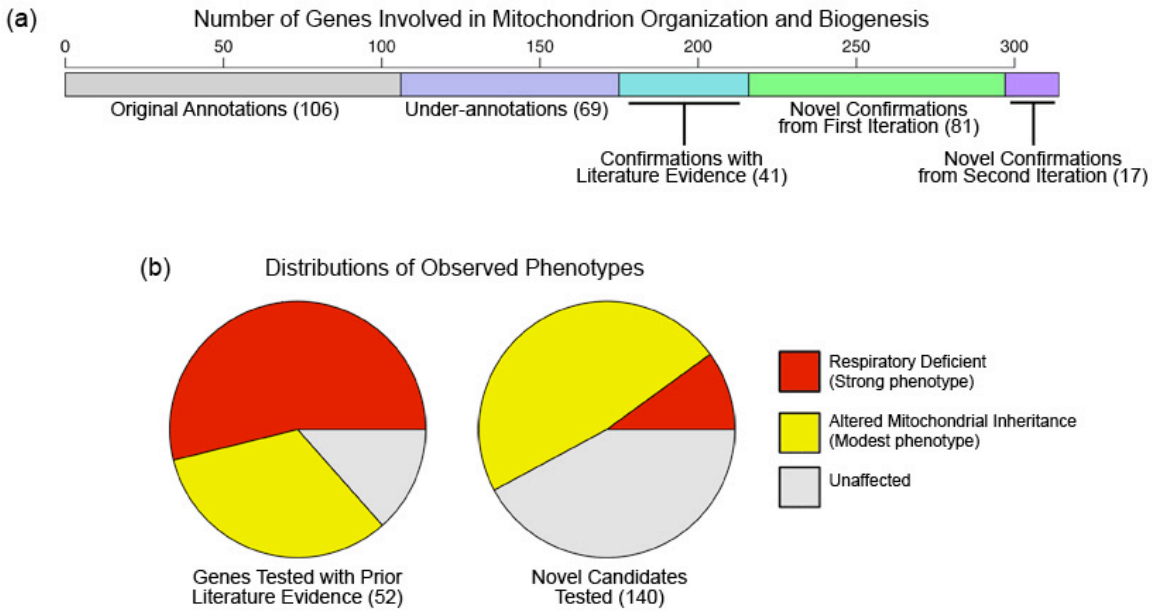


Figure 6.2: Annotations and phenotypic results for mitochondrion organization and biogenesis. Our study began with the 106 genes annotated to the GO term, 'mitochondrion organization and biogenesis.' Though the first round of our iterative computational prediction and laboratory experimentation, we confirmed 122 additional genes. Upon further investigation, 41 of these confirmations have previously existing literature evidence for involvement in this process, leaving 81 entirely novel discoveries during the first iteration. Based on continued literature searches we found an additional 69 genes with previous strong literature evidence for inclusion in this term. During our second iteration of testing, we confirmed an additional 17 predictions. (a) shows the sources of our current knowledge of associations. (b) shows the results of our petite frequency assay for genes with previous literature evidence (left) and our entirely novel first iteration predictions (right). Note that the majority of novel confirmations exhibited the more modest phenotype of "altered mitochondrial inheritance," whereas the majority of previously known genes exhibit the easier to find phenotype of "respiratory deficient."

In addition to a greatly increased discovery rate, our confirmed computational predictions are more integral to mitochondrial maintenance than the rare positives resulting from our random screen. Many of the genes involved in mitochondrial organization will localize either to the mitochondrion itself or to the actin cytoskeleton, as mitochondria associate with actin cables during inheritance [13]. Among phenotypically positive genes where localization data is available, 80% of our computational predictions are localized to either the mitochondrion or to actin, while only 36% from the random screen are localized. The increased enrichment of our computational confirmations with localization data is likely due to secondary effects among some of the genes in the random screen. As mitochondrion are vital for cellular respiration, our assays focused on discovering respiratory defects in single gene knockouts which is a strong indicator that the gene tested plays a role in mitochondrial processes. However, it is possible for secondary effects of mutations to result in similar phenotypes. For example, one of the randomly selected genes tested, *HTA1*, is a histone whose deletion is known to cause pleiotropic effects on transcriptional regulation of carbon metabolism [30]. Consequently our testing of an *hta1Δ* knockout strain resulted in a phenotype indicating involvement in mitochondrial organization and biogenesis, even though the true cause of this phenotype is likely a secondary effect due to a gross perturbation of carbon metabolism.

While enrichment for localization to the mitochondrion is an excellent indicator that our computational predictions are directly involved in mitochondrial maintenance, it is important to note that such localization is not a precondition for

involvement. Among all of our tested computational predictions, 65% are known to localize to the mitochondrion or actin cytoskeleton, and of these, 67% were confirmed. However, among the predictions not known to localize the accuracy is still quite high, at 45%. Thus, if a study examined only genes known to localize to the mitochondrion, it would fail to discover many of the verified genes that resulted from our use of computational predictions. By utilizing computational data integration techniques, we can leverage not only localization data, but also a wealth of other available knowledge to successfully direct our experimental efforts to the most promising targets.

6.2.3 Novel computationally-aided discoveries are likely to exhibit modest phenotypes

Another important finding in this study is the observation that many of the proteins we newly identified as related to mitochondrial organization exhibit a much more modest phenotypes than the previously known genes. Classic genetic screens and whole genome screens easily identify strong phenotypes, thus most of the single gene knockouts exhibiting a dramatic phenotype have already been discovered. However, characterizing genes with modest phenotypes will be central to our understanding of diseases, as it is often the case that mutations causing modest effects can be tolerated by the organism and present as diseases; whereas mutations causing severe defects in essential processes tend to be embryonic lethal. Identifying genes responsible for modest phenotypic effects requires robust, and time-consuming laboratory assays.

Specifically, based on our petite frequency assay, we classified single gene knockout mutants into three classes: respiratory deficient, altered

mitochondrial inheritance, and unaffected. Mutants classified as respiratory deficient were unable to grow on non-fermentable carbon sources, indicating that they completely lack functional mitochondria; while mutants exhibiting an altered mitochondrial inheritance rate produced some colonies capable of normal respiration, but at a rate significantly different than wild type. This second class is a more modest phenotype that could easily be overlooked by genetic screens. Among the 52 gene with prior literature evidence examined in this study, 45 exhibited a significant phenotype in the petite frequency assay, and of these, 62% were respiratory deficient, while 38% displayed altered mitochondrial inheritance. However, among our novel predictions that exhibited a significant phenotype, only 20% were respiratory deficient, while 80% had altered rates of mitochondrial inheritance (Figure 6.2b).

Thus, the majority of previously characterized genes exhibit a strong phenotype that could be found in a genetic screen, whereas nearly all of our novel discoveries exhibit a more modest phenotypic variance. As time-consuming laboratory assays are required to measure modest effects, it is unlikely that whole-genome screens for modest phenotypes can be performed in the near future. However, our results demonstrate that utilizing computational approaches can direct experimental efforts toward promising targets in order to discover modest effects within reasonable time frames.

6.2.4 Diverse, accurate predictions are made by different computational approaches

This study also demonstrates several important biological concerns that should be taken into account by computational biologists attempting to predict

gene/protein function. Each of the three function prediction methods used in this study achieved similarly high rates of phenotypic positives (Figure 6.3a). However, there was relatively small overlap between the most confident predictions of each method (Figure 6.3b). High true positive rates were achieved both among genes predicted by multiple methods, as well as genes predicted by just one method, indicating that each computational approach was accurately predicting disparate aspects of mitochondrial organization and biogenesis. This variation can be accounted for by differences in the underlying data as well as algorithmic differences among the computational approaches, and these differences are highly informative for computational biologists to consider when developing new methods or applying their methods to biological situations.

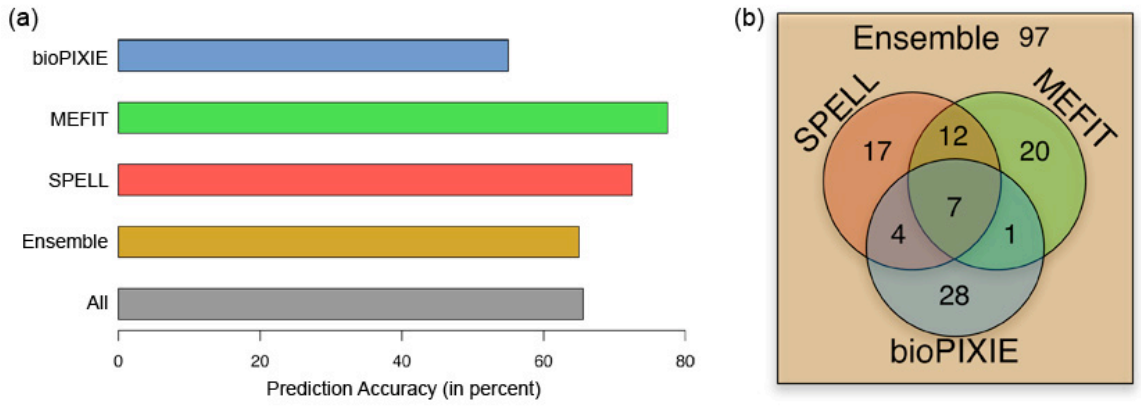


Figure 6.3: Individual method accuracy and overlap. Three computational methods and an ensemble of those methods were used to select candidates for evaluation. Of the 186 predictions evaluated in our first iteration, 89 were chosen from the top results of at least one individual method, while the remaining 97 were selected from the ensemble of all three. (a) shows the accuracy of the predictions chosen from each method, from genes selected by the ensemble, and the overall accuracy for all candidates. (b) shows the overlap between the candidates selected from the individual methods.

6.2.4.1 Underlying data affects the broad biological nature of predictions

Two of the methods used are based entirely on microarray data (MEFIT [37] and SPELL [32]), while the third method (bioPIXIE [61, 62]) utilizes not only microarray data, but also a wide variety of additional sources of information such as affinity precipitation, two-hybrid screens, sequence information, synthetic genetic interactions, etc. As such, the types of predictions made by the microarray-based methods were very different from the predictions produced by the method based on more diverse data. However, both sets of predictions achieved similar true positive rates during laboratory validation.

bioPIXIE	SPELL	MEFIT
mitochondrial part (GO:0044429)	mitochondrial part (GO:0044429)	mitochondrial part (GO:0044429)
actin cytoskeleton (GO:0015629)	--	--
mitochondrial distribution (GO:0048311)	--	--
--	mitochondrial ribosome (GO:0005761)	mitochondrial ribosome (GO:0005761)
--	translation (GO:0006412)	translation (GO:0006412)

Table 6.1: GO term enrichment among top predictions of each method. We evaluated the statistical enrichment of GO terms for the 50 most confident predictions of each computational method used in this study. This table shows significant, characteristic results from this analysis. GO terms common across methods are highlighted.

We examined the most confident predictions from each method and performed GO term enrichment analysis to classify the types of proteins each method was focused upon (Table 6.1). While all of the methods' predictions were enriched for GO terms such as 'mitochondrial part,' there were significant differences as well. The microarray-based methods were uniquely enriched for

categories related to the mitochondrial ribosome and translation, which reflects the ability of microarray data to easily capture ribosomal variation. However, the method based on more diverse data, including physical binding data, was uniquely enriched for genes related to the actin cytoskeleton and mitochondrial distribution, which is a direct result of the data used to generate these predictions.

We have further characterized the importance of underlying data by looking at sub-groups of genes known to be involved in mitochondrial organization and determining which groups are best captured through the computational prediction methods (Figure 6.4). The microarray-based methods clearly best capture information regarding 'mitochondrial ribosome and translation' which is consistent with other studies that have observed a strong ribosomal bias among microarray data [60]. The method based on diverse data best captured information about 'mitochondrial distribution' and 'mitochondrial fission and fusion.' This is likely due to the use of physical binding data, which allows this method to discover proteins involved in mitochondrial structure and motility. Another significant difference occurs in the area of 'mitochondrial complex assembly,' where the microarray-based methods are more successful than the method based on diverse data. This is likely due to the fact that many of the proteins involved in this process are membrane bound, which many sources of physical interaction data (e.g. yeast two-hybrid, affinity precipitation) are unable to assay due to technological limitations. However, as mitochondrial number strongly responds to environmental conditions, there is a transcriptional

signal associated with these proteins, which can explain the utility of microarray data for this area.

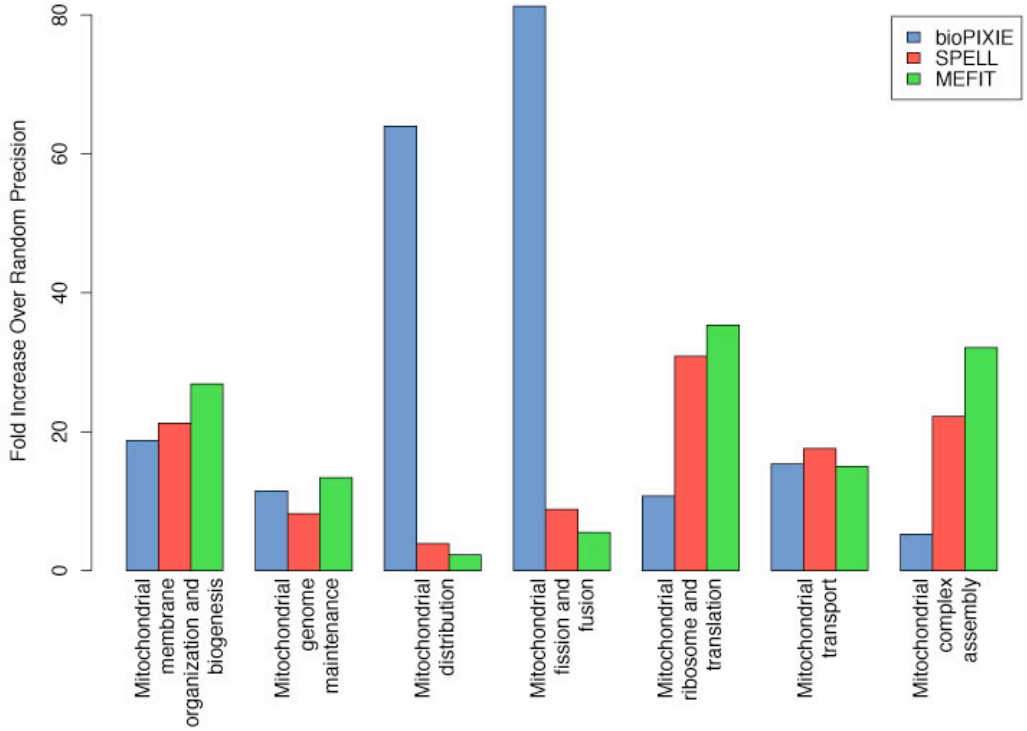


Figure 6.4: Biological differences between the three computational prediction methods. We evaluated which aspect of mitochondrial biology each computational function prediction method was targeting. Even though all three methods were trained and/or evaluated using the same training set of genes, the methods differ in which sub-group of mitochondrial biology they focused on. SPELL and MEFIT are both based solely on gene expression microarray data, which explains their strong coverage of the mitochondrial ribosome and translation sub-group. bioPIXIE is based on diverse data, including physical binding data, which explains its strong coverage of sub-groups involving mitochondrial motility and physical interactions.

Clearly, the types of predictions made by each of these methods are highly dependent on the data underlying each of the methods. It is important for computational biologists to understand what processes and functions data can be reasonably expected to capture, and to utilize that information in their

methods. Further, when evaluating the results of computational prediction methods, it is important to consider the specific aspects of biology captured, rather than simply rely on aggregate measures such as raw precision calculated over all biological classes [60].

6.2.4.2 Algorithmic differences affect specific computational predictions

Even among methods based on the same underlying data, very different predictions can be made depending on the computational approach used to analyze the data. We observed significant differences between the two microarray-based methods used in this study, even though they are based on the same input data and each method achieved similarly high levels of biological accuracy (Figure 6.3a). While the biological aspects of the predictions made by each of these methods are similar, the particular genes selected as top predictions differ.

There are several potential reasons for algorithms utilizing the same data to generate disparate predictions. Data normalization, distance metrics, training sets, evaluation metrics, algorithm type, and parameter choices can greatly affect the predictions generated by a computational approach. For this study we used the same set of roughly 120 microarray datasets normalized in the same manner, using the same distance metric, and evaluated on the same training set. Despite these similarities, we still observed several differences.

Each of the two microarray-based methods utilized in this study effectively assign a reliability weight to each dataset based on the biological context examined (in this case, mitochondrial organization and biogenesis). However,

their algorithmic aims and methodologies are quite different. MEFIT performs a Bayesian integration of microarray data trained on known genes. During the training process, the ability of each dataset to predict all known mitochondrial organization genes is assessed; then during evaluation of other genes, more reliable datasets are given greater weight. Predictions are obtained by examining the connectivity of other genes to a graph centered on the known players. Contrastingly, SPELL is an unsupervised search algorithm that obtains predictions by searching for candidates that behave similarly to known mitochondrial organization genes in the datasets where the set of known genes co-express. This algorithm was evaluated using many subsets of known genes as queries, and the results were averaged together. While MEFIT trusts datasets based on their ability to classify known mitochondrial organization genes as a whole, SPELL makes local decisions about the reliability of datasets for each of the queries utilized. Thus, MEFIT is likely to gain a better sense of the global transcriptional response of this process, while SPELL is more likely to find smaller groups of co-regulated genes.

The dataset reliability weights obtained by each method are significantly correlated ($r=0.55$, p-value$=6\times10^{-10}$), however there are also large differences. These differences can be accounted for by the details of each method's algorithm and their separate focuses on global or local patterns. Largely due to these differences in dataset reliability, each method generated very different specific predictions (Figure 6.3b), yet each method performed with roughly the same degree of precision. As both global and local perspectives achieved

impressive results in this case, it may be beneficial to utilize both types of approaches.

6.2.4.3 An ensemble of diverse prediction methods broadens the scope of results

Our study greatly benefited from the use of multiple, complementary functional prediction techniques. As described above, the three methods utilized in this study produced diverse, yet uniformly accurate, predictions spanning many biological aspects of mitochondrial organization and biogenesis. In addition to testing the most confident predictions of each method individually, we also combined the results of each method together into an ensemble predictor. We selected additional test candidates from this ensemble where moderately confident predictions made by multiple methods resulted in much higher aggregate confidence.

Approximately half (97) of the predictions tested in this study did not occur among the highest confidence predictions of any individual method, but were selected on the basis of the ensemble of all three methods. The accuracy of these ensemble predictions is roughly the same (65%) as the predictions made by any of the individual methods (Figure 6.3a). Thus by harnessing the diversity and complementarity of our computational prediction methods we were able to expand the biological breadth and scope of our investigation.

6.2.5 Iterative approaches converge on comprehensive prediction sets

In addition to predicting and validating novel functions for genes based on the current state of biological knowledge, we have begun to iterate the

prediction/validation process. Initially, we selected 186 testing candidates, 122 of which were verified through laboratory testing as likely involved in mitochondrial organization and biogenesis. In addition, we found that 41 of our verified candidates had strong existing support in the literature, which led us to identify 69 further genes with previously published literature evidence for inclusion in this process. After this first round of testing, we created a new training and evaluation standard for the computational methods including the original 106 annotated genes, the 69 genes with strong literature support, and the 122 experimentally verified players. Using this updated training set, we selected an additional 48 novel testing candidates with no previous literature evidence, 35% of which demonstrated a significant phenotype in the lab, resulting in our total of 139 gene function associations. Additionally, we made several important observations through the iteration process.

While the prediction methods differ with regard to which aspects of mitochondrial biology they best cover, the methods are beginning to converge after just one round of re-training. The correlation between the predictions of each method increased greatly (Figure 6.5), and the dataset weights used by the two microarray-based methods also became more correlated after one round of iteration. The convergence of the results of our computational prediction methods indicates that we are both expanding our knowledge of this area enough such that different approaches can arrive at the same conclusions, and that by cross-training the methods we are avoiding bias toward any one

functional aspect of the mitochondria, which often results from the application of an individual method.

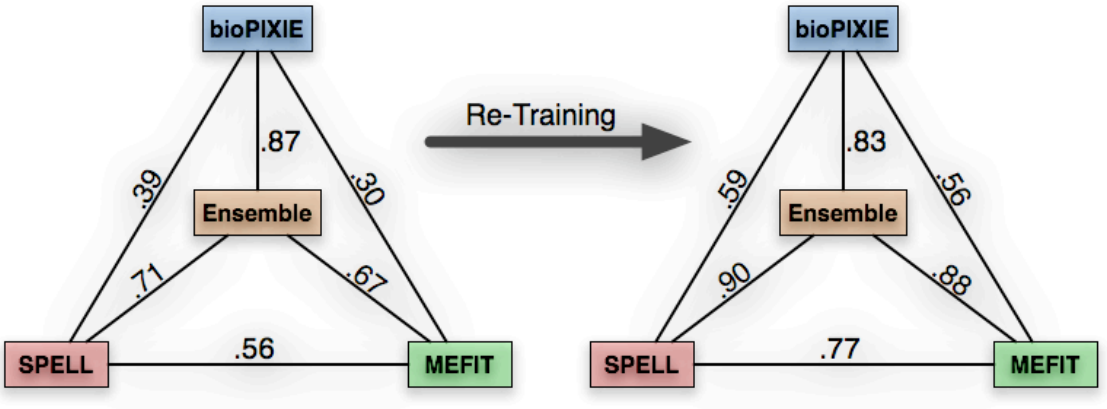


Figure 6.5: Convergence of computational predictions during iteration. We calculated the Pearson correlation between the estimated precisions of all genes between each of the three computational prediction methods and the ensemble predictor. These values are shown for the first iteration results (left) and after one round of retraining (right). After just one iteration of our prediction and evaluation cycle, the methods are beginning to converge.

This type of iterative learning is especially important as we move to less well-studied areas of biology and to less well understood organisms. By beginning with relatively little information, iterative applications of computational analysis and directed experimentation can enable the accurate annotation of a significant number of novel players by alternately refining the set of novel predictions and increasing the amount of information used for training.

6.3 Methods

A high level overview of our iterative prediction/experimentation/validation approach is shown in Figure 6.1. This section briefly details each of the steps involved in this process.

6.3.1 Computational prediction methodologies

We utilized three complementary, diverse computational gene function prediction methods in this study (bioPIXIE[61, 62], MEFIT[37], and SPELL[32]). Each of the methods generated predictions of genes involved in the GO biological process 'mitochondrion organization and biogenesis' (GO:0007005). All methods were initially trained and/or evaluated using the 106 annotations to this process as of April 15th, 2007. Full details of these methods can be found in their respective publications. Here we present a brief summary of each approach and a description of how each method was used to produce computational function predictions.

bioPIXIE utilizes a suite of context-specific Bayesian networks to predict pair-wise functional relationships between genes, which are then used to create fully, connected graphs weighted by confidence of functional interaction. This method integrates a wide variety of data sources, including physical interaction data (e.g. yeast two-hybrid, affinity precipitation, etc.), genetic interaction data (e.g. synthetic lethality, SLAM, etc.), gene expression data, and sequence data (e.g. coding and regulatory sequence similarity). One Bayesian classifier was trained per biological context of interest, where in this case, each context was an individual GO term. A positive standard generated from GO was used to learn conditional probability tables specific to mitochondrial organization and biogenesis. Predicted annotations to this term were derived from the resulting weighted interaction network by finding the significance of each gene's connectivity to known mitochondrial genes:

$$c_M = \left\{ \sum_{i \in M} \sum_{j \in G} w(i,j) \right\}, c_G = \left\{ \sum_{i \in G} \sum_{j \in G} w(i,j) \right\}$$

$$c_i = -\log\left[1 - HG\left(\left\{ \sum_{j \in M} w(i,j) \right\}, \left\{ \sum_{j \in G} w(i,j) \right\}, c_M, c_G \right) \right]$$

where c_i is gene *i*'s confidence of mitochondrial function, *M* is the set of genes known to be involved in mitochondrial organization, *G* is the set of all genes in the genome, $w(i, j)$ is the predicted probability of functional relationship between genes *i* and *j*, $HG(w, x, y, z)$ denotes the hypergeometric cumulative distribution function (CDF), and $\{x\}$ indicates that *x* is rounded to the nearest integer.

MEFIT also predicts pair-wise functional relationships using a collection of GO-trained naïve Bayesian classifiers; however, it is based entirely on gene expression data. Both MEFIT and SPELL (below) integrate roughly 2400 microarray conditions which are grouped into ~120 datasets by publication and further sub-divided by biological process examined. A ranked list of predictions was derived from the mitochondrial organization and biogenesis-specific network by calculating each gene's ratio of connectivity to known mitochondrial genes:

$$c_i = \frac{|G| \sum_{j \in M} w(i,j)}{|M| \sum_{j \in |G|} w(i,j)}$$

where c_i, *M*, *G*, and $w(i, j)$ are as above.

SPELL utilizes the same gene expression microarray data as MEFIT, but uses a query-driven search algorithm to identify novel players. While SPELL is not trained in a supervised fashion, it assigns a reliability weight to each dataset based on the co-regulation of a specified set of query genes and then orders the

rest of the genome based on their weighted co-expression with the query set. SPELL generated predictions by using all possible subset pairs of known mitochondrial organization and biogenesis genes as queries, and then averaged these rank orders together to produce a final prediction list.

Each of these methods generated a ranked list of the genome in order of confidence of involvement in mitochondrial organization and biogenesis. We assigned an estimated precision level to each gene in each list by calculating the fraction of genes with a higher confidence level that were already annotated to this GO term (disregarding genes with no biological process annotation or with annotations to the mitochondrial ribosome due to unusually strong expression co-regulation). We created a simple ensemble of the three methods by averaging these estimated precision levels for each gene. In this way each prediction method contributed to the ensemble based on its reliability to recapitulate known biology. Further, this ensemble allows a gene with moderate confidence from multiple methods to rise in the overall rankings.

6.3.2 Identification of "under-annotated" genes

Our initial evaluation of the computational predictions led us to discover that 41 of our experimentally confirmed predictions were "under-annotated" – meaning that they already had strong literature evidence for their involvement in mitochondrial organization and biogenesis, but were not yet annotated to the corresponding GO term. In most of these cases the information was already curated by SGD in the form of annotations to other GO terms, such as 'integral to the mitochondrial membrane' or 'mitochondrial protein import.' However, due to

the structure of the GO hierarchy, these terms are not directly linked to our process of interest, 'mitochondrial organization and biogenesis.' Beginning with these 41 genes, we identified an additional 69 genes that we believe have enough literature evidence to warrant their inclusion in this process without further laboratory testing. We have notified SGD of all 110 of these genes, and they are in the process of restructuring the GO hierarchy and making additional annotations. As of submission of this manuscript, SGD has already updated the annotations for 54 of these genes.

6.3.3 Selection of candidates for experimental testing

Novel candidates for laboratory evaluation were chosen on the basis of both the three individual computational approaches as well as the ensemble of their predictions. As our experimental methodologies (described below) are based on assessing phenotypes exhibited by single gene knockout mutants, we limited ourselves to consider only those genes with viable knockouts available in the heterozygous deletion collection. Additionally, we chose to evaluate both genes with no previously known association to a biological process as well as genes known to be involved in an area other than mitochondrial organization and biogenesis. Thus, we divided the predictions into genes of entirely unknown function and genes with existing biological process annotations.

We selected the 20 most confident genes of unknown function and the 20 most confident genes with existing annotations from each of the three individual methods for testing. Due to overlaps between the methods, this resulted in the selection of 89 genes as novel candidates (the overlap between methods is

shown in Figure 6.3b). We then chose an additional 97 genes from the ensemble list of predictions (38 from genes of unknown function and 59 from genes with known non-mitochondrial function) to arrive at our total of 186 test candidates in our first round of laboratory evaluation.

Of these predictions chosen for testing, we identified 46 as "under-annotated" (the remaining 140 predictions have no previous literature evidence for involvement in mitochondrial maintenance). We selected 6 additional test candidates from the existing annotations to mitochondrial organization and biogenesis, resulting in a total of 52 genes with prior literature evidence for involvement in this process. We also chose 48 genes at random from the set of all viable single gene knockouts in order to establish baseline rates of phenotypic positives. It should also be noted that by chance we would expect some overlap between our random selection of genes and our novel candidates; in our case, 3 genes are in common between these two groups.

6.3.4 Experimental methodologies and evaluation of results

We utilized two experimental approaches to assess a gene's involvement in mitochondrial organization and biogenesis. Both of these methods quantitatively measure a single gene knockout phenotype in comparison to the same phenotype for matched with type strains. Also, these methods were performed in replicate for each candidate examined such that robust statistical analysis could be performed on the results.

6.3.4.1 Strain preparation

For all of the genes examined, eight independent isolates of complete ORF deletions were obtained from freshly sporulated strains from the yeast heterozygous deletion collection [29, 90]. These isolates were catalogued and frozen until needed for testing.

6.3.4.2 Petite frequency assay

Yeast is able to grow and proliferate even without functional mitochondria on fermentable carbon sources. As such, yeast cells occasionally fail to pass working mitochondria on to daughter cells, but these cells can continue to proliferate. Cells lacking functional mitochondria are called petite cells. In this assay we assessed the rate at which single gene knockout strains produced petite offspring.

For each mutant strain tested, we grew several replicates of the strain for 48 hours using glycerol as a carbon source. Strains severely deficient in their ability to maintain functional mitochondrial cannot grow on glycerol as a carbon source. Strains unable to grow on glycerol were classified as respiration deficient in this first stage. Strains able to grow on glycerol were diluted and plated for single colonies on rich media, which releases the requirement for functional mitochondria. Thus, as colonies formed, cells without functional mitochondria were generated. When the colony is fully formed it is a mixture cells with functional mitochondria and cells without functional mitochondria. We measured this ratio by re-suspending a colony and plating a dilution of this re-suspension such that 100-300 colonies are formed on a plate. By overlaying with

soft agar containing tetrazolium, cells with functional mitochondria were stained red, while cells without functional mitochondria remained white. The ratio of white cells to total cells gives the petite frequency. Eight independent petite frequencies were measured for each strain tested. The distribution of these frequencies was compared to the frequency of petite generation in wild-type yeast. Strains identified as having the altered mitochondrial inheritance phenotype in this assay exhibit at least a 20% change in petite frequency from wild type, and have a p-value of less than 0.05 when comparing the petite frequency distributions of that strain to the wild-type petite frequency distribution, using a Mann-Whitney U test.

6.3.4.3 Growth rate assay

This assay measured the growth rate of mutant strains both in rich media as a control and in a non-fermentable carbon source (glycerol) to assess the ability of the mutants to perform respiration. Growth curves were performed in a 96-well plate format that tests 12 genes per run. For each gene tested, 6 independent deletion mutants of that gene were grown in separate wells. Twenty-four replicate wild-type strains were also present in each 96-well plate format. Plates were grown and measured using a Tecan GENios plate incubator and reader, which recorded densities every 15 minutes for 42 hours to generate growth curves.

Growth rates were derived from these curves by using Matlab to fit an exponential model:

$$y = a2^{bx}$$

For each well, this model was fit over the entire curve, the first 2/3, and the first half; whichever yielded the best fit was used in downstream analysis (to avoid plateau effects and to model only exponential growth). Wells with an adjusted R^2<0.9 were marked as non-growing, and growth rates for the remaining wells were determined by subtracting the row, column, and plate means for each well from the exponential parameter *b*, yielding a rate *b′* for each well. These *b′* parameters were tested for significance against the wild type population using a Mann-Whitney U test.

To detect colonies growing exponentially but with significant differences in fitness, smoothed maximum densities *d* were calculated for all wells deemed exponential. From these, plate, row, and column averages were subtracted from each well, generating adjusted maxima *d′*. These *d′* values for each mutant were again compared with the wild type values using a Mann-Whitney U test. Combined with the exponential rate tests, this assigned each mutant phenotypes in rich media and glycerol of no effect, no growth, or significant sickness. For, a mutant to be classified as having a respiratory growth defect, that defect was required to be specific to the glycerol media. If the mutant grew slowly in both glycerol and rich media then it was not considered to have a defect in respiration.

6.3.5 Assessing the comparative accuracy of the computational methods

In order to compare which aspect of mitochondrial biology was best captured by each of the computational methods, we created a breakdown of known mitochondrial biology into several sub-groups. Based on the 106 original annotations and the literature evidence for the 110 “under-annotations” we

created 7 more specific groups of genes shown in Figure 6.4. Given the prediction ordering of each computational method from our first iteration (i.e. using the original 106 genes as the training set) we calculated the average precision for each of the 7 more specific groups for each of the three computational approaches. The average precision was calculated for each sub-group, *G*, as

$$AP_G = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{i}{rank_i},$$

where $rank_i$ is the rank order of the i^{th} gene appearing from the sub-group in the ordered prediction list. For display in Figure 4, the average precisions were normalized by the expected average precision if the genome were ordered randomly, which corresponds to the number of genes in each sub-group divided by the number of genes in the genome.

6.3.6 Iterative re-training, prediction, and verification

After our first round of testing, 122 of the 186 predictions were found to have a significant phenotype strongly indicating involvement in mitochondrial organization and biogenesis. Combined with the original 106 annotated genes and the 69 genes identified as "under-annotated," this results in a total of 297 genes. Each of the three computational methods was re-applied using this updated training set of 297 genes and the same procedure was used to form an updated ensemble list of predictions. We selected 48 of the genes with the highest confidence from the updated results that were not previously tested for a second round of laboratory investigation. The same experimental assays and

evaluation procedures were used, and an additional 17 genes demonstrated a significant phenotype, resulting in a total of 139 out of 234 total predictions indicating involvement.

6.4 Conclusion

In order to fulfill the broad promise of computational functional genomics, we must undertake large-scale, iterative efforts to predict, evaluate, and verify novel gene functions through computationally directed experimentation. Our study demonstrates the potential utility of these types of approaches and in doing so has nearly tripled the number of genes annotated to the process of mitochondrial organization and biogenesis in *S. cerevisiae*. Additionally, our results demonstrate that computational methods must be designed and utilized on the basis of the biological nature of the data they consume and the types of interactions a particular algorithm is able to predict. Further, careful and rigorous experimental methods are required in order to confidently measure modest phenotypes. Finally, the iterative use of complementary computational methods combined with solid laboratory testing are able to greatly expand our knowledge of biology.

Methodologies such as ours are especially important given that the amount of functional genomics data generated is outpacing the amount of novel biological knowledge that is gained from these experiments. By directing experimental efforts to more promising targets, we can reduce the amount of laboratory effort required to discover new biology, including functional discoveries of unknown proteins. Of the genes this study newly identifies as involved in

mitochondrial organization and biogenesis, 60 had no previously known function. Thus, through the course of this single study, we have discovered a functional annotation for roughly 4% of the remaining genes of unknown function in *S. cerevisiae*. The intelligent combination and iteration of computational and experimental work can significantly speed our rates of discovery, not only in yeast mitochondrial biology, but also in a broad range of processes and organisms.

Chapter 7

Conclusions and Future Work

Understanding the functions and roles of genes and proteins within cells and within organisms is a vital step towards the larger goal of understanding and treating disease. Recent years have seen a dramatic increase in the amount of data generated to attempt to solve this problem. In particular, billions of gene expression measurements have been taken in a variety of organisms, from bacteria to humans, studying a wide range of experimental, environmental, and clinical perturbations. However, the information contained within this data remains largely obscure from most researchers.

We have introduced several new approaches for exploring and understanding the myriad of expression data available to biologists. Each of these methods addresses key concerns in modern computational biology. The incorporation of expert biological knowledge in the early phases of analysis, the meaningful presentation of available data to researchers, and the deep integration between computational and laboratory methods are all techniques that are vital to discovering novel biology.

Specifically, first, we have developed a query-driven similarity search algorithm for use with large collections of microarray data. This approach utilizes the biological diversity of published expression studies to identify novel genes involved in process or pathways. Additionally, through this work we have characterized the current state of available gene expression data for *S.*

cerevisiae, including a catalogue of biological areas that remain to be studied using microarray technology. This methodology reveals important information that was previously hidden within expression databases quickly and accurately. Further, this context-specific, user-driven search paradigm is very powerful and is applicable to higher organisms and additional types of data.

Second, we developed several novel visualization schemes aimed at incorporating statistical information into the visual representation of expression data. This work allows researchers to more accurately assess the results of analysis methods, such as clustering. These types of methods are especially important for examining microarray data, as so much analysis is performed in a visual manner. This work demonstrates the general necessity and utility of tailoring visualization approaches to the desired task.

Third, we have designed a new visualization paradigm that enables the simultaneous exploration of multiple expression datasets in a biologically meaningful fashion. This technique provides researchers with the ability to understand their own datasets within the broader biological context of other available data. As the amount of available biological data continues to grow, it will become increasingly important to comprehend the relationships between datasets, rather than considering each study within a vacuum.

Fourth, we have integrated our analysis and visualization approaches together into a unified platform for comprehensive exploration of large collections of gene expression microarray data. This platform demonstrates the value of fusing powerful analysis methodologies within the context of a meaningful

visualization interface. As the majority of researchers analyze expression data visually, the ability to perform multiple analyses interactively greatly increases the amount and quality of information that can be discovered. Further, the unification of statistical analysis methods and targeted visualizations is a compelling paradigm for the exploration of many other types of data.

Last, and perhaps most important, we have validated these approaches in a large-scale collaboration with laboratory biologists that is the first of its kind. This unique project based on our computational function prediction methods revealed a novel involvement in mitochondrial organization for nearly 100 proteins, including 5% of all the remaining proteins of unknown function in yeast. Of these genes, more than half are conserved through human, and more than 1 in 5 of these orthologous proteins are known to be involved in human diseases. Novel discoveries of this scale are rare in molecular biology, and this result strongly confirms the validity and usefulness of our computational analysis approaches. Further, this study led to many important observations concerning the role and applicability of computational predictions for functional investigations. Our results affirm that both laboratory biology and computational biology must inform and guide each other in order to solve important problems.

While the majority of this work has been focused on functional discovery in *S. cerevisiae* through the analysis of gene expression microarray data, these methods and algorithms are foundational for future work in analyzing additional data sources and in studying higher eukaryotes, especially human. Extending these methods to include additional organisms will range from trivial to very

complex. Many of the visualization-based paradigms presented here have already been used in a variety of organisms with great success. However, methods such as our query-driven similarity search will need to take into account developmental phases, tissue-specific variations, and the increased genetic complexity of higher organisms in order to be successful. While there are many new challenges that must be addressed, the basic principles of this work – incorporating expert knowledge with analysis of large data compendia, creating statistically meaningful visual exploration approaches, and integrating computational and laboratory analyses – will endure.

Although there is still a great deal of work to be done in functional genomics, understanding and treating human diseases will require further development of the field of computational biology. In the coming years, bioinformatics will expand its focus beyond understanding specific gene functions to a more comprehensive study of systems level biology, disease pathways, and clinical treatment options. This transition will require methods incorporating not only high-throughput genomic data, but also metabolomics, proteomics, and clinical data. Analysis of these data sources to achieve these new goals will bear many similarities to the work presented here. The general paradigm of statistically robust, scalable, and integrated analysis and visualization techniques for data exploration will remain a cornerstone of successful bioinformatics approaches.

Appendix A

Datasets used in the SPELL search engine

Microarray data was collected from a variety of sources to create our compendium, including NCBI's Gene Expression Omnibus [22], EBI's ArrayExpress [15], the Stanford Microarray Database [81], and several other publication and laboratory web pages. These data from 81 publications, totaling 2394 array hybridizations, were broken down into their smallest logical groupings of conditions. For example, the stress response dataset from Gasch *et al.*, 2000 [27] originally consisted of 142 hybridizations corresponding to several different types of induced stress and growth phases. We have separated this dataset in a manner similar to the authors' analyses, resulting in 21 logical datasets such as "hydrogen peroxide exposure," "osmotic shock," and "heat shock from 25° to 37°."

In order to make valid comparisons between the datasets collected, all data was normalized in a similar manner. First, suspect values were removed (i.e. missing values were inserted) in all data based on the information available in the original publication where possible, or in a manner appropriate to the microarray platform used. After identifying missing values, any genes present in less than 50% of the conditions in a dataset were removed from that dataset. Remaining missing values were imputed using the KNN impute algorithm [91] with K=10 using Euclidean distance to identify nearest neighbors. After the

imputation process, technical replicates were averaged together, resulting in data files of complete matrices with one entry per gene appearing in the dataset.

Most of the data collected falls into two main categories: dual-color competitive hybridization data and single-channel data. Dual-color data was typically found in log ratio format or was transformed into this format. Single-channel data was typically from Affymetrix platforms and was log transformed as a final step in normalization. Other types of data were transformed into a format as close as possible to these sources.

Table A.1: SPELL microarray data collection list. This table contains the full list of publications and datasets collected, and the subsequent breakdown of datasets into logical units for our functional coverage analysis and search engine.

Brief Description	# of Cond-itions	First Author	PubMed ID
rsc3/rsc30 knockouts	8	Angus-Hill ML	11336698
slt2/swi4/swi6/bck1 knockouts	5	Baetz K	11533240
splitomicin exposure and sir2 mutants	7	Bedalov A	11752457
Oxidative stress and glutaredoxin 5-deficient mutant	9	Belli G	14722110
Histone deacetylase (rpd3/sin3/hda1 deletions)	6	Bernstein BE	11095743
Histone deacetylase (sin3/sap30/ume6/hda1/hos2/hos3 deletions)	7	Bernstein BE	11095743
Trichostatin A treatment time course	5	Bernstein BE	11095743
leu3 mutant expression profiles	12	Boer VM	15949974
Diauxic shift time course (Batch1)	13	Brauer MJ	15758028
Diauxic shift time course (Batch2)	7	Brauer MJ	15758028
Transcriptional regulation (I)	40	Brem RB	11923494
Transcriptional regulation (I)(dye swap)	40	Brem RB	11923494
Transcriptional regulation (II)	12	Brem RB	11923494
Transcriptional regulation (II)(dye swap)	11	Brem RB	11923494
Genetic variation in gene expression among parents and progenies (dye-swap)	131	Brem RB	15659551
Genetic variation in gene expression among parents and progenies	131	Brem RB	15659551
Lithium response	7	Bro C	12791685
Chitin synthesis	11	Bulik DA	14555471

Genotoxic stress	24	Caba E	15878181
H2O2 exposure to wt and Deltatrr1 knockout	15	Carmel-Harel O	11169101
pho85 inhibition	12	Carroll AS	11675494
Unfolded protein response	2	Casagrande R	10882108
acid response	11	Causton HC	11179418
alkali response	8	Causton HC	11179418
heat response	7	Causton HC	11179418
NaCl response	6	Causton HC	11179418
peroxide response	7	Causton HC	11179418
Sorbitol response	6	Causton HC	11179418
TBP inhibition	20	Chitikila C	12419230
mitotic cell cycle	17	Cho RJ	9702192
Sporulation time course	7	Chu S	9784122
mRNA processing factors and splicing (dye swap)	17	Clark TA	11988574
mRNA processing factors and splicing	17	Clark TA	11988574
yap1 and yap2 knockouts with peroxide and cadmium added	11	Cohen BA	12006656
Osmotic stress	12	De Nadal E	14737171
Diauxic shift time course	7	DeRisi JL	9381177
post heat shock, delayed rapamycin exposure time course	20	Duvel K	12820961
post heat shock, immediate rapamycin exposure time course	10	Duvel K	12820961
Mitochondrial dysfunction	11	Epstein CB	11179416
SPT10 global transcription regulator null mutant	6	Eriksson PR	16199888
Evolved strains	4	Ferea TL	10449761
Hydrostatic pressure response	2	Fernandes PM	14706843
proteasome inhibition with exposure to PS-341	30	Fleming JA	11830665
Aging in yeast	8	Fry RC	12875747
Amino acid, adenine starvation	5	Gasch AP	11102521
Carbon sources	6	Gasch AP	11102521
Diamide treatment time course	8	Gasch AP	11102521
Dithiothrietol exposure time course (y13)	8	Gasch AP	11102521
Dithiothrietol exposure time course (y14)	7	Gasch AP	11102521
Hydrogen peroxide response	2	Gasch AP	11102521
Hydrogen peroxide response time course	9	Gasch AP	11102521
Heat Shock 25C to 37C time course	8	Gasch AP	11102521
Heat Shock 29C to 33C time course	4	Gasch AP	11102521
Heat Shock 30C to 37C time course	5	Gasch AP	11102521
Heat Shock 37C to 25C	5	Gasch AP	11102521
Mild Heat Shock	6	Gasch AP	11102521

Heat Shock from various temp to 37C	6	Gasch AP	11102521
Hyper-osmotic shock time course	6	Gasch AP	11102521
Hypo-osmotic shock time course	5	Gasch AP	11102521
Menadione exposure time course	9	Gasch AP	11102521
Nitrogen depletion time course	9	Gasch AP	11102521
Stationary phase time course (y12)	10	Gasch AP	11102521
Stationary phase time course (y14)	9	Gasch AP	11102521
Steady-state temperature (y13)	5	Gasch AP	11102521
Steady-state temperature (y14)	8	Gasch AP	11102521
Copper regulon	6	Gross C	10922376
rapamycin exposure	14	Hardwick JS	10611304
pho85 related knockouts	20	Huang D	12077337
diverse knockout mutants	300	Hughes TR	10929718
GAL mutants	21	Ideker T	11340206
SBF-MBF genomic distribution (ORF_intergenic_v1.0) (I)	2	Iyer VR	11206552
SBF-MBF genomic distribution (ORF_intergenic_v1.0) (II)	2	Iyer VR	11206552
SBF-MBF genomic distribution (intergenic_v1.0) (I)	6	Iyer VR	11206552
SBF-MBF genomic distribution (intergenic_v1.0) (II)	5	Iyer VR	11206552
Exposure to alkylating, oxidizing agents, ionizing radiation	28	Jelinsky SA	11027285
Xylose metabolism	6	Jin YS	15528549
Ras/cAMP signal transduction pathway (dye swap)	5	Jones DL	14570984
Ras/cAMP signal transduction pathway	5	Jones DL	14570984
Haa1 analysis	4	Keller G	11504737
Carbon source shift	3	Kuhn KM	11154278
Unfolded protein response and HAC1 transcription	13	Leber JH	15314654
rnt1 null mutant expression profile	9	Lee A	15989963
gcr1 mutant, glucose exposure	17	Lopez MC	10940042
Zinc homeostatis, zap1	9	Lyons TJ	10884426
MAPK mutants	11	Madhani HD	10535956
TOR2-controlled transcription	12	Martin DE	15476558
immunosuppressant response	7	Marton MJ	9809554
abf1-1 mutant at 36C	4	Miyake T	15192094
Phosphate-regulated pathway (I)	5	Ogawa N	11102525
Phosphate-regulated pathway (II)	3	Ogawa N	11102525
Fermentation time course	12	Olesen K	12702272
Deubiquitinating enzyme UBP10 inactivation	4	Orlandi I	14623890
HOG MAPK pathway	133	O'Rourke SM	14595107
Phosphomannose isomerase PMI40 deletion strain response to excess mannose	15	Pitkanen JP	15520001

Sporulation of two strains	24	Primig M	11101837
Filamentous-form growth on solid media time course	10	Prinz S	14993204
Iron concentration and AFT1 overexpression	4	Protchenko O	11673473
Pheremone response	56	Roberts CJ	10657304
TPK1, TPK2, TPK3 mutants	12	Robertson LS	10811893
sus1 mutant	6	Rodriguez-Navarro S	14718168
fhl1 and ifh1 deletion mutants	6	Rudra D	15692568
Iron homeostasis	2	Rutherford JC	11734641
Histone deacetylase RPD3 deletion and histone mutations	18	Sabet N	15456858
limitation by Leucine	29	Saldanha AJ	15240820
limitation by Phosphate	30	Saldanha AJ	15240820
limitation by Sulfate	21	Saldanha AJ	15240820
limitation by Uracil	20	Saldanha AJ	15240820
comparison of limitation by Ura, Sul, Pho, and Leu	24	Saldanha AJ	15240820
Pre-mRNA splicing factor mutants at restrictive temperature time course	24	Sapra AK	15452114
IFH1 overexpression: time course	24	Schawalder SB	15616569
Heat Shock, kin82 mutant	10	Segal E	12740579
Hypo-osmotic shock, ppt1 mutant	10	Segal E	12740579
Stationary phase, ypl230w mutant	12	Segal E	12740579
Iron deprivation	6	Shakoury-Elizeh M	14668481
oxidative stress responses	70	Shapira M	15371544
Cell cycle, alpha-factor block-release	16	Spellman PT	9843569
Cell cycle, cdc15 block-release	25	Spellman PT	9843569
Cyclin overexpression	2	Spellman PT	9843569
Cell cycle, elutriation	14	Spellman PT	9843569
Snf/Swi mutants (v1_2.2)	2	Sudarsanam P	10725359
Snf/Swi mutants (384_F_v1.0)	8	Sudarsanam P	10725359
Nutrient limitation under aerobic and anaerobic conditions	24	Tai SL	15496405
Ssl1 mutant for a subunit of TFIIH response to methyl methanesulfonate	12	Takagi Y	15837426
Ume6 regulon (Ye6100subA)	8	Williams RM	12370439
Ume6 regulon (Ye6100subB)	8	Williams RM	12370439
Ume6 regulon (Ye6100subC)	8	Williams RM	12370439
Ume6 regulon (Ye6100subD)	8	Williams RM	12370439
Heat shock transcription factor 1 mutant response to heat stress	4	Yamamoto A	15647283
Transcription factor deletions	7	Yeang CH	15998451
Ca(2+) exposure	24	Yoshimoto H	12058033
Na(+) exposure	16	Yoshimoto H	12058033
Iron uptake	4	Yun CW	10744769

Trans-acting regulatory variation (dye swap)	90	Yvert G	12897782
Trans-acting regulatory variation	90	Yvert G	12897782
Pseudohyphal Growth	26	Zhu G	10894548

Appendix B

Details of the functional coverage analysis of the S. cerevisiae microarray compendium

Full functional coverage results are available in the online supplement to the SPELL search engine[84]. The full table of functional coverage consists of a matrix containing pseudo p-values (based on the z-test for significance) for each combination of dataset and GO biological process examined. These files are available as a tab-delimited text file of p-values, a (very large) image, and a hierarchically clustered version compatible with JavaTreeView for browsing. Note that a p-value of ~10-10 corresponds to the Bonferroni corrected p-value of 10-4 which was used for significance testing in Figure 2.6.

In addition to the z-test results, we have also calculated significance based on the non-parametric two-sample Kolmogorov-Smirnov test. The results of the KS-test show significance for generally the same GO term/dataset pairings, however several more pairs are also found to be significant. This is due to the fact that the KS-test can judge distributions significantly different if the shapes are sufficiently different, while the means are very similar. As the z-test is based on differences in means, it would not consider such distributions to be significantly different. The full results of the KS-test are also available at the supplementary website in a tab-delimited text file.

Highly-represented (significant in >15 datasets)	Moderately-represented (significant in <15 but >3 datasets)	Under-represented (significant in <3 datasets)
tricarboxylic acid cycle	response to oxidative stress	MAPKKK cascade (A)
DNA repair	amino acid transport	protein kinase cascade (A)
glycolysis	exocytosis	Ras protein signal transduction (A)
phosphate metabolism	vesicle fusion	mating type switching (B)
chromosome segregation	meiotic recombination	invasive growth (B)
DNA replication	arginine biosynthesis	pseudohyphal growth (B)
electron transport	steroid metabolism	response to salt stress (C)
ubiquitin-dependent protein catabolism	alcohol metabolism	heme biosynthesis (C)
ribosome assembly	double-strand break repair	mitochondrial genome maintenance (C)
amino acid metabolism	filamentous growth	telomerase-dependent telomere maintenance (C)

Table B.1: Functional coverage classes. Our analysis of the functional coverage of existing gene expression microarray data for *S. cerevisiae* characterizes both which biological processes are represented in each dataset and which biological processes are represented in existing data as a whole. This table shows a selection of processes that are significant in many datasets (left column), significant in some, but not many datasets (center column), and significant in very few datasets (right column). Of those processes under-represented in the compendium, there are three major explanations identified: (A) non-transcriptionally regulated processes, (B) processes not occurring in many common laboratory strains, and (C) specific processes not yet targeted by existing gene expression microarray data. Biological processes in the later category may be areas that warrant further investigation.

Appendix C

Details of SPELL Biological Performance Evaluation

Precision-recall curves were created by traversing the ordered list of results for each method for each GO term examined and calculating precision, recall pairs at each step. Precision is calculated as the ratio of true positive (TP) predictions to the sum of TP and false positive (FP) predictions. Recall is measured as the number of TPs recovered for individual GO terms, or as the proportion of TPs to the total number of possible TPs (TP + FN [false negatives]) for results averaged over multiple GO terms. Average precision was used as a summary statistic for comparing the performance of different methods in a more straightforward way. In addition to the summary comparison available in Figure 2.7 of chapter 2, the individual results for all 126 GO terms analyzed are available at the SPELL supplementary website [84].

Table C.1: List of 126 GO terms used in SPELL evaluation. These GO terms were used for comparative evaluations of the SPELL search algorithm.

GOID	Term Name
GO:0000074	regulation of progression through cell cycle
GO:0000160	two-component signal transduction system
GO:0000278	mitotic cell cycle
GO:0000279	M phase
GO:0000746	conjugation
GO:0000902	cellular morphogenesis
GO:0001510	RNA methylation
GO:0005975	carbohydrate metabolism
GO:0006056	mannoprotein metabolism
GO:0006066	alcohol metabolism
GO:0006081	aldehyde metabolism
GO:0006082	organic acid metabolism

GO:0006112	energy reserve metabolism
GO:0006113	fermentation
GO:0006118	electron transport
GO:0006260	DNA replication
GO:0006308	DNA catabolism
GO:0006310	DNA recombination
GO:0006323	DNA packaging
GO:0006352	transcription initiation
GO:0006353	transcription termination
GO:0006354	RNA elongation
GO:0006360	transcription from RNA polymerase I promoter
GO:0006366	transcription from RNA polymerase II promoter
GO:0006383	transcription from RNA polymerase III promoter
GO:0006399	tRNA metabolism
GO:0006401	RNA catabolism
GO:0006417	regulation of protein biosynthesis
GO:0006457	protein folding
GO:0006461	protein complex assembly
GO:0006473	protein amino acid acetylation
GO:0006476	protein amino acid deacetylation
GO:0006508	proteolysis
GO:0006512	ubiquitin cycle
GO:0006519	amino acid and derivative metabolism
GO:0006629	lipid metabolism
GO:0006725	aromatic compound metabolism
GO:0006730	one-carbon compound metabolism
GO:0006766	vitamin metabolism
GO:0006790	sulfur metabolism
GO:0006793	phosphorus metabolism
GO:0006800	oxygen and reactive oxygen species metabolism
GO:0006807	nitrogen compound metabolism
GO:0006811	ion transport
GO:0006818	hydrogen transport
GO:0006839	mitochondrial transport
GO:0006869	lipid transport
GO:0006913	nucleocytoplasmic transport
GO:0006914	autophagy
GO:0006944	membrane fusion
GO:0006970	response to osmotic stress
GO:0006974	response to DNA damage stimulus
GO:0006986	response to unfolded protein
GO:0006997	nuclear organization and biogenesis
GO:0007005	mitochondrion organization and biogenesis
GO:0007010	cytoskeleton organization and biogenesis
GO:0007031	peroxisome organization and biogenesis
GO:0007033	vacuole organization and biogenesis
GO:0007034	vacuolar transport
GO:0007046	ribosome biogenesis
GO:0007047	cell wall organization and biogenesis
GO:0007059	chromosome segregation

GO:0007155	cell adhesion
GO:0007166	cell surface receptor linked signal transduction
GO:0007243	protein kinase cascade
GO:0007264	small GTPase mediated signal transduction
GO:0007530	sex determination
GO:0007568	aging
GO:0008213	protein amino acid alkylation
GO:0008219	cell death
GO:0008298	intracellular mRNA localization
GO:0008380	RNA splicing
GO:0008643	carbohydrate transport
GO:0009100	glycoprotein metabolism
GO:0009116	nucleoside metabolism
GO:0009117	nucleotide metabolism
GO:0009266	response to temperature stimulus
GO:0009308	amine metabolism
GO:0009415	response to water
GO:0010035	response to inorganic substance
GO:0015837	amine transport
GO:0015849	organic acid transport
GO:0015893	drug transport
GO:0015931	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
GO:0016071	mRNA metabolism
GO:0016072	rRNA metabolism
GO:0016192	vesicle-mediated transport
GO:0016458	gene silencing
GO:0016481	negative regulation of transcription
GO:0016485	protein processing
GO:0018193	peptidyl-amino acid modification
GO:0019236	response to pheromone
GO:0019748	secondary metabolism
GO:0019932	second-messenger-mediated signaling
GO:0019953	sexual reproduction
GO:0019954	asexual reproduction
GO:0030261	chromosome condensation
GO:0030435	sporulation
GO:0030447	filamentous growth
GO:0030705	cytoskeleton-dependent intracellular transport
GO:0031023	microtubule organizing center organization and biogenesis
GO:0031123	RNA 3'-end processing
GO:0040029	regulation of gene expression, epigenetic
GO:0042157	lipoprotein metabolism
GO:0042594	response to starvation
GO:0043094	metabolic compound salvage
GO:0043284	biopolymer biosynthesis
GO:0045184	establishment of protein localization

GO:0045185	maintenance of protein localization
GO:0045333	cellular respiration
GO:0045454	cell redox homeostasis
GO:0045941	positive regulation of transcription
GO:0046483	heterocycle metabolism
GO:0048284	organelle fusion
GO:0048308	organelle inheritance
GO:0050790	regulation of enzyme activity
GO:0050801	ion homeostasis
GO:0051052	regulation of DNA metabolism
GO:0051169	nuclear transport
GO:0051186	cofactor metabolism
GO:0051236	establishment of RNA localization
GO:0051301	cell division
GO:0051321	meiotic cell cycle
GO:0051325	interphase

Appendix D

Details of *ARP8* Predictions and Validations

Table D.1: Functions predicted for ARP8 by SPELL. SPELL predicts that the un-annotated gene, *ARP8* is involved in the following 13 biological processes which break down into 3 main classes

Predicted GO term for Arp8	Class
mitotic cell cycle	Cell Cycle
interphase	Cell Cycle
regulation of progression through cell cycle	Cell Cycle
cell division	Cell Cycle
asexual reproduction	Cell Cycle
transcription from RNA polymerase II promoter	Transcription
negative regulation of transcription	Transcription
positive regulation of transcription	Transcription
transcription initiation	Transcription
mRNA metabolism	Transcription
cellular morphogenesis	Morphology
cytoskeleton organization and biogenesis	Morphology
response to osmotic stress	Other

For verification of the cellular morphology defect, cell volume was determined using the Z2 automated cell counter (Beckman Coulter, Fullerton, California, United States). Culture was diluted into Isotone II buffer for the measurement. Cell morphology was determined using a 40x objective on a Zeiss Axioskop (Germany). The entire field of view is shown for both wild-type yeast and the *arp8* deletion allowing for direct comparison of the images in Figure 2.9.

Bibliography

[1] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A,* 97(18):10101-6, 2000.

[2] Amar, Stasko. A knowledge task-based framework for design and evaluation of information visualizations. *IEEE Symposium on Information Visualization,* 143-150, 2004.

[3] Andreoli C, Prokisch H, Hörtnagel K, Mueller JC, Münsterkötter M, Scharfe C, Meitinger T. MitoP2, an integrated database on mitochondrial proteins in yeast and man. *Nucleic Acids Res,* 32(Database issue):D459-62, 2004.

[4] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25(1):25-9, 2000.

[5] Babcock M, de Silva D, Oaks R, Davis-Kaplan S, Jiralerspong S, Montermini L, Pandolfo M, Kaplan J. Regulation of mitochondrial iron accumulation by Yfh1p, a putative homolog of frataxin. *Science,* 276(5319):1709-12, 1997.

[6] Baehrecke EH, Dang N, Babaria K, Shneiderman B. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics,* 584, 2004.

[7] Baldonado MQW, Woodruff A, Kuchinsky A. Guidelines for using multiple views in information visualization. *Proceedings of the working conference on Advanced visual interfaces,* 110-119, 2000.

[8] Baldwin DN, Vanchinathan V, Brown PO, Theriot JA. A gene-expression program reflecting the innate immune response of cultured intestinal epithelial cells to infection by Listeria monocytogenes. *Genome Biol,* 4(1):R2, 2003.

[9] Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. *Bioinformatics,* 22(7):830-6, 2006.

[10] Bederson, Grosjean, Meyer. Toolkit Design for Interactive Structured Graphics. *IEEE Trans Software Eng,* 30(8):535-546, 2004.

[11] Benjamini, Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics,* 29(4):1165-1188, 2001.

[12] Berrar P, Dubitzky, Granzow, eds. *A Practical Approach to Microarray Data Analysis* Kluwer Academic Publishers, Boston, MA (2003).

[13] Boldogh I, Vojtov N, Karmon S, Pon LA. Interaction between mitochondria and the actin cytoskeleton in budding yeast requires two integral mitochondrial outer membrane proteins, Mmm1p and Mdm10p. *J Cell Biol,* 141(6):1371-81, 1998.

[14] Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics,* 20(18):3710-5, 2004.

[15] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res,* 31(1):68-71, 2003.

[16] Cairns BR. Chromatin remodeling complexes: strength in diversity, precision through specialization. *Curr Opin Genet Dev,* 15(2):185-90, 2005.

[17] Cheng Y, Church GM. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol,* 893-103, 2000.

[18] Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. SGD: Saccharomyces Genome Database. *Nucleic Acids Res,* 26(1):73-9, 1998.

[19] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell,* 2(1):65-73, 1998.

[20] Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics,* 19(4):459-66, 2003.

[21] DiMauro S, Schon EA. Nuclear power and mitochondrial disease. *Nat Genet,* 19(3):214-5, 1998.

[22] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res,* 30(1):207-10, 2002.

[23] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A,* 95(25):14863-8, 1998.

[24] Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika,* 10507-521, 1915.

[25] Foury F. Human genetic diseases: a cross-talk between man and yeast. *Gene,* 195(1):1-10, 1997.

[26] Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A,* 98(24):13784-9, 2001.

[27] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell,* 11(12):4241-57, 2000.

[28] Genespring. *http://www.silicongenetics.com*.

[29] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K,

Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. Functional profiling of the Saccharomyces cerevisiae genome. *Nature,* 418(6896):387-91, 2002.

[30] Grunstein M. Histone function in transcription. *Annu Rev Cell Biol,* 6643-78, 1990.

[31] Hibbs MA, Dirksen NC, Li K, Troyanskaya OG. Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics,* 6115, 2005.

[32] Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics,* 23(20):2692-9, 2007.

[33] Hibbs, Wallace, Dunham, Li, Troyanskaya. Viewing the Larger Context of Genomic Data through Horizontal Integration. *Proceedings of IEEE-CS 11th International Conference on Information Visualization (IV'07),* , 2007.

[34] Hochheiser H, Baehrecke EH, Mount SM, Shneiderman B. Dynamic querying for pattern identification in microarray and genomic data.

Proceedings of the International Conference on Multimedia and Expo, 3, 2003.

[35] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. *Cell,* 102(1):109-26, 2000.

[36] Hughes TR, Robinson MD, Mitsakakis N, Johnston M. The promise of functional genomics: completing the encyclopedia of a cell. *Curr Opin Microbiol,* 7(5):546-54, 2004.

[37] Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray data sets. *Bioinformatics,* 22(23):2890-7, 2006.

[38] Huttenhower C, Troyanskaya OG. Bayesian data integration: a functional perspective. *Comput Syst Bioinformatics Conf,* 341-51, 2006.

[39] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. *Nat Genet,* 31(4):370-7, 2002.

[40] Inadome H, Noda Y, Adachi H, Yoda K. Immunoisolaton of the yeast Golgi subcompartments and characterization of a novel membrane protein, Svp26, discovered in the Sed5-containing compartments. *Mol Cell Biol,* 25(17):7696-710, 2005.

[41] Java3D. *http://java.sun.com/products/java-media/3D*.

[42] Jaimovich A, Elidan G, Margalit H, Friedman N. Towards an integrated protein-protein interaction network: a relational Markov network approach. *J Comput Biol,* 13(2):145-64, 2006.

[43] JAva MAtrix Package (JAMA). *http://math.nist.gov/javanumerics/jama*.

[44] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* 302(5644):449-53, 2003.

[45] Johnson JE, Stromvik MV, Silverstein KA, Crow JA, Shoop E, Retzel EF. TableView: portable genomic data visualization. *Bioinformatics,* 19(10):1292-3, 2003.

[46] Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci U S A,* 98(16):8961-5, 2001.

[47] Kitano H. Computational systems biology. *Nature,* 420(6912):206-10, 2002.

[48] Koutnikova H, Campuzano V, Foury F, Dollé P, Cazzalini O, Koenig M. Studies of human, mouse and yeast homologues indicate a mitochondrial function for frataxin. *Nat Genet,* 16(4):345-51, 1997.

[49] Krantz M, Nordlander B, Valadi H, Johansson M, Gustafsson L, Hohmann S. Anaerobicity prepares Saccharomyces cerevisiae cells for faster adaptation to osmotic shock. *Eukaryot Cell,* 3(6):1381-90, 2004.

[50] Lanckriet GR, Deng M, Cristianini N, Jordan MI, Noble WS. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput,* 300-11, 2004.

[51] Le Crom S, Devaux F, Jacq C, Marc P. yMGV: helping biologists with yeast microarray data mining. *Nucleic Acids Res,* 30(1):76-9, 2002.

[52] Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science,* 306(5701):1555-8, 2004.

[53] Li K, Chen H, Clark D, Cook P, Damianakis S, Essl G, Finkelstein A, Funkhouser T, Klein A, Liu Z, Praun E, Samanta R, Shedd B, Singh JP, Tzanetakis G, Zheng J. Building and Using a Scalable Display Wall System. *IEEE Comput Graph Appl,* 20(4):29-37, 2000.

[54] Madeira C, Oliveira L. A Linear Time Biclustering Algorithm for Time Series Gene Expression Data. *Proc. of the 5th Workshop on Algorithms in Bioinformatics (WABI'05),* 39-52, 2005.

[55] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform,* 1(1):24-45, 2004.

[56] Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res,* 25(1):28-30, 1997.

[57] Méndez MA, Hödar C, Vulpe C, González M, Cambiazo V. Discriminant analysis to evaluate clustering of gene expression data. *FEBS Lett,* 522(1-3):24-8, 2002.

[58] *Engineering Statistics* John Wiley & Sons, Inc., New York (2001).

[59] Morrison JL, Breitling R, Higham DJ, Gilbert DR. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics,* 6233, 2005.

[60] Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding function: evaluation methods for functional genomic data. *BMC Genomics,* 7(1):187, 2006.

[61] Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG. Discovery of biological networks from diverse functional genomic data. *Genome Biol,* 6(13):R114, 2005.

[62] Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. *Bioinformatics,* 23(17):2322-30, 2007.

[63] Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics,* 21 Suppl 1i302-10, 2005.

[64] North C, Schneiderman B. A Taxonomy of Multiple Window Coordinations. *Technical Report,* CS-TR-3854, 1998.

[65] Owen AB, Stuart J, Mach K, Villeneuve AM, Kim S. A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome Res,* 13(8):1828-37, 2003.

[66] Pavlidis P, Weston J, Cai J, Noble WS. Learning gene functional classifications from multiple data types. *J Comput Biol,* 9(2):401-11, 2002.

[67] Peña-Castillo L, Hughes TR. Why are there still over 1000 uncharacterized yeast genes?. *Genetics,* 176(1):7-14, 2007.

[68] Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE. The core meiotic transcriptome in budding yeasts. *Nat Genet,* 26(4):415-23, 2000.

[69] Prokisch H, Scharfe C, Camp DG, Xiao W, David L, Andreoli C, Monroe ME, Moore RJ, Gritsenko MA, Kozany C, Hixson KK, Mottaz HM, Zischka H, Ueffing M, Herman ZS, Davis RW, Meitinger T, Oefner PJ, Smith RD, Steinmetz LM. Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol,* 2(6):e160, 2004.

[70] Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet,* 2(6):418-27, 2001.

[71] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput,* 455-66, 2000.

[72] Rees CA, Demeter J, Matese JC, Botstein D, Sherlock G. GeneXplorer: an interactive web application for microarray data visualization and analysis. *BMC Bioinformatics,* 5141, 2004.

[73] Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques,* 34(2):374-8, 2003.

[74] Saldanha AJ, Brauer MJ, Botstein D. Nutritional homeostasis in batch and steady-state culture of yeast. *Mol Biol Cell,* 15(9):4089-104, 2004.

[75] Saldanha AJ. Java Treeview--extensible visualization of microarray data. *Bioinformatics,* 20(17):3246-8, 2004.

[76] Saraiya P, North C, Duca K. An Evaluation of Microarray Visualization Tools for Biological Insight. *The IEEE Symposium on Information Visualization,* 1-8, 2004.

[77] Sealfon RS, Hibbs MA, Huttenhower C, Myers CL, Troyanskaya OG. GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics,* 7443, 2006.

[78] Seo, Shneiderman. Interactively Exploring Hierarchical Clustering Results. *IEEE Computer,* 35(7):80-86, 2002.

[79] Sharan R, Maron-Katz A, Shamir R. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics,* 19(14):1787-99, 2003.

[80] Shen X, Mizuguchi G, Hamiche A, Wu C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature,* 406(6795):541-4, 2000.

[81] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. *Nucleic Acids Res,* 29(1):152-5, 2001.

[82] Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schönfisch B, Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N,

Meisinger C. The proteome of Saccharomyces cerevisiae mitochondria. *Proc Natl Acad Sci U S A,* 100(23):13207-12, 2003.

[83] Singh KK. The Saccharomyces cerevisiae Sln1p-Ssk1p two-component system mediates response to oxidative stress and in an oxidant-specific fashion. *Free Radic Biol Med,* 29(10):1043-50, 2000.

[84] SPELL Supplementary Materials. *http://function.princeton.edu/SPELL/supplement/*.

[85] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell,* 9(12):3273-97, 1998.

[86] Spotfire. *http://www.spotfire.com*.

[87] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW. Systematic screen for human disease genes in yeast. *Nat Genet,* 31(4):400-4, 2002.

[88] Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. *Bioinformatics,* 18(1):207-8, 2002.

[89] Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics,* 18 Suppl 1S136-44, 2002.

[90] Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C.

Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science,* 294(5550):2364-8, 2001.

[91] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics,* 17(6):520-5, 2001.

[92] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. 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, 2003.

[93] Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO. Targets of the cyclin-dependent kinase Cdk1. *Nature,* 425(6960):859-64, 2003.

[94] van Attikum H, Gasser SM. ATP-dependent chromatin remodeling and DNA double-strand break repair. *Cell Cycle,* 4(8):1011-4, 2005.

[95] Wall E, Rechtsteiner, Rocha M. Singular Value Decomposition and Principal Component Analysis, in *A Practical Approach to Microarray Data Analysis* (2003).

[96] Wallace G, Anshus OJ, Bi P, Chen H, Chen Y, Clark D, Cook P, Finkelstein A, Funkhouser T, Gupta A, Hibbs M, Li K, Liu Z, Samanta R, Sukthankar R, Troyanskaya O. Tools and applications for large-scale display walls. *IEEE Comput Graph Appl,* 25(4):24-33, 2005.

[97] Wei B, Silva C, Koutsofios E, Krishnan S, North S. Visualization Research with Large Displays. *IEEE Comput Graph Appl,* 20(4):50-54, 2000.

[98] Werner-Washburne M, Wylie B, Boyack K, Fuge E, Galbraith J, Weber J, Davidson G. Comparative analysis of multiple genome-scale data sets. *Genome Res,* 12(10):1564-73, 2002.

[99] Wysocka M, Rytka J, Kurlandzka A. Saccharomyces cerevisiae CSM1 gene encoding a protein influencing chromosome segregation in meiosis I interacts with elements of the DNA replication complex. *Exp Cell Res,* 294(2):592-602, 2004.

[100] Xiao B, Wilson JR, Gamblin SJ. SET domains and histone methylation. *Curr Opin Struct Biol,* 13(6):699-705, 2003.

[101] Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics,* 17(4):309-18, 2001.