Sackler Faculty of Exact Sciences, Blavatnik School of Computer Science

Network-based algorithms for analysis of heterogeneous biomedical data

THESIS SUBMITTED FOR THE DEGREE OF

"DOCTOR OF PHILOSOPHY"

by

Igor Ulitsky

The work on this thesis has been carried out

under the supervision of **Prof. Ron Shamir**

Submitted to the Senate of Tel-Aviv University

July 2009

Acknowledgments

This thesis summarizes five very pleasant and productive years in my life. I would like to thank several people for making them so.

First and foremost, I would like to express my deepest gratitude to my advisor Ron Shamir for his leading me though the labyrinths of science with the optimal conditions and for allowing me to benefit from his ideas, knowledge and experience. Most importantly, for teaching me to conduct research with utmost integrity and thoroughness.

I had the pleasure and the honor of collaborating with a large number of gifted researches and great friends that I would like to thank. To Irit Gat-Viks, Yonit Halperin, Chaim Linhart, Ofer Lavi, Adi-Maron Katz, Seagull Shavit and Rani Elkon from Ron's group for teaching me so much, being great friends and for the pleasant morning trips to get coffee. To my lab mates Daniela Raijman, Michael Gutkin, Michal Ozery-Plato, Israel Steinfeld, Amos Tanay, Michal Ziv-Ukelson, Falk Hueffner, Sharon Bruckner, Guy Karlebach, Shahar Kidron and Lior Mechlovich. Thanks for the fruitful conversations, advices, and especially for the laughs, arguments, and the great atmosphere in the lab. To Dudu Burstein for being a great friend and for the walks around campus. To Benny Chor, for teaching me that there is always another point of view. Last but not least, to the rest of my collaborators, from all of which I learnt so much, Louise Laurent, Franzef Müller, Jeanne Loring (Scripps Institute, La Jolla), Gal Romano, Martin Kupiec (Life Sciences, TAU), Tomer Shlomi (Technion), Tal Elkan, Keren Avraham and Yossi Shiloh (School of Medicine, TAU).

I would like to thank my family for their unconditional and endless support. My wife and best friend Liad without whom none of this would have been possible, my son Mattan for giving me so much happiness, and my parents for always being proud.

Finally, I deeply thank the Edmond J. Safra foundation for their financial support over the past four years. In addition, my research was partly supported by the Legacy Heritage Fund and the European Commission.

Preface

This thesis is based on the following collection of seven articles that were published throughout the PhD period in scientific journals and in refereed proceedings of conferences.

1. **Identification of Functional Modules using Network Topology and High-Throughput Data**

Igor Ulitsky and Ron Shamir

Published in *BMC Systems Biology* [1].

2. **Regulatory Networks Define Phenotypic Classes of Human Stem Cell Lines**

Franz-Josef Müller, Louise C. Laurent, Denis Kostka, Igor Ulitsky, Ron Williams, Cristina Lu, Mahendra S. Rao, Ron Shamir, Philip H. Schwartz, Nils O. Schmidt, Jeanne F. Loring

Published in *Nature* [2].

3. **Detecting Disease-Specific Dysregulated Pathways via Analysis of Clinical Expression Profiles**

Igor Ulitsky, Richard M. Karp and Ron Shamir

Published in *Proceedings of the Eleventh Annual International Conference on Research in Computational Molecular Biology* (*RECOMB 2008*) [3].

4. **Detecting Pathways Transcriptionally Correlated with Clinical Parameters**

Igor Ulitsky and R. Shamir

Published in *Proceedings of Computational Systems Bioinformatics (CSB) 2008* [4].

5. **Identifying Functional Modules Using Expression Profiles and Confidence-Scored Protein Interactions**

Igor Ulitsky and Ron Shamir

Published in *Bioinformatics* [5].

6. **Pathway Redundancy and Protein Essentiality Revealed in the *S. cerevisiae* Interaction Networks**

Igor Ulitsky and Ron Shamir

Published in *Molecular Systems Biology* [6].

7. From E-MAPs to Module Maps: Dissecting Quantitative Genetic Interactions Using Physical Interactions

Igor Ulitsky, Tomer Shlomi , Martin Kupiec and Ron Shamir

Published in *Molecular Systems Biology* [7].

Abstract

Technological breakthroughs in the past decade have enabled the collection of biomedical data on an unprecedented scale. Several methods are used to map diverse interactions among genes or gene products and their results can be collected into genome-wide networks. Other methods measure the activity or abundance of genes across different conditions. The data obtained by most available techniques are noisy, heterogeneous and difficult to interpret, making data integration essential. In addition, most cellular functions rely on the coordinated action of the products of multiple genes, often referred to as functional modules. A major goal of computational systems biology is to enable the delineation of these modules and to facilitate extraction of biological insights from the deduced modular structures. In this thesis we describe several methods for extraction of functional modules using heterogeneous data. We specifically focus on analysis of protein and genetic interaction networks and gene expression data, but our computational methodologies are suitable for handling additional data types. We demonstrated the effectiveness of these methods on a variety of biological systems in yeast and human. Our results include predictions of functions for genes and gene groups, delineation of novel pathways and complexes and the relationships between them, and prediction of functionally important interactions. In the context of human disease, the modules we identify provide a signature of the disease potentially useful for diagnosis, pinpoint possible pathways affected by the disease, and suggest targets for drug intervention.

Contents

1. Introduction

1.1 Systems biology and modularity

Biological research has undergone a paradigm shift over the last decade. Dramatic improvements in DNA sequencing enabled the determination of whole genome sequences, allowing us to identify the biological components serving as building blocks for cells and organisms. In parallel, other biotechnological breakthroughs allowed high-throughput screening of the abundance, localization and inter-relations of these building blocks. These developments enable researchers to inspect the cellular machinery at a genome-wide scale with an increasingly fine resolution, introducing systems biology [8, 9] as a new interdisciplinary science. Systems biology takes a holistic view of the biological system, trying to determine its global principles in order to understand and predict its behavior.

One of the most extensively studied features of biological systems is their modularity. Most cellular functions rely on the coordinated action of the products of multiple genes, often referred to as *functional modules* [10]. The physiological functions of cells and organisms can be viewed as the coordinated and integrated activity of multiple genetic circuits. A major goal of computational systems biology is to enable the delineation of these modules and to facilitate extraction of biological insights from the deduced modular structures.

1.2 Biological networks

Diverse biological information can be represented as networks of interactions among genes or their protein products. In principle, if our knowledge of the cellular biology was complete, we could construct and use a genome-wide regulatory network that could predict, given the extracellular stimuli, the abundance, localization and activity of each molecule in the cell at any time point. The accuracy of these predictions would in theory be limited only by the inherent biological noise. Unfortunately, such a network is not yet available for any cell type, and we are still probably decades or even centuries away from being able to construct such a network for mammalian cells. In the meantime, many types of data derived using functional genomics technologies can be represented as networks, most of which are partial and noisy fragments of the complete gene regulatory network. Such networks can be constructed using both traditional low-throughput methods and novel post-genomic techniques. These networks summarize the knowledge on interactions between pairs of genes or

their protein products. The combination of results from high- and low-throughput experiments leads to the most comprehensive networks available in the public domain today. The following types of interactions are typically measured and studied today:

- **Protein-Protein interactions (PPIs)**: Two proteins have an interaction if there is a physical contact between them. Such interactions can now be derived using genome-wide screens, such as yeast two-hybrid or co-immunoprecipitation [11]. High-throughput PPI screens were performed in several organisms, with the largest focus on the yeast *S. cerevisiae* [12-14] and human [15-17]. In addition, PPIs from small-scale experiments curated from the literature can be combined into large scale networks [18, 19]. High-throughput methods for measuring PPIs are very useful, but notorious for their high rates of false positive and false negative calls [20, 21].
- **Protein-DNA interactions (PDIs)**: Interactions reflecting physical contact between a transcription factor (TF) protein and a DNA sequence (typically the promoter sequence of a gene) can also be analyzed in a genome-wide fashion [22]. The established method for high-throughput measurements of such interactions is ChIP-chip [23, 24], but novel technologies such as ChIP-seq [25] and protein binding microarrays (PBMs) [26] are also becoming widely used. Experimental evidence about PDIs is currently available on a large scale in the yeast *S. cerevisiae* and for a limited number of TFs in mammals. These interactions can also be predicted computationally using the known binding preferences of transcription factors [27]. Note that unlike PPIs, PDIs are inherently asymmetric.
- **Metabolic interactions**: The most comprehensive biological networks today are metabolic networks. Basic cellular metabolism has been extensively studied for decades and information about the interactions within diverse pathways is deposited in highly curated databases, such as KEGG [28]. A metabolic interaction between genes typically connects two enzymes catalyzing successive reaction steps in some metabolic pathway. An alternative representation shows the inputs and outputs of a reaction (reactants) as nodes connected by an edge and the enzymes regulating the reaction.
- **Signaling interactions**: One of the key goals of systems biology is the reconstruction of the combinatorial regulation determining the

abundance and activity of each gene. PDIs form just one layer of signaling interactions. Mapping of other layers is also carried out, including kinase-substrate [29, 30] and microRNA-target networks [31]. Combination of the layers leads to multi-level signaling networks, a useful resource for studying cellular responses [32, 33].

An additional class of biological networks, genetic interaction networks, will be described in detail in section 1.4.

1.2.1 Computational methods for analysis of gene networks

The intriguing properties and the vast potential of biological networks triggered the development of numerous novel computational methods. Initial studies focused on basic network properties and studied various topological coefficients [34], abundance of small network motifs [35] and network evolution [36]. Dozens of computational methods were developed for detection of dense subnetworks in PPI networks, aiming to detect novel protein complexes and predict protein function (reviewed in [37]). Another set of methods aimed to detect paths in PPI networks that may correspond to linear signaling cascades [38, 39]. Importantly, gene networks have been repeatedly shown to be highly useful for interpretation of other genomic data. The use of PPI and PDI networks for the interpretation of gene expression and genetic interactions is described in section 1.4.1. In addition, gene networks were successfully used for analysis of genotypes [40, 41], deletion phenotypes [42, 43] and human disease [44, 45].

One of the obstacles to exploiting networks based on high-throughput experiments in general, and PPI networks in particular, is their high rate of false positive and false negative interactions [20, 46]. To better handle uncertainty in PPIs, several works devised probabilistic schemes to estimate the confidence of individual interactions [46-50].

1.3 High-throughput transcriptome profiling

One of the most dramatic developments in molecular biology over the past 15 years is the introduction of the microarray technology [51, 52]. Abstractly, a microarray is a dense array of oligonucleotides that can be used as probes for measuring the abundance of nucleic acids. To date, this technology has been mainly used for high-throughput profiling of mRNA abundance in the cell. A typical microarray profile gives the expression levels of several thousands of genes under a particular condition. As of April 2009, nearly 300,000 profiles were available in NCBI Gene Expression Omnibus (GEO) database [53].

High-throughput profiling of mRNA abundance can be used for answering diverse biological questions. In one type of experiment, typically called *time course experiment*, cells are exposed to diverse treatments and mRNA levels are measured at several time points after the treatment. Classical studies used this approach to study progression through cell cycle [54, 55] and diverse stress responses [56-59]. In another type of experiment, expression profiles of cells taken from different tissues or different populations are compared [60, 61]. Clinical microarray studies typically compare profiles of tissues taken from individuals with different pathological status [62-65].

1.3.1 Computational analysis of microarray data

The high volume of microarray data (in a typical human study the levels of ~20,000 transcripts are measured in tens to hundreds of samples) and the technical and biological noise in them require dedicated computational analysis tools. Hundreds of statistical and computational methods were developed for various subtasks of microarray analysis [66]. The low level analysis tasks include image analysis, the normalization of the data and removal of cross-hybridization effects. The high-level analysis tasks can be crudely divided into two types: unsupervised and supervised. In an *unsupervised* analysis, no prior knowledge about the nature of the samples is assumed, and clustering [67] or biclustering [68] methods are used to find groups of genes and/or samples that exhibit similar expression patterns. In a typical *supervised* analysis, two or more sample groups are compared, and the goal is to identify genes whose expression pattern can distinguish among those groups [69]. In other cases, the goal is to identify genes significantly related to some quantitative parameter describing the samples. As it is frequently difficult to suggest novel hypotheses based on individual genes with good accuracy, it is also desirable to identify differentially expressed sets of genes, or pathways. By considering together the whole pathway, correlations that would have been missed if we tested each gene separately can be revealed. One approach to this problem uses predefined gene sets describing pathways and quantifies the change in their expression levels [70-72]. Additional works proposed measures for scoring expression activity and co-expression in metabolic pathways [73, 74], complexes [75] and network neighborhoods [76]. Vert and Kanehisa [77] used kernel methods to identify expression patterns that characterize gene sets matching pathways in a given network. The drawback of all these approaches is that pathway boundaries are often difficult to assign, and in many cases only a fraction of the pathway is altered during disease. Moreover, activity changes in unknown pathways are

impossible to detect using this approach. As we shall describe below, using gene networks, one can address both of these problems to some extent.

1.3.2 Combining results from microarray studies in the context of human disease

To date, the main goal of the majority of microarray studies focused on human disease has been to identify a set of biomarkers that can be useful for diagnosis or prognosis [62]. In addition, these studies have the potential to improve the mechanistic understanding of the disease on the molecular level [78]. The current methods for analyzing disease data are mostly focused on comparing expression profiles derived in a single experiment, usually comparing several samples derived from diseased specimens and from normal specimens. A large variety of techniques from machine learning were applied to such data, with a primary focus on feature selection, i.e., detection of a gene subset that best distinguishes between the healthy and the diseased tissues [62, 79]. While the classification and feature selection problems in a specific study are of obvious interest, less effort was devoted to systematic computational analysis of disease-related data aiming to shed light on the molecular characteristics of the disease in question. Such analysis is especially compelling when multiple independently derived datasets relevant to the same disease are available.

How can one combine results of multiple studies comparing the same sample groups? The simplest approach is to compare the lists of ``interesting'' genes identified in different studies [80]. The prevalent methodology for identifying biological phenomena by combining expression profiles from multiple studies and platforms obtained under similar conditions is *meta-analysis*, which involves identifying the genes studied in all the experiments and combining their respective significance from all the experiments in a statistical framework [81, 82].

Despite the success of these analyses, they are generally only applicable when all the datasets compare similar sets of conditions, and the variance between the datasets stems mainly from using different platforms or from the differences between the populations profiled in each experiment. Thus, they cannot handle situations where the data are coming from diverse tissues, diverse diseases and different organisms. A more detailed modeling of the data is required in these cases.

Segal et al. [83] described a methodology for combining a large number of independent cancer-related datasets to identify modules of genes significantly

altered in a subset of human cancer subtypes. The proposed method starts from a large collection of biologically-relevant gene sets and iteratively refines them alongside the selection of the condition subset relevant to the module. A similar approach was recently proposed by Tomlins et al. [84], where 14,000 distinct gene sets from different perspectives and origins were used to analyze gene expression studies related to prostate cancer.

In contrast to using predefined gene sets, several works focused on identifying gene sets *de novo* from a large compendium of expression data [85-87]. Such a framework usually involves applying a common normalization to all the expression datasets and subsequent application of data mining algorithms for identification of modules, consisting of subsets of genes alongside corresponding subsets of the experimental conditions in which these genes show a coherent expression pattern.

Another available approach is to focus on the co-expression exhibited by gene pairs in the different experiments, postulating that functionally related genes are likely to exhibit co-expression in a significant number of experiments. 2nd-order expression analysis, described in [88], follows this paradigm. First, pairs of genes tightly co-expressed in several datasets, termed *doublets*, are identified. Then, pairs of such doublets exhibiting simultaneous high or low correlations across multiple datasets are sought, through the inspection of the 2nd-order correlations. This method can be generalized to find gene sets of arbitrary size, rather than gene pairs. In the CONDENSE algorithm [89], the transcription data from each study are converted into a co-expression network. These networks are then aggregated together in a single network, and dense patterns are sought in it.

1.3.3 Combined analysis of interaction networks and expression profiles

As described above, many fruitful algorithmic approaches were developed for dissection of network and expression data separately. However, methods analyzing together both information types have the potential to be more powerful, e.g., by highlighting biologically relevant expression changes that are too weak to be detected using only expression data, but emerge when that data are projected on the network structure. It has been established that genes connected by a PPI are more likely to be co-expressed [90, 91]. Exploiting this interconnection, co-expression is frequently used together with other genomic evidence for predicting PPIs (cf. [92]). PPI networks were shown to be very useful in interpreting gene expression data by improving sample classification using microarray data [93-95] and improving detection of differentially expressed genes [96-98].

One of the most heavily studied ways of combining network and gene expression data is detection of subnetworks with a particular expression behavior. These approaches can be crudely divided into four groups:

- **Detection of subnetworks active in a single condition**: Topological properties of interaction networks induced by genes active in one specific condition were studied [34, 99-101], and highlighted several basic biological principles such as just-in-time complex assembly [99], hub transience [100] and high centrality of cancer-related genes [101].
- **Subnetworks active in a subset of the studied conditions:** Ideker et al. [102] introduced a successful algorithm for identification of *active subnetworks*, i.e., connected regions of the network that show significant changes in expression over a particular subset of the conditions. The same methodology was recently used in [103], utilizing shortest-path algorithms for module finding. Other groups described additional extensions of this method [104-106].
- **Detection of subnetworks with correlated expression across all the conditions**: Several approaches sought modules by jointly analyzing network information with co-expression across the entire dataset. The *Co-clustering* methodology [107] uses a distance function that combines similarity of gene expression profiles with network topology. The network distance between two nodes is an edge-weighted version of their topological distance in the network. The expression distance is one minus the Pearson correlation between the expression patterns. The two distances are combined into a similarity score, and standard hierarchical algorithms [108] are used for clustering. While generally successful, this approach sometimes produces clusters corresponding to highly disconnected subnetworks, since the network is only used as one of the sources of distance information, without requiring connectivity. Guo et al. [109] defined an expression similarity score, based on Pearson correlation of the expression patterns among interacting genes. The significance of a subnetwork was evaluated by comparing its score to random subnetworks with the same number of edges. Because the score of Guo et al. takes into account only expression similarity among interacting genes, some genes (those far from each other in the subnetwork) may exhibit low co-expression. Segal et al. [110] provided an interesting formulation of the integration problem, in which a module is expected to contain a significant portion of the possible interactions. A

probabilistic graphical model was used to extract a prespecified number of modules from gene expression measurements combined with a protein interaction dataset. Finally, the problem of clustering attribute data with connectivity constraints has also been studied in non-biological context [111].

- **Subnetworks distinguishing among sample groups**: These methods take as an input, in addition to the network and microarray data, two groups of samples, and the goal is to identify subnetworks whose expression distinguishes among them. Breitling et al. [112] proposed a simple method named GiGA which receives a list of genes ordered by differential expression scores (e.g., *t*-test *p*-values) and extracts subnetworks corresponding to the top scoring genes. Nacu et al. [113] proposed Gene eXpression Network Analysis (GXNA) which uses a *t*-statistic based score and a search heuristic similar to that used in [102] to identify subnetworks containing genes that, on average, exhibit significant differential expression. A more complex statistical approach was recently proposed by Dittrich et al. [114]. In this method, weights are assigned for individual nodes based on the significance of their differential expression, and integer linear programming (ILP) is used to identify heavy subnetworks.

1.4 Exploring the fitness of single and double deletion mutants

Recent years have seen the emergence of novel technologies for high-throughput measurements of the fitness of single deletion mutants under diverse experimental conditions. These studies mostly focused on the yeast *S. cerevisiae* [115-118]. Such experimental essays are capable of determining the marginal contributions of individual genes towards the susceptibility of the organisms to changes in their environment. These studies have shown that only ~18% of *S. cerevisiae* genes are essential for growth on a rich medium [115]. Consequently, genetic *buffering*, in which partial redundancy between genes masks the effect of deleting one of them, is believed to be abundant in eukaryotes [119].

Genetic interactions (GIs) convey information about the phenotype of a double mutant in comparison to the phenotypes of single mutants. GIs can be crudely classified into positive (alleviating), neutral and negative (aggravating) interactions [120, 121]. In a *negative interaction*, the fitness of the double mutant is lower than expected given that of the single mutants. The most extreme example of a negative interaction is *synthetic lethality*, in which the joint deletion of two nonessential genes leads to a lethal phenotype. In a *positive interaction*, on

the other hand, the double mutant is healthier than expected. The 'expected' fitness is usually defined using a multiplicative model, as the product of the fitnesses of the single mutants [120, 122, 123].

The first high-throughput screens for double knock-outs in S. cerevisiae used the SGA [124, 125] and dSLAM [126] methods and identified many events of synthetic lethality and synthetic sickness, which is a qualitative term for a non-lethal, but significant, negative GI. Additional recent technological advances allowed high-throughput quantitative measurements of both negative and positive interactions [123, 127-129].

1.4.1 Computational analysis of GI networks

Initial studies have shown that proteins in the same region of the GI network are slightly more likely to physically interact [124, 125], and that a protein with many PPIs is likely to have also many GIs [130]. These findings suggested that integration of physical and genetic networks can lead to novel biological insights. The term *physical interactions* (PIs) usually refers to both PPIs and PDIs. Kelley and Ideker [131] defined a module (or pathway) as a group of proteins that are densely interconnected in the PI network, and studied the frequency of GIs within and between such modules. In a systematic analysis of large scale GI and PI data they concluded that between-pathway explanations of GIs are ≈ 3.5 times more abundant than within-pathway explanations, and that GIs mostly bridge redundant processes. Further arguments for the prevalence of between-pathway GIs were given by Ye et al. [132], who postulated that genes in the same pathway are expected to share common GI partners, and used similarity of GI patterns for successful function prediction. A module-based approach was also proposed for identifying groups of genes sharing a set of common negative GI partners [133]. Finally, Zhang et al. [134] and Le Meur and Gentleman [135] identified pairs of known complexes with many negative GIs between them.

While the majority of negative interactions occur between partially redundant pathways, within-pathway negative interactions also exist: mutations in one of the two subunits of the same complex may have only a mild phenotype, as long as the complex survives. However, deletion of both subunits may lead to a complex failure and to an aggravating phenotype. Positive interactions were shown to occur mostly within pathways [127]. Most of the positive are probably the result of a drastic effect of any of the single deletions on pathway activity, which abolishes the effects of additional deletions.

Standard hierarchical clustering was initially used for the analysis of quantitative GI data [123, 127, 136]. Two studies have proposed dedicated approaches for analyzing such data. Bandyopadhyay et al. proposed an agglomerative clustering technique for clustering together PPI data and quantitative GIs [137], and Pu et al. proposed a biclustering algorithm for quantitative GIs, that is capable of identifying overlapping gene modules [138]. Other computational methods used the GI network to predict of genetic novel GIs [139-143], to predict protein function [132] and to predict of genes targeted by chemical compounds [144].

1.5 Summary of articles included in this thesis

1. **Identification of Functional Modules using Network Topology and High-Throughput Data**

Igor Ulitsky and Ron Shamir

Published in *BMC Systems Biology* [1].

> Usually, separate and different analysis methodologies are applied to interaction networks and gene expression data. An integrated investigation of both data types can improve the quality of the analysis by accounting simultaneously for topological network properties alongside intrinsic features of the high-throughput data. We describe a novel algorithmic framework for this challenge. We first transform the high-throughput data into similarity values, (e.g., by computing pairwise similarity of gene expression patterns from microarray data). Then, given a network of genes or proteins and similarity values between some of them, we seek connected sub-networks (or modules) that manifest high similarity. We develop algorithms for this problem and evaluate their performance on the osmotic shock response network in *S. cerevisiae* and on the human cell cycle network. We demonstrate that focused, biologically meaningful and relevant functional modules are obtained. In comparison with extant algorithms, our approach has higher sensitivity and higher specificity.

2. **Regulatory Networks Define Phenotypic Classes of Human Stem Cell Lines**

Franz-Josef Müller, Louise C. Laurent, Denis Kostka, Igor Ulitsky, Ron Williams, Cristina Lu, Mahendra S. Rao, Ron Shamir, Philip H. Schwartz, Nils O. Schmidt, Jeanne F. Loring

Published in *Nature* [2].

> Stem cells are defined as self-renewing cell populations that can differentiate into multiple distinct cell types. However, hundreds of different human cell lines from embryonic, fetal and adult sources have been called stem cells, even though they range from pluripotent cells-typified by embryonic stem cells, which are capable of virtually unlimited proliferation and differentiation-to adult stem cell lines, which can generate a far more limited repertoire of differentiated cell types. The rapid increase in reports of new sources of stem cells and their anticipated value to regenerative medicine has highlighted the need for a general, reproducible method for classification of these cells. We

created and analyzed of a database of global gene expression profiles (which we call the 'stem cell matrix') that enables the classification of cultured human stem cells in the context of a wide variety of pluripotent, multipotent and differentiated cell types. Using an unsupervised clustering method to categorize a collection of approximately 150 cell samples, we discovered that pluripotent stem cell lines group together, whereas other cell types, including brain-derived neural stem cell lines, are very diverse. Using further bioinformatic analysis we uncovered a protein-protein network (PluriNet) that is shared by the pluripotent cells (embryonic stem cells, embryonal carcinomas and induced pluripotent cells). Analysis of published data showed that the PluriNet seems to be a common characteristic of pluripotent cells, including mouse embryonic stem and induced pluripotent cells and human oocytes. Our results offer a new strategy for classifying stem cells and support the idea that pluripotency and self-renewal are under tight control by specific molecular networks.

3. **Detecting Disease-Specific Dysregulated Pathways via Analysis of Clinical Expression Profiles**

Igor Ulitsky, Richard M. Karp and Ron Shamir

Published in *Proceedings of the Eleventh Annual International Conference on Research in Computational Molecular Biology* (*RECOMB 2008*) [3].

We present a method for identifying connected gene subnetworks significantly enriched for genes that are dysregulated in specimens of a disease. These subnetworks provide a signature of the disease potentially useful for diagnosis, pinpoint possible pathways affected by the disease, and suggest targets for drug intervention. Our method uses microarray gene expression profiles derived in clinical case-control studies to identify genes significantly dysregulated in disease specimens, combined with protein interaction data to identify connected sets of genes. Our core algorithm searches for minimal connected subnetworks in which the number of dysregulated genes in each diseased sample exceeds a given threshold. We have applied the method in a study of Huntington's disease caudate nucleus expression profiles and in a meta-analysis of breast cancer studies. In both cases the results were statistically significant and appeared to home in on compact pathways enriched with hallmarks of the diseases.

4. **Detecting Pathways Transcriptionally Correlated with Clinical Parameters**

Igor Ulitsky and R. Shamir

Published in *Proceedings of Computational Systems Bioinformatics (CSB) 2008* [4].

We describe a novel methodology for extraction of connected network modules with coherent gene expression patterns that are correlated with a specific clinical parameter. Our approach suits both numerical (e.g., age or tumor size) and logical parameters (e.g., gender or mutation status). We demonstrate the method on a large breast cancer dataset, where we identify biologically-relevant modules related to nine clinical parameters including patient age, tumor size, and metastasis-free survival. Our method is capable of detecting disease-relevant pathways that could not be found using other methods. Our results support some previous hypotheses regarding the molecular pathways underlying diversity of breast tumors and suggest novel ones.

5. **Identifying Functional Modules Using Expression Profiles and Confidence-Scored Protein Interactions**

Igor Ulitsky and Ron Shamir

Published in *Bioinformatics* [5].

The analysis of expression data can be improved by its integration with protein interaction networks, but the performance of these analyses has been hampered by the uneven quality of the interaction data. We present CEZANNE, a confidence-based method for extraction of functionally coherent co-expressed gene sets. CEZANNE uses probabilities for individual interactions, which can be computed by any available method. We propose a probabilistic model and a weighting scheme in which the likelihood of the connectivity of a subnetwork is related to the weight of its minimum cut. Applying CEZANNE to an expression dataset of DNA damage response in *S. cerevisiae*, we recover both known and novel modules and predict novel protein functions. We show that CEZANNE outperforms previous methods for analysis of expression and interaction data.

6. **Pathway Redundancy and Protein Essentiality Revealed in the *S. cerevisiae* Interaction Networks**

Igor Ulitsky and Ron Shamir

Published in *Molecular Systems Biology* [6].

In this study we devised novel analytic tools for interpreting genetic interactions in a physical context. We extend the model of Kelley and Ideker to analyze together genetic and physical networks, which explains many of the known genetic interactions as linking different pathways in the physical network. Applying these tools on a large-scale *Saccharomyces cerevisiae* data set, our analysis revealed 140 between-pathway models that explain 3,765 genetic interactions, roughly doubling those that were previously explained. Model genes tend to have short mRNA half-lives and many phosphorylation sites, suggesting that their stringent regulation is linked to pathway redundancy. We also identify 'pivot' proteins that have many physical interactions with both pathways in our models, and show that pivots tend to be essential and highly conserved. Our analysis of models and pivots sheds light on the organization of the cellular machinery as well as on the roles of individual proteins.

7. From E-MAPs to Module Maps: Dissecting Quantitative Genetic Interactions Using Physical Interactions

Igor Ulitsky, Tomer Shlomi , Martin Kupiec and Ron Shamir

Published in *Molecular Systems Biology* [7].

Here, we extended the model used in the previous study in two ways: the new method can identify a collection of functional modules rather than module pairs, and it can use quantitative genetic interaction data, using both positive and negative interactions. We used the method to build a module map of yeast chromosome biology and show how it provides clues for the elucidation of function both at the level of individual genes and at the level of functional modules.

2. Identification of Functional Modules using Network Topology and High-Throughput Data

BMC Systems Biology

BioMed Central

Research article

Identification of functional modules using network topology and high-throughput data

Igor Ulitsky and Ron Shamir*

Address: School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Email: Igor Ulitsky - ulitskyi@tau.ac.il; Ron Shamir* - rshamir@tau.ac.il

* Corresponding author

Published: 26 January 2007

BMC Systems Biology 2007, **1**:8 doi:10.1186/1752-0509-1-8

Received: 18 September 2006

Accepted: 26 January 2007

This article is available from: http://www.biomedcentral.com/1752-0509/1/8

Abstract

Background: With the advent of systems biology, biological knowledge is often represented today by networks. These include regulatory and metabolic networks, protein-protein interaction networks, and many others. At the same time, high-throughput genomics and proteomics techniques generate very large data sets, which require sophisticated computational analysis. Usually, separate and different analysis methodologies are applied to each of the two data types. An integrated investigation of network and high-throughput information together can improve the quality of the analysis by accounting simultaneously for topological network properties alongside intrinsic features of the high-throughput data.

Results: We describe a novel algorithmic framework for this challenge. We first transform the high-throughput data into similarity values, (e.g., by computing pairwise similarity of gene expression patterns from microarray data). Then, given a network of genes or proteins and similarity values between some of them, we seek connected sub-networks (or modules) that manifest high similarity. We develop algorithms for this problem and evaluate their performance on the osmotic shock response network in *S. cerevisiae* and on the human cell cycle network. We demonstrate that focused, biologically meaningful and relevant functional modules are obtained. In comparison with extant algorithms, our approach has higher sensitivity and higher specificity.

Conclusion: We have demonstrated that our method can accurately identify functional modules. Hence, it carries the promise to be highly useful in analysis of high throughput data.

Background

The accumulation of large-scale interaction data on multiple organisms, such as protein-protein and protein-DNA interactions, requires novel computational techniques that will be able to analyze these data together with information collected through other means. Such methods should enable thorough dissection of the data, whose dimensions have already extended far beyond the scope that is amenable to traditional analysis and manual interpretation. An important class of such biological information can be represented in the form of similarity relations. Quantitative molecular data, such as mRNA expression profiles, are often analyzed in this context through clustering algorithms. Similarity between genes can also be defined on other levels, such as function [1] or transcription factor binding patterns [2].

Although many fruitful algorithmic approaches have been developed for dissection of network and similarity data separately, methods analyzing together both information sources hold much promise. Several works have established the interconnection between expression profile similarity and protein interactions [3,4]. To exploit this interconnection, pairwise gene expression similarities

have been used together with other data sources for predicting pairwise protein interactions (e.g., [5]). Topological properties of interaction networks induced by genes active in different conditions were studied [6-9]. Several software tools allow the visual inspection of the clustering results in a network context [10]. However, ignoring the network information in the clustering process and using the rich and constantly growing network information solely for cluster evaluation seems suboptimal, as the network information can improve the cluster identification process. The prevalence of modularity in molecular cell biology has been widely recognized in the last decade. By *functional module* one typically means a group of cellular components and their interactions that can be attributed a specific biological function [11]. Several approaches sought modules by jointly analyzing network information with gene expression data. Initial works [12,13] proposed measures for scoring expression activity in metabolic pathways (e.g. KEGG database [14]) and complexes [15]. Vert and Kanehisa [16] used kernel methods to identify expression patterns that characterize gene sets matching pathways in a given network.

The *Co-clustering* methodology [17] uses a distance function that combines similarity of gene expression profiles with network topology. The network distance between two nodes is an edge-weighted version of their topological distance in the network. The expression distance is one minus the Pearson correlation between the expression patterns. The two distances are combined into a similarity score, and standard hierarchical algorithms [18] are used for clustering. While generally successful, this approach sometimes produces clusters corresponding to highly disconnected subnetworks, since the network is only used as one of the sources of distance information, without requiring connectivity.

Ideker *et al.* [19] introduced a successful algorithm for identification of *active subnetworks*, i.e., connected regions of the network that show significant changes in expression over a particular subset of the conditions. Unfortunately, this method can be used only when one has an activity p-value for every measurement, a situation which is rather uncommon. In addition, the method cannot handle pairwise gene similarity input. The same methodology was recently used in [20], utilizing shortest-path algorithms for module finding. Segal *et al.* [21] provided another interesting formulation of the integration problem, in which a module is expected to contain a significant portion of the possible interactions. A probabilistic graphical model was used to extract a prespecified number of modules from gene expression measurements combined with a protein interaction dataset.

In this study we seek functional modules by identifying connected subnetworks in the interaction data that exhibit high average internal similarity. We call such a module a *Jointly Active Connected Subnetwork (JACS)*. By imposing network topology constraints on clusters of expression data, the biological interpretation of the clusters becomes easier, and, as we shall see, one can detect weaker signals that were indistinguishable by extant methods.

We develop a novel computational method for efficient detection and analysis of JACSs, implemented in a program called MATISSE (Module Analysis via Topology of Interactions and Similarity SEts). The proposed methodology has a statistical basis, which allows confidence estimation of the results. The algorithm assumes no prior knowledge on the number of JACSs, and allows imposing constraints on their size. We do not require precalculation of the statistical significance of expression values. The methodology is general enough to suit any type of network data overlaid with pairwise similarities.

Our algorithm detects JACSs by identifying heavy subgraphs in an edge-weighted similarity graph while maintaining connectivity in the interaction network. By transforming edge weights to attain probabilistic meaning, we are actually seeking subnetworks of maximum likelihood. We show that this problem is computationally hard, devise several heuristic methods and analyze their practical performance.

When using gene expression similarity, analysis of known pathways in yeast has shown that only a fraction of the genes in a pathway are usually coherently regulated at the transcription level (and thus highly similar) [22]. Our method allows assignment of different priors to different genes, reflecting their probability to be regulated at the transcription level. We believe this is the first study to allow such flexibility. In addition, the goal of our approach is to detect non-overlapping JACSs rather than to partition all the genes into clusters.

We first evaluate the performance of our algorithm on synthetic data with planted modules, and verify its ability to recover planted modules with high accuracy. Then, we analyze two real systems for which large datasets are available: the osmotic shock response of *S. cerevisiae*, and the cell cycle in human HeLa cells. For *S. cerevisiae*, we compiled and carefully annotated from diverse sources a protein-protein and protein-DNA interaction network consisting of 6,230 nodes and 89,327 interactions. The performance of MATISSE is shown to exceed that of extant analysis schemes in terms of the ability to retrieve biologically relevant groups, as analyzed by four different anno-

tation datasets. We identify specific subnetworks relevant to different processes that are known to be activated and repressed by the MAPK cascades following osmotic shock, such as ergosterol biosynthesis and pheromone response. In addition, we identify novel pathways, such as pyridoxine metabolism, as differentially expressed during osmotic shock. Detailed analysis shows that some of the involved processes can not be detected based on the expression data alone. The human network contains 9,135 nodes and 25,086 protein-protein interactions collected from several sources, including recently published studies [23,24]. Our analysis identifies subnetworks active in specific phases of the human cell cycle. These results underly the ability of our approach to provide novel, previously undetected biological insights. The inspection of "hubs" in the subnetworks delineated by MATISSE reveals key regulators of the cell cycle.

Results and discussion

A framework for detection of jointly active subnetworks

Let us first state our problem abstractly. We are given an undirected *constraint graph* $G^C = (V, E)$, a subset $V_{sim} \subseteq V$ and a symmetric matrix S where S_{ij} is the *similarity* between $v_i, v_j \in V_{sim}$. The goal is to find disjoint subsets $U_1, U_2, \ldots, U_m \subseteq V$ called *JACSs*, so that each JACS induces a connected subgraph in G^C and contains elements that share high similarity values. We call the nodes in V_{sim} *front nodes* and nodes in $V \backslash V_{sim}$ *back nodes*.

In the biological context, V represents genes or gene products (we shall use the term 'gene' for brevity), and E represents interactions between them. These can be known protein-protein or protein-DNA interactions or alternatively can originate from a known regulatory network where arc orientations are ignored. S_{ij} measures the similarity between genes i and j, e.g., the Pearson correlation between their gene expression patterns. The set V_{sim} may be smaller than V as some of the genes may be absent from the array, and others may show insignificant expression patterns across the tested conditions and thus excluded. Hence, a JACS aims to capture a set of genes that have highly similar behavior, and are also topologically connected, and thus may share a common function, e.g., belong to a single complex or pathway. As elaborated in Methods, we formulate the problem of JACS identification as a hypothesis testing question. In this approach statistically significant JACSs correspond to heavy subnetworks in a similarity graph, with nodes inducing a connected subgraph in G^C (Figure 1). The probabilistic model we propose also accommodates the use of gene-specific priors, reflecting our confidence that they are transcriptionally regulated in the studied conditions.

As exact optimization is intractable, we designed and tested several heuristics for solving the problem (see Methods). The version that performed best on real biological data had the following three phases: (1) detection of relatively small, high-scoring gene sets, or *seeds*; for each node, the set consisting of it along with the neighboring nodes that are connected to it via positive-weighted edges was a candidate seed; (2) seed improvement, and (3) significance-based filtering (see Methods for full details). This version, which we call MATISSE, was used in subsequent analysis.

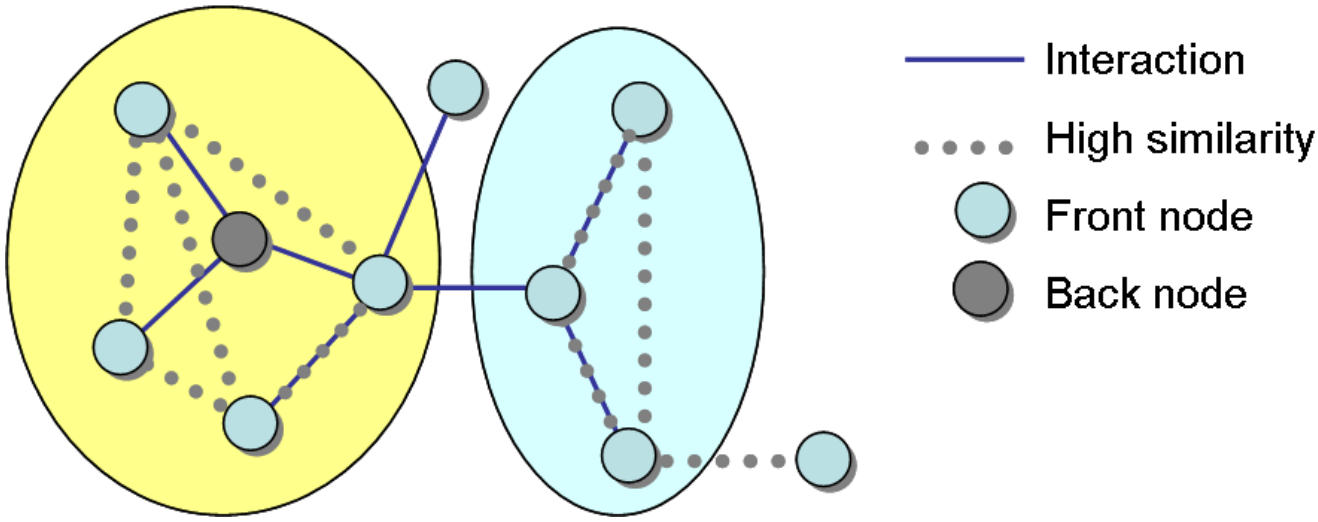


Figure 1

Toy input example. A toy example of an input problem with two distinct JACSs and with front and back nodes. Both JACSs (circled) are connected in the interaction network and heavy in the similarity graph. Note that the four front nodes in the left JACS form a connected subgraph only after the addition of the back node.

Analysis of performance using simulated similarity values

In order to evaluate the ability of our method to detect subnetworks of high pairwise similarity, we first tested its performance on simulated similarity data. The simulation used a connected subnetwork of 2,000 nodes from the *S. cerevisiae* interaction network (described below) as the constraint graph. The similarity data were generated by "planting" a collection of JACSs with several defining parameters in the network, using two similarity value distributions, where members of the same JACS tend to have higher similarity, as described in Methods.

In order to test the effect of each parameter on the performance of the different module finding algorithms, we carried out simulations in which one parameter was varied while keeping the rest at their default values. We also tested simple clustering of the similarity data with the *K*-means algorithm and with the Co-clustering approach of Hanisch *et al*. [17], which proposes a distance measure based on topology and expression. Since the latter method does not readily provide clusters, we used that measure with a *K*-means-like algorithm (with $K = 15$, and moving genes between clusters based on average distance). Other methods (e.g., [21]) were not readily available for comparison.

We evaluated the ability of the methods to recover the planted components using Jaccard coefficient. The coefficient ranges between 0 and 1 with 1 indicating perfect recovery (see Methods). The results are presented in Figure 2. MATISSE is able to retrieve the planted components with good precision when there is a plausible separation between the two similarity value distributions (above 1.3 standard deviations) and the fraction of the front nodes exceeds 0.8. The performance of MATISSE exceeds that of other methods for most of the parameter range.

Response to osmotic stress in S. cerevisiae

We generated a comprehensive *S. cerevisiae* protein-protein and protein-DNA interaction network by combining information from the interaction databases SGD, BioGRID and BIND and recent high-throughput studies (e.g., [25], see our website for a complete list). This resulted in a network containing 6,230 nodes and 89,327 interactions. We also used 133 expression profiles of *S. cerevisiae* under different perturbations and different environmental conditions focused on the osmotic stress response [26]. The 2,000 genes whose patterns exhibit the highest variation in the data were designated as front nodes. We used Pearson correlation for scoring similarities between expression patterns. The parameters of the probabilistic model were assigned as described in Methods. Maps of the subnetworks produced by MATISSE are provided on our website and in the supplement [see Additional file 1].

Comparison of the modules produced by each method

We compared the performance of MATISSE to Co-clustering and to clustering based solely on the gene expression data. We used the CLICK algorithm [27] for clustering, as it was shown to outperform several extant gene expression clustering algorithms, and since it can determine the number of clusters and also leave some vertices unclustered. The Ideker *et al*. method [19] could not be tested in this setting, since measurement p-values could not be computed.

Table 1 compares the properties of the modules produced by every method. Expression homogeneity is calculated as the average Pearson correlation between genes within the same module. The *edge density* of a subgraph is the number of edges it contains as a fraction of all its node pairs. The *clustering coefficient* of a node is the fraction of its neighbor pairs that are connected in the network [28]. The clustering coefficient of a module is the average coefficient in the subgraph induced by the module. In the "Random connected" and "Random" solutions, modules were randomly sampled gene groups with and without the requirement for network connectivity, respectively. The sizes of the random groups were matched to the sizes obtained by MATISSE.

Expression homogeneity

As expected, the most homogeneous clusters in terms of expression similarity are obtained by CLICK, which optimized this type of similarity. The homogeneity of the MATISSE JACSs is higher than that of co-clusters. As previously reported [3], the expression homogeneity of a random connected set is higher than that of a random arbitrary set (average coherence of 0.063 for the random connected solution, vs. 0.033 for random arbitrary solution).

Topological descriptors

MATISSE is designed to produce connected subnetworks. The significance of this criterion is evident from the comparison to the other algorithms. In contrast to MATISSE, both CLICK and Co-clustering produce modules that are highly disconnected (averaging 80–90 components per module). Interestingly, the subnetworks produced by MATISSE are not denser than random connected components in the network. This observation can be explained by the fact that the network contains several dense complexes that do not participate in the solutions, as their components are not homogeneously expressed under the examined conditions.

Functional enrichment

In order to compare the functional relevance of the modules found by the different methods we used four annotation databases: (a) GO "biological process" ontology (level 7; 474 categories) [29]; (b) GO complexes annotation (subterms of "protein complex" term, 213 com-

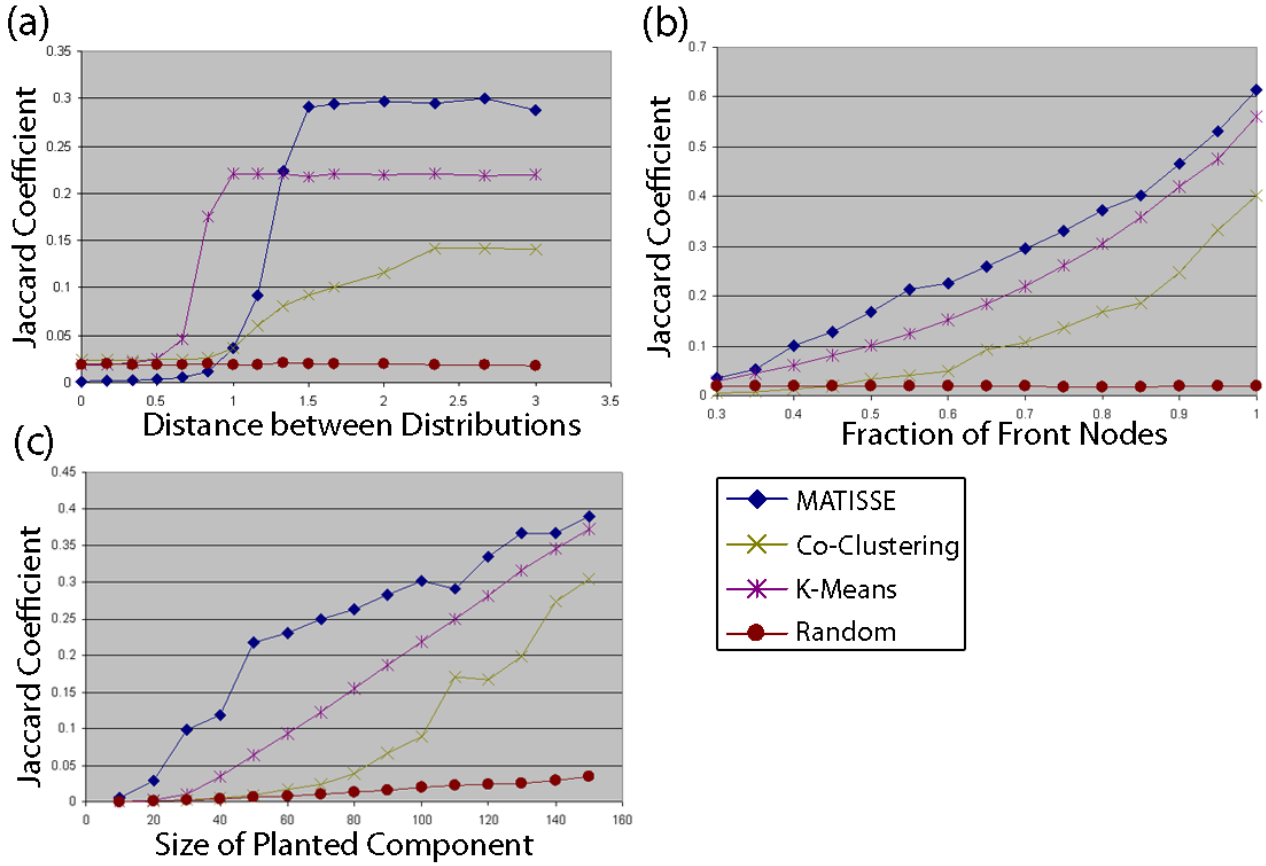


Figure 2

Performance of different module finding procedures on simulated data. Co-clustering: clustering based on the distance metric of [17]. K-Means: clustering of the similarity data. Random: random sampling of connected subnetworks matched in size and number to the planted components. The quality of solutions produced by the different procedures is evaluated by the Jaccard coefficient, (a) Performance as a function of the distance between the means of the mates and the non-mates distributions (μ_m). (b) Performance as a function of the fraction of front nodes (p_f). (c) Performance as a function of planted component size (k).

Table 1: Performance of the different module finding algorithms on the S. cerevisiae osmotic shock data

Solution	No. of modules	Total nodes	Average size	Expression homogeneity	Clustering coefficient	Edge density	No. of connected components
MATISSE	20	2107	105.35	0.361	0.073	0.035	1.00
Co-clustering	19	1991	104.79	0.354	0.035	0.010	89.67
CLICK	20	1988	99.40	0.438	0.030	0.011	77.61
Random connected	20	2107	105.35	0.063	0.050	0.036	1.00
Random	20	2105	105.35	0.033	0.004	0.003	89.78

Numbers in columns 4–8 are averages over all the modules in each solution.

plexes); (c) MIPS deletion phenotype annotations [30] (181 phenotypes); (d) KEGG molecular pathways (310 pathways) [14]. A relatively wide selection of annotations was used to encompass diverse biological functions. Note that the GO "molecular function" categories are not relevant here, as the identified sets of genes are not expected to have similar molecular mechanisms.

For each annotation and for each group of genes produced by every method, the hypergeometric p-value was computed (without correcting for multiple testing, see below). We analyzed the percentage of the modules (Figure 3a) and of the categories (Figure 3b) enriched with p-value ≤ 10^{-3} in each solution. MATISSE exhibits high performance in functional terms and in most cases the produced JACSs show higher enrichment than expression clusters and co-clusters. Co-clustering and CLICK perform slightly better than MATISSE in covering KEGG categories. This is probably due to the overrepresentation of metabolic pathways in KEGG. Metabolic pathways are generally poor in direct protein-protein and protein-DNA interactions, and thus less likely to be recognized by MATISSE, which relies also on direct interactions, than by a clustering algorithm based on expression alone.

As an additional comparison between MATISSE and Co-clustering, we compared the p-values obtained by each solution on each GO biological process (level 7) class attaining enrichment of $p \leq 0.01$ in at least one of the solutions. The MATISSE modules gave better significance to 238 functions, while only 116 functions had higher significance in the Co-clustering solution.

In order to check the added value of incorporating network constraints over using only expression profiles, we compared the results to clustering of the expression profiles with CLICK. In the same pairwise comparison, 223 MATISSE functions exhibited a higher enrichment, compared to 146 in CLICK. Several relevant functions, such as pyridoxine metabolism, cellular response to phosphate starvation, protein ubiquitination and post-Golgi transport, were enriched with $p < 10^{-5}$ in MATISSE, but were not significantly enriched in any CLICK cluster. When seeking functions enriched by the other clustering methods, the only function enriched was "NAD biosynthesis" ($p < 10^{-5}$) discovered by CLICK. The six genes in our dataset that are annotated with this category do not contain any interactions between them and the average length of the shortest path between them is 7.

Functional subnetworks identified by MATISSE

In the previous analysis we did not correct for multiple testing since our goal was the comparison of the different methods. To address the multiple testing problem, we performed a GO functional enrichment analysis using the TANGO algorithm [31]. The algorithm considers all levels of the GO hierarchy and provides p-values corrected for multiple testing and for category dependency using resampling (see Methods).

21 distinct functional terms were found to be enriched ($p < 0.05$) in 14 distinct modules. The complete list of the enriched functions and their respective JACSs is shown in Table 2. Interactive maps of these JACSs can be found at our website along with the corresponding expression data. Note that JACSs were artificially limited to contain no more than 120 nodes in order to provide a better separation between pathways with slightly similar expression patterns. Nevertheless, it appears that this bound does not cause substantial fragmentation of the true clusters, as almost all the JACSs were enriched with distinct functions. Reassuringly, most of the enriched functions are highly

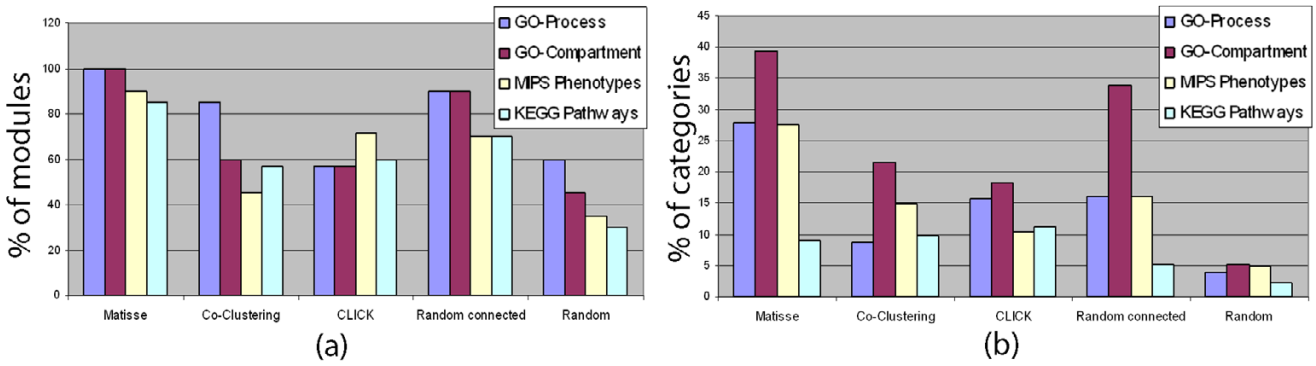


Figure 3

Performance of different module finding algorithmson S. cerevisiae osmotic shock data. (a) The fraction of the modules for which at least one category was enriched, (b) The fraction of the categories enriched in at least one module. Enrichment was defined as attaining hypergeometric p-value ≤ 10^{-3}. Annotation sets: *GO-Process*: Level 7 of the GO "biological process" ontology; *GO-Complex*: subterms of "protein complex" term, GO:0043234; *MIPS Phenotypes*: MIPS deletion phenotype annotations; *KEGG Pathways*: KEGG molecular pathway participation.

relevant to the conditions and the perturbations in the data [32]. These include stress responses, such as repression of the translational machinery (JACSs 1–3) as well as general stress response genes (JACS 11 and 17). In addition, a specific subnetwork relevant to the activation of the pheromone response pathway following osmotic shock in *hog1* strain [32] was identified (JACS 5). Indeed, since the HOG pathway shares protein kinases and phosphatases with other MAPK pathways, it was demonstrated that perturbations in Pbs2 or Hog1 lead to osmostress-induced stimulation of the pheromone response pathway [33].

JACS 7 contains seven genes from the yeast membrane ergosterol biosynthesis pathway which is strongly repressed following osmotic shock in the WT strain but not in *hog1* strains. Lower levels of ergosterol make the membrane more compact and less flexible and hence lead to diminished transmembrane flux of glycerol, which is important for recovery from both hyper-osmotic and hypo-osmotic shock [32].

JACS 16 contains 19 genes members of the proteosome complex. 9 of these are back nodes, underlying the ability of MATISSE to use the network for linking co-activated genes with biologically relevant partners. Inspection of the expression data reveals a slight induction of the proteolysis genes following osmotic shock. This subtle response is missed when clustering solely the expression data, as no more than seven proteolysis genes are clustered together in the CLICK solution. Ubiquitin-dependent proteolytic mechanisms were linked to osmotic responses before [32], and our findings support this hypothesis.

Figure 4 shows JACSs 5 and 16. These subnetworks demonstrate the use of different interaction types by MATISSE: JACS 5 is dominated by protein-DNA interactions, involving the transcription factors (TFs) Tec1, Kss1 and Dig1; JACS 16 is dominated by the protein interactions within the proteosome and the mitochondrial ribosome complexes. This subnetwork contains multiple back nodes linking front nodes. In fact, Table 2 shows that some JACSs make extensive use of nodes with no similarity data.

Table 2: Functionally enriched modules found in the yeast osmotic shock data

JACS	Size	Front	Enriched GO terms	p-value	TFs	p-value
1	120	119	processing of 20S pre-rRNA	< 0.001	Fhl1	$4.82 \cdot 10^{-16}$
			rRNA processing	< 0.001	Rap1	$2.89 \cdot 10^{-11}$
			35S primary transcript processing	< 0.001	Sfp1	$2.98 \cdot 10^{-8}$
			ribosomal large subunit assembly and maintenance	0.019		
			rRNA modification	< 0.001		
			ribosome biogenesis	0.029		
2	120	118	translational elongation	< 0.001	Fhl1	$1.03 \cdot 10^{-5}$
3	120	118	processing of 20S pre-rRNA	< 0.001		
			rRNA processing	0.030		
			35S primary transcript processing	0.011		
			ribosomal large subunit assembly and maintenance	0.019		
			ribosomal large subunit biogenesis	< 0.001		
5	120	112	signal transduction during filamentous growth	0.010	Ste12	$5.41 \cdot 10^{-13}$
			conjugation with cellular fusion	< 0.001	Dig1	$5.41 \cdot 10^{-13}$
6	120	99	transcription from RNA polymerase III promoter	< 0.001		
			transcription from RNA polymerase I promoter	0.006		
7	120	107	ergosterol biosynthesis	< 0.001		
			hexose transport	0.019		
8	114	85	chromatin remodeling	0.050		
11	120	114	pseudohyphal growth	0.010	Msn2	$3.17 \cdot 10^{-4}$
			response to stress	< 0.001	Msn4	$1.82 \cdot 10^{-12}$
14	120	102	ubiquitin-dependent protein catabolism	0.047		
15	120	96	nuclear mRNA splicing, via spliceosome	< 0.001		
16	89	61	ubiquitin-dependent protein catabolism	< 0.001	Rpn4	$6.44 \cdot 10^{-6}$
17	120	109	response to stress	< 0.001	Msn4	$1.74 \cdot 10^{-3}$
			mitochondrial electron transport	< 0.001		
18	87	59	nuclear mRNA splicing, via spliceosome	0.012		
20	46	35	pyridoxine metabolism	0.045		

The GO p-value was adjusted for multiple testing using the TANGO algorithm (see Methods). Enriched TF binding site motifs were detected using the PRIMA algorithm [35]. TF p-values were Bonferroni corrected for multiple testing.

For several pathways, such as pyridoxine biosynthesis, intracellular transport and chromatin-related complexes (mainly SAGA, Cdc73, COMPASS and RSC) that were linked by MATISSE to osmotic shock in *S. cerevisiae*, this linking is novel. Pyridoxine was recently linked to osmotic shock response in *A. thaliana* [34]. These findings underlie the ability of MATISSE to produce testable hypotheses and novel insights.

Promoter analysis

Based on the assumption that genes that exhibit similar expression pattern over multiple conditions are likely to be co-regulated and to share common *cis*-regulatory elements in their promoters, we searched for over-representation of known transcription factor binding site motifs in the promoters of the genes in each JACS. When using the PRIMA motif finding tool [35], six subnetworks showed significant enrichment ($p < 10^{-5}$) for at least one TF (Table 2). All the TFs corresponded to known regulators of the processes enriched in the subnetworks. For example, JACS 5, enriched for pheromone response pathway genes, was enriched with putative targets of Dig1 and Ste12, known regulators of these pathways [36]. Subnetwork 11, associated with general stress response, contained multiple targets of the Msn2 and Msn4 stress TFs [37]. We validated that these motif enrichments are not a byproduct of the functional enrichment in the JACSs ($p < 10^{-4}$, by random sampling of gene groups with the same fraction of genes from the corresponding functional category as in the JACS). This analysis suggests that the JACS we obtained indeed correspond to gene modules with a common transcriptional regulation.

Cell cycle in human

We constructed a human protein-protein interaction network by combining information from the BIND and HPRD databases and from two recent large-scale yeast two-hybrid studies on human cells [23,24]. The resulting network contains 9,135 nodes and 25,086 interactions. Expression profiles of the synchronized HeLa cell lines from [38] were used. Only the 19 point time series obtained for synchronization by thymidine-nocodazole block was selected for the analysis, as it contains the fewest missing values. Genes for which the maximal fold change across the conditions was below 2 were filtered, leaving 1,536 genes (front nodes).

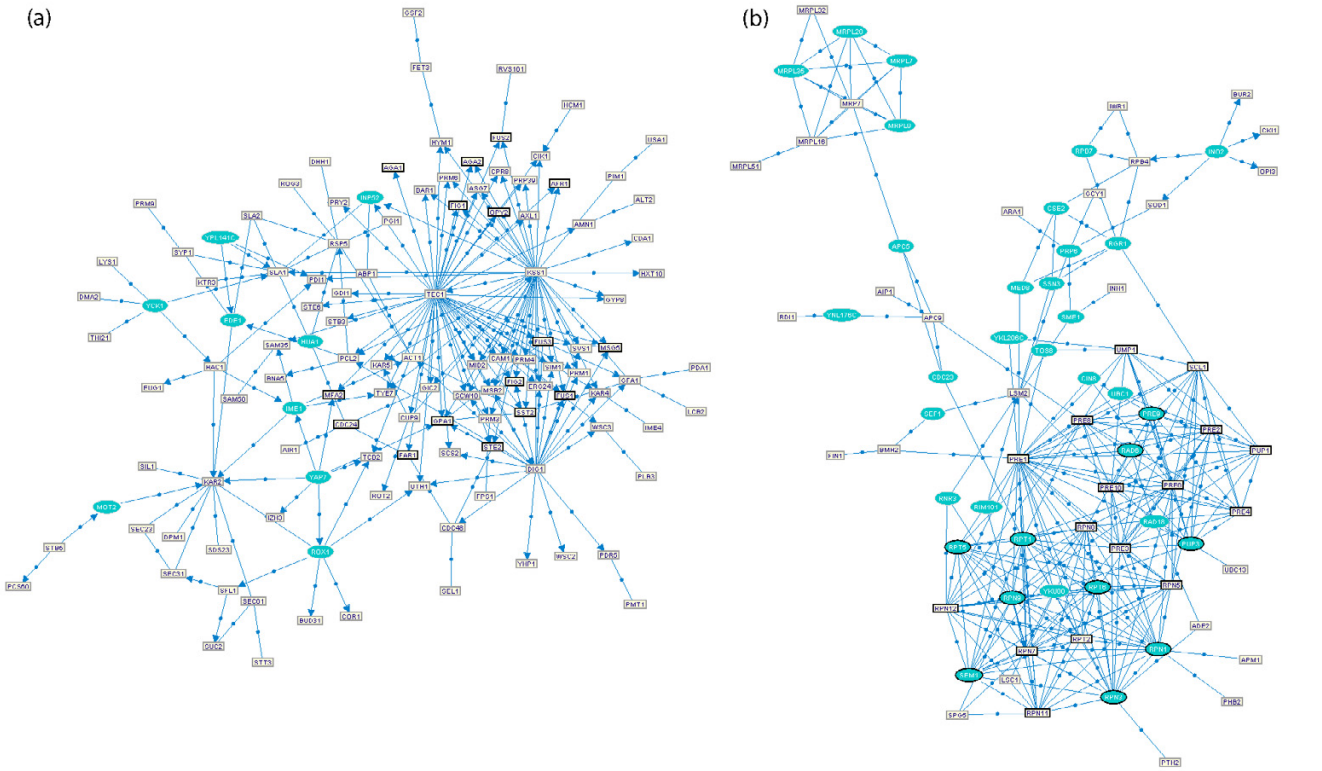


Figure 4

Two of the JACSs identified in the S. cerevisiae analysis. (a) The pheromone response subnetwork, (b) The proteolysis subnetwork. The front nodes are the yellow (light gray) rectangles and the back nodes and the blue (dark gray) ovals. The genes annotated with pheromone response (a) and proteolysis (b) are drawn with thicker border. Gene lists, expression matrices and interactive display of all the subnetworks are available at the supplementary website.

We performed MATISSE analysis using the All-Neighbors heuristic, and the same parameters as in the previous section, and obtained 14 significant JACSs. Maps of these subnetworks are provided on our website and in the supplement [see Additional file 1]. To check the ability to discover subnetworks active at different cell cycle phases, we analyzed the overlap between the JACSs and annotations of specific cell-cycle phases as provided in [38]. Indeed, seven modules were enriched for specific phases of the cell cycle with $p < 0.05$ after Bonferroni correction. The module with the highest cell cycle enrichment (JACS 5, $p = 2.85 \cdot 10^{-17}$) is shown in Figure 5a.

The advantage of MATISSE is evident when comparing the modules most enriched for the GO "cell cycle" category in the MATISSE and the Co-clustering solutions. While the MATISSE module is a single connected component of 120 genes, the corresponding co-cluster contains 110 connected components and 519 genes, and thus is much less amenable to interpretation in terms of the functional connections between its genes.

Subnetwork hub analysis

We hypothesized that the topology of the JACSs obtained by MATISSE can provide clues to the key players in the regulation of the cell cycle machinery. To test this, we looked for "subnetwork hubs" in the JACSs, i.e., genes whose degrees in a JACS were high both absolutely and relatively to their network degree (see Methods). This analysis on the 14 JACSs identified 52 hubs, 18 of them with "cell cycle" annotation ($p = 5.13 \cdot 10^{-11}$). This set contained many cell cycle master regulators such as p53, ATM, E2F1, TGFβR, CDK4 and CDC42. Remarkably, 36 out of 52 hubs form a single connected subnetwork, displayed in Figure 5b. This demonstrates that subnetwork hubs represent key regulators relevant to the experimental conditions tested. The interactions between the subnetwork hubs are putative regulatory interactions governing the progression of the cell cycle. As only 33 of the 52 hubs are

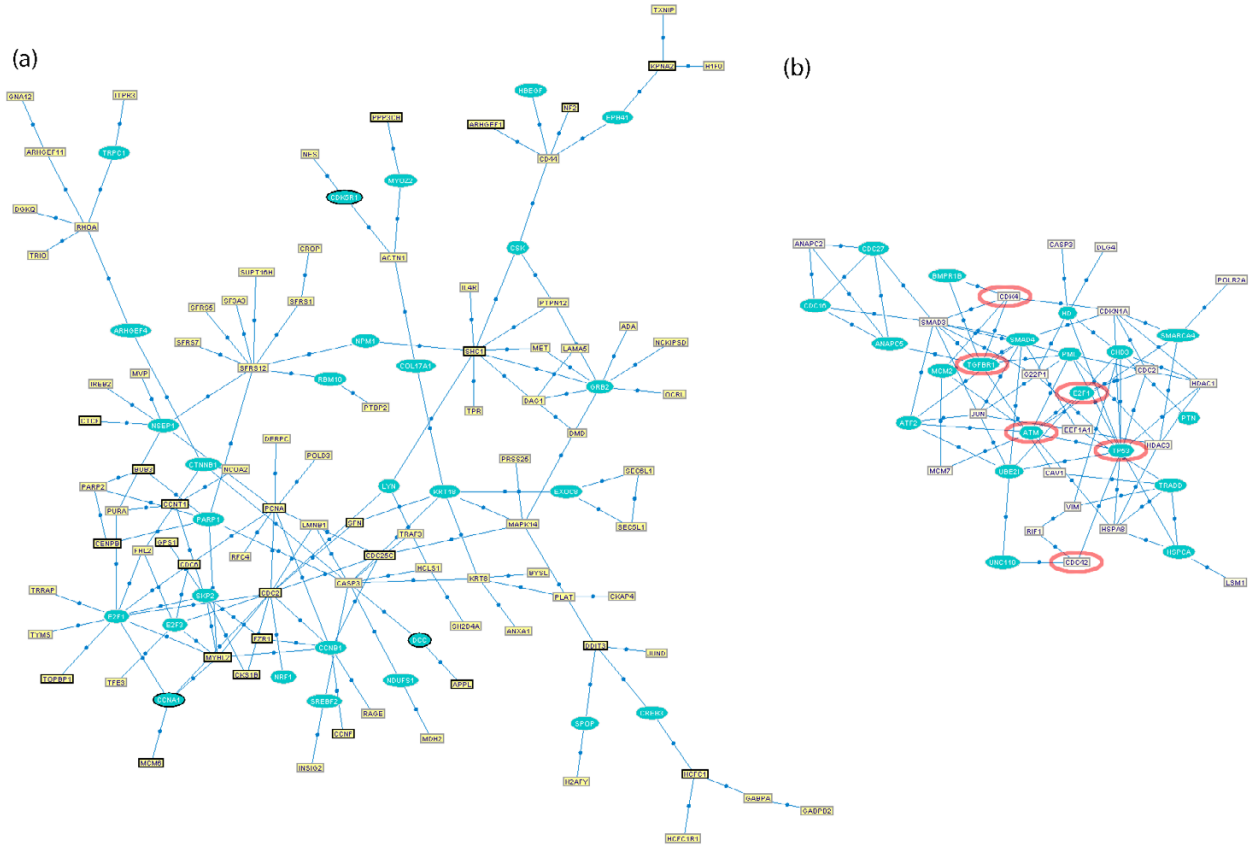


Figure 5

Examples of the MATISSE analysis in the cell cycle data of human HeLa cells. Front nodes and back nodes are as indicated in Figure 4. (a) The highest scoring cell-cycle related JACS identified. The genes annotated with "cell cycle" are drawn with thicker border. Gene lists, expression matrices and interactive display of all the subnetworks are available at the supplementary website, (b) Subnetwork hubs. The figure shows 36 nodes in the JACSs that were identified as subnetwork hubs and induced a connected component in the network. 16 additional hubs that had no interactions with other hubs are not shown. The known master regulators p53, ATM, E2F1, TGFβR, CDK4 and CDC42 are circled.

front nodes in their respective JACS, this set could not be identified using expression data alone.

Conclusion

We have developed a novel computational technique for the integrated analysis of network and similarity data. The method is aimed to dissect together topological properties of gene or protein networks and other high-throughput data. We used the method to analyze large-scale protein interaction networks and genome-wide transcription profiles in yeast and human. The method was shown to identify functionally sound modules, i.e., connected subnetworks with highly coherent expression showing significant functional enrichment. In comparison to the extant Co-clustering method, which aims to integrate similar data, our method demonstrated substantial improvement in solution quality. Comparison to solutions produced by clustering highlights the advantage of utilizing topological connectivity in the hunt for functionally sound modules. By construction, our method is specifically powerful in detection of regulatory modules, and less fit for detection of metabolic modules. Our technique, implemented in the program MATISSE, is efficient and can analyze genome-scale interaction and expression data within minutes.

The proposed algorithm is very flexible and – unlike Co-clustering – can handle situations where not all genes in the network have similarity information or expression patterns. In particular, MATISSE can determine the subset on which similarity is computed using various criteria, e.g., initial probe filtering, differential expression confidence values, etc. As we demonstrate, even when only a modest fraction of the overall network genes have expression/similarity information, the method finds meaningful modules successfully.

The requirement for network connectivity as proposed in our method can be viewed as problematic due to high rate of false negative interactions. A natural extension of MATISSE which we intend to pursue is to take into account the interaction confidence. As a first step towards this goal, we assessed the composition of the interactions in the reported subnetworks as follows: we compared the observed and expected number of interactions within the subnetworks, from each of the publications used as interaction sources in the *S. cerevisiae* interactions network. We found a clear enrichment for interactions from recent experiments, such as [39] and [40], opposed to an underrepresentation of interactions from older works, such as [41,42] and [43] (see supplementary table). As currently the coverage of the protein interaction network is limited, we suggest performing MATISSE analysis in addition to standard clustering analysis.

The framework described in this work is directly applicable to any kind of pairwise similarity data where the probabilistic assumptions hold. While this study focused on protein interaction networks and gene expression, the approach is general enough to treat many other data types. These include other types of interactions, such as genetic interactions, regulation and protein-DNA binding patterns, and other similarity measures, such as functional similarity or similarity in protein-DNA binding profiles [2]. We intend to extend MATISSE to these types of data as well.

While the rapidly expanding resource of microarray data is currently analyzed primarily using diverse clustering techniques, methods for the analysis of network-type data describing interrelations of genes and proteins are less mature, and methods for joint analysis of the two data types are in nascent stage. We expect the proposed method to become widely used for dissecting expression data in light of the interaction knowledge. Our initial results show that despite the high complexity and the relatively low coverage of the human interactome, biologically relevant modules can be found in the human protein interaction network through integrative analysis.

Methods

The probabilistic model

Recall that we formalize the problem as finding disjoint node sets that induce connected subgraphs in the constraint graph and manifest high internal similarity. We formulate this problem as a hypothesis testing question. For this, we define a probabilistic model for the similarity data, using ideas from [27] and [44]. Given a set U of k genes, we compare two hypotheses: the *null hypothesis* H_0: U is a set of unrelated genes; and the *JACS hypothesis* H_1: U is a JACS. We assume that the observed pairwise similarity values are a mixture of two Gaussian distributions: one for pairs of genes that are highly co-expressed (such pairs are called *mates*) and another for the rest. Let M_{ij} denote the event that i and j are mates. The similarity values between mates ($P(S_{ij}|M_{ij})$) are normally distributed with mean μ_m and variance σ_m^2. The similarity levels of all non-mates are distributed normally with the parameters μ_n and σ_n^2. These assumptions are theoretically justified in certain situations [27]. Empirically, analysis using normal quantile plots [45] indicates that they are valid for the biological data analyzed in this paper (results not shown). We also assume that the probability that a pair of genes are mates is high if they belong to the same JACS and low otherwise.

Differential regulation

Not all genes within the interaction network are regulated on the expression level. Thus, when working with expression profiles, we would like the model to allow lower similarity levels between genes that are not necessarily regulated on the expression level, while penalizing heavily for low similarity between transcriptionally regulated genes. This allows flexibility on two levels in our setting. First, the genes can be filtered prior to computing similarities (e.g., only genes passing a threshold of observed fold change or variation level are included in V_{sim}). Note that genes that fail to pass the filter remain in the interaction network and can be incorporated into a JACS, while not used for its scoring. Second, a prior can be assigned to the likelihood that a gene is regulated: we define R_i as the event that gene i is regulated on the expression level under the conditions studied and let $P(R_i)$ designate the probability of that event.

The likelihood score

We assume that JACSs contain a much higher proportion of mates than gene pairs that do not belong to the same JACS. Specifically, we assume that a large fraction β_m (e.g. 0.9) of the pairs of transcriptionally regulated genes within the JACS are mates and thus their similarity levels are distributed $N(\mu_m, \sigma_m)$. Then $P(M_{ij}|R_i \wedge R_j, H_1) = \beta_m$. We make the simplifying approximation that the scores of different gene pairs are independent. Consequently, the likelihood of a JACS U is decomposable on every pair of genes in it:

$$P(S_{U\times U} \mid H_1) = \prod_{(i,j)\in U\times U} P(S_{ij} \mid H_1)$$

Let $\gamma_{ij}^m = \beta_m P(R_i)P(R_j)$. Then:

$$P(S_{ij}|H_1) = \gamma_{ij}^m P(S_{ij}|M_{ij}) + (1 - \gamma_{ij}^m)P(S_{ij}|\overline{M_{ij}})$$

The null hypothesis (H_0) is that the fraction of mates in U is not surprising: every two transcriptionally regulated genes are mates with the probability expected from the relative portion of mates among all the regulated genes, denoted p_m. Let $\gamma_{ij}^n = p_m P(R_i)P(R_j)$. The likelihood ratio between the two hypotheses $\left(\frac{P(Data \mid H_1)}{P(Data \mid H_0)}\right)$ is:

$$\frac{\prod_{(i,j)\in U\times U}\gamma_{ij}^m P(S_{ij} \mid M_{ij}) + (1-\gamma_{ij}^m)P(S_{ij} \mid \overline{M_{ij}})}{\prod_{(i,j)\in U\times U}\gamma_{ij}^n P(S_{ij} \mid M_{ij}) + (1-\gamma_{ij}^n)P(S_{ij} \mid \overline{M_{ij}})} = \prod_{(i,j)\in U\times U} \frac{\gamma_{ij}^m P(S_{ij} \mid M_{ij}) + (1-\gamma_{ij}^m)P(S_{ij} \mid \overline{M_{ij}})}{\gamma_{ij}^n P(S_{ij} \mid M_{ij}) + (1-\gamma_{ij}^n)P(S_{ij} \mid \overline{M_{ij}})}$$

Define the *similarity graph*, $G^S = (V_{sim}, E^S)$, where $E^S = (V_{sim} \times V_{sim})$ and set $w_{ij} = \log \frac{\gamma_{ij}^m P(S_{ij} \mid M_{ij}) + (1-\gamma_{ij}^m)P(S_{ij} \mid \overline{M_{ij}})}{\gamma_{ij}^n P(S_{ij} \mid M_{ij}) + (1-\gamma_{ij}^n)P(S_{ij} \mid \overline{M_{ij}})}$ as the weight of the edge (v_i, v_j). The log-likelihood score for a given U translates to the total edge weight of the subgraph induced by U in G^S.

JACS finding algorithm

Our goal is to find disjoint sets $U_1, U_2, \ldots, U_m$ that induce connected subgraphs in G^C and heavy subgraphs in G^S. When weights can be both positive and negative (as is the case in our formulation), even the problem of finding a single heavy subgraph is NP-Hard (by a simple reduction from Max-Clique using a complete constraint graph). Hence, exact optimization is intractable, and we experimented with several heuristic algorithms for solving the problem. All the schemes share the following three phases: (1) detection of relatively small, high-scoring gene sets, or *seeds*, (2) seed improvement, and (3) significance-based filtering.

Identifying seeds

We tested three different methods for generating high scoring seeds. In all the methods a large set of non-overlapping potential seeds is first generated, and only seeds passing a certain score threshold are passed to the next phase.

Best-neighbors

In this method, high scoring seeds of a predefined size k are constructed. The nodes of the graph are ranked based on their total incident edge weights in G^S (their *weighted degree*). The algorithm repeatedly creates a seed and removes its nodes from the graph. The seed generating step picks the highest ranking node v, and selects a set of k - 1 neighbors of v in G^S that maximize the seed score. The optimal neighbor set can be found through exhaustive enumeration (enumeration is needed since the score for different neighbor sets depends also on the weights of the edges between them). When enumeration is computationally prohibitive, a heuristic that picks nodes with the highest weighted degree within the immediate neighborhood of v is utilized. Specifically, let N_v be the set of all the immediate neighbors of v. For $i \in N_v$ define $w_i^v = \sum_{v_j \in N_v} w_{ij}$. The heuristic selects k - 1 nodes with the highest w^v values.

All-neighbors

This method is similar to Best-Neighbors, but instead of selecting k - 1 neighbors for a potential seed, in this version, all the neighbors of v with a non-negative edge score (including neighboring back nodes with zero score) enter the seed.

Heaviest-subnet

This method is inspired by Charikar's 2-approximation algorithm for the densest subgraph problem [46]. An *articulation node* in a connected graph is one whose removal disconnects the graph. The following algorithm is executed independently on each connected component in the constraint graph. The algorithm works in a "destructive" fashion: starting from the original constraint graph, nodes are removed from the graph one at a time until none remain. The next node to be removed is one with the smallest weighted degree in the current similarity graph that is not an articulation node in the current constraint graph. It is easy to see that such a node always exists. After each node removal, the overall score of the remaining graph is recorded. After all nodes are removed, the highest-scoring (possibly size-constrained) subgraph that was encountered is selected as the seed. That subgraph is then removed from the graph and the next seed is sought.

Seed optimization

Once a set of high-scoring seeds is established, a greedy algorithm aims to optimize all the seeds simultaneously. In our tests, this strategy worked better than optimizing each seed separately, as it produced more diverse JACSs. The algorithm keeps a set of disjoint subnetworks at every iteration and considers the following moves (Figure 6):

Node addition

Addition of an unassigned node to an existing JACS.

Node removal

Removal of a node from a JACS.

Assignment change

Exchange of a node between JACSs.

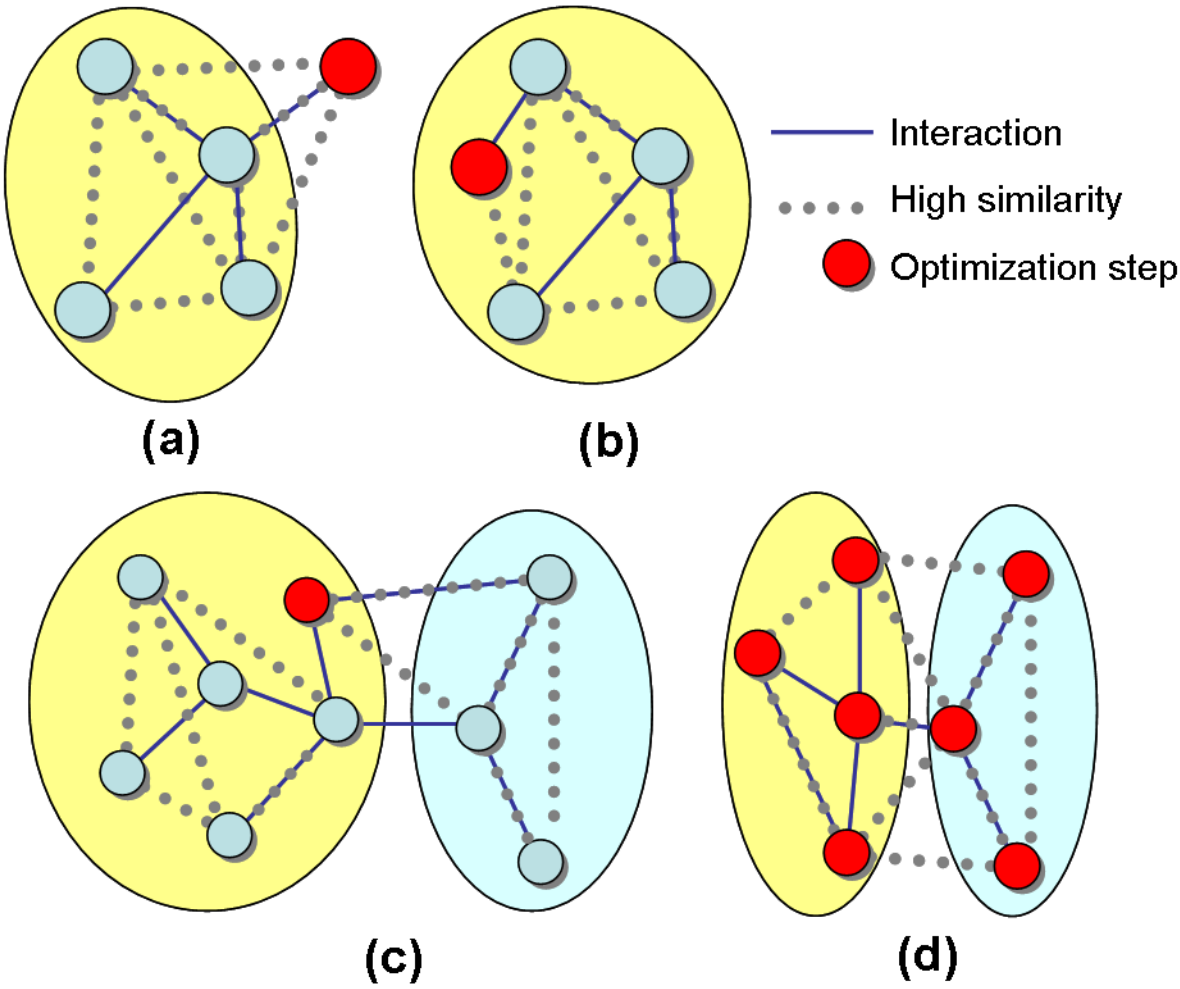


Figure 6
Toy examples of the moves performed by the optimization algorithm. (a) Node addition; (b) Node removal; (c) Assignment change; (d) JACS merge. In each case the affected nodes are in red (black).

JACS merge

A new JACS is formed by taking the union of the nodes in two existing JACSs. This step is particularly beneficial when the original seeds are relatively small.

At every step a move is selected only if (1) it improves the overall score of the solution, i.e., the sum of the weights of all the JACSs and (2) the move maintains the connectivity of the JACSs. If no such step exists, a "cleanup" procedure iteratively removes from every JACS non-articulation back nodes that are not found on any simple path between front nodes. If the clean-up step does not remove any nodes, the optimization halts. Note that the algorithm is guaranteed to converge, as the global score is monotonically increasing. In addition, in order to obtain biologically meaningful JACSs, an upper bound on the size of a JACS can be employed throughout the optimization. If a JACS reaches this upper bound in the course of the optimization, any node added to it causes a removal of a low-scoring node, maintaining the JACS size. Note that this procedure can add only front nodes.

Filtering

After a collection of putative JACSs is obtained, it is filtered based on the significance of the JACS score. For that purpose, for every candidate JACS, an empirical p-value of its score is calculated using sampling randomly gene groups of the same size. Only candidate JACSs with p-value below a threshold p pass the filtering stage ($p = 0.05$ after Bonferroni correction was used). In a second step, to avoid possible bias in the score, we empirically test the JACS significance using only expression similarity scores. The same sampling procedure is performed using the average raw expression pairwise similarity values, and JACSs whose average similarity is not sufficiently high compared to the sampled sets of the same size are removed. An efficient computation of this step is done as suggested in [15].

Implementation issues

For efficient implementation, several slight modifications were made to the algorithm described above:

Removal of non-contributing nodes

As in our framework only front nodes are used for JACS scoring, back nodes will be incorporated into the subnetwork only if they appear on some path between two front nodes. Thus, prior to algorithm execution we remove from G^C all back nodes that are leaves (nodes with degree smaller than 2). The procedure is iterated until no such leaves remain in the graph. In practice, due to the nature of the protein interaction network used, this step significantly reduces the size of the network, without influencing the quality of the solution.

Similarity graph adjustment

When finding Heaviest-Subnet seeds, low edge density in the graph is crucial for efficiency. We therefore remove edges with low absolute weight from the graph, as their contribution to the overall JACS score is small. All the edges are used in the subsequent phases.

Finding heaviest-subnet seeds

Efficient implementation of this algorithm can be done using a data structure similar to the one developed for the dynamic connectivity problem [47]. This would take $O(|V|\log^4|V|)$ time per seed. Instead, we used a simple algorithm for detection of articulation nodes in each iteration. Articulation nodes can be detected during a depth-first traversal of the graph, by calculating the "lowpoint" values of every node (cf. [48]).

This implementation required complexity of $O(|V||E^S|)$ time per seed. Since this time can be too long for very large graphs, we use a sampling approach when the component contains more than 1,500 nodes: a connected subgraph of a more modest size is randomly sampled (as described in [49]) and then used for seed finding. This sampling is repeated several times, with the highest scoring seed used for further optimization.

Implementation

MATISSE was implemented as a Java stand-alone application. In addition to the algorithmic engine, it contains a visualization tool allowing flexible inspection of the obtained subnetworks and diverse post-process analyses. Running times are efficient enough to accommodate large interaction networks and gene expression datasets. For example, on a constraint graph of 4, 543 nodes and 1, 996 expression profiles, the processing took less than 15 minutes for All-Neighbors and Best-Neighbors methods and 78 minutes for Heaviest-Subnet, on a Pentium 4 3 GHz machine with 2 GB memory. About 10 – 20% of the time is needed to learn the parameters using EM, and this time is saved in all subsequent runs on the same data. The running time depends sublinearly on the bound on the maximum size of the JACS (Figure 7). The application will soon be available at [50].

Simulation setup

Our simulations used the real connected network of 2,000 yeast proteins described in Results, and synthetic similarity values, generated as follows. First, a set of m disjoint connected subnetworks $P_1,\ldots, P_m$ of equal size k was randomly selected as in [49]. Then, from each subnetwork a subset of size $k \cdot p_f$ was randomly selected to be included in V_{sim} (front nodes). The resulting V_{sim} was expanded by additional randomly selected nodes, to contain n_{sim} nodes in total. Similarity values were generated as in [27] using two Gaussian distributions - N_m with parameters μ_m, σ_m for similarity between mates and N_n with parameters μ_n, σ_n for all other pairs.

Similarity values were determined independently for each node pair, as follows: If the two nodes reside in the same JACS, the value was drawn from N_m with probability β_m

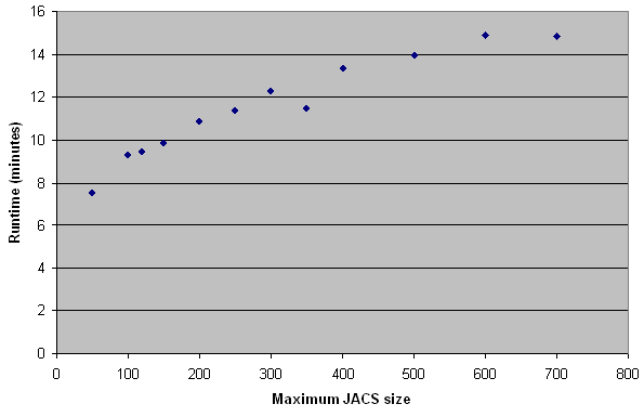


Figure 7
Dependence of the running time on the size of the JACS. The running time of MATISSE with different maximum JACS size parameters. The execution did not include the weight calculation step, as it is not dependent on the JACS size.

and from N_n with probability 1 - β_m. Otherwise, the value was drawn from N_m with probability p_m.

The default values for the simulations were set to n_{sim} = 1, 000 (out of $|V|$ = 2, 000);

$m = 6; k = 100; p_f = 0.7; \mu_m = 0.5; \mu_n = 0; \sigma_m = \sigma_n = 0.3; \beta_m = 0.95; p_m = 0.01$.

Evaluating performance

The success of an algorithm in recovering the planted components was measured using the Jaccard coefficient [51]. It is defined as $\frac{n_{11}}{n_{11}+n_{10}+n_{01}}$, where n_{11} is the number of node pairs included both in the same planted component and in the same JACS, n_{10} is the number of pairs included in the same planted component but not in the same JACS, and n_{01} is the number of pairs in the same JACS but not in the same planted component. Hence, a perfect fit of the two solutions would get a score of 1, and lower scores indicate reduced fit.

Parameter estimation

To obtain meaningful results, a good assessment of the parameters of the probabilistic model is prerequisite. We tested different schemes for assessing $P(R_i)$, and selected the following scheme. We ranked the genes based on the variation observed across their expression patterns and then applied a logistic function to the normalized ranks to obtain: $P(R_i) = \alpha + (1 - \alpha)\frac{1}{1+e^{-\beta(x_i-\gamma)}}$, where x_i is the normalized rank of gene i. The logistic parameters were empirically set to $\alpha = 0.6$, $\beta = 24$ and $\gamma = 0.25$. To evaluate the effect of the specific form of the prior on the results, we reran the JACS finding algorithms with different logistic parameter settings ($\alpha = 0.4..0.8$, $\beta = 1..24$, $\gamma = 0.2..0.7$). The average expression homogeneity and the average functional homogeneity of the produced JACSs (computed as described in [1]) of the JACSs did not change by more than 6%.

We adjusted the standard EM algorithm used for learning a mixture of Gaussians (cf. [52]) in order to estimate μ_m, σ_n, μ_n, σ_n and p_m. A detailed description of the EM algorithm can be found at our website ([50]). The produced JACSs were constrained to the size range of 5–120 and β_m was set to 0.9. We verified that the reported results are robust to changes in the value of β_m by varying it between 0.75 and 0.99 and analyzing the obtained solutions. We found that both the average expression homogeneity and the average functional homogeneity did not change by more than 3% across this parameter range.

Comparison of the heuristics

We evaluated the three proposed heuristics both in our simulation setting and on the osmotic shock response in *S. cerevisiae*. The results of the comparison on simulation data are presented in Figure 8. Overall, as can be seen in Figure 8, all three MATISSE variants show similar performance. All the methods exhibit poor performance in detection of small planted components (k < 50). Best-Neighbors seems to be the preferred method on the simulated data. Best-Neighbors and All-Neighbors is that Best-Neighbors does not incorporate back nodes at all, while All-Neighbors may include some. As we shall show below, using back nodes is in fact advantageous in real biological data. The performance of the Heaviest-Subnet seeding is highly variable, probably due to its relatively significant dependency on the structure of the similarity graph.

The results of the comparison on simulation data are presented in Figure 9. The Best-Neighbors variant performs slightly better than All-Neighbors in terms of the fraction of enriched modules, but All-Neighbors performs significantly better in terms of category coverage, due to its inclusion of back nodes. We therefore carried out all subsequent analysis using the modules produced with the All-Neighbors variant.

Functional enrichment analysis

We used the TANGO algorithm [31] for finding GO terms enriched in the JACSs. The algorithm considers all levels of GO and corrects p-values for multiple testing and for category dependency using resampling. Briefly, TANGO repeatedly selects random sets of genes to compute an

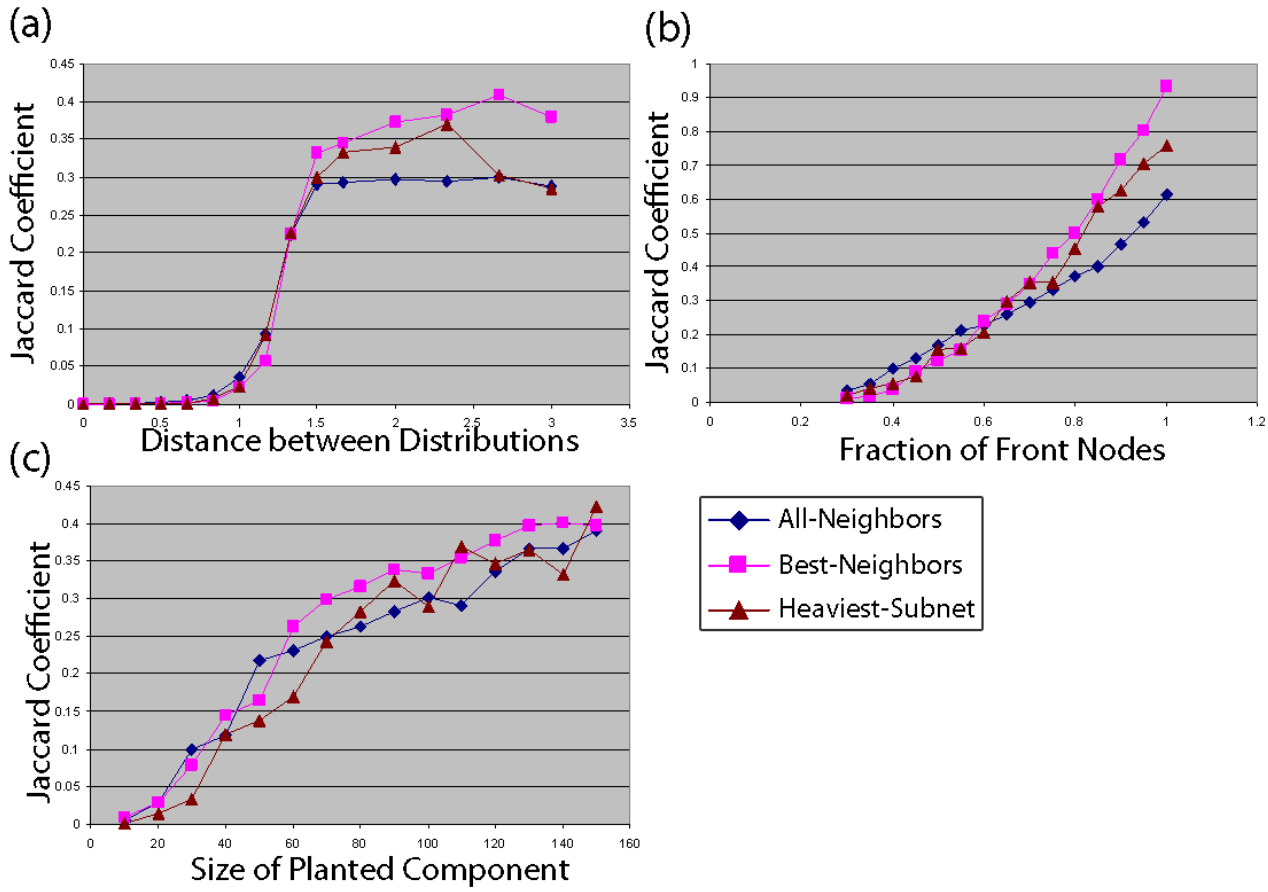


Figure 8
Performance of the three proposed heuristics on simulated data. See Figure 2 for further details.

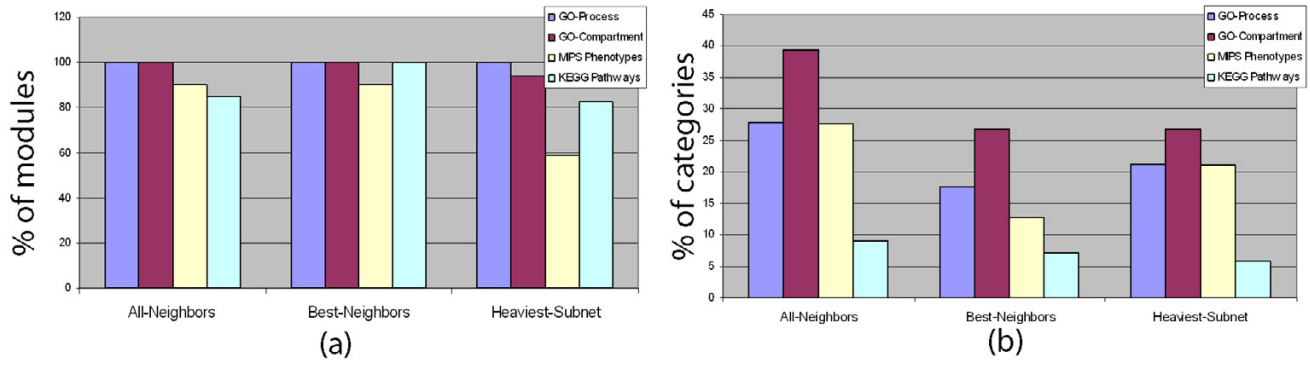


Figure 9
Performance of the three proposed heuristic in terms of annotation enrichment. See Figure 3 for further details.

empirical distribution of maximum p-values for functional enrichment obtained across a random sample of sets that maintain the same size characteristics of the ones analyzed. TANGO uses this empirical distribution to determine thresholds for significant enrichment on the true clusters. The algorithm filters out redundant categories by performing conditional enrichment tests that ensure that all the reported enriched categories are statistically significant even after taking into account the enrichment of their ancestor and children nodes in the tree.

Extraction of subnetwork hubs

Given a JACS J, $v \in J$ was called a *hub* if it satisfied three requirements: (a) the degree of v within the subnetwork J exceeds 7; (b) the degree of v in J is among the five highest in J; (c) the degree of v in J is significantly high given its degree in the whole network ($p < 0.05$ using hypergeometric distribution). Note that back nodes can also be hubs.

Authors' contributions

IU and RS designed the study. IU developed MATISSE and performed the statistical analysis. IU and RS wrote the manuscript. Both authors read and approved the final manuscript.

Additional material

Additional File 1

JACSs identified by MATISSE. *Images of the subnetworks identified by MATISSE in the osmotic shock response and the cell cycle datasets.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-1-8-S1.pdf]

Acknowledgements

We thank Irit Gat-Viks, Chaim Linhart, Daniela Raijman, Israel Steinfeld and Amos Tanay for helpful discussions. IU is supported in part by a fellowship from the Safra Foundation. RS was supported in part by the Wolfson foundation, and by the EMI-CD project that is funded by the European Commission within its FP6 Programme, under the thematic area "Life Sciences, genomics and biotechnology for health", contract number LSHG-CT-2003-503269.

References

1. Lord P, Stevens R, Brass A, Goble C: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19(10):**1275-83.
2. Kim R, Ji J, Wong W: **An improved distance measure between the expression profiles linking co-expression and co-regulation in mouse.** *BMC Bioinformatics* 2006, **7:**44.
3. Ge H, Liu Z, Church G, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.** *Nat Genet* 2001, **29(4):**482-486.
4. Hahn A, Rahnenführer J, Talwar P, Lengauer T: **Confirmation of human protein interaction data by human expression data.** *BMC Bioinformatics* 2005, **6:**112.
5. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644):**.
6. de Lichtenberg U, Jensen L, Brunak S, Bork P: **Dynamic complex formation during the yeast cell cycle.** *Science* 2005, **307(5710):**.
7. Luscombe N, Babu M, Yu H, Snyder M, Teichmann S, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431(7006):**.
8. Wachi S, Yoneda K, Wu R: **Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues.** *Bioinformatics* 2005, **21(23):**4205-4208.
9. Balazsi G, Barabasi A, Olvai Z: **Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli.** *PNAS* 2005, **102(22):**7841-7846.
10. van Helden J, Gilbert D, Wernisch L, Schroeder M, Wodak S: **Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data.** In *Proc JOBIM '00* London, UK: Springer-Verlag; 2000:147-164.
11. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl):**.
12. Zien A, Kuffner R, Zimmer R, Lengauer T: **Analysis of Gene Expression Data with Pathway Scores.** *Proc ISMB '00* 2000:407-417.
13. Kurhekar M, Adak S, Jhunjhunwala S, Raghupathy K: **Genome-wide pathway analysis and visualization using gene expression data.** In *Proc PSB '02* Springer-Verlag; 2002:462-73.
14. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
15. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Research* 2002, **12:**37-46.
16. Vert J, Kanehisa M: **Extracting active pathways from gene expression data.** *Bioinformatics* 2003, **19:**II238-II244.
17. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data.** *Bioinformatics* 2002, **18:**S145-54.
18. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *PNAS* 1998, **95:**14863-14868.
19. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks.** *Bioinformatics* 2002, **18:**S233-S240.
20. Cabusora L, Sutton E, Fulmer A, Forst C: **Differential network expression during drug and stress response.** *Bioinformatics* 2005, **21(12):**2898-2905.
21. Segal E, Wang H, Koller D: **Discovering molecular pathways from protein interaction and gene expression data.** *Bioinformatics* 2003, **19 Suppl 1:**i264-71.
22. Ihmels J, Levy R, Barkai N: **Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae.** *Nat Biotechnol* 2003, **22:**86-92.
23. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6):**957-968.
24. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062):**1173-1178.
25. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RE, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M: **Global analysis of protein phosphorylation in yeast.** *Nature* 2005, **438(7068):**679-84.
26. O'Rourke S, Herskowitz I: **Unique and redundant roles for Hog MAPK pathway components as revealed by whole-genome expression analysis.** *Mol Biol Cell* 2004, **15:**532-42.
27. Sharan R, Shamir R: **CLICK: A clustering algorithm with applications to gene expression analysis.** In *Proc Int Conf Intell Syst Mol Biol Volume 8*. AAAI Press; 2000:307-316.

28. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393(6684):**440-442.
29. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Jea Eppig: **Gene ontology: Tool for the unification of biology.** *Nature Genetics* 2000, **25:**25-29.
30. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2002, **30(1):**31-4.
31. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R: **EXPANDER: an integrative suite for microarray data analysis.** *BMC Bioinformatics* 2005, **6(232):**.
32. Hohmann S: **Osmotic stress signaling and osmoadaptation in yeasts.** *Microbiol Mol Biol Rev* 2002, **66(2):**300-72.
33. O'Rourke SM, Herskowitz I: **The Hog1 MAPK prevents cross talk between the HOG and pheromone response MAPK pathways in Saccharomyces cerevisiae.** *Genes Dev* 1998, **12(18):**2874-2886.
34. Chen H, Xiong L: **Pyridoxine is required for post-embryonic root development and tolerance to osmotic and oxidative stresses.** *Plant Journal* 2005, **44(3):**396-408.
35. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-Wide In Silico Identification of Transcriptional Regulators Controlling the Cell Cycle in Human Cells.** *Genome Research* 2003, **13(5):**773-780.
36. Olson KA, Nelson C, Tai G, Hung W, Yong C, Astell C, Sadowski I: **Two regulators of Ste12p inhibit pheromone-responsive transcription by separate mechanisms.** *Mol Cell Biol* 2000, **20(12):**4199-209.
37. Martinez-Pastor MT, Marchler G, Schuller C, Marchler-Bauer A, Ruis H, Estruch F: **The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE).** *EMBO J* 1996, **15(9):**.
38. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D: **Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors.** *Molecular Biology of the Cell* 2002, **13:**1977-2000.
39. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440(7084):**631-6.
40. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440(7084):**637-643.
41. Ito T, Chiba T, Yoshida M: **Exploring the protein interactome using comprehensive two-hybrid projects.** *Trends Biotechnol* 2001, **19:**S23-S27.
42. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RG, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415:**180-3.
43. Uetz P, Giot L, Cagney G, Mansfield TA, Judson R, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403(6770):**623-7.
44. Sharan R, Ideker T, Kelley B, Shamir R, Karp R: **Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data.** *Journal of Computational Biology* 2005, **12:**835-846.
45. Sokal RR, Rohlf FJ: *Biometry: The Principles and Practice of Statistics in Biological Research* W.H. Freeman and company; 1995.
46. Charikar M: **Greedy Approximation Algorithms for Finding Dense Components in a Graph.** *Lecture Notes in Computer Science* 2000, **1913:**84-95.
47. Holm J, de Lichtenberg K, Thorup M: **Poly-logarithmic deterministic fully-dynamic algorithms for connectivity, minimum spanning tree, 2-edge, and biconnectivity.** In *Proc STOC '98* New York, NY, USA: ACM Press; 1998:79-89.
48. Even S: *Graph Algorithms* Potomac, Maryland: Computer Science Press; 1979.
49. Kashtan N, Itzkovitz S, Milo R, Alon U: **Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs.** *Bioinformatics* 2004, **20(11):**1746-58.
50. **MATISSE web page** [http://www.cs.tau.ac.il/~rshamir/matisse/]
51. Everitt B: *Cluster analysis* third edition. London: Edward Arnold; 1993.
52. McLachlan GJ, Krishnan T: *The EM Algorithm and Extensions* John Wiley and Sons, inc; 1997.

3. Regulatory Networks Define Phenotypic Classes of Human Stem Cell Lines

LETTERS

Regulatory networks define phenotypic classes of human stem cell lines

Franz-Josef Müller[1,2], Louise C. Laurent[1,3], Dennis Kostka[4]†, Igor Ulitsky[5], Roy Williams[6], Christina Lu[1], In-Hyun Park[7], Mahendra S. Rao[8,9], Ron Shamir[5], Philip H. Schwartz[10,11], Nils O. Schmidt[12] & Jeanne F. Loring[1,6]

Stem cells are defined as self-renewing cell populations that can differentiate into multiple distinct cell types. However, hundreds of different human cell lines from embryonic, fetal and adult sources have been called stem cells, even though they range from pluripotent cells—typified by embryonic stem cells, which are capable of virtually unlimited proliferation and differentiation—to adult stem cell lines, which can generate a far more limited repertoire of differentiated cell types. The rapid increase in reports of new sources of stem cells and their anticipated value to regenerative medicine[1,2] has highlighted the need for a general, reproducible method for classification of these cells[3]. We report here the creation and analysis of a database of global gene expression profiles (which we call the 'stem cell matrix') that enables the classification of cultured human stem cells in the context of a wide variety of pluripotent, multipotent and differentiated cell types. Using an unsupervised clustering method[4,5] to categorize a collection of ~150 cell samples, we discovered that pluripotent stem cell lines group together, whereas other cell types, including brain-derived neural stem cell lines, are very diverse. Using further bioinformatic analysis[6] we uncovered a protein–protein network (PluriNet) that is shared by the pluripotent cells (embryonic stem cells, embryonal carcinomas and induced pluripotent cells). Analysis of published data showed that the PluriNet seems to be a common characteristic of pluripotent cells, including mouse embryonic stem and induced pluripotent cells and human oocytes. Our results offer a new strategy for classifying stem cells and support the idea that pluripotency and self-renewal are under tight control by specific molecular networks.

Cultured cell populations are traditionally classified as having the qualities of stem cells by their expression of immunocytochemical or PCR markers[7]. This approach can often be misleading if these markers are used to categorize novel stem cell preparations or predict inherent multipotent or pluripotent features[8]. To develop a more robust classification system, we created a framework for identifying putative novel stem cell preparations by their whole-genome messenger RNA expression phenotypes (Fig. 1). The core reference data set, which we call the 'stem cell matrix', includes cultures of human cells that have been reported to have either stem cell or progenitor qualities, including human embryonic stem cells, mesenchymal stem cells and neural stem cells. To provide the context in which to place the stem cells, we included non-stem-cell samples such as fibroblasts and differentiated embryonic stem cell derivatives. To avoid biasing the classification methods, it was critical that we designated the input cell types with terminology that carried as little preconception about

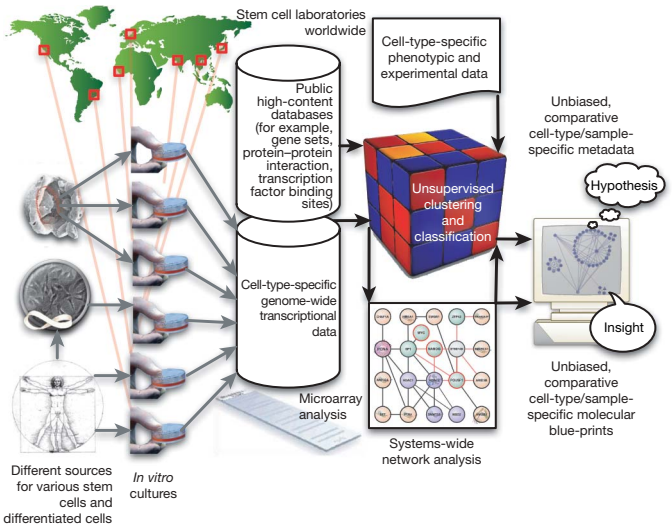


Figure 1 | Sample collection and analysis for the stem cell matrix. Cell preparations for the stem cell matrix are cultured in the authors' laboratories or collected from other sources worldwide. Samples are assigned source codes that capture their biological origin and a relatively unbiased description of the cell type (such as BNLin for brain-derived neural lineage). Samples are collected and processed at a central laboratory for microarray analysis on a single Illumina BeadStation instrument. The genomics data are processed by unsupervised algorithms that are capable of grouping the samples based on non-obvious expression patterns encoded in transcriptional phenotypes. For pathway discovery, existing high-content databases with experimental data (for example, protein–protein interaction data or gene sets) are combined with our transcriptional database, a priori assumed identity of cell types and bootstrapped sparse non-negative matrix factorization (sample clustering) to produce metadata that can be mined with GSA software and topology-based gene set discovery methods (systems-wide network analysis). Web-based, computer-aided visualization methodologies can be used by researchers to formulate testable hypotheses and generate results and insights in stem cell biology. Two exemplary results we report in this paper are the classification of novel stem cell types in the context of other better understood stem cell preparations, and a molecular map of interacting proteins that appear to function together in pluripotent stem cells.

[1]Center for Regenerative Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. [2]Center for Psychiatry, ZIP-Kiel, University Hospital Schleswig Holstein, Niemannsweg 147, D-24105 Kiel, Germany. [3]University of California, San Diego, Department of Reproductive Medicine, 200 West Arbor Drive, San Diego, California 92035, USA. [4]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin, Germany. [5]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [6]The Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, California 92037, USA. [7]Division of Pediatric Hematology/Oncology, Children's Hospital Boston and Dana Farber Cancer Institute, Boston, Massachusetts 02115, USA. [8]Invitrogen Co, 3705 Executive Way, Frederick, Maryland 21704, USA. [9]Center for Stem Cell Biology, Buck Institute on Aging, 8001 Redwood Boulevard, Novato, California 94945, USA. [10]Center for Neuroscience Research, Children's Hospital of Orange County Research Institute, 455 South Main Street, Orange, California 92868, USA. [11]Developmental Biology Center, University of California, Irvine, 4205 McGaugh Hall, Irvine, California 92697, USA. [12]Department for Neurosurgery University Medical Center Hamburg-Eppendorf, Martinistrasse 52, D-20246 Hamburg, Germany. †Present address: Genome and Biomedical Sciences Facility and Department of Statistics, University of California, Davis 451 Health Sciences Drive, Davis, California 95616, USA.

their identity as possible. Our nomenclature ('source code') has two components: the first is the tissue or cultured cell line of origin. The second term captures a description of the culture itself. Supplementary Tables 1–8 summarize the descriptions of the core samples and their assigned source codes.

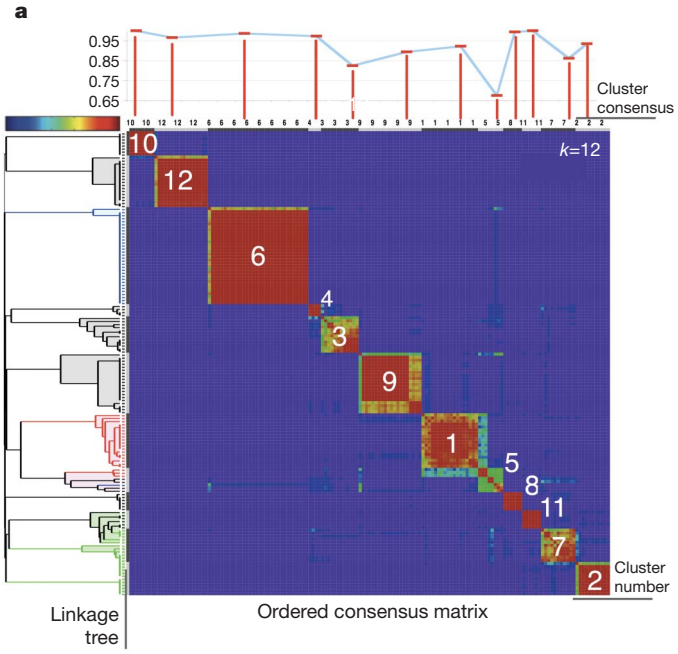
To sort the cell types we used an unsupervised machine learning approach to cluster transcriptional profiles of the cell preparations into stable distinct groups. Sparse non-negative matrix factorization (sNMF) was adjusted for this task by implementing a bootstrapping algorithm to find the most stable groupings (see also Supplementary Discussion 1)[4,5]. The stability of the clustering[9] indicated that the data set most likely contained about 12 different types of samples (Fig. 2a and Supplementary Methods 2). The composition of the stable clusters revealed both predictable and unpredicted groupings of a priori designations (Fig. 2b and Supplementary Fig. 1). The 20 samples identified as undifferentiated human pluripotent stem cell (PSC) preparations were grouped together in one dominant cluster (Fig. 2, cluster 1) and one secondary cluster (Fig. 2, cluster 5). Sixty-two of the samples were brain-derived cells that were described as neural stem or progenitor cells based on their source, culture methods and classical markers. Most of the designated neural stem cells were distributed among multiple clusters, indicating a great deal of diversity in neural stem cell preparations. But one group of the brain-derived lines, those derived from surgical specimens from living patients (HANSE cells, see below), remained together throughout the iterative clusterings (Fig. 2, cluster 6; see also Supplementary Fig. 3 and Supplementary Methods 1). The HANSE cell group consisted of transcriptional profiles that were derived from neurosurgical specimens following published protocols for multipotent neural progenitor derivation and propagation[10,11]. These cells expressed markers that are commonly used to identify neural stem cells[12] (see Supplementary Fig. 4), but the clustering clearly separated them from the other samples that had been derived from post-mortem brains of prematurely born infants (SC23 and SC30, see Fig. 2b)[10,11].

We tested the ability of our data set to categorize additional preparations by adding 66 samples comprising new cultures derived from PSC lines that were already in the matrix, preparations that were not yet included (but their presumptive cell type was already represented), or new cell types. We chose two new types of cells: a differentiated cell type (umbilical vein endothelial cells (HUVECs)) and a recently developed new source of pluripotent cells called induced pluripotent stem cells[13–16] (iPSCs, Supplementary Table 9). iPSCs have been generated from somatic cells, including adult fibroblasts, by genetic manipulation of certain transcription factors[13,15–17]. We re-computed clustering results including the test data set (Supplementary Table 10). All of the HUVEC samples clustered together and formed a distinct group. Most of the additional PSC lines (human embryonic stem cells (embryonic PSCs; ePSCs) and iPSCs) from several different laboratories were placed into a context that contained solely PSC lines. Three additional germ cell tumour lines clustered together with the tumour-derived pluripotent stem cell (tPSC) line 2102Ep and samples of three human embryonic stem (ES) cell lines: BG01v (ref. 18), Hues7 (ref. 19) and Hues13 (ref. 19). BG01v is an established aneuploid variant line and the two Hues lines are aneuploid variants of the originally euploid lines (not shown).

We used a combination of analysis tools to explore the basis of the unsupervised classification of the samples in the core data set. Gene Set Analysis[20] (GSA) is a means to identify the underlying themes in transcriptional data in terms of their biological relevance.

GSA uses lists of genes[20] that are related in some way; the common criterion is that the relationships among the genes in the lists are supported by empirical evidence[20]. GSA highlighted numerous significant differences among the computationally defined categories. (See Supplementary Fig. 2, Supplementary Table 11, Supplementary Methods and http://www.stemcellmatrix.org).

Although GSA is valuable for discovering specific differences among sample groups, it is limited to curated gene lists and cannot be used to discover new regulatory networks. The MATISSE algorithm[6] (http://acgt.cs.tau.ac.il/matisse) takes predefined protein–protein interactions (for example, from yeast two-hybrid screens) and seeks connected subnetworks that manifest high similarity in sample subsets. The modified version used in this analysis is capable of extracting subnetworks that are co-expressed in many samples but also significantly upregulated or downregulated in a specific sample cluster.



b

Cluster	No.	No.	Source code	No.	Cell lines	No.	Cell lines
1	18	17	ePSC-UN	6	WA09	2	Miz5
				1	BG01	2	Miz4
				2	BG03	2	Miz6
				1	hES1	1	Hues7
		1	tPSC-UN	1	2102Ep		
2	11	11	B-NLin	4	SM-1/-2		
				2	SC23		
				2	THD-hFB17/-hWB15		
				2	HFB2050		
				1	HFT13		
3	12	6	BM-MC	2	SC41AMSC		
				2	SC31AMSC		
				2	BMSC-21/-25		
		4	B-AS	2	SC01Glia	2	Astros
		2	ePSC-XE	2	PEL-cells		
4	4	4	CT-Fib	2	SC33Fib		
				2	SC30Fib		
5	5	3	ePSC-UN	2	Hues13		
				1	Hues7		
		4	ePSC-NLin	2	NSWA09-4		
				2	ES-NSC-N5		
		1	B-NLin	1	HANSE8		
6	32	32	B-NLin	6	HANSE1¶	2	HANSE6¶
				4	HANSE2¶	1	HANSE8¶
				2	HANSE3¶		
				8	HANSE4¶		
				7	HANSE5¶		

Cluster	No.	No.	Source code	No.	Cell lines
7	6	6	B-NLin	2	HNSC-CTX
				2	HNSC-HIP
				2	HOPC
8	6	6	ePSC-EB	2	Miz5
				2	Miz4
				2	Miz6
9	20	20	ePSC-NLin	4	NSWA09-1/-2
				5	R-ES-NSC
				2	SNU16-NS
				5	ES-NSC-N5/-N6/-N7
				4	ES-NSC-N8/-N9
10	8	8	ePSC-NLin	2	ES-NSC-N1
				2	ES-NSC-N2
				2	ES-NSC-N3
				2	ES-NSC-N4
11	11	8	B-NLin	4	SC30
				3	APC
				1	HFT13
		2	CT-Fib	2	HS27
		1	ePSC-NLin	1	NSWA09-1
12	17	17	tPSC-NLin	8	NT2
				6	NT2CM
				3	NT2PA6

Figure 2 | Clusters of samples based on machine learning algorithm. Samples were distributed on the basis of their transcriptional profiles into consensus clusters using sNMF. **a**, Consensus matrix from consensus clustering results (centre matrix plot). The consensus matrix is a visual representation of the clustering results and the separation of the sample clusters from each other. Blue indicates no consensus; red indicates very high consensus. The numbers (1–12) on the diagonal row of clusters indicate the number assigned to the cluster by sNMF. These numbers (cluster 1 to cluster 12) are used throughout the text to indicate the group of samples in that cluster. The bar graph above the consensus matrix plot shows the summary statistics assessing the overall quality of each cluster. The cluster consensus value (0–1) is plotted above the corresponding cluster in the matrix plot. Note that most clusters (clusters 10, 12, 6, 4, 9, 1, 8, 11, 7 and 2) have a high-quality measurement. To the left of the consensus matrix is another view of the consensus data, visualized as a dendrogram. This is a representation of the hierarchical clustering tree of the consensus matrix. **b**, The content of the sample clusters resulting from the same sNMF run are displayed. Numbers are the same cluster numbers assigned by the consensus clustering algorithm that are used throughout the text and figures. For more information on samples, source code and references see Supplementary Tables 1–10. No., number of samples. The symbol '¶' indicates that samples were derived from adult brain specimens.

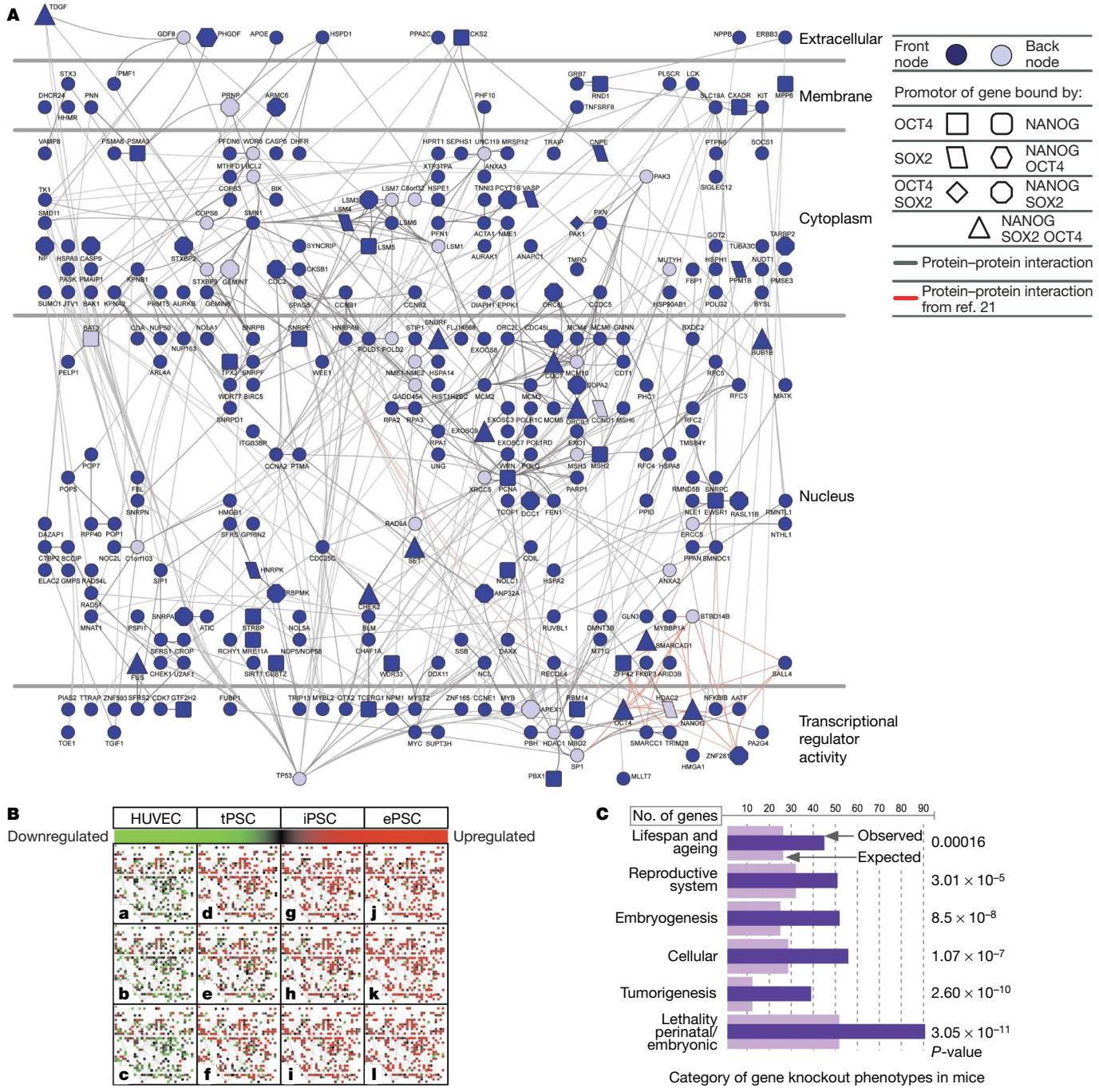


Figure 3 | Pluripotent stem-cell-specific protein–protein interaction network detected by MATISSE. Clusters from the sNMF $k = 12$ analysis were used in combination with the transcriptional database to identify protein–protein interaction networks enhanced in PSCs. **A**, A large differentially expressed connected subnetwork (PluriNet) shows the dominance of cell cycle regulatory networks in PSCs (see legend). All of the dark blue symbols are genes that are highly expressed in most PSCs compared to the other cell samples in the data set. Front nodes, as represented by stem cell matrix expression data, and back nodes, as inferred by MATISSE, are displayed with different colour shades[6]. Highlighted in red are the interactions of a group of proteins associated with pluripotency in murine ePSCs[21]. This subnetwork shows a significant enrichment in genes that are targeted in the genome by the transcription factors NANOG ($P = 5.88 \times 10^{-4}$), SOX2 ($P = 0.058$) and E2F ($P = 1.29 \times 10^{-16}$, all *P*-values are Bonferroni corrected). For an interactive visualization of PluriNet, see http://www.stemcellmatrix.org. **B**, Heat-map-like visualization of PluriNet genes for samples from the test data set: HUVECs (UC-EC, **a**–**c**, derived from three independent individuals), germ cell tumour-derived pluripotent stem cells (tPSC-UN, **d**–**f**, lines GCT-C4, GCT-72, GCT-27X, derived from three independent individuals), induced pluripotent stem cells (iPSC-UN, **g**–**i**, BJ1-iPS12, MSC-iPS1, hFib2-iPS5, three independently derived lines from different somatic sources) and embryonic stem cells (ePSC-UN, **j**–**l**, lines Hues22, HSF6, ES2, derived from three independent blastocysts in three independent laboratories). Most PluriNet genes are markedly upregulated in iPSC-UN and ePSC-UN cells. tPSC-UN cells show a less consistent expression pattern. UC-EC cells show lower expression levels of most PluriNet genes. See Supplementary Fig. 5 for a larger version of the same heat maps. **C**, Analysis of genes from PluriNet in the context of phenotypes that have been reported to result from specific genetic manipulations (for example, gene knockout) in mice in the MGI 3.6 phenotype ontology database (http://www.informatics.jax.org/). We find significant over-representation of phenotypes 'lethality (perinatal/ embryonic)', 'tumorigenesis', 'cellular', 'embryogenesis', 'reproductive system' and 'lifespan and ageing' among the genes in PluriNet. Although these broad categories might be rather unspecific surrogate markers for PSC function in mammals, this analysis might point towards PluriNet's role *in vivo*. For more details, see also Supplementary Fig. 6 and Supplementary Table 12.

Table 1 | PluriNet expression patterns in various model systems for pluripotency

a Expression of PluriNet genes in murine model systems

Cell type	Upregulated/downregulated
MII oocytes	Upregulated*
Zygote	Upregulated*
Embryo (two-cell blastocyst)	Upregulated*
ePSC	Upregulated†
EpiSC	Upregulated†
iPSC	Upregulated†
Fibroblasts (normal)	Downregulated†
Fibroblasts (transformed)	Downregulated†

b Successful PluriNet-based, post-hoc classification in murine model systems

Cell type	Upregulated/downregulated	Pluripotency (PAM)	Germline transmission (PAM)
ePSC	Upregulated	Yes‡	Yes‡
EpiSC	Upregulated	Yes‡	Yes‡
iPSC	Upregulated	Yes‡	Yes‡
Fibroblasts (normal)	Downregulated	Yes‡	Yes‡
Fibroblasts (transformed)	Downregulated	Yes‡	Yes‡

c Expression of PluriNet genes in human model systems

Cell type	Upregulated/downregulated
MII oocytes	Upregulated§
tPSC	Upregulated‖
ePSC	Upregulated‖¶
iPSC	Upregulated‖¶
ePSC-derived cell types	Downregulated‖
Somatic cell types	Downregulated‖¶
Somatic cancer cell line (HeLa)	Downregulated#

d Successful PluriNet-based, post-hoc classification in human model systems

Cell type	Upregulated/downregulated	Pluripotency (PAM)
tPSC	Upregulated	Yes**
ePSC	Upregulated	Yes**
iPSC	Upregulated	Yes**
ePSC-derived cell types	Downregulated	Yes**
Somatic cell types	Downregulated	Yes**

This table summarizes the expression patterns of PluriNet in various model systems of pluripotency and differentiation. More details on the specific tests and explanations of the data sources for the results can be found as indicated below. EpiSC, epiblast-derived stem cells[24]; PAM, prediction analysis of microarray, classifier with leave-one-out cross validation[27]. 'Yes' in parts **b** and **d** indicates correct classification of pluripotent state (pluripotent or not pluripotent) in >90% of samples.

* For more details see Supplementary Figs 8 and 9.
† For more details see Supplementary Fig. 10.
‡ For more details see Supplementary Fig. 10.
§ For more details see Supplementary Fig. 7.
‖ For more details see Fig. 3B and Supplementary Figs 5 and 12.
¶ For more details see Supplementary Fig. 11.
# For more details see Supplementary Discussion 2.
** For more details see Supplementary Fig. 12.

Because the PSC preparations were consistently clustered together we used MATISSE to look for distinctive molecular networks that might be associated with the unique PSC qualities of pluripotency and self-renewal. A Nanog-associated regulatory network has been outlined in mouse embryonic PSCs[21], and we looked for the elements of this network in human PSCs using our unbiased algorithm. We found that the algorithm predicts that human PSCs possess a similar NANOG-linked network (Fig. 3A; elements labelled in red). However, we also discovered that the human NANOG network seems to be integrated as a small component of a much larger protein–protein interaction network that is upregulated in human PSCs (Fig. 3). Notably, this PSC-specific network (termed pluripotency-associated network, PluriNet) contains key regulators that are involved in the control of cell cycle, DNA replication, DNA repair, DNA methylation, SUMOylation, RNA processing, histone modification and nucleosome positioning (see also Supplementary Discussion 2 and http://www.openstemcellwiki.org). Many of the genes in the PluriNet have been linked to embryogenesis, tumorigenesis and ageing (Fig. 3C and Supplementary Fig. 6). We further explored the hypothesis that pluripotency is closely linked to PluriNet expression by analysing published gene expression data sets from human oocytes, various types of PSCs and murine embryos (see Table 1 for a summary of our findings in various model systems). Analysis of a microarray data set[22] that spans development from murine oocytes to the late blastocyst stage revealed that the PluriNet expression is dynamic and upregulated during early mammalian embryogenesis (Table 1 and Supplementary Figs 7–9)[23]. Also, our preliminary analyses indicate that the PluriNet is strongly upregulated in mouse PSCs, mouse iPSCs and mouse epiblast-derived stem cells[24] when compared to somatic cells. Therefore the PluriNet may be useful as a biologically inspired gauge for classifying both murine and human PSC phenotypes (Table 1 and Supplementary Figs 10–13).

Our data indicate that an unbiased global molecular profiling approach combined with a transcriptional phenotype collection using suitable machine learning algorithms can be used to understand and codify the phenotypes of stem cells[4,5,25]. Although it is more extensive than any stem cell data set reported so far, we consider our database and the PluriNet to be a work in progress. As more direct evidence for protein–protein interactions in human cells becomes available, it will be possible to refine the networks we have defined and make them more useful for testing hypotheses about the nature of stem cell pluripotency and multipotency. Also, our sample collection is limited to pluri- and multipotent stem cell types that grow well in culture, and does not include some of the most well studied lineages, such as haematopoietic stem cells. Resolution and reliability of a context-based unsupervised classification can be expected to grow with the breadth and depth of the database content[26]. Even with these limitations, we have shown that the data set and PluriNet have already proved useful for categorizing cell types using unbiased criteria. As more stem cell populations become available, cultured by new methods, isolated from new sources, or induced by new methods, we will use the PluriNet and the stem cell matrix as a reference system for phenotyping the cells and comparing them with existing cell lines.

METHODS SUMMARY

For an overview of the general workflow, please also refer to Fig. 1. A detailed list of the samples, culture methods and reference publications is provided in Supplementary Information[11]. Generally, RNA from each sample was prepared from approximately 1×10^6 cultured cells. Sample amplification, labelling and hybridization on Illumina WG8 and WG6 Sentrix BeadChips were performed for all arrays in this study according to the manufacturer's instructions (http://www.illumina.com) at a single Illumina BeadStation facility. We used the Consensus Clustering framework[9] to cluster transcription profiles and to assess stability of the results. As the algorithm, we used sparse non-negative matrix factorization[5]. For data perturbation, 30 subsampling runs were performed for each considered number of clusters (k). In each run, 80% of the data was subjected to ten random restarts. The R-script can be downloaded at http://www.stemcellmatrix.org. Details on the application of GSA[20], PAM[27], MATISSE[6] as well as publicly available data sets used in this study can be found in the Methods section. We modified the MATISSE[6] computational framework to fit the goals of this study. For the present analysis we used the human physical interaction network that we had previously assembled[6] and augmented it with additional interactions from recent publications[21,28,29]. The 64 interactions in ref. 21 were mapped to the corresponding human orthologues using the NCBI HomoloGene database.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 December 2007; accepted 26 June 2008.
Published online 24 August 2008.

1. Müller, F. J., Snyder, E. Y. & Loring, J. F. Gene therapy: can neural stem cells deliver? *Nature Rev. Neurosci.* **7**, 75–84 (2006).
2. Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132**, 661–680 (2008).
3. Adewumi, O. *et al.* Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnol.* **25**, 803–816 (2007).
4. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).

5. Gao, Y. & Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **21,** 3970–3975 (2005).
6. Ulitsky, I. & Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* **1,** 8 (2007).
7. Carpenter, M. K., Rosler, E. & Rao, M. S. Characterization and differentiation of human embryonic stem cells. *Cloning Stem Cells* **5,** 79–88 (2003).
8. Goldman, B. Magic marker myths. *Nature Reports Stem Cells*. doi:10.1038/stemcells.2008.26 (2008).
9. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52,** 91–118 (2003).
10. Palmer, T. D. *et al.* Cell culture. Progenitor cells from human brain after death. *Nature* **411,** 42–43 (2001).
11. Schwartz, P. H. *et al.* Isolation and characterization of neural progenitor cells from post-mortem human cortex. *J. Neurosci. Res.* **74,** 838–851 (2003).
12. Kornblum, H. I. & Geschwind, D. H. Molecular markers in CNS stem cell research: hitting a moving target. *Nature Rev. Neurosci.* **2,** 843–846 (2001).
13. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
14. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
15. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–1920 (2007).
16. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451,** 141–146 (2008).
17. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448,** 313–317 (2007).
18. Zeng, X. *et al.* BG01V: a variant human embryonic stem cell line which exhibits rapid growth after passaging and reliable dopaminergic differentiation. *Restor. Neurol. Neurosci.* **22,** 421–428 (2004).
19. Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N. Engl. J. Med.* **350,** 1353–1356 (2004).
20. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1,** 107–129 (2007).
21. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444,** 364–368 (2006).
22. Wang, Q. T. *et al.* A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev. Cell* **6,** 133–144 (2004).
23. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450,** 1230–1234 (2007).
24. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448,** 196–199 (2007).
25. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
26. Donoho, D. & Stodden, V. When does non-negative matrix factorization give correct decomposition into parts? *Proc. NIPS* (2003) ⟨http://books.nips.cc/papers/files/nips16/NIPS2003_LT10.ps.gz⟩.
27. Lacayo, N. J. *et al.* Gene expression profiles at diagnosis in *de novo* childhood AML patients identify FLT3 mutations with good clinical outcomes. *Blood* **104,** 2646–2654 (2004).
28. Ewing, R. M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89 (2007).
29. Mishra, G. R. *et al.* Human protein reference database–2006 update. *Nucleic Acids Res.* **34,** D411–D414 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Stubban, H. Dittmer, S. Zapf and H. Meissner for their work with various cell cultures. We are grateful to D. Wakeman, R. Gonzalez, S. McKercher, J. P. Lee, H.-S. Park and S. Y. Moon for sharing their cell preparations for the type collection. We are also grateful to R. Wesselschmidt and M. Pera for their unique GCT lines and G. Daley for providing human iPSCs. A. M. Kocabas and J. Cibelli shared their human oocyte expression data with us. A. Barsky let us use the Cerebral 2.0 plug-in before its publication. M. Rosentraeger helped to compile the cell culture metadata. We thank J. Aldenhoff, D. Hinze-Selch, M. Westphal, K. Lamszus, U. Kehler, D. Barker and A. Fritz for their support and discussions of this project. This study has been supported by the following grants and awards: Christian-Abrechts University Young Investigator Award (F.-J.M.), SFB-654/C5 Sleep and Plasticity (F.-J.M. and D. Hinze-Selch), Hamburger Krebsgesellschaft Grant (N.O.S.), Edmond J. Safra Bioinformatics program fellowship at Tel-Aviv University (I.U.), Converging Technologies Program of The Israel Science Foundation Grant No 1767.07 (R.S.), Raymond and Beverly Sackler Chair in Bioinformatics (R.S.), Reproductive Scientist Development Program Scholar Award K12 5K12HD000849-20 (L.C.L.), California Institute for Regenerative Medicine Clinical Scholar Award (L.C.L.), NIH P20 GM075059-01 (J.F.L.), the Alzheimer's Association (J.F.L.), and anonymous donations in support of stem cell research.

Author Contributions J.F.L. and F.-J.M. designed the study and wrote the manuscript; I.U., R.W., D.K., R.S., L.C.L. and F.-J.M. designed and conducted the bioinformatics analysis; L.C.L., C.L., P.H.S., M.S.R., I.-H.P., F.-J.M. and N.O.S. conducted experiments and provided essential materials for this study.

Author Information The microarray data have been deposited at NCBI GEO (accession number GSE11508) and can also be accessed, processed and downloaded at http://www.stemcellmesa.org. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.-J.M. (fj.mueller@zip-kiel.de) or J.F.L. (jloring@scripps.edu).

METHODS

Compilation of type collection. Samples were either grown in our own laboratory or provided by collaborators. Each sample was prepared from approximately 1×10^6 cultured cells, which were mechanically harvested, pelleted and snap frozen in liquid nitrogen. Biological replicates were produced for almost all samples. Details on the included cell lines and culture methods can be found in the Supplementary Tables 3–8.

Neural progenitor cultures (HANSE) from neurosurgical specimens. Brain tissue samples were obtained from patients who underwent surgery for intractable temporal lobe epilepsy at the Department of Neurosurgery, University Medical Center Hamburg-Eppendorf, Germany ($n = 6$; 4 males and 2 females; mean age 33). All procedures were performed with informed consent and in accordance with institutional human tissue handling guidelines. We used modifications of reported protocols for establishing neural progenitor cultures from fetal and postmortem brain tissue[10,30]. A more detailed description can be found in Supplementary Methods 1.

Whole-genome gene expression. All RNA was purified in our laboratory using standard methods. Sample amplification, labelling and hybridization on Illumina WG8 and WG6 Sentrix BeadChips were performed for all arrays in this study according to the manufacturer's instructions (Illumina) using an Illumina BeadStation (Burnham Institute Microarray Core).

Microarray data pre-processing. Raw data extraction was performed with BeadStudio v1.5 and probes with a detection score of less than 0.99 in all of the samples were discarded. The resulting probes were then quantile-normalized to correct for between-sample variation[31]. The sample data were quality controlled before normalization using the quality parameters provided by BeadStudio software. Before and after normalization the arrays were inspected with signal distribution box plots and by using the maCorrPlot package[32].

Parameters for unsupervised classification. The data sets and the sparseness factor λ were adjusted for the unsupervised clustering task following previous reports[4,5]. Parameters we have used for this study are: SCM core data set (153 samples), $\lambda = 0.01$; SCM test data set (219 samples), $\lambda = 0.1$. The pre-processed data sets used can be downloaded at http://www.stemcellmatrix.info.

Gene expression and gene set analysis. To screen for differentially expressed groups of genes between computationally defined sample clusters we used the Gene Set Analysis (GSA) methods proposed previously[33,34]. GSA was chosen because it uses a stringent max-mean algorithm to identify significantly differentially regulated gene sets. The cutoff *P*-value was adjusted to accommodate a false discovery rate (FDR) of 10%. A translation file was built to use GSA with Illumina expression data. We collected gene lists from recent publications and public repositories (MolSigDB2, Stanford repository). These files can be downloaded from http://www.stemcellmatrix.org. To screen for differentially expressed genes between computationally defined sample clusters we used standard *t*-test-based methods implemented in the R Bioconductor package[35]. The cutoff *P*-value was adjusted to accommodate a FDR of 5%.

Detection of cluster-specific subnetworks using MATISSE. MATISSE[6] (http://acgt.cs.tau.ac.il/matisse) was adjusted to detect differentially expressed connected subnetworks (DECSs), corresponding to connected subnetworks in a physical interaction network that show a significant co-expression pattern. The physical network used by MATISSE contains vertices corresponding to genes and edges corresponding to protein–protein and protein–DNA interactions. For the present analysis we used the human physical interaction network that we had previously assembled[6] and augmented it with additional interactions from recent publications[21,28,29]. In total, the network contained 34,212 interactions among 9,355 proteins.

Originally, MATISSE used the Pearson correlation coefficient as a measure of similarity between the expression patterns of gene pairs. These similarity values serve as a starting point for the computation of pair-wise weights using a probabilistic model. The Pearson correlation between a pair of genes captures a global similarity trend between their expression patterns. In this work we were interested in extracting groups of genes that are not only similar across the experimental conditions, but also show significantly high or significantly low expression values in a specific subset of the samples, identified using the sNMF clustering scheme. To this end we devised a hybrid similarity score that captures two features: (1) both genes show differential expression; (2) the genes have similar expression patterns, regardless of their differential expression.

We denote the expression pattern of gene i by $x^i = (x^i_1, x^i_2, \ldots, x^i_m)$. Assume that we are interested in DECSs upregulated in a condition subset $A \subseteq \{1,\ldots,m\}$. To address goal (1), we use an 'ideal' expression profile $p = (p_1, p_2, \ldots, p_m)$ where $p_i = 1$ if $i \in A$ and $p_i = -1$ otherwise. The signs are reversed if we are interested in a DECS downregulated in A. r_{kp} is the Pearson correlation coefficient between x^k and p. Intuitively, r_{kp} is close to 1 if the corresponding transcript is strongly upregulated in A compared to the other conditions, and close to -1 if it is strongly downregulated in A. This measure has been suggested as an aparametric differential expression score[36]. Note that the Pearson correlation is invariant under normalization of the patterns to zero mean and standard deviation of 1. For every gene pair (i,j) we compute $S_{\mathrm{diff}}(i,j) = (r_{ip} + r_{jp})/2$. To address goal (2) we use the partial correlation coefficient between the gene patterns conditioned on the ideal profile. Formally, $S_{\mathrm{part}}(i,j) = \dfrac{r_{x^i,x^j} - r_{x^i,p} r_{x^j,p}}{\sqrt{(1-r^2_{x^i,p})(1-r^2_{x^j,p})}}$, where r_{yz} is the Pearson correlation coefficient between the profiles y and z. Intuitively, S_{part} conveys the information about how similar x^i and x^j are, regardless of their differential expression in A. Finally, we use the similarity score $S = \lambda S_{\mathrm{diff}} + S_{\mathrm{part}}$, where λ is a trade-off parameter setting the relative importance of the differential expression in the similarity score. We used $\lambda = 3$ for the analysis described in this paper. These S scores are then modelled using the probabilistic model described previously[6]. The advantage of using this pair-wise scoring scheme over the use of gene-specific differential expression scores, such as those proposed by others[37], is that it will prefer gene groups that are not only differentially expressed in the specified condition subset, but also have coherent expression profiles.

To diminish the effect of the size difference between the clusters, we reduced the number of conditions in clusters 1, 2, 3, 6, 9, 10 and 12, by including fewer replicates. Overall, 105 samples were used in the MATISSE analysis and can be downloaded at http://www.stemcellmatrix.org. We executed this MATISSE variant iteratively, each time setting A to contain all the samples of a single cluster or a cluster pair. The upper bound on module size was set to 300 and the rest of the parameters were as previously reported[6]. We then filtered the resulting networks by removing DECSs that overlapped more than 50% with other, higher scoring DECSs. The full set of the DECSs is available at http://www.stemcellmatrix.org.

Visualization. For visualization of the selected DECS we used Cytoscape 2.5 (ref. 38) and Cerebral 2.0 (ref. 39). Localization data from HRPD and the GO-Molecular function categories were also used[29]. NANOG, POU5F1/OCT4 and SOX2 promoter binding information was used to code the ESC-specific regulation of nodes[40]. Permutmatrix was used for heat maps[41]. Data for the analysis of human oocytes were accessed on the authors' or the journals' website[42]. For analysis of iPSCs induced with LIN28, OCT4, NANOG and SOX2, the data set was obtained from the Thomson laboratory[15].

Classification based on PluriNet. We used the 299 genes from DECS (Up(1,5)A) (PluriNet) with the PAM[20] software package. Class probabilities were re-computed 10,000 times; average scores are reported in Supplementary Figs 10 and 12. We translated the human genes into their murine orthologues from PluriNet using the NCBI HomoloGene database for re-analysing murine expression profiles. The expression array data from murine fibroblasts, induced pluripotent cells, epiblast-derived stem cells and murine embryonic stem cells were downloaded from NCBI GEO[21–24].

30. Imitola, J. *et al.* Directed migration of neural stem cells to sites of CNS injury by the stromal cell-derived factor 1α/CXC chemokine receptor 4 pathway. *Proc. Natl Acad. Sci. USA* **101,** 18117–18122 (2004).
31. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. & Pavlidis, P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* **33,** 5914–5923 (2005).
32. Ploner, A., Miller, L. D., Hall, P., Bergh, J. & Pawitan, Y. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics* **6,** 80 (2005).
33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
34. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1,** 107–129 (2007).
35. R Development Core Team, R. A language and environment for statistical computing, help files. ⟨http://www.bioconductor.org/⟩ (2007).
36. Troyanskaya, O., Garber, M., Brown, P., Botstein, D. & Altman, R. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18,** 1454–1461 (2002).
37. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** (suppl. 1), S233–S240 (2002).
38. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2,** 2366–2382 (2007).
39. Barsky, A., Gardy, J. L., Hancock, R. E. & Munzner, T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23,** 1040–1042 (2007).
40. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
41. Caraux, G. & Pinloche, S. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* **21,** 1280–1281 (2005).
42. Kocabas, A. *et al.* The transcriptome of human oocytes. *Proc. Natl Acad. Sci. USA* **103,** 14027–14032 (2006).

4. Detecting Disease-Specific Dysregulated Pathways via Analysis of Clinical Expression Profiles

Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles

Igor Ulitsky[1], Richard M. Karp[2], and Ron Shamir[1]

[1] School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel
({ulitskyi,rshamir}@post.tau.ac.il.)
[2] International Computer Science Institute, 1947 Center St., Berkeley, CA 94704
(karp@icsi.berkeley.edu)

Abstract

We present a method for identifying connected gene subnetworks significantly enriched for genes that are dysregulated in specimens of a disease. These subnetworks provide a signature of the disease potentially useful for diagnosis, pinpoint possible pathways affected by the disease, and suggest targets for drug intervention. Our method uses microarray gene expression profiles derived in clinical case-control studies to identify genes significantly dysregulated in disease specimens, combined with protein interaction data to identify connected sets of genes. Our core algorithm searches for minimal connected subnetworks in which the number of dysregulated genes in each diseased sample exceeds a given threshold. We have applied the method in a study of Huntington's disease caudate nucleus expression profiles and in a meta-analysis of breast cancer studies. In both cases the results were statistically significant and appeared to home in on compact pathways enriched with hallmarks of the diseases.

1 Introduction

Systems biology has the potential to revolutionize the diagnosis and treatment of complex disease by offering a comprehensive view of the molecular mechanisms underlying the pathology. To achieve these goals, a computational analysis extracting mechanistic understanding from the masses of available data is needed. To date, such data include mainly microarray measurements of genome-wide expression profiles, with over 160,000 profiles stored in GEO alone as of August 2007. A wide variety of approaches for elucidating molecular mechanisms from expression data have been suggested [1]. However, most of these methods are effective only when using expression profiles obtained under diverse conditions and perturbations, while the bulk of data currently available from clinical studies are expression profiles of groups of diseased individuals and matched controls. The standard "pipeline" for analysis of such datasets involves the application of statistical and machine learning methods for identification of the genes that best predict the pathological status of the samples [2]. While these methods are successful in identifying potent signatures for classification purposes, the insights that can be obtained from examining the gene lists they produce are frequently limited [3].

It is thus desirable to develop computational tools that can extract more knowledge from clinical case-control gene expression studies. A challenge of particular interest is to identify the pathways involved in the disease, as such knowledge can expedite

development of directed drug treatments. One strategy of solution to this problem uses predefined gene sets describing pathways and quantifies the change in their expression levels [4]. The drawback of this approach is that pathway boundaries are often difficult to assign, and in many cases only part of the pathway is altered during disease. To overcome these problems, the use of gene networks has been suggested [5]. The appeal of using network information increases as the quality and scale of experimental data on such interaction networks improve.

Several approaches for integrating microarray measurements with network knowledge were described in the literature. Some (including us) proposed computational methods for detection of subnetworks that show correlated expression [6, 7]. A successful method for detection of 'active subnetworks' was proposed by Ideker *et al.* and extended by other groups [8–12]. These methods are based on assigning a significance score to every gene in every sample and looking for subnetworks with statistically significant combined scores. Breitling *et al.* proposed a simple method named GiGA which receives a list of genes ordered by their differential expression significance and extracts subnetworks corresponding to the most differentially expressed genes [13]. Other tools use network and expression information together for classification purposes [5, 14].

Methods based on correlated expression patterns do not use the sample labels, and thus their applicability for case-control data is limited, as correlation between transcript levels can stem from numerous confounding factors not directly related to the disease (e.g., age or gender). The extant methods that do use the sample labels rely on the assumption that the same genes in the pathway are differentially expressed in all the samples (an exception is jActiveModules which can identify a subset of the conditions in which the subnetwork is active [8]). This assumption may hold in simple organisms (e.g., yeast or bacteria) or in cell line studies. However, in human disease studies, the samples are expected to exhibit intrinsic differences due to genetic background, environmental effects, tissue heterogeneity, disease grade and other confounding factors. Here we propose a new viewpoint for analysis of clinical gene expression samples in the context of interaction networks, which avoids the above assumption.

Our approach aims to detect subnetworks in which multiple genes are dysregulated in the diseased specimens, while allowing for distinct affected gene sets in each patient. We call such modules *dysregulated pathways* (DPs). Specifically, we look for minimal connected subnetworks in which the number of dysregulated genes in each diseased sample exceeds a given threshold. By comparing to statistics of randomized networks, we can identify statistically significant DPs. As finding such modules is NP-hard, we propose heuristics and algorithms with provable approximation ratios and study their performance on real and simulated data. Our approach has several important advantages over the existing methods: (a) the dysregulated genes in a DP can vary between patients; (b) the method is robust to outliers (i.e., patients with unusual profiles); (c) the DPs can contain relevant genes based on their interaction pattern, even if they are not dysregulated; (d) it has only two parameters, both of which have an intuitive biological interpretation; (e) while not guaranteeing optimality, the algorithmic backbone of the method has a provable performance guarantee.

We first tested the performance of our method on simulated data. We then used it to dissect the gene expression profiles of samples taken from the caudate nucleus of

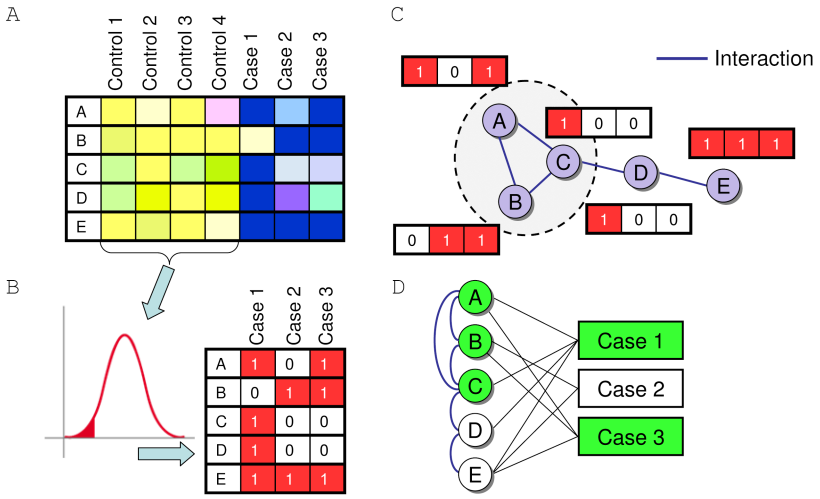


Fig. 1. From case-control profiles to dysregulated pathways. (A) The first input to our method is the gene expression matrix where the columns correspond to samples taken from case/control subjects and rows correspond to genes. (B) In a preprocessing step, differential expression is analyzed and, for each gene, the set of cases in which it is differentially expressed (up-regulated, down-regulated or both) is extracted. (C) A second input is a protein interaction network with nodes corresponding to genes and edges to interactions. The row next to each gene is its dysregulation pattern (its row from B). The goal is to find a smallest possible subnetwork in which, in all but l cases, at least k genes are differentially expressed. In this example, the circled subnetwork satisfies the condition with $k = 2, l = 1$: (i) A and C are dysregulated in case 1; (ii) A and B are dysregulated in case 3. (D) The bipartite graph representation of the data. Genes (left) are connected to the cases (right) in which they are differentially expressed. Edges between genes constitute the protein interaction network. The genes of the minimal cover and the samples covered by them are in green.

Huntington's Disease (HD) patients. We reveal specific subnetworks that are up and down regulated in cases in comparison to controls, and show that they are significantly enriched with known HD-related genes. Finally, we performed a network-based meta-analysis of six breast cancer datasets, extracting DPs associated with good and poor outcome of the disease. In all cases, the DPs are significantly enriched with genes from relevant pathways and contain both known and novel potential drug targets.

For lack of space, some details and proofs are not included in this manuscript.

2 Methods

2.1 Problem formulation

In this section we describe the theoretical foundations of our methodology (**Fig. 1**). The known gene network is presented as an undirected graph, where each node (gene) has a corresponding set of elements (samples) in which it is differentially expressed. Our goal is to detect a DP, which is a minimal connected subnetwork with at least k nodes differentially expressed in all but l analyzed samples (l thus denotes of the number of allowed 'outliers').

We formalize these notions as follows. We are given an undirected graph $G = (V, E)$ and a collection of sets $\{S_v\}_{v \in V}$ over the universe of elements U, with $|U| =$

n. For ease of representation, we will use, in addition to G, a bipartite graph $B = (V, U, E^B)$ where $(v, u) \in E^B, v \in V, u \in U$ if and only if $u \in S_v$ (**Fig. 1D**). A set $C \subseteq V$ is a *connected* (k, l)*-cover* (denoted $CC(k, l)$) if C induces a connected component in G and a subset $U' \subseteq U$ exists such that $|U'| = n - l$ and for all $u' \in U'$, $|N(u') \cap C| \geq k$, i.e., in the induced subgraph (C, U') the minimal degree of nodes in U' is at least k ($N(x)$ is the set of neighbors of x in B). We are interested in finding a $CC(k, l)$ of the smallest cardinality. We denote this minimization problem by ***MCC(k,l)***.

2.2 Similar problems and previous work

If G is a clique, $MCC(1, 0)$ is equivalent to the Set Cover problem [15]. For this classical NP-hard problem, Johnson proposed a simple greedy algorithm with approximation ratio $O(ln(n))$ [15]. If $k > 1$ and G is a clique, the $MCC(k, 0)$ problem is equivalent to the *set multicover problem*, also known as the *set k-cover problem*, a variant of the Set Cover problem in which every element has to be covered k times. The set multicover problem can be approximated to factor of $O(p)$, where p is the number of sets covering the element that appears in the largest number of sets [15]. The greedy algorithm for set multicover was shown to achieve an approximation ratio of $O(\log(n))$ [16]. See [15] for a comprehensive review of the available approximation results on set cover and set multicover problems.

For a general G, $MCC(1, 0)$ is the *Connected Set Cover problem*, which has been recently studied in the context of wavelength assignment of broadcast connections in optical networks [17]. It was shown to be NP-Hard even if at most one vertex of G has degree greater than two, and approximation algorithms were suggested for the cases where G is a line graph or a spider graph. Both of these special cases are not applicable in our biological context.

2.3 Greedy algorithms for $MCC(k, l)$

We tested two variants of the classical greedy approximation for Set Cover. For simplicity we will describe them for $MCC(1, 0)$. The first algorithm, *ExpandingGreedy* works as follows: Given a partial cover $W \subseteq V$ and the set of corresponding covered elements $X \subseteq U$, the algorithm picks a node $v \in V$ that is adjacent to W and that covers the largest number of elements of $U \setminus X$, adds v to the cover and adds $N(v) \cap U$ to X. Initially $W = \emptyset$, $X = \emptyset$ and the first node is picked without connectivity constraints. Unfortunately, *ExpandingGreedy* can be shown to give a solution that is $O(|V|)$ times the optimal solution. Specifically, it runs into difficulties in cases where all the nodes in the immediate neighborhood of the current solution have equal benefit, and the next addition to the cover is difficult to pick. The second algorithm, *ConnectingGreedy*, first uses the simple greedy algorithm [15] to find a set cover that ignores the connectivity constraints and then augments it with additional nodes in order to obtain a proper cover. The *diameter* of a graph is the maximum length of a shortest path between a pair of nodes in V. It can be shown that *ConnectingGreedy* guarantees an approximation ratio of $O(D \log n)$ for $MCC(1, 0)$, where D is the diameter of G.

2.4 The CUSP algorithm

We next describe an algorithm called Covering Using Shortest Paths (*CUSP*). Let $d(v, w)$ be the distance in edges between v and w in G. For each *root* node r and for each element $u \in U$ the algorithm computes distances $(M[r, u]_1, ..., M[r, u]_k)$ and pointers $(P[r, u]_1, ..., P[r, u]_k)$ to the k nodes closest to r that cover u. This can be done by computing the distances from r to all the nodes in V that cover u, and then retrieving the k closest nodes, which is an instance of the *selection problem* and can be solved in expected linear time [18]. Now take X_r, the union of the paths to the nodes covering the $n - l$ elements for which $max_q\{d(r, P[r, u]_q), 1 \leq q \leq k\}$ are the smallest. X_r is a proper $CC(k, l)$: (a) it is a subtree of T and thus induces a connected component in G; (b) $n - l$ elements of U are covered k times by the corresponding $\{P[r, u]_i\}$. The final solution is $X = \arg\min_v |X_v|$. This algorithm can be proved to give a $k(n - l)$-approximation for $MCC(k, l)$.

In terms of computational complexity, the total amount of work for each choice of r is $O(|V| + |E| + |E^B|)$ and the overall complexity is $O(|V|(|V| + |E| + |E^B|))$. Note that it is not necessary to execute the algorithm from every root node, but only from the $l + 1$ nodes that cover elements from $U' \subseteq U$ for which $max_{u' \in U'}|N(u')|$ is minimal.

2.5 Practical heuristics and implementation details

In order to improve the performance of the proposed algorithms, we implemented several practical heuristics.

CUSP* - starting from high coverage cores: A drawback of CUSP is that it ignores the number of elements covered by each node, and treats the coverage of every element separately. We therefore also implemented the CUSP* heuristic: For each root, it uses dynamic programming to identify a subnetwork of k nodes that offers a good coverage of the elements, and then extends it to a proper $CC(k, l)$ as in CUSP.

Clean-up: The DPs produced by all the described algorithms may contain superfluous nodes that are not necessary neither for the cover requirements nor for subnetwork connectivity. In all algorithms we therefore perform a clean-up step that iteratively removes such nodes until no further reduction is possible.

Shortest path tree construction: While the approximation bound of CUSP holds regardless of the shortest paths used, some sets of such paths may eventually give rise to smaller covers than others. We used the following heuristic in the BFS algorithm: at each level of the constructed BFS tree, we sort the nodes in descending order based on the added coverage they offer. The nodes are then scanned in this order and the next level of the tree is built.

Starting points: The performance of the algorithms depends on the number of starting points/seeds used. In all the results described here we executed all algorithms starting from the 30 nodes that had the highest degrees in B.

Assessment of DP significance: CUSP produces a set of DPs for a range of k values. To select the most significant DP, 200 random networks were generated by degree-preserving randomization [19]. CUSP was executed on each network, for a range of k values, and an empirical p-value was computed. The k value for which the size of the DP was most significant was subsequently used. In case of a tie, a normal distribution

was fitted to the random scores, and k yielding the subnetwork with the most significant z-score was selected.

Finding multiple DPs: After recovering the first DP V_1, we seek additional DPs by removing all the edges adjacent to V_1 from E^B and reapplying the search procedure. This is repeated until no significant DP is found.

Our algorithms were implemented in Java, and source code of the implementation is available upon request. A user-friendly graphical interface for the algorithms described here is currently in development.

3 Results

Human protein interaction network: We compiled a human protein-protein interaction network encompassing 7,384 nodes corresponding to Entrez Gene identifiers and 23,462 interactions. The interactions are based mostly on small-scale experiments and were obtained from several interaction databases. The network and the sources information are available at our website `http://acgt.cs.tau.ac.il/clean`.

3.1 Simulation

We first evaluated the algorithms on simulated data in which a single DP is planted. We used the human protein interaction network as G, created a biclique between a connected subgraph of G and a specified number of elements in U and added noise to B by randomly removing and inserting edges. In the simulations (results not shown) *ExpandingGreedy* generally found the smallest covers. The results produced by CUSP and CUSP* were only slightly inferior. However, the covers produced by CUSP and CUSP* were much more compact, giving a much lower mean shortest path length between nodes in the cover.

3.2 Analysis of Huntington's disease caudate nucleus expression profiles

Huntington's disease (HD) is a devastating autosomal dominant neurological disorder caused by an expansion of glutamine repeats in the ubiquitously expressed huntingtin (*htt*) protein. HD pathology is well understood at a histological level but its effect on the molecular level in the human brain is poorly understood. Recent studies have shown that mutant huntingtin interferes with the function of widely expressed transcription factors, suggesting that gene expression may be altered in a variety of tissues in HD. Hodges *et al.* reported gene expression profiles in grade 0-2 HD brains obtained using oligonucleotide arrays [20]. We focused our analysis on 38 patient samples and 32 unaffected control samples from that study, all taken from the caudate nucleus region of the brain, as this is the region where the disease is manifested the most. For every sample (patient), differentially expressed genes were selected based on comparison to the controls. The expression pattern of each gene was first standardized to mean 0 and standard deviation of 1. For every gene v, a normal distribution was fitted to its expression values in the control group, and for every HD sample u, a one-tailed p-value p_v^u was computed. We

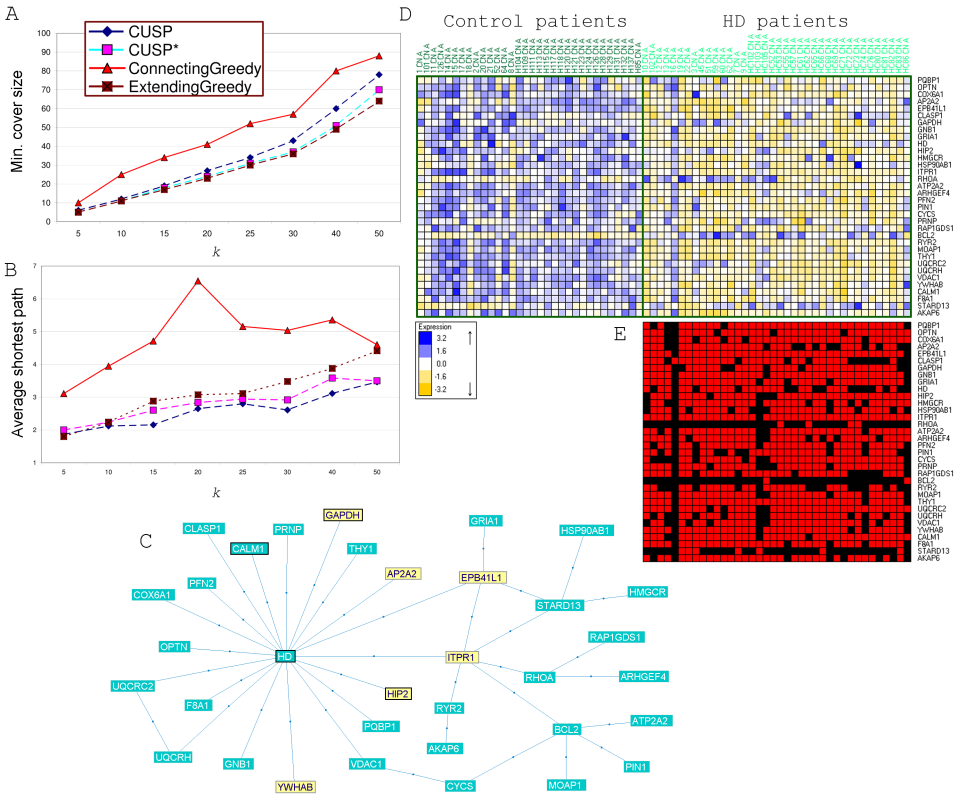


Fig. 2. Subnetwork identified by the CUSP algorithm as down-regulated in the caudate nucleus of HD patients. (A) Comparison of the minimal cover size obtained by the greedy and the CUSP algorithms. (B) Comparison of the average shortest path length between nodes in the minimal cover obtained by the greedy and the CUSP algorithms. (C) The subnetwork obtained for $k = 25$ and $l = 8$. HD modifiers described in [21] are in yellow. KEGG HD pathway genes are drawn with thick border. Note that HD is the official name of huntingtin (*htt*). (D) Heat map of the normalized expression values of the subnetwork genes in the control and HD groups. (E) The subnetwork genes and their differential expression in each HD samples. Red cells correspond to significantly down-regulated genes.

then introduced an edge (v, u) to E^B if and only if $p_v^u < 0.05$. At this significance level, 1,073 (1,696) genes were selected as down (up) regulated in a sample on average.

We first describe the results on down-regulation (**Fig. 2**), using $l = 8$. While CUSP, CUSP* and *ExpandingGreedy* found minimal covers of similar size (**Fig. 2A**), the covers found by CUSP were the most compact, as evident from the average shortest path length between a pair of nodes in the subnetwork (**Fig. 2B**). As compact and dense subnetworks are more likely to correspond to real biological pathways, we used the results of CUSP in further analysis.

Our significance evaluation of the results showed that for values of k between 10 and 40 the cover found was significantly smaller than the one obtained at random, indicating that genes dysregulated in HD are indeed clustered in the network. The most significant DP was obtained for $k = 25$ ($p < 0.005$). It contained 34 genes (**Fig. 2C-E**),

	CUSP	GiGA	jActiveModules	t-test top	t-test FDR < 0.05
Number of genes	34	34	282	34	1762
Contains Huntingtin?	**Yes**	No	No	No	**Yes**
HD modifiers	6 ($7.7 \cdot 10^{-10}$)	3 ($1.55 \cdot 10^{-4}$)	12 ($\mathbf{3.15 \cdot 10^{-11}}$)	2 (0.001)	16 ($3.47 \cdot 10^{-5}$)
HD relevant	7 ($\mathbf{4.29 \cdot 10^{-11}}$)	2 (0.008)	14 ($1.42 \cdot 10^{-9}$)	1 (0.124)	18 ($6.06 \cdot 10^{-5}$)
KEGG HD pathway	4 ($\mathbf{7.95 \cdot 10^{-7}}$)	0	4 (0.003)	0	8 (0.03)
Calcium signaling	6 ($9.23 \cdot 10^{-7}$)	5 ($1.99 \cdot 10^{-5}$)	10 ($5.68 \cdot 10^{-4}$)	3 (0.005)	49 ($\mathbf{2.97 \cdot 10^{-12}}$)

Table 1. Comparison of gene sets identified as down-regulated in HD caudate nucleus using different methods. GiGA was implemented as described in [13] and used to produce a subnetwork of 34 nodes. jActiveModules [8] was executed from Cytoscape and yielded five subnetworks. The reported results are for the highest scoring subnetwork. 't-test top' refers to the 34 down regulated genes with the most significant t-scores. HD modifiers are taken from [21]. HD relevant genes are taken from [23]. Calcium signalling genes are taken from MSigDB [4].

with the *htt* protein as the major hub. Indeed, mutations in *htt* are the cause of the HD pathology. Moreover, the network contains six additional genes identified as genetic modifiers of the HD phenotype in a fly model of the disease [21] (the modifiers are highlighted in **Fig. 2C**). The network is also enriched with genes from the KEGG HD pathway ($p = 7.95 \cdot 10^{-7}$). Furthermore, the network contains at least six genes related to regulation of calcium levels (data taken from MSigDB [4], $p = 9.23 \cdot 10^{-7}$), which is known to be intimately related to HD [22]. An inspection of the expression patterns (**Fig. 2D**) indicates the importance of the outlier parameter l. A few of the samples (patients 16,103,86) have profiles that differ from those of the other patients, but this fact does not affect the algorithm.

A comparison of the DP we identified with gene sets identified using other methods (**Table 1**) reveals that the subnetwork produced by our method is more significantly enriched with most hallmarks of HD. The subnetwork identified by jActiveModules is also enriched for these hallmarks, but this subnetwork is an order of magnitude larger, and thus less focused. The output of jActiveModules consists of (i) the 'active' subnetwork; and (ii) the samples in which the subnetwork is active. In this dataset, the active subnetwork produced by this algorithm was based on a single sample, and thus it does not reflect common dysregulation across most patients in the study.

The running time on this dataset, for $k = 25$, was 10.6 seconds on a PC with two 2.67GHz processors and 4GB of memory. A search for additional down-regulated DPs (see Methods) did not produce significant networks.

Similar analysis of genes up-regulated in HD samples identified a marginally significant subnetwork ($k = 10, p = 0.11$) of 14 nodes centered at BRCA1 and p53, which are master regulators of DNA damage response, and are known to be hyperactive in HD affected cells [24]. Interestingly, p53 and BRCA1 are not differentially expressed in most HD samples, and the functional category 'DNA damage response' is not enriched in the 100 genes most significantly up-regulated in the HD samples (as obtained by a t-test). This further underlines the ability of our method to extract relevant pathways even if only part of the pathway is differentially expressed in diseased specimens. Another hub in this focused subnetwork is HDAC1, a histone deacetylase known to be elevated in HD neurons [25]. Sodium phenylbutyrate, a histone deacetylase inhibitor, is currently tested as a potent drug for HD [26], and was shown to revert HD transcriptional dysreg-

ulation in mouse and human brain and blood tissues [27, 28]. Hence, the inclusion of HDAC1 in a focused subnetwork identified as up-regulated in diseased caudate nuclei demonstrates the ability of our method to detect potential therapeutic targets.

3.3 Meta-analysis of breast cancer studies

In order to test our methodology on other diseases and on inter-study comparisons we performed meta-analysis of six breast cancer studies, spanning together expression profiles of 1,004 patients. Full details on the studies are available at our website. These studies compared breast cancer tumor samples, for which follow-up outcome information was available. We focused on comparison of tumors with good and poor prognosis (defined as development of distant metastases within five years [2]). In each study, using a one-tailed t-test, we extracted a set of differentially expressed genes between good and poor prognosis patients ($p = 0.05$ was used as a threshold). Here we applied CUSP to the genes vs. studies matrix. The most significant DP up-regulated in poor prognosis cancers is shown in **Fig. 3A** ($k = 40, l = 2, p < 0.005$). This network is highly enriched with cell-cycle genes (28 out of 51 genes are associated with cell-cycle in GO, $p = 2.44 \cdot 10^{-26}$). Cell cycle and proliferation genes are known to be associated with higher grade, poor prognosis tumors in numerous studies (see [29] and the references therein). In addition, this DP contains 15 genes shown to be regulated by YY1 (as found in [30], $p = 2.42 \cdot 10^{-16}$), known to be associated with overexpression of the ERBB2 oncogene and with poor prognosis of breast cancer [31]. We recovered an additional significant DP which is described on our website.

The most significant DP down-regulated in poor prognosis cancers ($k = 25, p < 0.005$, **Fig. 3B**) is enriched with genes associated with drug resistance and metabolism (Source:MSigDB, $p = 3.54 \cdot 10^{-9}$), p53 signalling ($p = 3.54 \cdot 10^{-9}$) and the JAK-STAT signalling pathway ($p = 3.68 \cdot 10^{-4}$). The latter pathway mediates the signals of a wide range of cytokines, growth factors and hormones, making its aberrant activation prone to lead to malignancy. This pathway was also linked specifically to breast cancer [32]. Our results indicate the down-regulation of this pathway on the expression level is associated with cancers with poor prognosis. Interestingly, this subnetwork, but not the up-regulated one, was enriched with genes that are frequently mutated in cancer in general ($p = 1.14 \cdot 10^{-7}$) and in breast cancer in particular ($p = 3.2 \cdot 10^{-4}$, both sets taken from [33]). A search for additional DPs did not yield significant results.

4 Discussion

We have developed a novel computational technique for network-based analysis of clinical gene expression data. The method is aimed at identifying pathways in the interaction network that exhibit ample evidence of disruption of transcription that is specific to diseased patients. Application of the method to a large-scale human protein-protein interaction network and a Huntington's disease study as well as meta-analysis of six breast cancer studies has shown its potential in outlining subnetworks with a high relevance to the mechanisms of pathogenesis. Comparison to extant techniques for analysis

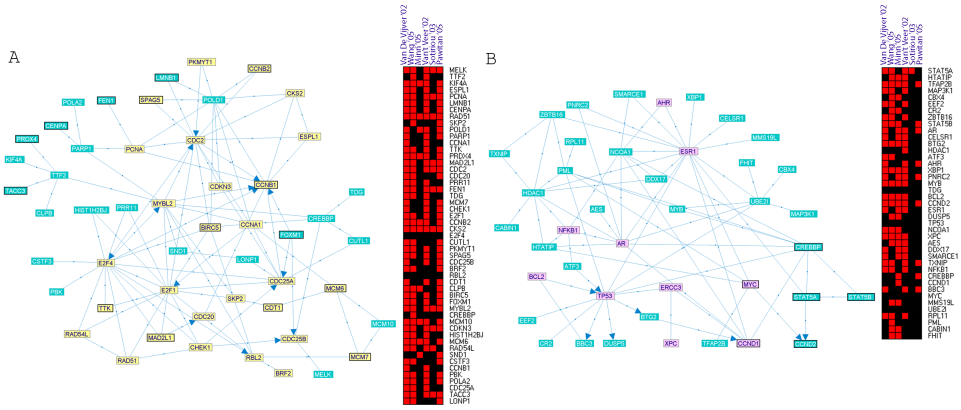


Fig. 3. **DPs identified in breast cancer meta-analysis**. In the differential expression maps (right) red cells correspond to differentially expressed genes. (A) a DP up-regulated in poor prognosis breast cancers ($k = 40, p < 0.005$). Cell cycle genes (from GO) are in yellow. YY1 regulated genes are drawn with thick border. (B) DP with a lower expression in poor prognosis breast cancers ($k = 25$). Drug resistance pathway genes appear in pink. JAK-STAT signalling pathway genes are drawn with thick border.

of gene expression data highlights the advantages of our approach in identifying clinically sound pathways.

While the results presented here are encouraging, there is certainly room for further development of these methods. Currently, we look for multiple subnetworks by iteratively finding and removing the most significant DP from the network. Better methods are needed to detect overlapping DPs. Furthermore, one can obtain significance scores for individual nodes in the DPs using established statistical methods such as bootstrapping [34].

Our problem formulation used a fixed k value, thus requiring that the same least number of genes is altered in all patients (or studies). All the algorithms and proofs presented are generalizable to the scenario where different samples have different thresholds. This case can be attractive if, for example, the number of differentially expressed genes varies significantly among patients or studies, and the goal is to detect subnetworks covering a fixed percentage of the differentially expressed genes. The value of l used in the examples presented here was set to 20% of the elements (cases or studies) in the dataset. While we observed that our method is rather robust to l values in the range of 15-40% of the cases, the methodology for a more rigorous selection of the l value is also an interesting subject for further research.

One of the main goals of case-control studies using microarrays is the detection of biomarkers, leading to an improved characterization of the pathologies of each patient. We believe that the fact that the subnetworks that we identified for HD and breast cancer contain numerous established therapeutic targets carries the promise that an integrative analysis of such studies with complementary molecular datasets can also indicate specific points for medical intervention.

Acknowledgements

We thank David Burstein, Yonit Halperin and Chaim Linhart for helpful discussions. IU is a fellow of the Edmond J. Safra Bioinformatics Program at Tel-Aviv University. This research was supported by the GENEPARK project which is funded by the European Commission within its FP6 Programme (contract number EU-LSHB-CT-2006-037544).

References

1. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.: How to infer gene networks from expression profiles. Molecular Systems Biology **3** (2007) 78
2. van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415** (2002) 530–536
3. Segal, E., Friedman, N., Kaminski, N., Regev, A., Koller, D.: From signatures to models: understanding cancer using microarrays. Nat Genet **37 Suppl** (2005) S38–45
4. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., Mesirov, J.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS **102** (2005) 15545–15550
5. Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., Vert, J.: Classification of microarray data using gene networks. BMC Bioinformatics **8** (2007) 35
6. Ulitsky, I., Shamir, R.: Identification of functional modules using network topology and high-throughput data. BMC Systems Biology **1** (2007)
7. Segal, E., Wang, H., Koller, D.: Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics **19** (2003) I264–I272
8. Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F.: Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics **18** (2002) S233–S240
9. Rajagopalan, D., Agarwal, P.: Inferring pathways from gene lists using a literature-derived network of biological relationships. Bioinformatics **21** (2005) 788–793
10. Cabusora, L., Sutton, E., Fulmer, A., Forst, C.: Differential network expression during drug and stress response. Bioinformatics **21** (2005) 2898–2905
11. Nacu, S., Critchley-Thorne, R., Lee, P., Holmes, S.: Gene expression network analysis and applications to immunology. Bioinformatics **23** (2007) 850
12. Liu, M., Liberzon, A., Kong, S., Lai, W., Park, P., Kohane, I., Kasif, S.: Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genetics **3** (2007) e96+
13. Breitling, R., Amtmann, A., Herzyk, P.: Graph-based iterative group analysis enhances microarray interpretation. BMC Bioinformatics **5** (2004) 100
14. Chuang, H., Lee, E., Liu, Y., Lee, D., Ideker, T.: Network-based classification of breast cancer metastasis. Mol Syst Biol **3** (2007)
15. Hochbaum, D.S.: Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems. In Hochbaum, D.S., ed.: Approximation algorithms for NP-hard problems. PWS, Boston (1997) 94–143
16. Dobson, G.: Worst-case analysis of greedy heuristics for integer programming with non-negative data. Mathematics of Operations Research **7** (1982) 515–531
17. Shuai, T., Hu, X.: Connected set cover problem and its applications. In Cheng, S., Poon, C., eds.: AAIM. Volume 4041 of Lecture Notes in Computer Science., Springer (2006) 243–254

18. Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms. MIT Press, Cambridge, MA (1990)
19. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **298** (2002) 824–827
20. Hodges, A., Strand, A., Aragaki, A., Kuhn, A., Sengstag, T., Hughes, G., Elliston, L., Hartog, C., Goldstein, D., Thu, D., et al.: Regional and cellular gene expression changes in human Huntington's disease brain. Human Molecular Genetics **15** (2006) 965
21. Kaltenbach, L., Romero, E., et al.: Huntingtin interacting proteins are genetic modifiers of neurodegeneration. PLoS Genet **3** (2007) e82
22. Rockabrand, E., Slepko, N., Pantalone, A., Nukala, V., Kazantsev, A., Marsh, J., Sullivan, P., Steffan, J., Sensi, S., Thompson, L.: The first 17 amino acids of Huntingtin modulate its sub-cellular localization, aggregation and effects on calcium homeostasis. Human Molecular Genetics **16** (2007) 61
23. Borrell-Pagès, M., Zala, D., Humbert, S., Saudou, F.: Huntington's disease: from huntingtin function and dysfunction to therapeutic strategies. Cellular and Molecular Life Sciences (CMLS) **63** (2006) 2642–2660
24. Giuliano, P., De Cristofaro, T., et al.: DNA damage induced by polyglutamine-expanded proteins. Human Molecular Genetics **12** (2003) 2301–2309
25. Hoshino, M., Tagawa, K., et al.: Histone deacetylase activity is retained in primary neurons expressing mutant huntingtin protein. J Neurochem **87** (2003) 257–67
26. Butler, R., Bates, G.: Histone deacetylase inhibitors as therapeutics for polyglutamine disorders. Nat Rev Neurosci **7** (2006) 784–96
27. Ferrante, R., Kubilus, J., Lee, J., Ryu, H., Beesen, A., Zucker, B., Smith, K., Kowall, N., Ratan, R., Luthi-Carter, R., et al.: Histone deacetylase inhibition by sodium butyrate chemotherapy ameliorates the neurodegenerative phenotype in Huntington's disease mice. Journal of Neuroscience **23** (2003) 9418–9427
28. Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H.D., Hersch, S.M., Hogarth, P., Bouzou, B., Jensen, R.V., Krainc, D.: Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proc Natl Acad Sci U S A **102** (2005) 11023–8
29. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst **98** (2006) 262–72
30. Affar, E., Gay, F., Shi, Y., Liu, H., Huarte, M., Wu, S., Collins, T., Li, E., Shi, Y.: Essential Dosage-Dependent Functions of the Transcription Factor Yin Yang 1 in Late Embryonic Development and Cell Cycle Progression. Molecular and Cellular Biology **26** (2006) 3565–3581
31. Begon, D., Delacroix, L., Vernimmen, D., Jackers, P., Winkler, R.: Yin Yang 1 Cooperates with Activator Protein 2 to Stimulate ERBB2 Gene Expression in Mammary Cancer Cells. Journal of Biological Chemistry **280** (2005) 24428–24434
32. Li, L., Shaw, P.: Autocrine-mediated activation of STAT3 correlates with cell proliferation in breast carcinoma lines. Journal of Biological Chemistry **277** (2002) 17397–17405
33. Futreal, P., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.: A census of human cancer genes. Nature Reviews Cancer **4** (2004) 177–183
34. Efron, B., Tibshirani, R.: An introduction to the bootstrap. Chapman & Hall New York (1993)

5. Detecting Pathways Transcriptionally Correlated with Clinical Parameters

DETECTING PATHWAYS TRANSCRIPTIONALLY CORRELATED WITH CLINICAL PARAMETERS

Igor Ulitsky and Ron Shamir

School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Email: {ulitskyi,rshamir}@post.tau.ac.il

The recent explosion in the number of clinical studies involving microarray data calls for novel computational methods for their dissection. Human protein interaction networks are rapidly growing and can assist in the extraction of functional modules from microarray data. We describe a novel methodology for extraction of connected network modules with coherent gene expression patterns that are correlated with a specific clinical parameter. Our approach suits both numerical (e.g., age or tumor size) and logical parameters (e.g., gender or mutation status). We demonstrate the method on a large breast cancer dataset, where we identify biologically-relevant modules related to nine clinical parameters including patient age, tumor size, and metastasis-free survival. Our method is capable of detecting disease-relevant pathways that could not be found using other methods. Our results support some previous hypotheses regarding the molecular pathways underlying diversity of breast tumors and suggest novel ones.

1. INTRODUCTION

Systems biology has the potential to improve the diagnosis and management of complex diseases by offering a comprehensive view of the molecular basis behind the clinical pathology. To achieve this, a computational analysis extracting mechanistic understanding from the available data is required. Such data include many thousands of genome-wide expression profiles obtained using the microarray technology. A wide variety of approaches have been suggested for reverse engineering of mechanistic molecular networks from expression data[1-3]. However, most of these methods are effective only when using expression profiles obtained under diverse conditions and perturbations, while the bulk of data currently available on human clinical studies are expression profiles of groups of individuals sampled from the natural population. The standard methodologies for analysis of such datasets usually include: (a) unsupervised clustering of the samples to reveal the basic correlation structure, and (b) focus on a specific clinical parameter and the application of statistical methods for identification of a gene signature that best predicts it. While these methods are successful in identifying potent signatures for classification purposes[4,5], the insights that can be obtained from examining the gene lists they produce are frequently limited.

It is thus desirable to develop novel computational tools that will utilize additional information in order to extract more knowledge from gene expression studies. Various parameters are commonly recorded in such studies, and they can be classified into two types: (a) logical parameters (e.g., gender or tumor subtype) and (b) numerical parameters (e.g., patient age or tumor grade). A key question is how to identify genes significantly related to a specific clinical parameter. As it is frequently difficult to suggest novel hypotheses based on individual genes, it is desirable to identify the *pathways* that are correlated with a clinical parameter. By considering together the whole pathway, correlations that would have been missed if we tested each gene separately can be revealed. One approach to this problem uses predefined gene sets describing pathways and quantifies the change in their expression levels[6-8]. The drawback of this approach is that pathway boundaries are often difficult to assign, and in many cases only part of the pathway is altered during disease. Moreover, unknown pathways are harder to find in this approach. To overcome these problems, the use of gene networks was suggested. Several approaches for integrating microarray measurements with network knowledge have been

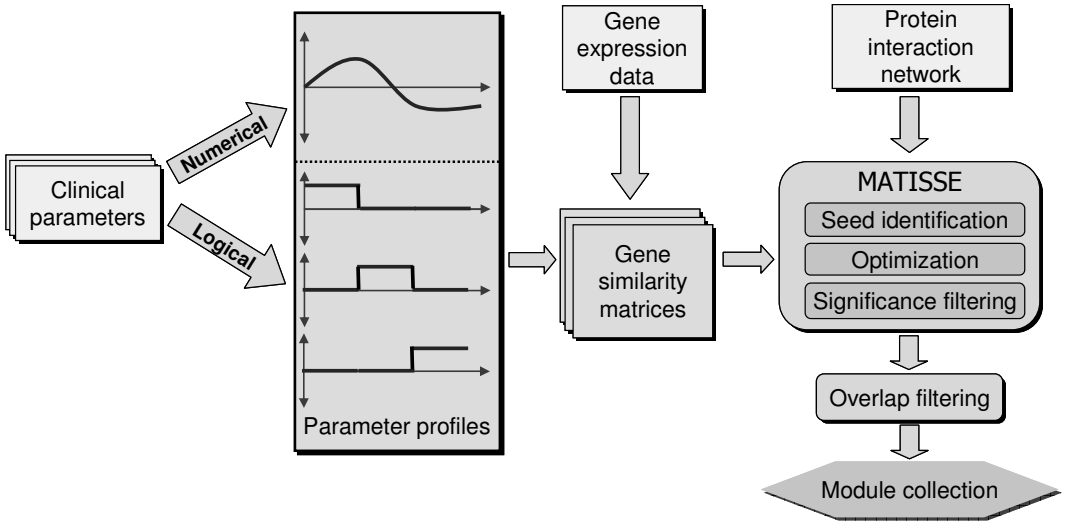


Fig. 1. Study outline. Clinical parameters are used to generate a collection of parameter profiles. The parameter profiles are used, together with gene expression data, to generate gene similarity scores. These scores, together with a protein interaction network serve as an input to MATISSE, which identifies a set of modules for each parameter. The modules are then filtered and a collection of non-redundant modules is produced.

proposed, some of which can be applied also for binary clinical parameters. Some proposed computational methods for detection of subnetworks that show correlated expression[9-11]. A successful method for detection of `active subnetworks' was proposed by Ideker *et al.*[12] and extended by other groups[13-16]. These methods are based on assigning a significance score to every gene in every sample and looking for subnetworks with statistically significant combined scores. Breitling *et al.*[17] proposed a simple method named GiGA which receives a list of genes ordered by expression relevance and extracts subnetworks corresponding to the most relevant genes. Other tools use network and expression information together, but for sample classification[18,19].

The most basic parameter in clinical studies is the binary disease status (case vs. control). Other studies provide more clinical information in the form of additional parameters. For example, in the breast cancer expression data published by Minn *et al.*[20], each sample was accompanied by up to 10 different parameters (Table 1). These parameters include general characteristics of the patients (e.g., age), pathological status of the tumor and follow-up information. Given such data, we wish to identify pathways whose transcription is dysregulated in a manner that is consistent with a particular clinical parameter. This information can then be used both for predictive purposes and for improving our understanding of the biology underlying the disease progression. This requires identifying subnetworks with expression patterns correlated to numerical or multi-valued logical parameters with more than two possible values.

We have previously developed the MATISSE algorithm for extraction of functional modules from expression and network data[9]. It receives as input a protein interaction (PI) network alongside a collection genome-wide mRNA expression profiles. The output of MATISSE is a collection of modules: connected subnetworks in the PI graph, whose corresponding mRNAs exhibit significantly correlated expression patterns. Here we describe an extension of the MATISSE algorithm aimed at extraction of modules of genes whose expression profiles are not only correlated to one another, but also correlated with one of the clinical parameters. These two requirements aim to identify subnetworks that constitute functional modules in the cell and are involved with a specific clinical phenotype.

We used a human PI network consisting of 10,033 nodes and 41,633 interactions (see Methods) and applied our algorithm to 99 breast cancer samples (BC dataset[20]) in conjunction with 10 numerical and logical parameters (Figure 1). This analysis identified several modules significantly correlated with various parameters such as patient age, tumor size, Her2 status and metastases-free survival period length.

Table 1. Parameters from the breast cancer dataset that were used in this study.

Parameter	Samples*	Type	Distribution
Age at diagnosis	99	Numerical	55.80±13.6
Tumor Size (cm)	99	Numerical	3.62±1.7
Positive Lymph Nodes	99	Numerical	3.59±6.3
Estrogen receptor (ER) status	99	Logical	
Progesterone receptor (PR) status	98	Logical	
Her2 staining (grade)	88	Numerical	0.53±0.98
Metastasis after 5 years?	68	Logical	
Metastasis free survival (years)	82	Numerical	5.17±2.3
Lung metastasis free survival (years)	82	Numerical	5.50±2.3
Bone metastasis free survival (years)	82	Numerical	5.34±2.3

* Number of samples for which the parameter was available

Importantly, our results provide support for the correlation between the expression levels of several pathways, such as the ribosomal proteins and the patient prognosis. However, this is not always the case, as we did not find support for the correlation between survival and the levels of the unfolded protein response pathway genes. Finally, we show that the specific disease-related insights suggested by our method can not be picked up using existing alternative methods.

2. METHODS

2.1. The basic methodology

Our approach builds on the MATISSE methodology for identifying co-expressed subnetworks[9]. We first outline that methodology here. The input to MATISSE includes an undirected *constraint graph* G^C = (V, E), a subset $V_{sim} \subseteq V$ and a symmetric matrix S where S_{ij} is the similarity between $v_i, v_j \in V_{sim}$. The goal is to find disjoint subsets $U_1, U_2, ..., U_k \subseteq V$ called *modules*, so that each subset induces a connected subgraph in G^C and contains elements that share high similarity values. We call the nodes in V_{sim} *front nodes* and nodes in $V \backslash V_{sim}$ *back nodes*.

In the biological context, V represents genes or gene products (we shall use the term 'gene' for brevity), and E represents interactions between them. S_{ij} measures the similarity between genes i and j. Originally, we used the Pearson correlation between gene expression patterns as a similarity metric[9]. The set V_{sim} is smaller than V in several cases. For example, when using mRNA microarrays, some of the genes may be absent from the array, and others may be excluded due to insignificant expression changes across the tested conditions. Hence, a module aims to capture a set of genes that have highly similar behavior, and are also topologically connected, and thus may belong to a single complex or pathway. The quantification of gene similarity is obtained by formulating the problem as a hypothesis testing question. In this approach statistically significant modules correspond to heavy subnetworks in a similarity graph, with nodes inducing a connected subgraph in G^C. A three-stage heuristic is used to obtain high-scoring modules.

2.2. Identifying modules correlated with clinical parameters

Here, we are interested in extracting groups of genes that are not only similar across the experimental conditions, but also exhibit significant correlation with one of the clinical parameters. To this end we devised a hybrid similarity score that reflects these two phenomena. Importantly, our scheme can handle both numerical and logical parameters. The advantage of a uniform scheme is that the modules identified for different parameters are directly comparable, and in case of overlaps, the more significant module can be picked.

Formally, we are given a set of parameters $P_1, \ldots, P_m$ (numerical and logical) and we wish to quantify, for each gene pair (i,j), the extent to which the genes are correlated to P_k and to each other. For each parameter we first discard the samples for which the value of the parameter is not available. Let m be the number of samples that survived this filter. Then, we compute one or more *parameter profiles* $p_{ij} = (p_{ij}^1, p_{ij}^2, ..., p_{ij}^m)$. If P_i is a numeric parameter, it is assigned a single parameter profile vector p_{i1},

and p_{i1}^k equals the value of P_i in sample k. If P_i is a logical parameter that attains with k different values $c_i^1, c_i^2, ..., c_i^l$, then for each $1 \le j \le l$ we compute a 0/1 parameter profile vector $p_{ij} = (p_{ij}^1, p_{ij}^2, ..., p_{ij}^m)$ where $p_{ij}^k = 1$ if the value of P_i in sample k is c_j and 0 otherwise.

We denote the expression pattern of gene k by $x_k = (x_k^1, x_k^2, ..., x_k^m)$. We are interested in quantifying the similarity between p_{ij} and x_k. Let r_{ijk} be the Pearson correlation coefficient between p_{ij} and x_k. If P is numerical, then r_{j1k} is close to 1 if the transcript and the parameter are strongly correlated. If P is logical, r_{ijk} is close to 1 if the transcript levels are high when the value of P_i is c_j and low otherwise. Transcript correlation to such 0/1 profiles was previously used successfully as a differential gene expression score[21].

Recall that we are interested in gene pairs a,b that are: (i) correlated with p_{ij} and (ii) correlated with each other. To address (i) we would like the similarity score of genes a and b to be high only if both a and b are correlated with the parameter. We thus first set $S_{diff}(i, j) = \min\{r_{ija}, r_{ijb}\}$. To address (ii) we use the partial correlation coefficient between the gene patterns conditioned on p_{ij}. Formally:

$$S_{corr}(a, b \mid p_{ij}) = \frac{r_{a,b} - r_{ija} r_{ijb}}{\sqrt{(1 - r_{ija}^2)(1 - r_{ijb}^2)}}$$

where $r_{a,b}$ is the Pearson correlation coefficient between the profiles of genes a and b. Intuitively, S_{corr} conveys the information about how similar a and b are, given their correlation to p_{ij}. Finally, we use the similarity score:

$$S = \frac{S_{diff} + \lambda \cdot S_{corr}}{1 + \lambda}$$

where λ is a tradeoff parameter setting the relative importance of the correlation with the clinical parameter. For each parameter profile S scores were computed for both positive and negative correlations with the parameter. Note that the values of S are always between -1 and 1. Note that standard Pearson correlation can also be used as S_{corr}. We chose to use partial correlation in this work, as it allows us to penalize gene pairs for which most of the correlation can be explained by their separate correlations with the clinical parameter. The S scores are then modeled using the probabilistic model described previously[9].

2.3. Finding high-scoring modules

MATISSE uses a three-step heuristic to identify high-scoring modules. The heuristic consists of (a) identification of small high-scoring seeds; (b) seed optimization using a greedy algorithm; (c) significance filtering. The seed finding step was performed as described previously[9]. The greedy algorithm was improved in this study. To allow improvement of modules that reached the maximum size limit, we added two additional operation types: (a) a "replace" operation in which a node is added to a module replacing the node that contributes least to the module score; (b) a "swap" operation, in which module assignments of two nodes are swapped. Both operations are performed only if they improve the total solution weight jeopardizing the connectivity of the modules.

In order to evaluate the statistical significance of the modules found in a dataset, we randomly shuffled the expression pattern of each gene and re-executed the algorithm. This process was repeated 100 times and the best score of a module in each run was recorded. These scores were then used to compute an empirical p-value for modules found in the real data. Only modules with p<0.1 were retained.

2.4. Filtering overlapping modules

We removed modules that overlapped by >50% with another module that was more significantly correlated with a clinical parameter.

2.5. MATISSE parameters

We used λ=4 for the analysis described in this paper. The upper bound on module size was set to 120. The rest of the parameters were set as described previously[9].

2.6. Network and expression data

A human PI network was compiled from the HPRD[22], BIND[23], BioGrid[24], HDBase (http://hdbase.org/), and SPIKE[25] databases. The resulting network consisted of 10,033 proteins (mapped to Entrez Gene entries) and 41,633 interactions.

The expression dataset was obtained from GEO (Accession GSE2603). We used the normalized expression values available in the respective GEO records. Affymetrix probe identifiers were mapped to

Entrez Gene. If several probes mapped to the same Entrez Gene, the highest intensity was used in every sample. Values <20 were set to 20 and values >20,000 were set to 20,000. 2,000 genes that showed the highest gene pattern variance were used as front nodes.

2.7. Module annotation

We annotated the modules using Gene Ontology (http://www.geneontology.org/) and MSigDB (http://www.broad.mit.edu/gsea/, "curated gene sets" collection[6]). Gene Ontology enrichment p-values were computed using TANGO[26], which uses resampling to correct for multiple testing and annotation overlap. All other p-values were Bonferroni corrected for multiple testing.

3. RESULTS

3.1. Breast cancer dataset

The breast cancer (BC) dataset contained 99 expression profiles of tumor samples from the MSKCC cohort[20]. 15 different parameters were available for each sample, some of which were not sufficiently clear or redundant. The 10 parameters we used are listed in Table 1. For 9 parameters at least one significant module was identified. After filtering module overlaps (see Methods) we identified 10 significant non-redundant modules, with sizes ranging from 84 to 118 (Table 2).

Using GO and MSigDB annotations (see Methods) we found that 6 modules (60%) were significantly enriched with at least one GO biological process and all 10 modules (100%) were enriched with at least one MSigDB category (Table 2). Seven modules (70%) were enriched with at least one of the 16 MSigDB gene sets related to breast cancer. Overall, eight of the breast cancer related gene sets were enriched in at least one module.

Module BC-1 was positively correlated with the age of the woman at the time of breast cancer diagnosis. Inspection of the expression data revealed that the module was particularly up-regulated in women above age 72 (Figure 2). The module did not show significant GO enrichment categories. When examining 27 MSigDB gene sets related to aging, we found a significant between BC-1 and the MSigDB category "AGED_RHESUS_UP" (8 genes, p=0.002), which contains genes identified as up-regulated in the muscles of aged rhesus monkeys when compared to young ones[27]. One of these eight genes is RELA, a transcription factor component of the NFκB complex. BC-1 contained two additional genes from the PKC pathway which activates NFκB – NFKBIA and PKCA (MSigDB gene set PKCPATHWAY, p=0.04). Increased activity of the NFκB pathway has been recently implicated in aging in a study utilizing diverse expression data and transcription factor binding motifs[28]. Adler *et al.* have also shown that blocking of this pathway can reverse the age-related transcriptional program. Note that our methodology connecting NFκB to aging is completely different: Adler *et al.* sought motifs over-represented in age-dependent genes in various microarray datasets, whereas we looked for connected PI subnetworks that are correlated with age on the expression level. Our results thus provide further support for the relationship between NFκB and age-dependent transcriptional changes.

BC-2 is an apoptosis-related module that is positively correlated to the size of the tumor. This module is also significantly enriched with genes related to unfolded protein response (UPR) and the TNF pathway. Accordingly, this module also significantly enriched with heat shock factor (HSF) targets (p=0.03) and genes localized to the ER (from GO, $p=6.81*10^{-9}$). Interestingly, heat shock protein level has been traditionally associated with poor breast cancer prognosis and higher metastasis likelihood[29]. However, BC-2 was only weakly correlated with metastases-free survival period in our dataset (r=0.038).

Two modules, BC-3 and BC-4, were identified as negatively correlated with tumor size. Both modules were enriched with genes previously associated with ER-positive tumors. However, the correlation of the module profiles with ER status was very weak in our dataset (r=0.001 and r=0.008, for BC-3 and BC-4, respectively). However, we did find a significant overlap between genes in BC-3 and the recently mapped targets of the estrogen receptor[30] ($p=1.13*10^{-4}$). Finally, estrogen receptors Esr1 and Esr2 both appeared in BC-3. This suggests that increased ER transcription factor activity could result in smaller tumors.

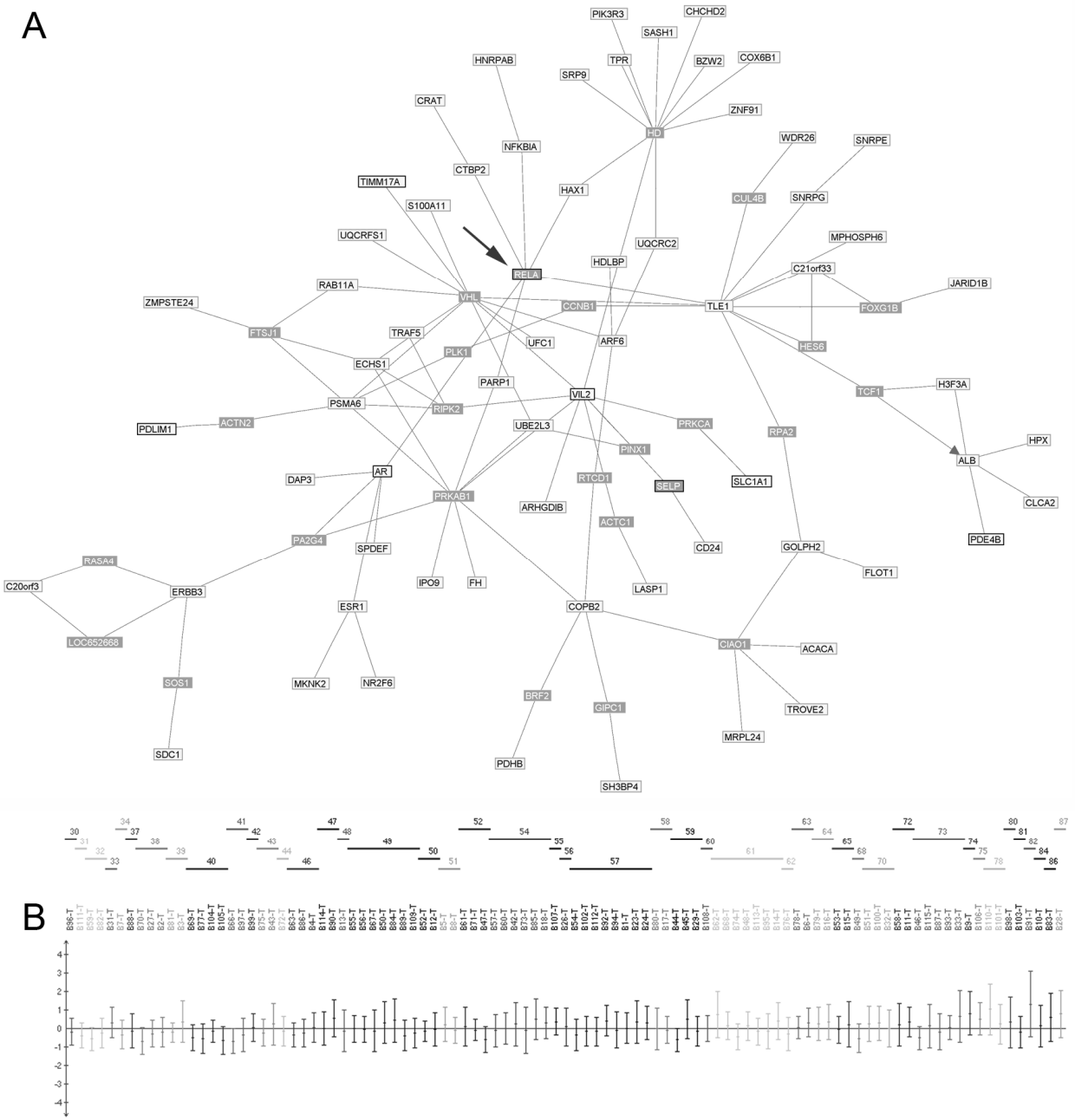


Fig. 2. BC-1 module related to age at diagnosis. (A) The subnetwork view of the module. Front nodes have a brighter background color. Gene overlapping the MSigDB RHESUS_AGING_UP category have thicker border. The arrow points at the RELA transcription factor. (B) Average expression levels of BC-1. Numbers on top indicate the age of diagnosis. The error bars represent ± one standard deviation.

Table 2. Modules identified in the breast cancer dataset of Minn *et al*. Front nodes are nodes for which expression data are used (see Methods). GO enrichment p-values were computed using TANGO. MSigDB enrichment p-values are Bonferroni corrected. For MSigDB, up to 5 most significantly enriched gene sets are shown.

Module	Parameter	Average correlation	Total nodes	Front nodes	Score FDR	GO biological process	p-value	MSigDB gene set	p-value
BC-1	Age at diagnosis	0.196	90	64	0.08			HUMAN_MITODB_6_2002	0.016
								MITOCHONDRIA	0.022
								BRCA_ER_POS	0.026
								PKCPATHWAY	0.04
BC-2	Tumor Size	0.188	118	82	<0.01	response to unfolded protein	<0.001	ST_TUMOR_NECROSIS_FACTOR_PATHWAY	9.36E-10
								BRCA_ER_NEG	8.76E-08
								STEMCELL_NEURAL_UP	9.11E-08
								APOPTOSIS	3.79E-07
								APOPTOSIS_GENMAPP	1.68E-06
BC-3	Tumor Size	-0.175	115	86	<0.01			BRCA_ER_POS	2.13E-09
								ALZHEIMERS_DISEASE_DN	1.92E-05
								BREASTCA_TWO_CLASSES	3.05E-04
								DRUG_RESISTANCE_AND_METABOLISM	9.96E-04
								CARM_ERPATHWAY	0.034
BC-4	Tumor Size	-0.157	97	60	0.09			BRCA_ER_POS	0.002
								AKAPCENTROSOMEPATHWAY	0.009
								P53PATHWAY	0.023
BC-5	Positive lymph nodes	-0.143	84	66	<0.01			BRCA_ER_NEG	1.32E-09
								STEMCELL_NEURAL_UP	1.41E-05
								TARTE_PLASMA_BLASTIC	7.84E-05
								PENG_GLUTAMINE_DN	8.87E-04
								ALZHEIMERS_DISEASE_DN	0.004
BC-6	Her2 staining	0.204	107	80	0.01	positive regulation of I-kappaB kinase/NF-kappaB cascade	0.009	ALZHEIMERS_DISEASE_DN	2.74E-08
								HUMAN_MITODB_6_2002	9.84E-05
								FLECHNER_KIDNEY_TRANSPLANT_REJECTION_DN	2.83E-04
								PGC	3.67E-04
								MITOCHONDRIA	9.48E-04
BC-7	Metastasis after 5 years?	-0.203	96	74	0.04	translation	0.004	RIBOSOMAL_PROTEINS	9.23E-33
								JISON_SICKLECELL_DIFF	3.86E-08
								FLOTHO_CASP8AP2_MRD_DIFF	3.32E-07
								HCC_SURVIVAL_GOOD_VS_POOR_DN	3.43E-04
								TRANSLATION_FACTORS	0.009
BC-8	Metastasis after 5 years?	0.224	116	86	0.02	antigen processing	<0.001	WIELAND_HEPATITIS_B_INDUCED	1.09E-11
						antigen presentation	<0.001	PROTEASOME	9.97E-11
						modification-dependent protein catabolism	<0.001	FLECHNER_KIDNEY_TRANSPLANT_WELL_UP	5.12E-08
								PROTEASOMEPATHWAY	7.40E-08
								TCRAPATHWAY	3.04E-06
BC-9	Mestassis free survival	0.191	118	91	<0.01	translation	0.02	RIBOSOMAL_PROTEINS	1.40E-33
								JISON_SICKLECELL_DIFF	4.30E-11
								FLOTHO_CASP8AP2_MRD_DIFF	2.22E-10
								MYC_TARGETS	6.95E-04
								HCC_SURVIVAL_GOOD_VS_POOR_DN	0.003
BC-10	Lung metastatis free survival	0.195	102	74	0.01	positive regulation of I-kappaB kinase/NF-kappaB cascade	<0.001	RIBOSOMAL_PROTEINS	7.08E-11
								NFKBPATHWAY	3.23E-06
								JISON_SICKLECELL_DIFF	7.28E-06
								ST_TUMOR_NECROSIS_FACTOR_PATHWAY	1.96E-05
								APOPTOSIS_GENMAPP	3.04E-04

Three modules (BC-7, BC-9 and BC-10) were significantly enriched with ribosomal proteins (RPs). Expression levels of these modules were correlated with Her2- and ER-positive longer metastases-free survival in the lungs and in the bone marrow. High expression of RPs is indicative of a higher metabolic rate within malignant cells. High levels of RP expression have been previously associated with Her2 overexpression in BC cell lines[31]. RP over-expression was also associated with less aggressive ovarian tumors[32]. Our results provide additional support for the notion that RP expression is positively correlated with longer survival. Surprisingly, two of the modules enriched for ribosomal proteins (BC-7 and BC-9) were enriched with the MSigDB class "HCC_SURVIVAL_GOOD_VS_POOR_DN" described as representing genes associated with poor survival in hepatocellular carcinoma. However, this class is not associated with any publication and BC-7 and BC-9 were not enriched with other gene sets related to survival in MSigDB, so further corroboration is required here.

BC-8 was significantly enriched with proteasomal genes and associated with shorter metastases-free survival periods. This module contained 16 different proteasomal subunits, all as front nodes. It also contained multiple genes associated with antigen representation and the immune response. Interestingly, this module was also significantly enriched with genes located on chromosome 6 (p=1.29*10^{-6}, the most significant module-chromosome association). Therefore, it is possible that the up-regulation results from aberrations of this chromosome in a subset of the tumors.

3.2. Comparison with other methods

We first compared the parameter-correlated modules (PCMs) to the modules obtained using the standard MATISSE algorithm with the same parameters. MATISSE identified 19 modules covering 996 genes. 8 of the modules (42%) were significantly enriched for a GO category and 11 (58%) were enriched for an MSigDB category (all 11 were enriched with at least one breast-cancer related category), indicating that a larger percentage of PCMs are functionally relevant compared to MATISSE modules. However, 18 GO annotations were enriched in the MATISSE solution only, compared to 5 in the parameter-correlated solution only (195 vs. 47 for MSigDB gene sets), indicating a trade-off between specificity and selectivity in functional module selection. As expected, the MATISSE module genes were more strongly correlated on the expression level (average r=0.3 vs. 0.14), whereas PCMs were more strongly correlated with clinical parameters (average maximum correlation of 0.14 per PCM, compared to 0.12 for MATISSE modules).

Some of the insights described above could not be revealed using MATISSE: only two small modules (9 genes each) were slightly correlated with age and they did not overlap the rhesus aging signature; (b) the MATISSE modules that were slightly correlated with tumor size were not enriched for the UPR pathway; (c) no MATISSE modules were enriched for ribosomal or other translation-related proteins; (d) the maximum enrichment for same-chromosome genes was significantly lower (p=0.002 vs. p=1.29*10^{-6}). Thus we conclude that while using expression correlation alone can lead to more diverse functional modules, using clinical parameter correlation enables detection of more specific disease-relevant modules that are missed otherwise.

The insights also could not be based on parameter correlation alone. When taking the 200 genes with the highest enrichment with the parameters: (a) the genes correlated with age at prognosis were not enriched with the rhesus gene set and did not contain RELA; (b) the genes correlated with tumor size were not enriched with UPR pathway genes; (c) the genes negatively correlated with tumor size were not enriched with ER targets; (c) the genes correlated with metastases-free survival were not enriched with ribosomal proteins.

Finally, logical parameters can be analyzed using GSEA[6]. GSEA found 130 (9) gene sets associated with poor (good) prognosis at FDR<0.1. 31 (3) were associated with negative (positive) ER status, none of them breast cancer related. No gene sets were significantly associated with PR status. Similar to our analysis, GSEA identified the correlation between survival and the levels of the ribosomal proteins and the proteasome. However, only one breast cancer related gene set appeared in the GSEA results (BRCA_ER_POS), and none of the pathways we identified using continuous parameters could be found using GSEA.

4. DISCUSSION

The increasing availability of network and expression data in multiple species led to development of several methods for detecting modular structures through joint analysis of network and expression data[9,11-17]. As the coverage and quality of the interaction networks improve, we expect that these tools will play a central part in the analysis of microarray data. A prominent current challenge is to enable these tools to use as much additional information as possible in order to produce more accurate and biologically relevant results. Clinical parameters of the profiled tissue can help in association of genes and pathways with clinical phenotypes.

To the best of our knowledge, the method we described here is the first capable of jointly analyzing interaction data, expression profiles and continuous numerical clinical parameters. A simple alternative for joint analysis of the three sources is to first apply a module finding algorithm to network and expression data, and then associate modules with parameters. As our results show, module finding algorithms are indeed successful at identifying multiple functional modules. However, clinically important pathways are missed if the clinical data are used only in the post-processing of the modules.

While the results we present are encouraging, there is certainly room for improvement. In particular, it would help to incorporate confidence levels for individual interactions[33] and to further improve our optimization algorithm. Our methodology for integrating parameter data currently analyzes each parameter in isolation, ignoring correlations between parameters. Another important frontier is to associate modules with combinations of different parameter values, e.g., up-regulation in poor prognosis and in ER-negative tumors.

Finally, we are currently developing a user-friendly interface to the methods described here that will allow analysis through the MATISSE software (http://acgt.cs.tau.ac.il/matisse).

Acknowledgements

IU is a fellow of the Edmond J Safra Bioinformatics Program at Tel-Aviv University. This research was supported in part by the "GENEPARK: GENomic Biomarkers for PARKinson's disease" project that is funded by the European Commission within its FP6 Programme (contract number EU-LSHB-CT-2006-037544), and by the Israel Science Foundation (grant no. 385/06).

References

1. Gat-Viks, I. & Shamir, R. Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res* **17**, 358-67 (2007).
2. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. How to infer gene networks from expression profiles. *Mol Syst Biol* **3**, 78 (2007).
3. Sprinzak, D. & Elowitz, M.B. Reconstruction of genetic circuits. *Nature* **438**, 443-8 (2005).
4. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).
5. Ben-Dor, A. et al. Tissue classification with gene expression profiles. *J Comput Biol* **7**, 559-83 (2000).
6. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
7. Kim, S.Y. & Volsky, D.J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**, 144 (2005).
8. Jiang, Z. & Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* **23**, 306-13 (2007).
9. Ulitsky, I. & Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol* **1**, 8 (2007).
10. Segal, E., Wang, H. & Koller, D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19 Suppl 1**, i264-71 (2003).
11. Hanisch, D., Zien, A., Zimmer, R. & Lengauer, T. Co-clustering of biological networks and gene expression data. *Bioinformatics* **18 Suppl 1**, S145-54 (2002).
12. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1**, S233-40 (2002).
13. Rajagopalan, D. & Agarwal, P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* **21**, 788-93 (2005).

14. Cabusora, L., Sutton, E., Fulmer, A. & Forst, C.V. Differential network expression during drug and stress response. *Bioinformatics* **21**, 2898-905 (2005).
15. Nacu, S., Critchley-Thorne, R., Lee, P. & Holmes, S. Gene expression network analysis and applications to immunology. *Bioinformatics* **23**, 850-8 (2007).
16. Liu, M. et al. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet* **3**, e96 (2007).
17. Breitling, R., Amtmann, A. & Herzyk, P. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics* **5**, 100 (2004).
18. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3**, 140 (2007).
19. Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E. & Vert, J.P. Classification of microarray data using gene networks. *BMC Bioinformatics* **8**, 35 (2007).
20. Minn, A.J. et al. Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518-24 (2005).
21. Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. & Altman, R.B. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18**, 1454-61 (2002).
22. Peri, S. et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, D497-501 (2004).
23. Bader, G.D. et al. BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res* **29**, 242-5 (2001).
24. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).
25. Elkon, R. et al. SPIKE - a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics* **9**, 110 (2008).
26. Shamir, R. et al. EXPANDER--an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**, 232 (2005).
27. Kayo, T., Allison, D.B., Weindruch, R. & Prolla, T.A. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc Natl Acad Sci U S A* **98**, 5093-8 (2001).
28. Adler, A.S. et al. Motif module map reveals enforcement of aging by continual NF-kappaB activity. *Genes Dev* **21**, 3244-57 (2007).
29. Calderwood, S.K., Khaleque, M.A., Sawyer, D.B. & Ciocca, D.R. Heat shock proteins in cancer: chaperones of tumorigenesis. *Trends Biochem Sci* **31**, 164-72 (2006).
30. Kwon, Y.S. et al. Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters. *Proc Natl Acad Sci U S A* **104**, 4852-7 (2007).
31. Oh, J.J., Grosshans, D.R., Wong, S.G. & Slamon, D.J. Identification of differentially expressed genes associated with HER-2/neu overexpression in human breast cancer cells. *Nucleic Acids Res* **27**, 4008-17 (1999).
32. Welsh, J.B. et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci U S A* **98**, 1176-81 (2001).
33. Suthram, S., Shlomi, T., Ruppin, E., Sharan, R. & Ideker, T. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7**, 360 (2006).

6. Identifying Functional Modules Using Expression Profiles and Confidence-Scored Protein Interactions

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 9 2009, pages 1158–1164
doi:10.1093/bioinformatics/btp118

Systems biology

Identifying functional modules using expression profiles and confidence-scored protein interactions

Igor Ulitsky and Ron Shamir*

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Received on September 18, 2008; revised on January 21, 2009; accepted on February 26, 2009

Advance Access publication March 17, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Microarray-based gene expression studies have great potential but are frequently difficult to interpret due to their overwhelming dimensions. Recent studies have shown that the analysis of expression data can be improved by its integration with protein interaction networks, but the performance of these analyses has been hampered by the uneven quality of the interaction data.

Results: We present Co-Expression Zone ANalysis using NEtworks (CEZANNE), a novel confidence-based method for extraction of functionally coherent co-expressed gene sets. CEZANNE uses probabilities for individual interactions, which can be computed by any available method. We propose a probabilistic model and a weighting scheme in which the likelihood of the connectivity of a subnetwork is related to the weight of its minimum cut. Applying CEZANNE to an expression dataset of DNA damage response in *Saccharomyces cerevisiae*, we recover both known and novel modules and predict novel protein functions. We show that CEZANNE outperforms previous methods for analysis of expression and interaction data.

Availability: CEZANNE is available as part of the MATISSE software at http://acgt.cs.tau.ac.il/matisse.

Contact: rshamir@tau.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The use of microarrays for gene expression profiling has recently become widespread in biomedical research. While microarray gene expression profiles can provide answers to many biological questions and suggest novel hypotheses, they are frequently difficult to interpret due to the large volumes of data and the noise inherent in the biological and experimental systems. Integration of microarray data with additional data sources can help overcome these problems.

Protein–protein interaction (PPI) networks were shown to be very useful in interpreting gene expression data by improving sample classification using microarray data (Chuang *et al*., 2007; Rapaport *et al*., 2007) and improving detection of differentially expressed genes (Li and Li, 2008; Ma *et al*., 2007; Wei and Pan, 2008). Here, we focus on using network information to enhance detection of modules of co-expressed genes. Ideker and colleagues pioneered this approach, proposing a method for detecting subnetworks active in a subset of the profiled samples (Ideker *et al*., 2002), an approach that was extended and improved by several groups (Cabusora *et al*., 2005; Guo *et al*., 2007; Liu, *et al*., 2007; Nacu *et al*., 2007; Rajagopalan and Agarwal, 2005). We and others proposed methods for identifying subnetworks co-expressed across all the sampled conditions (Hanisch *et al*., 2002; Segal *et al*., 2003; Ulitsky and Shamir, 2007). Our method, called MATISSE, has several important advantages: (i) it does not require the number of modules to be specified in advance; (ii) modules can incorporate genes that are not affected on the transcription level; (iii) it can handle not only expression profiles but also any type of data that can be represented as a similarity matrix. A slightly modified version of MATISSE was recently employed to identify a key subnetwork up-regulated in human pluripotent stem cells (Muller *et al*., 2008).

One of the obstacles to exploiting PPI networks is their high rate of false positive and false negative interactions (Suthram *et al*., 2006; von Mering *et al*., 2002). To better handle uncertainty in PPIs, several works devised probabilistic schemes to estimate the confidence of individual interactions (Collins *et al*., 2007; Li, *et al*., 2008; Rhodes *et al*., 2005; Suthram *et al*., 2006; von Mering, *et al*., 2007). To the best of our knowledge, none of the existing methods for identifying functional modules using network and expression data make use of these confidence scores. Here, we develop and employ CEZANNE (Co-Expression Zone ANalysis using NEtworks), a novel methodology for extracting subnetworks with correlated expression profiles (*co-expression modules*) that uses a confidence-based interaction network. CEZANNE builds upon MATISSE and extends it with a novel probabilistic model for subnetwork connectivity. We show that, with an appropriate edge weighting scheme, identifying modules connected with high confidence is equivalent to identifying subgraphs in which the weight of the minimum cut exceeds a threshold. We then show how to identify such modules efficiently. Our probabilistic model and methodology are general and can be employed with other methods that use network connectivity.

In order to evaluate its performance, we applied CEZANNE to a dataset of gene expression of *Saccharomyces cerevisiae* following treatment with various DNA damaging agents (Gasch *et al*., 2001). Our analysis identified well characterized co-expressed protein complexes, such as the ribosomes, as well as novel splicing and actin-related modules. In several cases, we were able to predict novel protein functions based on module assignment. A comparison with other methods showed that the use of confidence levels can

*To whom correspondence should be addressed.

significantly improve the integration of network and expression data for extraction of functional modules.

2 METHODS

2.1 The basic methodology

Our approach builds on the MATISSE methodology for identifying co-expressed subnetworks (Ulitsky and Shamir, 2007). We outline that methodology and describe the improvements in CEZANNE. A pseudocode of the algorithm appears in the Supplementary Material. The input to MATISSE includes an undirected *constraint graph* $G^C=(V, E)$, a subset $V_{sim} \subseteq V$ and a symmetric matrix S where S_{ij} is the similarity between v_i and v_j, where $v_i, v_j \in V_{sim}$. The goal is to find disjoint subsets $U_1, U_2, \ldots, U_m$, called *modules*, with each subset inducing a connected subgraph in G^C and containing elements that share high similarity values. We call the nodes in V_{sim} *front nodes* and the nodes in $V \backslash V_{sim}$ *back nodes*.

In the biological context, V represents genes or gene products (we use the term 'gene' for brevity), and E represents interactions between them. S_{ij} measures the similarity between genes i and j, e.g. the Pearson correlation between their gene expression patterns. The set V_{sim} may be smaller than V. For example, when using mRNA microarrays, some of the genes may be absent from the array, and others may show insignificant expression patterns across the tested conditions and therefore be excluded. Since a module is a set of genes that have highly similar behavior and also induce a connected component in the constraint graph, it should capture genes that belong to a single complex or pathway and therefore share a common function. The quantification of module similarity is obtained in MATISSE by formulating the problem as a hypothesis-testing question. This formulation leads to a full weighted similarity graph whose vertices correspond to V_{sim}. Statistically significant modules correspond to heavy subnetworks in this graph (i.e. subnetworks having high *co-expression score*), with nodes inducing a connected subgraph in G^C. This score is described in the Supplementary Material. A three-stage heuristic was developed in Ulitsky and Shamir (2007) to obtain high-scoring modules. Here, we use the same co-expression score, but replace the connectivity condition by the requirement that modules must be connected with high confidence. We will next describe a novel methodology for identifying such modules.

2.2 The probabilistic model for module connectivity

The following is a description of our model for using interaction confidence. In addition to the constraint graph $G^C=(V,E)$, we are given, for every edge $e \in E$, the probability that the edge exists $p(e) \in (0,1)$. Edge occurrences are assumed to be mutually independent. We can assume that G^C is a complete graph; otherwise, it can be completed by adding all the missing edges with zero probability. The key difference in our model here is that since edge occurrences are probabilistic, connectivity must also be accounted for in a probabilistic sense. We call a set of vertices $U \subseteq V$ *q-connected* if, for all $U' \subset U$, the probability that at least one edge connects U' with $U \backslash U'$ is at least q (Fig. 1). We now show the relationship between this characteristic and the weight of the minimum cut in the subgraph induced by the set. A *cut* in a graph is a partition of its nodes into two disjoint sets. A *minimum cut* in a graph is a cut for which the total weight of the edges between the two sets is minimal (see Supplementary Material for a formal definition). Let $G(U)$ be the subgraph induced by U in G. Let $E(U,W)$ denote the event that at least one edge connects a node from W with a node from $U \backslash W$. Then U is q-connected if and only if $P(E(U, W)) < 1-q$ for every $W \subset U$. Assuming edge appearances are independent, we get

$$P(\overline{E(U,W)}) = \prod_{e \in (W, U \backslash W)} \left(1 - p(e)\right).$$

Note that if we set $w(e) = -\log(1-p(e))$, then

$$P(\overline{E(U,W,q)}) < 1-q \Leftrightarrow \sum_{e \in (W, U \backslash W)} w(e) \geq -\log(1-q)$$

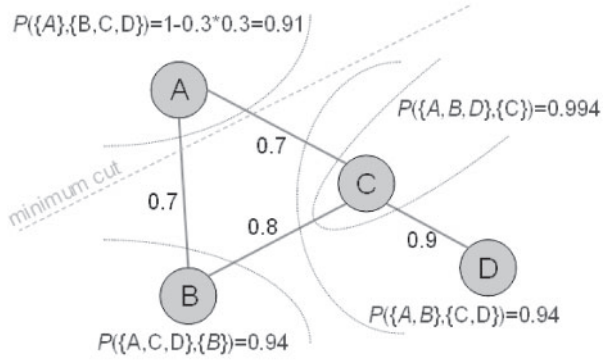


Fig. 1. A q-connected module for $q=0.9$. The numbers of the edges indicate edge probabilities. The probability of missing edges is 0. For every possible partition of the nodes into two sets, the probability that at least one true interaction connects the two sets exceeds 0.9. Four such partitions are shown.

When setting $w(e) = -\log(1-p(e))$, U is q-connected if the weight of every cut exceeds $T = -\log(1-q)$. Hence, it is enough to check that the weight of the *minimum cut* exceeds T. From this point on, we will refer to $-\log(1-p(e))$ as the *confidence weight* of an edge e.

2.3 Finding *q*-connected modules

CEZANNE is designed to identify modules that are q-connected and have maximum co-expression score. The CEZANNE framework consists of three basic steps: (i) identification of high-scoring seeds; (ii) greedy optimization; and (iii) significance filtering.

2.3.1 Seed identification Our tests show that modules consisting of single nodes provide poor starting points for a local search algorithm with minimum-cut constraints, such as the algorithm we use here (results not shown). Thus, we devised the following seed-finding algorithm. We first execute MATISSE on an unweighted graph containing only edges that pass a certain confidence threshold. This yields a collection of disjoint initial seeds. We then assign the confidence weights to the edges and extract q-connected seeds by recursively computing the minimum cut and using it to split the initial seed into two. This procedure is repeated until the weight of the minimum cut exceeds T. The resulting modules with more than two genes constitute the set of seeds for the optimization phase.

2.3.2 Optimization We use a greedy algorithm to optimize the initial seeds while maintaining their q-connectivity. The basic greedy algorithm described in Ulitsky and Shamir (2007) aims to optimize together a collection of sets (and singletons). It allows the following operations: (i) addition of a singleton to a module; (ii) removal of a node from a module; (iii) reassignment of a node from one module to another; and (iv) merging of two modules. The algorithm iteratively seeks the highest scoring operation and performs it. Here, unlike in Ulitsky and Shamir (2007), edge weights must be taken into account. In order to maintain q-connectivity throughout the optimization procedure, we must make sure that no operation causes the minimum cut in a module to drop below T. This problem is a *dynamic minimum cut* problem (Thorup, 2007) for a weighted graph. Its simple (but expensive) solution is to solve a new minimum cut problem for every tested operation. Instead, we use the following heuristic. We use an implementation of the Stoer–Wagner algorithm (Stoer and Wagner, 1997) for each minimum cut computation, which requires $O(mn+n\log n)$ on a graph with n nodes and m weighted edges. The observations below allow us to perform a relatively limited number of such computations, keeping the running time of the entire algorithm practical on a standard PC. Our optimization first considers all possible node additions and module merges. Node removal or reassignment is considered only if no such operation can improve the score.

Node addition and module merging. Let $C(U)$ be the weight of the minimum cut in the subgraph induced by the module U. We observe that, since the confidence weights are non-negative,

$$C(U\cup\{x\}) \geqslant \min\left\{C(U), \sum_{u\in U} w(x,u)\right\}$$

Suppose that U is q-connected, and we are considering adding x to U. If $\sum_{u\in U}(x,u)\geqslant T$, then $U\cup\{x\}$ will also be q-connected. The total weight of the edges between every node x and every module U can be easily maintained in $O(m)$ after each operation performed. Similarly, in module merging,

$$C(U_1\cup U_2) \geqslant \min\left\{C(U_1)+C(U_2), \sum_{x\in U_1, y\in U_2} w(x,y)\right\}.$$

In that case, it is enough to maintain the total weight of the edges between every pair of modules. This enables addition and merging operations to be checked efficiently without executing the full minimum cut computation.

Node removal or reassignment. Since $C(U\backslash\{x\})$ can be significantly smaller than $C(U)$, we must explicitly validate that node removal does not violate the q-connectivity of the module. We call a node $v\in M$ *min-cut essential* if $C(M\backslash\{v\})<T$. The set of min-cut essential nodes can be maintained throughout the optimization, and recomputed only when necessary using the Stoer–Wagner algorithm. Specifically, the min-cut essential nodes are recomputed every time the removal of any node v from module U can improve the score, unless U has not changed since the last time its minimum cut was computed.

2.3.3 Evaluation of statistical significance An empirical P-value for module significance was computed as follows: we randomly shuffled the expression pattern of each gene and re-ran the algorithm. This process was repeated 100 times and the highest co-expression score obtained in each run was recorded. Modules in the real dataset were given P-values according to the distribution of these recoded scores. Only modules with $P<0.1$ were retained.

2.4 Module annotation with gene functional categories

We used the TANGO algorithm (Shamir *et al*., 2005) to find annotations enriched in the modules. TANGO considers all levels of gene ontology (GO) and uses the standard hypergeometric test to compute raw enrichment P-values. It then uses resampling to correct these P-values for multiple testing and for category dependency. Briefly, TANGO repeatedly selects random sets of genes to compute an empirical distribution of maximum P-values for annotation enrichment obtained across a random sample of sets that maintain the same size characteristics as the ones analyzed. TANGO uses this empirical distribution to determine thresholds for significant enrichment on the true clusters. The algorithm filters out redundant categories by performing conditional enrichment tests.

3 RESULTS

3.1 DNA damage response in *S. cerevisiae*

Our method was applied to a dataset containing expression profiles measured over time in wild-type and mutant yeasts exposed to DNA damage caused by methylmethane sulfonate (MMS) or by ionizing radiation (IR) (Gasch *et al*., 2001). This dataset contained 47 expression profiles of 6167 genes. The 2074 genes that showed at least 2-fold change in the expression levels across the conditions were used as front nodes (Section 2). The network and confidence values were based on purification enrichment (PE) scores, as described by Collins *et al.* (2007). Importantly, the GO classifications we later used to compare CEZANNE to other methods were not used to calculate these scores. In order to enhance computational efficiency confidence values below 0.1 were set to 0. The distribution of confidence scores is shown in Supplementary Figure 1. Analysis of the data with CEZANNE resulted in 14 modules encompassing 471 genes (Table 1 and Supplementary File 1). The modules varied greatly in size, ranging from 3 to 346 genes (average 33.6 genes). By using confidence weights, we were not required to set an artificial upper limit on module size, which was necessary with MATISSE (Ulitsky and Shamir, 2007). Enrichment analysis using TANGO (Section 2) found significantly enriched 'biological process' categories in all 14 modules and 'molecular function' categories in 11 modules (79%). When using GO-slim protein complex annotations, 85.7% of the CEZANNE modules were enriched for at least one complex. The enriched GO annotations are listed in Table 1 and in Supplementary File 1.

3.2 Comparison to other methods

The modules obtained by CEZANNE were compared with those obtained on the same data by several other methods: MATISSE (which ignores the edge confidence values), co-clustering of network and expression data (Hanisch *et al*., 2002) and two clustering algorithms (which work only on the expression data): k-means and CLICK (Sharan and Shamir, 2000). Enrichment was computed using the standard hypergeometric test without correction (see Supplementary File 1 for P-values corrected for multiple testing). For each method, we measured the fraction of annotations that are enriched in at least one module at $P<10^{-4}$ (sensitivity) and the fraction of modules enriched with at least one annotation at $P<10^{-4}$ (specificity). We summarized the two terms using the F-measure defined as $F=2\times\text{Sensitivity}\times\text{Specificity}/(\text{Sensitivity}+\text{Specificity})$ (Van Rijsbergen, 1979). Modules extracted using CEZANNE were significantly superior to those extracted by other methods in terms of the enrichment significance for GO biological process, GO-slim complex annotations and MIPS complex annotations (Fig. 2 and Supplementary Fig. 2).

We also compared, for each annotation, the lowest P-value it got in any module identified by each algorithm. When both CEZANNE and a competing algorithm identified the same annotation enriched at $P<10^{-4}$, the enrichment P-values in CEZANNE modules tended to be more significant (sign test, $P<0.01$). The improved performance in comparison to clustering, which uses only expression data and is oblivious of the network, is expected, since it was observed that genes connected in the PPI network tend to be functionally related (Wu *et al*., 2006). This fact is also reflected in the better performance of network-based co-clustering method in comparison to k-means clustering. We verified that the performance comparisons are not biased by a single predominant module (Module 1), which is enriched for many functional categories (Supplementary Fig. 3). We got similar results when using another expression dataset, for the osmotic shock response in yeast (Supplementary Fig. 4).

3.3 DNA damage response modules

The modules found by CEZANNE identify both known and novel pathways involved in *S. cerevisiae* DNA damage response. The largest module, Module 1 with 346 genes, consists of

Table 1. Modules identified in the response of *S. cerevisiae* to DNA damage

Module (size)	GO biological process	*P*-value	GO-slim protein complexes	*P*-value
1 (346)	Ribosome biogenesis and assembly	$1.2{\cdot}10^{-117}$	Ribosome	$5.9{\cdot}10^{-91}$
	Translation	$1.0{\cdot}10^{-85}$	Eukaryotic 43S preinitiation complex	$3.8{\cdot}10^{-49}$
	rRNA processing	$7.5{\cdot}10^{-79}$	Small nucleolar ribonucleoprotein complex	$1.5{\cdot}10^{-41}$
	35S primary transcript processing	$4.6{\cdot}10^{-44}$	DNA-directed RNA polymerase III complex	$3.1{\cdot}10^{-17}$
	Ribosome assembly	$4.3{\cdot}10^{-39}$	Exosome (RNase complex)	$4.4{\cdot}10^{-15}$
	Ribosomal large subunit biogenesis	$9.2{\cdot}10^{-14}$	DNA-directed RNA polymerase I complex	$5.7{\cdot}10^{-14}$
	rRNA modification	$4.4{\cdot}10^{-12}$	Noc complex	$3.2{\cdot}10^{-6}$
2 (38)	Protein catabolism	$1.8{\cdot}10^{-46}$	Proteasome complex (sensu Eukaryota)	$5.7{\cdot}10^{-71}$
	Proteolysis	$9.0{\cdot}10^{-44}$	Proteasome core complex (sensu Eukaryota)	$9.4{\cdot}10^{-32}$
	Ubiquitin cycle	$1.1{\cdot}10^{-42}$		
3 (12)	Histone acetylation	$3.6{\cdot}10^{-13}$	Histone acetyltransferase complex	$2.1{\cdot}10^{-12}$
	Chromatin modification	$5.9{\cdot}10^{-11}$		
	Transcription from RNA polymerase II promoter	$1.4{\cdot}10^{-6}$		
4 (12)	Translation	$1.1{\cdot}10^{-14}$	Ribosome	$1.4{\cdot}10^{-15}$
5 (12)	Nuclear mRNA splicing, via spliceosome	$3.5{\cdot}10^{-21}$	Spliceosome complex	$3.5{\cdot}10^{-17}$
			Small nuclear ribonucleoprotein complex	$2.5{\cdot}10^{-15}$
6 (10)	Barbed-end actin filament capping	$4.8{\cdot}10^{-6}$	F-actin capping protein complex	$4.8{\cdot}10^{-6}$
	Endocytosis	$1.1{\cdot}10^{-5}$		
	Cytoskeleton organization and biogenesis	$2.8{\cdot}10^{-5}$		
7 (8)	Establishment and/or maintenance of chromatin architecture	$1.1{\cdot}10^{-5}$	Chromatin remodeling complex	$4.6{\cdot}10^{-6}$
8 (7)	Glycogen metabolism	$3.0{\cdot}10^{-8}$	Protein phosphatase type 1 complex	$3.3{\cdot}10^{-5}$
	Sporulation (sensu Fungi)	$2.0{\cdot}10^{-6}$		
9 (6)	Translation	$1.1{\cdot}10^{-7}$	Ribosome	$4.0{\cdot}10^{-8}$
10 (6)	tRNA processing	$2.5{\cdot}10^{-14}$	Ribonuclease P complex	$9.2{\cdot}10^{-8}$
	rRNA processing	$2.2{\cdot}10^{-9}$		
11 (4)	Trehalose biosynthesis	$6.8{\cdot}10^{-14}$	Alpha, alpha-trehalose-phosphate synthase complex (UDP-forming)	$6.8{\cdot}10^{-14}$
12 (4)	Ubiquitin-dependent protein catabolism	$5.2{\cdot}10^{-7}$		
13 (3)	Pseudohyphal growth	$9.8{\cdot}10^{-7}$	cAMP-dependent protein kinase complex	$9.6{\cdot}10^{-7}$
14 (3)	Proteasome assembly	$3.2{\cdot}10^{-6}$		
	Protein folding	$3.9{\cdot}10^{-6}$		

P-values listed in the table are raw hypergeometric enrichment scores. Corrected p-values, accounting for multiple testing, appear in Supplementary File 1. All the annotations in this table attained a corrected *P*-value <0.05. Only the seven most significantly enriched GO biological process categories are shown for Module 1.

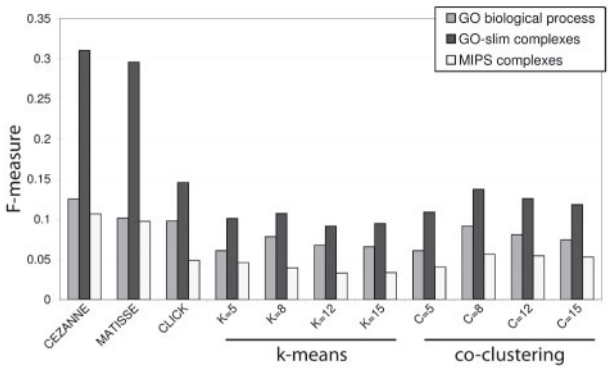


Fig. 2. Performance of several module finding methods. All GO annotations were used in the comparison. The *F*-measure evaluates sensitivity and specificity (see text).

ribosomal biosynthesis proteins, probably the best characterized transcription program in yeast (Gasch *et al*., 2000). These proteins are strongly downregulated in a Mec1-dependent way following both MMS and IR treatments. The second largest module, Module 2 (Fig. 3A), consists of the proteasome, a large complex strongly transcriptionally co-regulated by Rpn4 under various conditions, including DNA damage (London *et al*., 2004). The transcription levels of the genes in the module exhibit mild upregulation following DNA damage, which is stronger after MMS than after IR treatment.

Module 4 (Fig. 3B) consists of 11 known genes from the small subunit of the mitochondrial ribosome that are downregulated following mock irradiation. It also contains *SWS2*, which is a putative mitochondrial ribosomal protein (Gan *et al*., 2002). *SWS2* is significantly correlated to the other genes in the module on the expression level ($r = 0.46$ on average), but is not linked to them using MATISSE, CLICK or other approaches based on expression

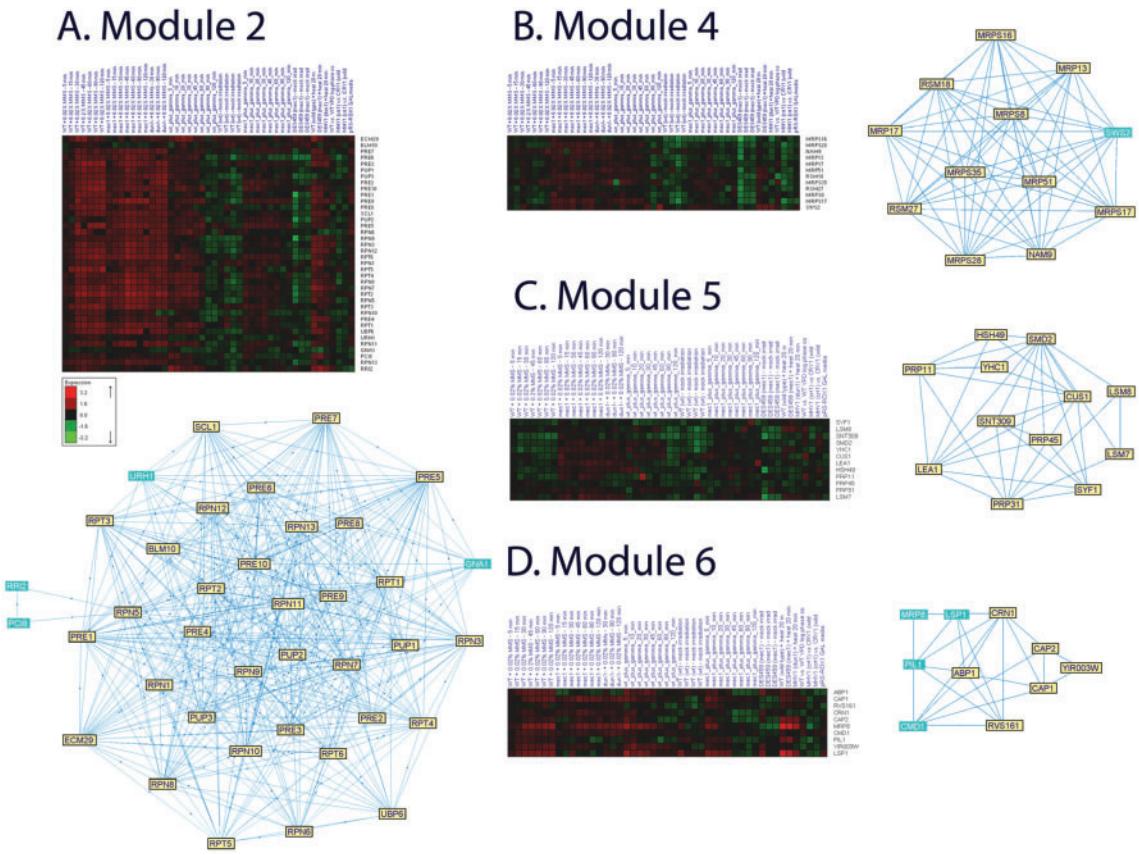


Fig. 3. Modules identified in *S. cerevisiae* response to DNA damage. For each module, the expression heat-map is presented together with the interaction network. In each subnetwork, the genes belonging to the dominant annotation are highlighted. (**A**) Members of the proteasome are in yellow. (**B**) Small mitochondrial ribosome subunit genes (from MIPS) are in yellow. (**C**) Genes annotated with 'nuclear mRNA splicing' in GO are in yellow. (**D**) Genes localized to actin in (Huh *et al*., 2003) are in yellow.

data (Tanay *et al*., 2005; Wapinski *et al*., 2007). Our analysis thus provides further support for the role of *SWS2* in the small subunit of the mitochondrial ribosome, adding to evidence based on localization (Huh *et al*., 2003), sequence (Gan *et al*., 2002) and deletion phenotypes (Steinmetz *et al*., 2002). Members of the large mitochondrial ribosomal subunit are enriched in a different module, Module 9.

Module 5 (Fig. 3C) consists of 12 spliceosome-related genes, whose transcription is weakly but consistently downregulated in a Mec1-dependent manner following DNA damage. This raises the interesting possibility of the spliceosome's involvement in the DNA damage response. Nine of the 12 genes in Module 5 are essential and therefore were not tested in systematic screens for MMS-affected genes. However, deletion of two of the non-essential genes, *LEA1* and *LSM7*, caused MMS sensitivity (Parsons *et al*., 2006).

Module 6 (Fig. 3D) is a 10-gene module strongly upregulated after DNA damage and other stresses, as evident in the Gasch *et al.* (2000) stress dataset. Module 6 contains members of two known complexes: two members of the F-actin capping protein complex and two of the eisosome complex. Interestingly, six of the module's genes are localized to actin (Huh *et al*., 2003) ($P = 4.8 \cdot 10^{-6}$), including *YIR003W*, a protein of unknown function. *CMD1* (calmodulin) is known to be required for actin organization (Desrivieres *et al*., 2002). Surprisingly, this module also contains *MRP8*, a putative mitochondrial ribosomal protein that was shown to have a different transcriptional program than the known mitochondrial ribosome proteins (Matsumoto *et al*., 2005). Our results further suggest that *MRP8* has a role unrelated to mitochondria, perhaps one involving cytoskeleton organization. Module 6 was strongly upregulated in response to treatment with a variety of DNA damaging agents, without dependence on Rpn4, in another DNA damage dataset (Jelinsky *et al*., 2000), and was strongly upregulated following a variety of other stresses in a stress dataset (Gasch *et al*., 2000). The phenotypic profile of the Δ*yir003w* strain in (Brown *et al*., 2006) was similar to that of the Δ*abp1* and Δ*cap1* strains (Pearson correlations of 0.49 and 0.12, respectively) in that all three deletion mutants show sensitivity to Calcofluor, a phenotype related to cell wall biosynthesis. Taken together, these findings suggest that Module 6 corresponds to a novel transcriptionally co-regulated complex or pathway with cellular localization at actin microfilaments.

These findings demonstrate the ability of CEZANNE to extract modules that correlate well with the known biology of transcriptional responses, and to point to novel functional associations between genes and processes.

3.4 Robustness to noise in the interaction network

In order to test the effect of noise in the network on the performance of CEZANNE, we randomly removed or added edges to the interaction network and reevaluated the sensitivity and specificity of the obtained modules using GO and MIPS gene annotations. The results are presented in Supplementary Figure 5. We find that removal of up to 20% of the edges or randomly doubling the number of edges degrades performance by not more than 20%. The better tolerance to edge addition compared to edge removal is probably due to CEZANNE's ability to ignore edges that do not connect co-expressed genes.

3.5 Implementation and user interface

A graphical user interface to CEZANNE is available as part of the MATISSE software (http://acgt.cs.tau.ac.il/matisse). It allows full setting of the methods parameters, execution on network and expression data from any organism, visualization of the network and expression data for each module and functional annotation of the obtained modules. The Java source code for CEZANNE is available upon request.

4 DISCUSSION

We have presented a novel approach that makes better use of PPI networks for the interpretation of microarray study results. Augmented with proper search algorithms, our methodology can be used to improve other methods involving network connectivity, such as those described in (Chuang *et al*., 2007; Ideker *et al*., 2002; Nacu *et al*., 2007; Ulitsky *et al*., 2008). The approach is not specific to PPI networks and can applied directly to other networks with differential interaction confidence, such as protein-DNA (Lee *et al*., 2002) and functional linkage (von Mering *et al*., 2007) networks.

We note that the interaction probabilities we use here correspond to the confidence in the existence of an interaction, and are not the probability that an interaction takes place in the cell at any particular time point. However, if information on the latter becomes available it can also be used by our method.

While the results of our method are promising, there is room for many algorithmic improvements. The greedy optimization algorithm we currently employ can converge to local minima, in terms of both the co-expression score and the minimum cut requirements. Our approach can be improved by better search initialization algorithms and by allowing more complex optimization moves (e.g. adding two nodes simultaneously). The latter approach will probably demand a more efficient optimization algorithm, one that requires less time per iteration for maintaining the minimum cut.

Which method should be used for future data analysis—MATISSE or CEZANNE? The answer depends on the availability and the quality of the interaction confidence data. Information on functional interactions for several species is available in the STRING database (von Mering *et al*., 2007). Confidence of individual PPI interactions is yet to be systematically assessed in most species. Given a confidence-based network for the studied organism, as our results show, CEZANNE should be the method of choice. In the absence of reliable confidence values MATISSE is more useful.

The modules found by CEZANNE in the DNA damage response of *S. cerevisiae* accurately identify complexes with known roles in DNA repair, such as the RPA and complexes whose regulation is known to be related to stress response in *S. cerevisiae*, such as the ribosomes and the proteasome. In addition, we identify rather large modules that were not previously associated with DNA damage response. This highlights the main goal of integrating network into gene expression analysis: achieving higher sensitivity in identifying transcriptional programs that are missed when the analysis if performed on the level of an individual gene. Together with the user-friendly interface that we provide, we hope that CEZANNE will be highly instrumental in the analysis of future microarray studies.

Funding: European Commission within its FP7 Programme (APOSYS project, contract number HEALTH-F4-2007-200767); Converging Technologies Program of the Israel Science Foundation (grant no 1767.07); Edmond J. Safra bioinformatics program at Tel-Aviv University (fellowship to I.U., in part).

Conflict of Interest: none declared.

REFERENCES

Brown,J.A. *et al*. (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol. Syst. Biol.*, **2**, 2006 0001.

Cabusora,L. *et al*. (2005) Differential network expression during drug and stress response. *Bioinformatics*, **21**, 2898–2905.

Chuang,H.Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Collins,S.R. *et al*. (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell Proteomics*, **6**, 439–450.

Desrivieres,S. *et al*. (2002) Calmodulin controls organization of the actin cytoskeleton via regulation of phosphatidylinositol (4,5)-bisphosphate synthesis in *Saccharomyces cerevisiae*. *Biochem. J*., **366**, 945–951.

Gan,X. *et al*. (2002) Tag-mediated isolation of yeast mitochondrial ribosome and mass spectrometric identification of its new components. *Eur. J. Biochem.*, **269**, 5203–5214.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

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

Guo,Z. *et al*. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, **23**, 2121–2128.

Hanisch,D. *et al*. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18** (**Suppl 1**), S145–S154.

Huh,W.K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Ideker,T. *et al*. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (**Suppl 1**), S233–S240.

Jelinsky,S.A. *et al*. (2000) Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell Biol.*, **20**, 8157–8167.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Li,D. *et al*. (2008) PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Mol. Cell Proteomics*, **7**, 1043–1052.

Liu,M. *et al*. (2007) Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet*, **3**, e96.

London,M.K. *et al*. (2004) Regulatory mechanisms controlling biogenesis of ubiquitin and the proteasome. *FEBS Lett*., **567**, 259–264.

Ma,X. *et al*. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.

Matsumoto,R. *et al*. (2005) The stress response against denatured proteins in the deletion of cytosolic chaperones SSA1/2 is different from heat-shock response in *Saccharomyces cerevisiae*. *BMC Genomics*, **6**, 141.

Müller,F.J. *et al*. (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**, 401–405.

Nacu,S. *et al*. (2007) Gene expression network analysis and applications to immunology. *Bioinformatics*, **23**, 850–858.

Parsons,A.B. *et al*. (2006) Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell*, **126**, 611–625.

Rajagopalan,D. and Agarwal,P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 788–793.

Rapaport,F. *et al*. (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Rhodes,D.R. *et al*. (2005) Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.

Segal,E. *et al*. (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19** (**Suppl 1**), i264–i271.

Shamir,R. *et al*. (2005) EXPANDER–an integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**, 232.

Sharan,R. and Shamir,R. (2000) CLICK: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 307–316.

Steinmetz,L.M. *et al*. (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.*, **31**, 400–404.

Stoer,M. and Wagner,F. (1997) A simple min-cut algorithm. *JACM*, **44**, 585–591.

Suthram,S. *et al*. (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**, 360.

Tanay,A. *et al*. (2005) Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Mol. Syst. Biol.*, **1**, 2005 0002.

Thorup,M. (2007) Fully-dynamic min-cut. *Combinatorica*, **27**, 91–127.

Ulitsky,I. and Shamir,R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol*., **1**, 8.

Ulitsky,I. *et al*. (2008) Detecting Disease-specific Dysregulated Pathways via Analysis of Clinical Expression Profiles. In *Proceedings of Research in Computational Molecular Biology (RECOMB) 2008*. Vol. 4955/2008, Springer, Berlin, pp. 347–359.

Van Rijsbergen,C.J. (1979) *Information Retrieval*. Butterworth-Heinemann Newton, MA, USA.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

von Mering,C. *et al*. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*., **35**, D358–D362.

Wapinski,I. *et al*. (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature*, **449**, 54–61.

Wei,P. and Pan,W. (2008) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, **24**, 404–411.

Wu,X. *et al*. (2006) Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res*., **34**, 2137–2150.

7. Pathway Redundancy and Protein Essentiality Revealed in the *S. cerevisiae* Interaction Networks

Molecular Systems Biology 3; Article number 104; doi:10.1038/msb4100144
Citation: ***Molecular Systems Biology*** **3:104**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks

Igor Ulitsky and Ron Shamir*

School of Computer Science, Tel Aviv University, Tel Aviv, Israel
* Corresponding author. School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. Tel.: +972 3 6405383; Fax: +972 3 6405384;
E-mail: rshamir@tau.ac.il

Received 8.1.07; accepted 11.2.07

The biological interpretation of genetic interactions is a major challenge. Recently, Kelley and Ideker proposed a method to analyze together genetic and physical networks, which explains many of the known genetic interactions as linking different pathways in the physical network. Here, we extend this method and devise novel analytic tools for interpreting genetic interactions in a physical context. Applying these tools on a large-scale *Saccharomyces cerevisiae* data set, our analysis reveals 140 between-pathway models that explain 3765 genetic interactions, roughly doubling those that were previously explained. Model genes tend to have short mRNA half-lives and many phosphorylation sites, suggesting that their stringent regulation is linked to pathway redundancy. We also identify 'pivot' proteins that have many physical interactions with both pathways in our models, and show that pivots tend to be essential and highly conserved. Our analysis of models and pivots sheds light on the organization of the cellular machinery as well as on the roles of individual proteins.
Molecular Systems Biology 17 April 2007; doi:10.1038/msb4100144
Subject Categories: metabolic and regulatory networks; computational methods
Keywords: essential genes; genetic interactions; pathway analysis; protein interactions; *S. cerevisiae*

Introduction

Gene knockout studies have shown that only ~18% of *Saccharomyces cerevisiae* genes are essential for growth on a rich medium (Giaever *et al*, 2002). Consequently, buffering on the genetic level is believed to be abundant in eukaryotes (Hartman *et al*, 2001). To better understand the role of nonessential genes, several large-scale studies performed double knockouts (Pan *et al*, 2004; Tong *et al*, 2004) and identified many events of *synthetic lethality*, where a mutant carrying deletions of two nonessential genes is lethal, and *synthetic sickness*, where the mutant shows a weaker phenotype. We will use here the term genetic interaction (GI) for the interaction of two genetic perturbations in affecting the phenotype, whether lethal or sick. The graph that has genes as nodes and edges corresponding to GIs is called the *GI network*.

Recent technologies also enable a systematic mapping of protein–protein (Ito *et al*, 2001) and protein–DNA (Lee *et al*, 2002) interactions (*physical* interactions (PIs)), yielding large PI networks. As the networks get larger, the need for computational tools for dissecting them is mounting. The integrated analysis of PI and GI networks is a compelling challenge, as they carry important and complementary biological signals. Initial studies have shown that proteins in the same region of the GI network are slightly more likely to interact physically (Tong *et al*, 2001, 2004), and that a protein with many PIs is likely to have also many GIs (Ozier *et al*, 2003).

The modular nature of the cellular organization has been widely recognized (Hartman *et al*, 2001). Many methods have been developed for detecting functional modules within PI networks. Such modules, often termed *pathways*, represent physically interacting proteins involved in carrying out a particular function. Depending on the detection method, pathways may represent molecular complexes (Bader and Hogue, 2003) or signaling cascades (Rives and Galitski, 2003). Kelley and Ideker (2005) defined a pathway as a group of proteins that are densely interconnected in the PI network, and studied the frequency of GIs within and between such pathways. In a systematic analysis of large-scale GI and PI data, they concluded that between-pathway explanations of GIs are ~3.5 times more abundant than within-pathway explanations, and concluded that GIs mostly bridge redundant processes. Further arguments for the prevalence of between-pathway GIs were given by Ye *et al* (2005), who postulated that genes in the same pathway are expected to share common GI partners, and used similarity of GI patterns in a successful function prediction. Similar results were established recently on the DNA damage system (Pan *et al*, 2006).

Results and discussion

Assembly of GI and PI networks

We assembled a GI network (Figure 1A) by taking from the BioGRID database version 2.0.19 (Stark *et al*, 2006) 13 632 synthetic lethality and synthetic fitness interactions for *S. cerevisiae*, covering 2682 genes. By focusing on genes with at least two interactions, we obtained a *GI network* of 1869 genes and 12 850 interactions. Our PI network, consisting of protein–protein and protein–DNA interactions from multiple sources (Supplementary information 1), contained 68 172 interactions covering 6184 proteins.

Pathway definitions and between-pathway models

Our starting point was the computational framework of Kelley and Ideker (2005) for detection of between-pathway interpretations for GIs. Kelley and Ideker define a 'pathway' as a densely connected set of proteins in the PI network, and a 'between-pathway model' as a disjoint pair of pathways that are densely interconnected in the GI network (Figure 1B). Models are defined probabilistically and are found using a greedy algorithm. While the requirement of high PI density is appropriate for complexes, many other known biological pathways (e.g. linear signaling cascades) do not induce dense subnetworks in the physical network. We therefore chose to employ an alternative definition, in which *a pathway is simply a connected subnetwork* in the PI network (a *connected pathway*, described in Materials and methods). An example of two sparse pathways is presented in Figure 2A. The buffering between the mechanism of DNA repair through homologous recombination and the response to oxidative stress is indeed only partially recovered when using the dense pathway definition (not shown). We define a *between-pathway model* (*BPM*) as in Kelley and Ideker (2005), but using the new notion of a pathway (Figure 1C). The scoring of models and the model detection algorithm are defined in Materials and methods. A comparison of our models with those found using dense pathways on the same interaction data (Supplementary information A) shows that we construct more between-pathway explanations of GIs (3765 versus 3117), while maintaining the significant functional content of models. Our models also allow more direct interpretation of specific buffering cases than congruence methods (Supplementary information B).

A comprehensive model map in *S. cerevisiae*

Our BPM finding approach generated 140 models and provided between-pathway explanations for 3765 GIs, a 2.7-fold increase from the 1377 interactions explained in Kelley and Ideker (2005). This is mainly due to the incorporation of more GIs (12 850 instead of 4125) and to a lesser extent due to using more PIs (68 172 versus 27 604), as we use those only

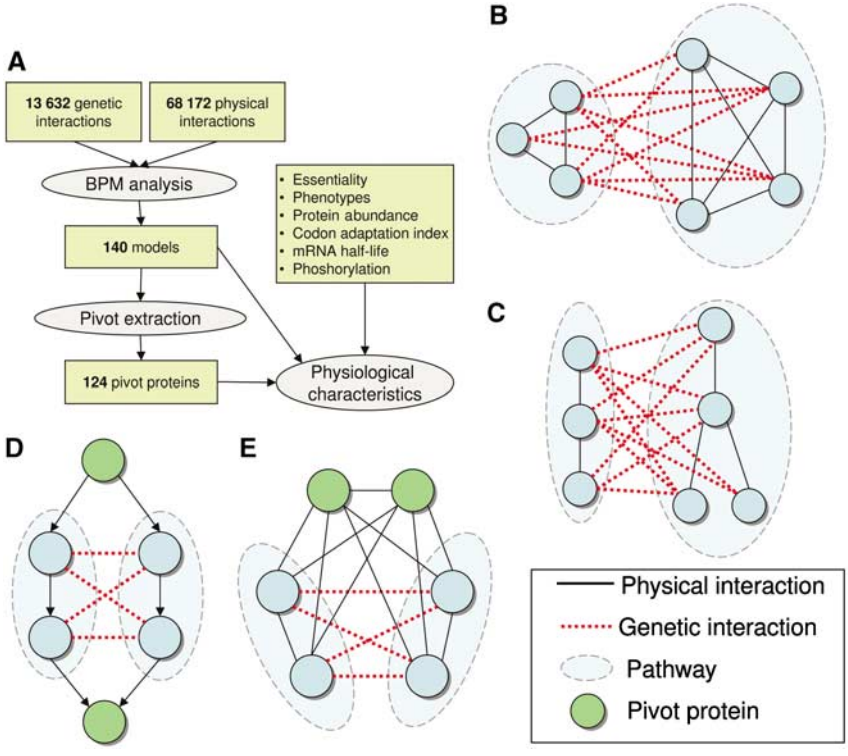


Figure 1 Study outline and methodology. (**A**) Overview of the analysis methods and the reported results. (**B**) A BPM constructed from two dense pathways in the PI network. (**C**) A BPM constructed from two connected pathways in the PI network. (**D, E**) Examples of two biological explanations for pivot proteins. In (D), the pivots correspond to shared members of two linear pathways, where the signal flow is indicated by the arrow directions, and in (E), they correspond to shared complex members. Note that the pivots in (E) are not redundant, as they are densely connected to both pathways with physical interactions and do not have a genetic interaction between them.

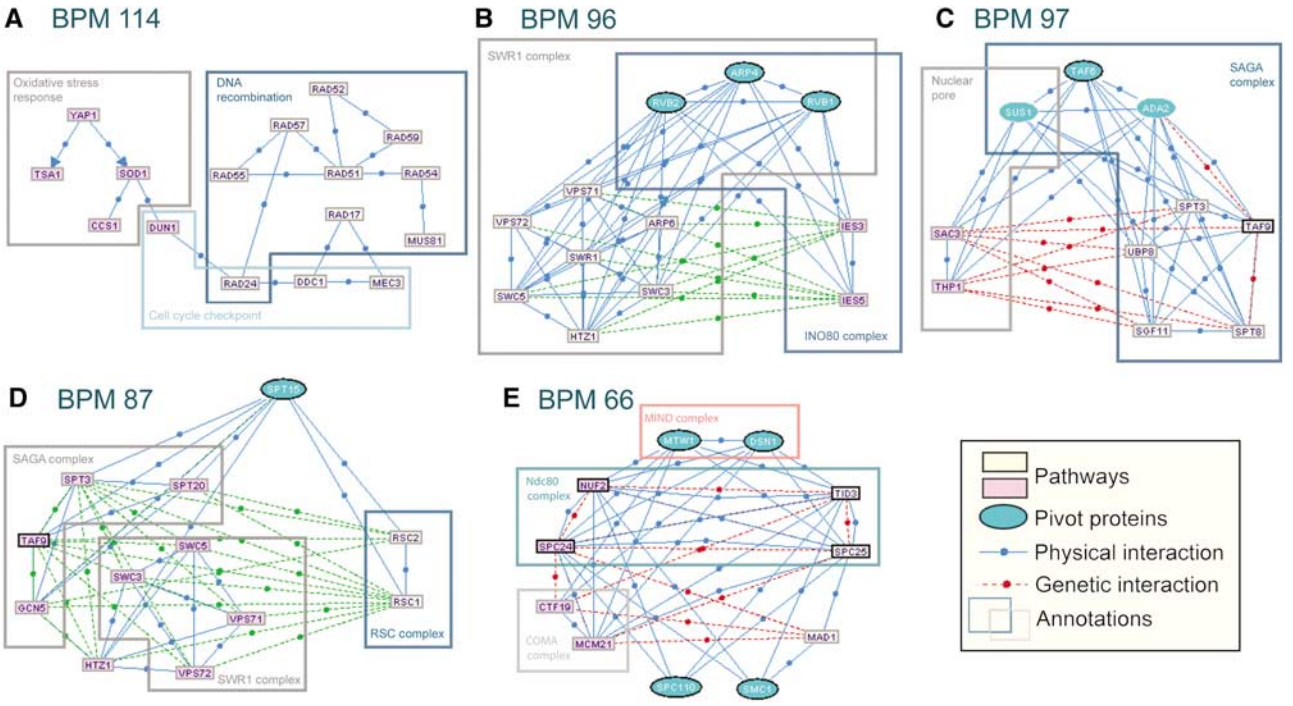


Figure 2 Model examples. Rectangles represent genes, and the two pathways are shown in different colors. The blue ovals are the pivot nodes. Essentials genes are drawn with thicker border. See main text for discussion of each of the BPMs (**A–E**). In (A), for clarity, 52 genetic interactions between the pathways are not shown.

to create 'scaffolds' of pathways and not for scoring models. The gene content of the models is available in a supplementary archive. A full description and an interactive visualization of the models are available at http://acgt.cs.tau.ac.il/bpm.

Functional enrichment of models

We utilized the TANGO algorithm (Shamir *et al*, 2005) in order to test the functional enrichment of the models in GO categories (Ashburner *et al*, 2000). A total of 71.4 and 69.3% of the BPMs were enriched with at least one functional category from the 'biological process' and the 'cellular compartment' ontologies, respectively (Supplementary Table S1). Of the complexes annotated in SGD GO-slim (Cherry *et al*, 1998), 46.3% were enriched in at least one BPM. Despite the low coverage of the GI network, the coverage of complexes is comparable to that achievable by direct analysis of the PI network (Supplementary information C).

Phenotypic coherence within and between pathways

To what extent do the two pathways in a BPM have the same function? To answer this question, we used the phenotypic contribution of non-essential genes in *S. cerevisiae*, as measured by the fitness of deletion mutants in diverse treatments (Brown *et al*, 2006). We found that a BPM pathway is significantly more coherent in its phenotypic response pattern than random connected groups of the same size in the physical network (Pearson correlation of 0.384 versus an average of 0.0382, $P<0.001$; Supplementary Figure S2). The correlation between the pathways in a BPM is also higher than expected (0.112 versus 0.0378 expected, $P<0.001$; Supplementary Figure S2).

Identification of pivot proteins

As we established that the pathways within models frequently represent coherent functional units, and as functional units within the cell sometimes share components (Krause *et al*, 2004), we tested the possibility of computationally detecting such shared components within the PI network. To this end, for each model, we sought proteins that are densely connected to both of its pathways (Figure 1D and E, Materials and methods), and called them the *pivot proteins* for the model. Altogether, we identified 124 distinct pivots in 40 models. On average, 1.09 pivots were found in each model, and each pivot appeared in 1.22 models.

We systematically analyzed the representation of proteins that are known to take part in several distinct processes in the group of pivots. To this end, we identified proteins participating in several complexes or pathways (see Materials and methods), and also used a curated set of multicomplexed genes (Krause *et al*, 2004). As summarized in Table I, the pivots were enriched in all three sets. One example of such overlap is in BPM 96 (Figure 2B). In a model containing as pathways parts of the SWR1 and Ino80 complexes involved in chromatin remodeling, we identified the pivot proteins Arp4, Rvb1 and Rvb2, three out of the four proteins known to participate in both the SWR1 and the Ino80 complexes (Shen *et al*, 2000; Krogan *et al*, 2003). In BPM 97 (Figure 2C), Sus1, which has been shown to take part both in the nuclear pore

Table I Multiple roles of pivot proteins

	No. of proteins	Pivots	Expected	Significance
GO complexes	206	21	4.35	$P=1.27\times10^{-9}$
KEGG pathways	390	11	8.24	$P=0.200$
Filtered KEGG pathways	71	6	1.50	$P=3.68\times10^{-3}$
Known complex overlaps	39	8	0.55	$P=7.49\times10^{-9}$

The enrichment of proteins known to participate in multiple physical pathways within the set of pivot proteins. GO complexes are taken from SGD 'macromolecular complex GO-slim' ontology. The filtered KEGG pathways are KEGG pathways in which at least 50% genes formed a connected component in our physical network. Known complex overlaps are taken from Krause *et al* (2004). Significance was evaluated using hypergeometric distribution.

and the SAGA complex (Rodriguez-Navarro *et al*, 2004) was identified as a pivot in a model representing GIs between the two pathways. When pivots do not correspond to known complex or pathway overlaps, they frequently represent general purpose genes cooperating with multiple pathways. For example, in BPM 87, the general transcription factor Spt15 was identified as a pivot of a model that contains components of the distinct transcription-related complexes RSC, SWR1 and SAGA (Figure 2D).

Essentiality and evolutionary retention of pivot proteins

The two pathways that form a model are often partially redundant in function, and as the pivots represent proteins that are active in both pathways, we hypothesized that the pivots will frequently correspond to essential genes. Indeed, 72 of the pivots were found to be essential, a highly significant fraction given the total number of essential genes in the network (22.6 essentials expected, $P=1.42\times10^{-23}$). Although we observed a general strong correlation between degree and essentiality ($P=5.87\times10^{-85}$ using rank-sum test), as previously reported (Jeong *et al*, 2001), the high prevalence of essential genes among pivots is far beyond what can be explained by their degrees alone (39.98 essentials expected, $P<10^{-5}$). The essential pivots also tend to have closer functions to their BPMs (Supplementary information D). Essentiality was recently shown to be connected to the evolutionary retention of genes in eukaryotes (Gustafson *et al*, 2006). Using the data from Gustafson *et al* (2006) we found that the pivots are significantly retained in evolution ($P=9.79\times10^{-9}$), even when controlling for the large fraction of essential genes ($P=0.029$).

Example: the spindle checkpoint model

An interesting example of using models and pivots for understanding cellular mechanisms is BPM 66 (Figure 2E). This model represents buffering between different components of the kinetochore, a complex bound to the centromeres of chromosomes during mitosis. Together with its pivots, the model is composed of members of three known subcomplexes of the inner kinetochore: Ndc80, COMA and MIND (De Wulf *et al*, 2003). Specifically, the pivots Mtw1 and Dsn1 correspond to a distinct unit of the MIND complex, which bridges different kinetochore subcomplexes (De Wulf *et al*, 2003; Westermann *et al*, 2003). The two subunits of the Ndc80 complex, Nuf2/Spc24 and Tid3/Spc25, which were shown to be at least partially redundant (McCleland *et al*, 2003), are placed in different pathways. Note that even though the four subunits of the Ndc80 complex show all possible GIs and PIs between pairs, the biologically correct partition of this complex into pathways was obtained by taking into consideration the other GIs in the model. For example, Mad1 physically connects only with the Tid3/Spc25 subunit and genetically interacts only with the Nuf2/Spc24 subunit. These additional external interactions cause the biologically correct partition to score higher than other alternatives.

Mad1, a highly conserved protein with a specific function in the spindle cell cycle checkpoint, is of additional interest. At the spindle checkpoint, the cells are arrested in metaphase until all chromosomes successfully attach to microtubules. Tid3 and Spc25, members of the Ndc80 complex, which appear in the pathway with Mad1, were specifically linked to the spindle checkpoint in several organisms (summarized in Bharadwaj *et al*, 2004). Moreover, the recruitment of Mad1 was shown to be dependent on Spc25 and Tid3 in *Xenopus* and human cells (McCleland *et al*, 2003; Bharadwaj *et al*, 2004) and the spindle checkpoint was shown to be defective in *spc25* mutants (Wigge and Kilmartin, 2001). How exactly does Mad1 attach to the kinetochore is currently not known. Although Mad1 shows a yeast two-hybrid interaction with Spc25 in *S. cerevisiae* (Newman *et al*, 2000) and with Tid3 in human cells (Martin-Lluesma *et al*, 2002), attempts to demonstrate a biochemical interaction between Ndc80 and Mad1 have been reported unsuccessful (Martin-Lluesma *et al*, 2002; McCleland *et al*, 2003).

In our model, Mad1 is linked to the pivot Smc1, a member of the cohesin complex, required for sister chromatid cohesion during mitosis. Smc1 was shown to localize to the kinetochore during meiosis and to interact with Tid3 in yeast and human cells (Zheng *et al*, 1999; Gregson *et al*, 2002). Furthermore, Smc1 was shown to be required for proper assembly of the mitotic spindle in human cells (Gregson *et al*, 2001), but its exact function in the metaphase is unknown. Our findings suggest that the connection of Mad1 to the kinetochore in general and to the Ndc80 complex in particular, is mediated through Smc1. Note how the use of pivots provides additional clues to BPM annotation and to the understanding of inter-pathway organization.

Physiological properties of models

Recently, the physiological properties of the PI network hubs were extensively analyzed (Batada *et al*, 2006). As many proteins in our models were such hubs, we asked whether

their physiological properties differed from those of other hubs. Here, we focus on the analysis of mRNA stability (Wang *et al*, 2002) and the number of putative phosphorylation sites (Obenauer *et al*, 2003), two properties manifesting the turnover and regulation of the gene. Detailed analysis, covering additional properties, is described in Supplementary information E and Supplementary Table S2.

Our tests show that BPM genes behave significantly unlike other genes. The average mRNA half-life of BPM genes was 21.4, versus 26.2 in all other genes (P=9.97 × 10^{-10}, rank-sum test). The average number of phosphorylation sites was 5.1, versus 4.0 for all other genes (P=3.13 × 10^{-9}). Both parameters were significantly correlated between the two pathways that constitute a BPM (intraclass correlations of 0.408 and 0.189, $P<10^{-4}$ and P=0.0184, respectively). This finding cannot be explained by the high degrees of model genes (Supplementary Table S2).

Note that the mRNA half-lives are experimentally derived, whereas the phosphorylation sites are computationally inferred from sequence. On all the 3552 genes for which both parameters are available, they are not correlated (ρ=−0.0182 P=0.276). However, the parameters are significantly correlated on the genes within BPMs (566 genes, ρ=−0.132, P=1.6 × 10^{-3}). These results remain highly significant when controlling for key functions enriched in the model participants (Supplementary Figure S4). These findings suggest that genes in BPMs might experience more stringent regulation. A possible hypothesis is that in pathways for which redundant mechanisms are available, a tighter regulation allows the cell to switch between the alternatives faster. However, this conclusion has to be revalidated when GIs covering a wider functional range become available.

Our results indicate that despite the limitations of today's GI and PI networks, their integrated analysis is a powerful approach for understanding the organization of the yeast cellular system. We expect that such analysis will provide insights into the large picture of genetic redundancy in higher eukaryotes as well.

Materials and methods

Scoring models

We build upon the probabilistic score described in Kelley and Ideker (2005) and Sharan *et al* (2005) to identify between-pathway explanations for GIs by finding BPMs within the GI and the PI networks. Let $G^G=(V,E^G)$ be the GI network and let $G^P=(V,E^P)$ be the PI network. Note that nodes in V represent both the genes and their products, depending on the context. A BPM is a pair of disjoint sets V_1, V_2, such that (a) $|V_1|, |V_2| \geqslant 2$, (b) each V_i induces a connected subgraph in G^P and (c) there are unusually many GIs between V_1 and V_2 (Figure 1C). We call each V_i a *pathway*. Property (c) reflects the assumption that genetic buffering implies a dense set of GIs between the pathways.

We now quantify property (c). We derive a log-odds score reflecting the likelihood that the density of GIs between the two pathways is unusually high. We compare two hypotheses: under the *BPM hypothesis*, every pair of genes, one from V_1 and the other from V_2, genetically interact with a high probability β, independently of all other gene pairs, and the likelihood of a model (V_1, V_2) is thus $\Pi_{(a,b)\in(V1\times V2)}\beta I(a,b)+(1-\beta)(1-I(a,b))$, where $I(a,b)$ equals 1 if there exists a GI between a and b and otherwise equals 0; in the *null hypothesis*, every pair (a,b) is connected with probability $r_{a,b}$, representing the chance of observing this interaction at random, given the degrees of a and b in G^G. We estimate $r_{a,b}$ by generating a random ensemble of networks with the same degree sequence and counting what fraction of them contain an interaction between a and b. The log-odds score is then

$$S(V_1,V_2)=\log\frac{P(V_1,V_2|M_{\mathrm{BPM}})}{P(V_1,V_2|M_{\mathrm{null}})}=\frac{\log\prod_{(a,b)\in V_1\times V_2}\beta \mathrm{I}(a,b)+(1-\beta)(1-\mathrm{I}(\mathrm{a},\mathrm{b}))}{\prod_{(a,b)\in V_1\times V_2} r_{a,b}I(a,b)+(1-r_{a,b})(1-I(a,b))}$$

The main difference between this score and the score described in Kelley and Ideker (2005) is in the structure imposed by the BPMs in the PI network. Here we do not score the model for density of PIs within each pathway, and instead require that each is connected in G^P.

Model finding algorithm

The model finding procedure described in Kelley and Ideker (2005) is a greedy network search algorithm that uses as seeds single GIs. We improve this procedure by initializing the algorithm with better seeds that are maximal bicliques in the GI network (a biclique is a disjoint pair of node sets such that every node in each set has edges to all nodes in the other set). The following procedure is performed for each u, v such that $(u,v)\in E^P$

1. Identify B—the set of nodes adjacent to both u and v in G^G. Proceed only if $|B|\geqslant k_{\min}$.
2. Partition B into connected components in G^P: $B_1, B_2, \ldots, B_l$.
3. For each B_i such that $|B_i|\geqslant k_{\min}$, identify the set A_i of nodes adjacent to all the nodes in B_i in G^G.
4. For each A_i, partition it into the connected components it induces in G^P, $A_i^1, A_i^2, \ldots, A_i^m$ and add (B_i, A_i^j) to the set of seeds if $|A_i^j|\geqslant k_{\min}$.

This algorithm produces maximal bicliques (V_1, V_2) in G^G, such that each V_i induces a connected component in G^P and has size $\geqslant k_{\min}$. The produced set of seeds is then filtered for overlaps. We used $k_{\min}$=2 in all tests. Owing to the relatively sparse nature of both interaction networks, this method is very efficient in practice.

The optimization phase starts with each seed as a candidate model, and greedily tries to improve the score of the current model through addition, removal or exchange of nodes between the two pathways while keeping each pathway connected in the PI network. In order to efficiently keep track of the connectivity requirement, we use the notion of articulation nodes. An *articulation node* in a connected graph is a node whose removal disconnects the graph. Articulation nodes can be efficiently detected during a depth-first search traversal of the graph, by calculating the 'lowpoint' values of every node (cf. Even, 1979). The algorithm maintains the connectivity of each pathway by dynamically updating, for each pathway, its set of articulation nodes. This set is used to ban optimization moves which disrupt connectivity. After the optimization, a filtering step removes models that overlap by $>50\%$ in both pathways to higher scoring models. An additional model filtering step computes an empiric P-value by sampling 1000 random gene groups of the same size, and retains only BPMs with $P<0.05$ (see Supplementary information F).

Identification of pivot proteins

Given a model (V_1, V_2), we seek nodes that are densely connected to both pathways in the physical network. Specifically, for every $v\in V$, denote by $N_i(v)$ the nodes in V_i that are adjacent to v in G^P. We call v a *pivot* if, for i=1,2, $|N_i(v)|\geqslant l_{\min}$ and $|N_i(v)|$ is significant, given the degree of v in G^P ($P<P_{\max}$, using hypergeometric test). In the actual analysis described in this paper, we used $l_{\min}$=2 and $P_{\max}$=0.05. In addition, to filter master regulator genes that are involved in many processes, such as protein folding chaperons, we only considered as pivots proteins with degree <250 in G^P.

Statistical analysis

Correlation analysis was performed using the non-parametric Spearman test, unless otherwise indicated. All P-values reported when

controlling for a specific gene class (e.g. essential genes) were obtained by random sampling of a large number of gene groups with the same fraction of genes from that class. The *P*-values reported when controlling for the degrees were calculated by first binning all the genes into 40 equal-size bins based on their degree, and then sampling genes from the bins, while maintaining the proportion of genes from each bin.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank the members of the Shamir laboratory for helpful discussions. IU is a fellow of the Edmond J Safra Bioinformatics Program at Tel-Aviv University. RS was supported in part by the Wolfson foundation, and by the EMI-CD project that is funded by the European Commission within its FP6 Programme, under the thematic area 'Life Sciences, genomics and biotechnology for health', contract number LSHG-CT-2003-503269.

References

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 (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

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

Batada NN, Hurst LD, Tyers M, Russell R (2006) Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol* **2:** e88

Bharadwaj R, Qi W, Yu H (2004) Identification of two novel components of the human NDC80 kinetochore complex. *J Biol Chem* **279:** 13076–13085

Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, Wu HI, McCann KE, Troyanskaya OG, Brown JM (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol Syst Biol* **2:**1

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D (1998) SGD: *Saccharomyces* genome database. *Nucleic Acids Res* **26:** 73–79

De Wulf P, McAinsh AD, Sorger PK (2003) Hierarchical assembly of the budding yeast kinetochore from multiple subcomplexes. *Genes Dev* **17:** 2902–2921

Even S (1979) *Graph Algorithms*. Potomac, MD: Computer Science Press

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre 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, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter 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 (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391

Gregson HC, Schmiesing JA, Kim JS, Kobayashi T, Zhou S, Yokomori K (2001) A potential role for human cohesin in mitotic spindle aster assembly. *J Biol Chem* **276:** 47575–47582

Gregson HC, Van Hooser AA, Ball Jr AR, Brinkley BR, Yokomori K (2002) Localization of human SMC1 protein at kinetochores. *Chromosome Res* **10:** 267–277

Gustafson AM, Snitkin ES, Parker SC, DeLisi C, Kasif S (2006) Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* **7:** 265

Hartman JLt, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. *Science* **291:** 1001–1004

Ito T, Chiba T, Yoshida M (2001) Exploring the protein interactome using comprehensive two-hybrid projects. *Trends Biotechnol* **19:** S23–S27

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41–42

Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23:** 561–566

Krause R, von Mering C, Bork P, Dandekar T (2004) Shared components of protein complexes—versatile building blocks or biochemical artefacts? *BioEssays* **26:** 1333–1343

Krogan NJ, Keogh MC, Datta N, Sawa C, Ryan OW, Ding H, Haw RA, Pootoolal J, Tong A, Canadien V, Richards DP, Wu X, Emili A, Hughes TR, Buratowski S, Greenblatt JF (2003) A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol Cell* **12:** 1565–1576

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804

Martin-Lluesma S, Stucke VM, Nigg EA (2002) Role of Hec1 in spindle checkpoint signaling and kinetochore recruitment of Mad1/Mad2. *Science* **297:** 2267–2270

McCleland ML, Gardner RD, Kallio MJ, Daum JR, Gorbsky GJ, Burke DJ, Stukenberg PT (2003) The highly conserved Ndc80 complex is required for kinetochore assembly, chromosome congression, and spindle checkpoint activity. *Genes Dev* **17:** 101–114

Newman JR, Wolf E, Kim PS (2000) A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **97:** 13203–13208

Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* **31:** 3635–3641

Ozier O, Amin N, Ideker T (2003) Global architecture of genetic interactions on the protein network. *Nat Biotechnol* **21:** 490–491

Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124:** 1069–1081

Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD (2004) A robust toolkit for functional profiling of the yeast genome. *Mol Cell* **16:** 487–496

Rives AW, Galitski T (2003) Modular organization of cellular networks. *Proc Natl Acad Sci USA* **100:** 1128–1133

Rodriguez-Navarro S, Fischer T, Luo MJ, Antunez O, Brettschneider S, Lechner J, Perez-Ortin JE, Reed R, Hurt E (2004) Sus1, a functional component of the SAGA histone acetylase complex and the nuclear pore-associated mRNA export machinery. *Cell* **116:** 75–86

Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R (2005) EXPANDER—an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6:** 232

Sharan R, Ideker T, Kelley B, Shamir R, Karp RM (2005) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol* **12:** 835–846

Shen X, Mizuguchi G, Hamiche A, Wu C (2000) A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406:** 541–544

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34:** D535–D539

Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294:** 2364–2368

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. *Science* **303:** 808–813

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA* **99:** 5860–5865

Westermann S, Cheeseman IM, Anderson S, Yates 3rd JR, Drubin DG, Barnes G (2003) Architecture of the budding yeast kinetochore reveals a conserved molecular core. *J Cell Biol* **163:** 215–222

Wigge PA, Kilmartin JV (2001) The Ndc80p complex from *Saccharomyces cerevisiae* contains conserved centromere components and has a function in chromosome segregation. *J Cell Biol* **152:** 349–360

Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* **1:**26

Zheng L, Chen Y, Lee WH (1999) Hec1p, an evolutionarily conserved coiled-coil protein, modulates chromosome segregation through interaction with SMC proteins. *Mol Cell Biol* **19:** 5417–5428

8. From E-MAPs to Module Maps: Dissecting Quantitative Genetic Interactions Using Physical Interactions

Molecular Systems Biology 4; Article number 209; doi:10.1038/msb.2008.42
Citation: ***Molecular Systems Biology*** **4:209**

www.molecularsystemsbiology.com

molecular systems biology

From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions

Igor Ulitsky[1], Tomer Shlomi[1], Martin Kupiec[2] and Ron Shamir[1,*]

[1] School of Computer Science, Tel Aviv University, Ramat Aviv, Israel and [2] Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel
* Corresponding author. School of Computer Science, Tel Aviv University, Ramat Aviv 69978, Israel. Tel.: +972 3 6405383; Fax: +972 3 6405384;
E-mail: rshamir@tau.ac.il

Received 20.3.08; accepted 28.5.08

Recent technological breakthroughs allow the quantification of hundreds of thousands of genetic interactions (GIs) in *Saccharomyces cerevisiae*. The interpretation of these data is often difficult, but it can be improved by the joint analysis of GIs along with complementary data types. Here, we describe a novel methodology that integrates genetic and physical interaction data. We use our method to identify a collection of functional modules related to chromosomal biology and to investigate the relations among them. We show how the resulting map of modules provides clues for the elucidation of function both at the level of individual genes and at the level of functional modules.
Molecular Systems Biology 15 July 2008; doi:10.1038/msb.2008.42
Subject Categories: bioinformatics; functional genomics
Keywords: data integration; gene modules; genetic interactions; protein interaction networks

Introduction

One of the central tasks of current cell biology is to reveal and understand the functional relationships between cell components. Physical interaction (PI) and genetic interaction (GI) data provide largely complementary functional information that can be used to elucidate these relationships. In particular, quantitative GIs can be a powerful source for understanding both functions of individual genes and the interplay between pathways in the cell.

GIs convey information about the phenotype of a double mutant in comparison to the phenotypes of single mutants. GIs can be crudely classified into alleviating, neutral and aggravating interactions (Segre *et al*, 2005; Beyer *et al*, 2007). In an *aggravating* interaction, the fitness of the double mutant is lower than expected given that of the single mutants. The most extreme example of an aggravating interaction is *synthetic lethality*, in which the joint deletion of two non-essential genes leads to a lethal phenotype. In an *alleviating* interaction, on the other hand, the double mutant is healthier than expected. The 'expected' fitness is usually defined using a multiplicative model, as the product of the fitnesses of the single mutants (Schuldiner *et al*, 2005; Segre *et al*, 2005; St Onge *et al*, 2007). High-throughput mapping of aggravating interactions, in particular synthetic lethality, has first been performed in *Saccharomyces cerevisiae* using the SGA (Tong *et al*, 2004) and dSLAM (Pan *et al*, 2006) methods. Recently, the exploration of GI data was pushed forward by the development of the Epistatic MiniArray (E-MAP) technology, building on SGA and allowing a quantitative estimation of both aggravating and alleviating information (Schuldiner *et al*, 2005; Collins *et al*, 2007b). The largest published E-MAP to date (Collins *et al*, 2007b) covers GIs between 743 *S. cerevisiae* genes involved in various aspects of chromosome biology (we will refer to this map as the ChromBio E-MAP). It was shown that the use of quantitative data can significantly increase the amount of information on gene function (Collins *et al*, 2007b).

The computational analysis of E-MAPs has to address several problems. First, due to technical and biological difficulties, the ChromBio E-MAP contains as many as 40% missing values. Imputation of these values is difficult, and the computational methods require the development of *ad hoc* techniques to handle missing data. Second, as the single deletion mutants are not measured in the same experiment, a multiplicative model cannot be directly fitted to the data and thus it is difficult to properly interpret every individual GI. For this reason, the insights derived from the E-MAP data were so far mostly based on correlations of GI profiles, and not on the GIs themselves (Schuldiner *et al*, 2005; Collins *et al*, 2007b; Ihmels *et al*, 2007).

The development of high-throughput GI assays has occurred in parallel to the development of methods for genome-wide mapping of protein–protein interactions (PPIs; Collins *et al*,

2007a). It was recently shown that joint analysis of GIs and PIs can shed additional light on the organization of cellular pathways. This integration is particularly appealing due to the complementarity of the two interaction types: PIs describe direct spatial association between molecules, whereas GIs refer to functional associations between genes, connecting the physical architecture to phenotypes (Beyer *et al*, 2007). The integration of genetic and physical data was used to classify GIs as occurring between or within different pathways (Kelley and Ideker, 2005). Between-pathway GIs usually indicate partial pathway redundancy, as deletion of a single gene affects only one of the pathways, while deletion of two genes from distinct pathways leads to the inactivation of both (Tucker and Fields, 2003). Accordingly, it was found that most aggravating interactions occur between pathways (Kelley and Ideker, 2005). Zhang *et al* (2005) mapped pairs of complexes with many aggravating GIs between them. We have previously extended the analysis of between-pathway explanations for GIs and shown that further physical evidence can shed light on additional properties of such pathway pairs (Ulitsky and Shamir, 2007b). However, within-pathway aggravating interactions also exist: mutations in one of the two subunits of the same complex may have only a mild phenotype, as long as the complex survives. However, deletion of both subunits may lead to a complex failure and to an aggravating phenotype. On the other hand, alleviating interactions were shown to occur mostly within pathways (Collins *et al*, 2007b). These are the result of a drastic effect of any of the single deletions on pathway activity, which abolishes the effects of additional deletions.

In this study, we propose a novel methodology for integrating GI and PI data. While extant methods (Kelley and Ideker, 2005; Ulitsky and Shamir, 2007b) have used GI data to characterize a single pathway or a pathway pair at a time, we propose a method for analyzing all the available data together and producing a set of modules identified in the data, alongside the module pairs that exhibit significant complementarity, as evidenced by the presence of multiple aggravating GIs (Figure 1). Our method can be viewed as a clustering algorithm that explicitly addresses the relation between each pair of modules (which can be complementary or unrelated). By extracting a collection of related modules, rather than a set of module pairs as in Ulitsky and Shamir (2007b), we are able to identify weaker signals in the data and extract a consistent set of modules. Similar ideas have been successfully used by Segre *et al* (2005) for *in silico* analysis of GIs.

Previous studies analyzed E-MAP data primarily using hierarchical clustering, and successfully recovered known and novel pathways and complexes (Schuldiner *et al*, 2005; Collins *et al*, 2007b). Our method has several advantages over hierarchical clustering: (a) it readily provides the pairs of modules exhibiting complementarity; (b) it produces a set of disjoint modules corresponding to putative pathways, rather than a tree; (c) the number of modules is determined by the algorithm and does not have to be determined by the user and (d) hierarchical clustering considers only similarity between pairs of gene profiles. By considering GIs between module pairs in addition to the gene similarity, our method can pick up modules based on a consistent module-wise GI pattern, even if gene profile similarity is relatively weak, e.g. due to missing

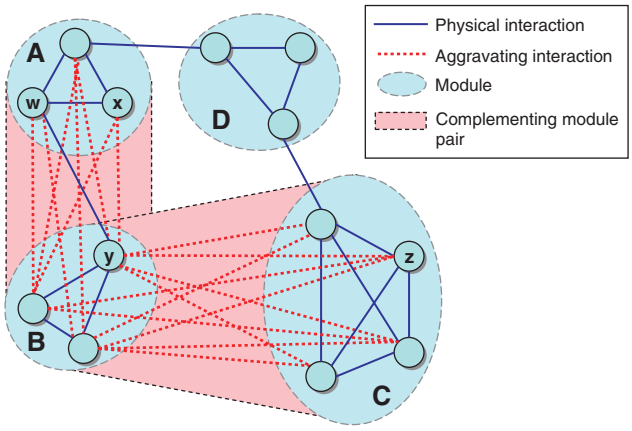


Figure 1 Toy example of a modular partition. The genes are partitioned into four modules. Each module induces a connected component in the PI network. Modules A and B have multiple aggravating GIs between them and are thus designated as a CMP. The same is true for modules B and C. Module D is not involved in any CMP. Genes w, x are siblings; genes w, y are cousins; genes y, z are cousins; genes x, z are strangers.

values. As we shall show, these theoretical advantages indeed yield practical advantage, as we are able to identify important module relations that cannot be identified using gene similarity alone.

We applied our method to the ChromBio E-MAP and obtained a collection of modules as well as a map of related module pairs. In particular, we provided the first comprehensive map of the relationships among ChromBio modules, which could not be obtained by prior means. The results improve over extant methods in terms of the functional enrichment of the obtained modules. Using a collection of single-deletion phenotypes we found that although the modules are based on GIs measured in rich medium, they remain cohesive functional units under other conditions, emphasizing the power of the E-MAP coupled with our methodology in recovering functional modules. We showed that the module map can be utilized for function prediction on several levels: to suggest with high confidence novel functions for individual genes, to identify novel functions of complete modules and to highlight interplay between modules. In particular, we provided genetic and physical evidence for (1) a new role for the nuclear pore in the mitotic spindle checkpoint; (2) a new role for proteolysis in mitosis and (3) an interplay between the THO complex and deubiquitination.

Results

A novel methodology for partitioning E-MAPs into functional module

We developed four methods for partitioning of E-MAPs into functional modules and identifying complementing module pairs (CMPs). The methods are described in detail in Materials and methods. The methods use models that differ in the way they treat inter-module GIs and in their use of PIs. There are two basic models, 'Alleviating' and 'Correlated'. Both prefer partitions in which GIs between CMP modules are mostly aggravating. The Alleviating model scores highly partitions in

which intra-module GIs are mostly alleviating. The Correlated model scores highly partitions in which the correlation between GI profiles are high within each module. The 'Connected' variants of the two basic models, termed 'AlleviatingConnected' and 'CorrelatedConnected', also require that each module induce a connected component in the PI network.

Analysis of the ChromBio E-MAP and comparison with other methods

We analyzed the E-MAP of GIs among 743 *S. cerevisiae* genes involved in chromosome biology (the ChromBio E-MAP; Collins *et al*, 2007b) alongside a network containing 2061 PIs between the genes contained in the E-MAP. The PIs were taken from SGD and BioGrid databases (Cherry *et al*, 1998; Stark *et al*, 2006) (Supplementary information). We excluded yeast two-hybrid interactions from the analysis as we found that this improved the results (results not shown).

We compared the results obtained under each of our four formulations and of other methods for extracting modules from these data types: hierarchical clustering of the GI profiles, clustering of the GI profiles using Markov clustering (MCL; Enright *et al*, 2002), clustering of the PI network using MCL and previous methods for combining binary GI and PI data (Kelley and Ideker, 2005; Ulitsky and Shamir, 2007b). MCL was chosen for clustering PI data as it was recently shown to outperform other alternatives for this task (Brohee and van Helden, 2006). Different parameter values were tested for MCL and hierarchical clustering (see Materials and methods). Results were measured in terms of the enrichment for (a) GO 'biological process' annotations, (b) MIPS complexes and (c) genes with similar phenotype (taken from SGD; Cherry *et al*, 1998). In all cases, we considered all the annotations that contained at least two genes in the ChromBio E-MAP (see Supplementary information for annotation lists). Statistics on the modules found by each method are given in Table I. The fraction of annotations enriched in at least one module (which we refer to as 'recall') and fraction of modules enriched with at least one annotation (which we refer to as 'precision') are shown in Figure 2.

We summarized recall and sensitivity using the *F*-measure (Van Rijsbergen, 1979), which is the weighted harmonic mean of precision and recall: $F=2\cdot(\text{precision}\cdot\text{recall})/(\text{precision}+\text{recall})$. The *F*-measures of the different methods are listed in Table I. It is evident that both 'Correlated' variants usually outperform the corresponding 'Alleviating' variants. An inspection of well-characterized yeast complexes (Supplementary Figure 2) reveals the reason for this superiority. Except for a few complexes (e.g., prefoldin and SWR1), pairs of genes within the same complex generally do not exhibit strong alleviating GIs. We found many cases in which the *S*-scores between members of the same complex were missing (e.g. in the mediator complex), neutral or aggravating (e.g., in the SAGA complex). Our results thus indicate that although positive *S*-scores (corresponding to alleviating GIs) do, to some extent, enable extraction of functional modules, correlations of *S*-score profiles are more helpful for this task.

As expected, it is also evident that using information on the PI network allows for a more biologically meaningful solution, as the 'CorrelatedConnected' formulation usually outperforms the 'Correlated' one (an exception is the phenotype analysis, where connectivity seems to worsen the results, see also Supplementary Figure 4). When considering all three benchmarks together, using GIs together with PIs improves upon using the PI data alone for module identification, as evident by higher *F*-measures of our methods when compared to MCL clustering of the PI network.

Table I Comparison of the modules found by different methods

Algorithm	Reference	Number of modules	Genes in modules	*F*-measure		
				GO biological process	MIPS complexes	SGD phenotypes
CorrelatedConnected	This study	62	313	**0.629**	**0.496**	0.233
AlleviatingConnected	This study	29	182	0.389	0.423	**0.276**
Connected	This study	53	446	0.420	0.316	0.262
Alleviating	This study	54	457	0.257	0.213	0.187
US	Ulitsky and Shamir (2007b)	46	229	0.559	0.381	0.188
KI	Kelley and Ideker (2005)	98	305	0.602	0.468	0.167
MCL:PPI $I=1.2$	Enright *et al* (2002)	22	597	0.397	0.202	0.113
MCL:PPI $I=2$	Enright *et al* (2002)	116	585	0.620	0.425	0.117
MCL:PPI $I=3$	Enright *et al* (2002)	154	552	0.574	0.333	0.114
MCL:PPI $I=4$	Enright *et al* (2002)	161	517	0.553	0.292	0.078
MCL:PPI $I=5$	Enright *et al* (2002)	158	477	0.528	0.259	0.082
MCL:E-MAP $I=3$	Enright *et al* (2002)	3	754	0.179	0.065	0.220
MCL:E-MAP $I=5$	Enright *et al* (2002)	10	750	0.326	0.211	0.249
MCL:E-MAP $I=7$	Enright *et al* (2002)	21	735	0.381	0.330	0.225
MCL:E-MAP $I=9$	Enright *et al* (2002)	33	690	0.425	0.284	0.196
MCL:E-MAP $I=11$	Enright *et al* (2002)	40	654	0.378	0.267	0.170
Hierarchical $t=0.2$	Collins *et al* (2007b)	110	736	0.407	0.212	0.210
Hierarchical $t=0.3$	Collins *et al* (2007b)	124	567	0.508	0.271	0.198
Hierarchical $t=0.4$	Collins *et al* (2007b)	90	384	0.547	0.314	0.209
Hierarchical $t=0.5$	Collins *et al* (2007b)	78	269	0.526	0.250	0.209
Hierarchical $t=0.6$	Collins *et al* (2007b)	52	167	0.429	0.198	0.105
Hierarchical $t=0.7$	Collins *et al* (2007b)	29	92	0.337	0.169	0.138

Only modules with at least two genes are considered. The highest *F*-measure (see Results) in each column is in bold. In MCL clustering, the *I* parameter is the 'inflation' parameter of the algorithm. In hierarchical clustering, the *t* parameter is the threshold used to extract modules from the clustering tree (see Materials and methods).

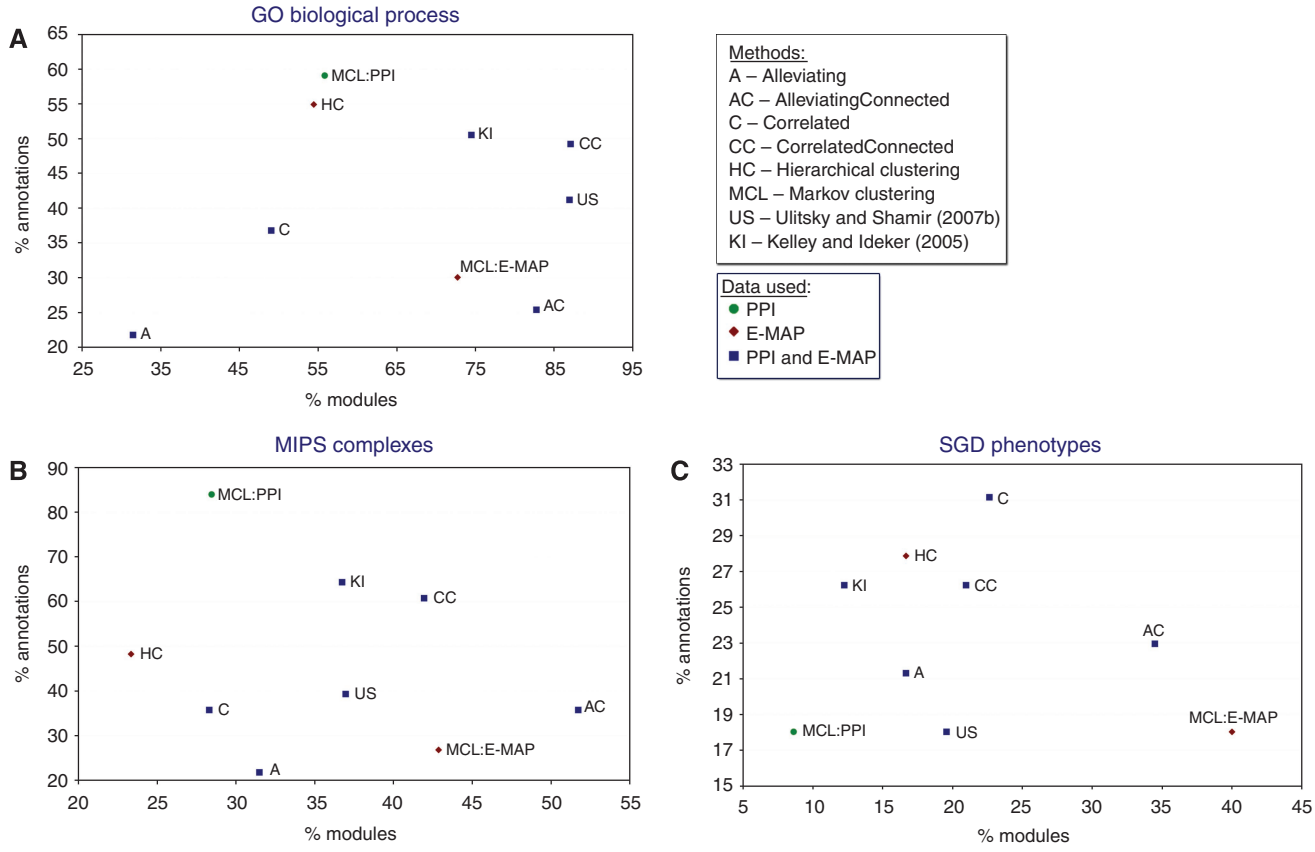


Figure 2 Comparison of the functional coherence of modules found by different methods. Only modules with at least two genes and categories with at least two genes in the E-MAP were considered. The methods are compared in terms of the fraction of annotations enriched with $P<0.001$ in at least one module and the fraction of modules enriched with $P<0.001$. 'US' is an implementation of a method that is similar to *CorrelatedConnected*, but looks for a single CMP pair at a time (Ulitsky and Shamir, 2007b). 'KI' is an implementation of the method proposed by Kelley and Ideker (2005) where edges in the GI graph appear between any pair of genes with an *S*-score below −3. MCL clustering of the PPI network and of the E-MAP correlations was executed with different parameters (see Table I). For clarity, only the execution with the highest *F*-measure is shown. Different symbols represent different data sources. All the methods were applied to the same E-MAP and PI data sets.

A comparison of the methods thus reveals that the 'CorrelatedConnected' formulation outperforms other alternatives. We therefore used the results of the CorrelatedConnected formulation (Figure 3) in all subsequent analysis. Figure 3 presents a 'heatmap' of the solution focusing on intra-module and inter-complementing module pairs (CMP) interactions. An alternative presentation showing all interactions is shown in Supplementary Figure 3. A searchable interface to the module collection obtained using this method is available at http://acgt.cs.tau.ac.il/emap/chromBio/.

Functional characterization of the modules

When correcting for multiple testing using TANGO (Shamir *et al*, 2005), we found that 27 out of 62 modules were significantly enriched ($P<0.05$) for GO 'biological process' and 32 were enriched for a GO 'cellular compartment' (looking only at subterms of 'protein complex'). Together, 45 modules (72.5%) were enriched with a known annotation. Manual inspection of the remaining 17 modules revealed that 11 of them in fact match known complexes, which are not annotated in GO. A full listing of the modules and their functions appears in Supplementary information. The fact that the vast majority of the modules (56 out of 62) correspond to known protein complexes demonstrates the ability of our approach to identify functionally cohesive units. In addition, as we show below, it appears that the main power of the modular approach is in identifying novel protein functions.

Protein function prediction

As our method can extract functionally coherent modules, it can reveal novel gene functions through 'guilt by association'. When a module is significantly enriched with a function, unannotated genes in the module can be predicted to have the same function. Using cross-validation (see Materials and methods), we estimate that this method can predict the correct function for a protein in 161 out of 204 (78.9%) of the cases. This figure is likely to be an underestimate of the specificity of our method, as even for some of the most studied proteins not all the functions are known. After manual evaluation of the obtained modules, we identified several cases where our predictions had some support from other experimental evidence:

- Gbp2 is a poly(A+) RNA-binding protein, involved in the export of mRNAs from the nucleus to the cytoplasm. It shares a module together with four members of the NuA4

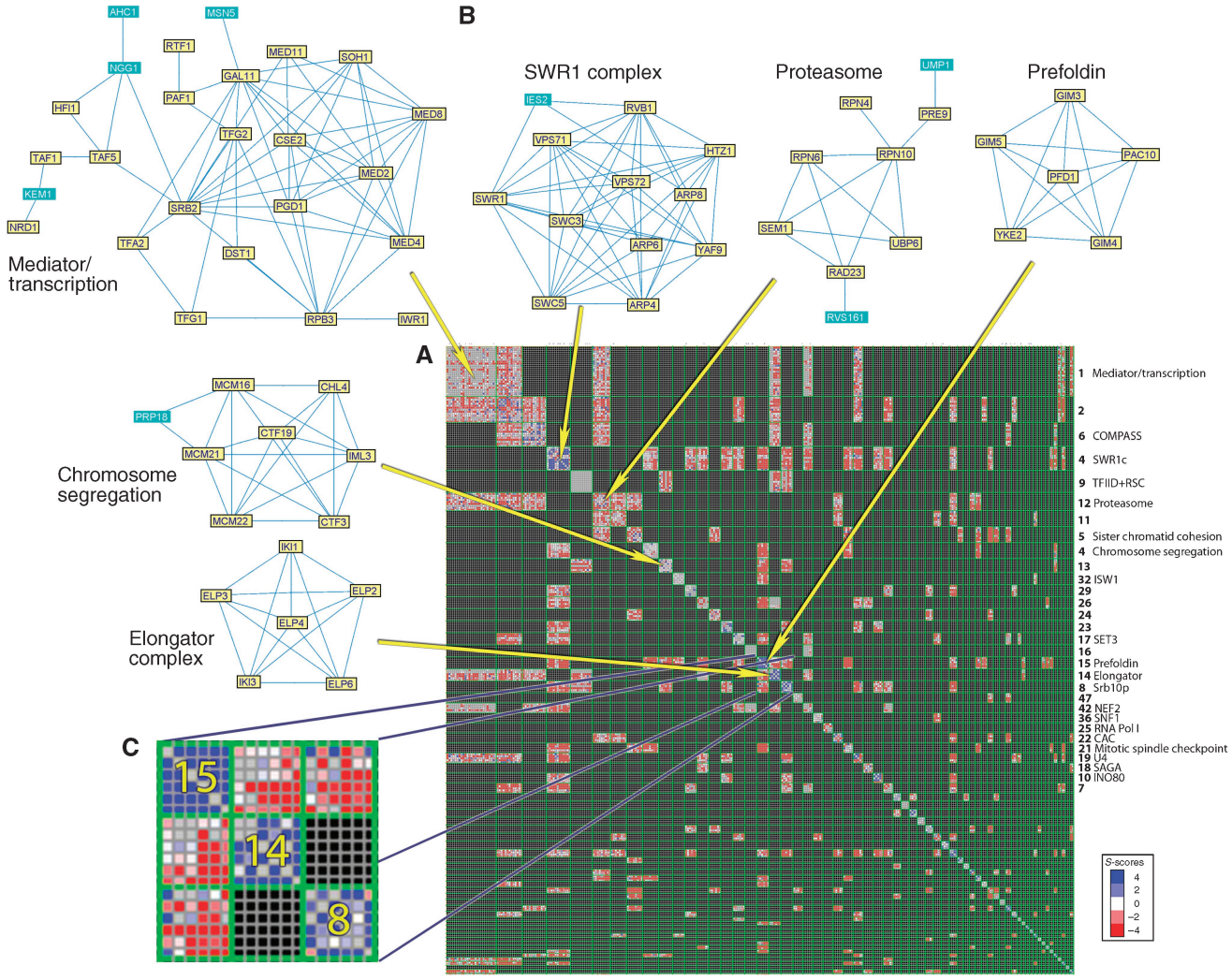


Figure 3 A summary map of the modules found in the ChromBio E-MAP. (**A**) The heat map shows the ChromBio *S*-scores between genes appearing in modules with at least two genes found using the CorrelatedConnected method. Rows and columns correspond to genes, ordered so that genes in the same module appear consecutively. Selected module names and functions are listed on the right. Green lines separate modules. Modules are sorted by their size. To facilitate recognition of CMPs, rectangles corresponding to NMPs are drawn in black. In rectangles corresponding to CMPs, the *S*-scores are colored by scale (blue—alleviating, red—aggravating, white—neutral). Missing values are in gray. (**B**) Module examples. The node labels correspond both to the gene and to the protein and therefore capitalised. Edges correspond to protein–protein interactions. In each module, the genes having the GO annotation most enriched in the module are in yellow. Module networks were drawn using MATISSE (Ulitsky and Shamir, 2007a). (**C**) A blowup of the submatrix showing the *S*-scores between genes in modules 15, 14 and 8. Modules 15 and 8 form a CMP; Modules 15 and 14 form a CMP; modules 14 and 8 do not.

histone acetyltransferase complex, as well as with a histone methyltransferase (Set2) and Rco1, part of the Rpd3S histone deacetylase complex (Figure 4A). Evidence for co-transcriptional processing of RNA has accumulated in the recent years, and it is becoming clear that RNA expression, stability and export from the nucleus are tightly regulated (Keene, 2007). Indeed, ChIP experiments have shown that Gbp2 is localized to the promoters of actively transcribed genes (Hurt *et al*, 2004). We thus propose that the interaction between Gbp2 and chromatin remodelers plays a role in the coupling of transcription with mRNA export.

- *YDL176W* is a non-essential gene of unknown function, which appears in module 17, together with five genes involved in the ubiquitination of fructose-1,6-bisphosphatase (FBPase), as part of the gluconeogenesis pathway (Figure 4B). Indeed, a structure-based study has recently suggested that this protein is involved in glycolysis and gluconeogenesis (Ferre and King, 2006). The fact that our method suggests the same function, using a completely different methodology and data, further supports the conjecture that *YDL176W* is involved in gluconeogenesis. The five genes in module 17 with a known role in FBPase degradation were identified using a genome-wide reverse genetics screen (Regelmann *et al*, 2003). We suggest that analysis of the stability of an FBPase-β-galacosidase fusion in strains deleted for *YDL176W* can be carried out to further analyze its function.
- Module 25 contains *YTA7* (*YGR270W*), an ATPase of unknown function, alongside five genes involved in chromatin silencing at the telomeres and other heterochromatic regions (Figure 4C). Indeed, it has been found that mutations in *YTA7* lead to shortened telomeres (Askree

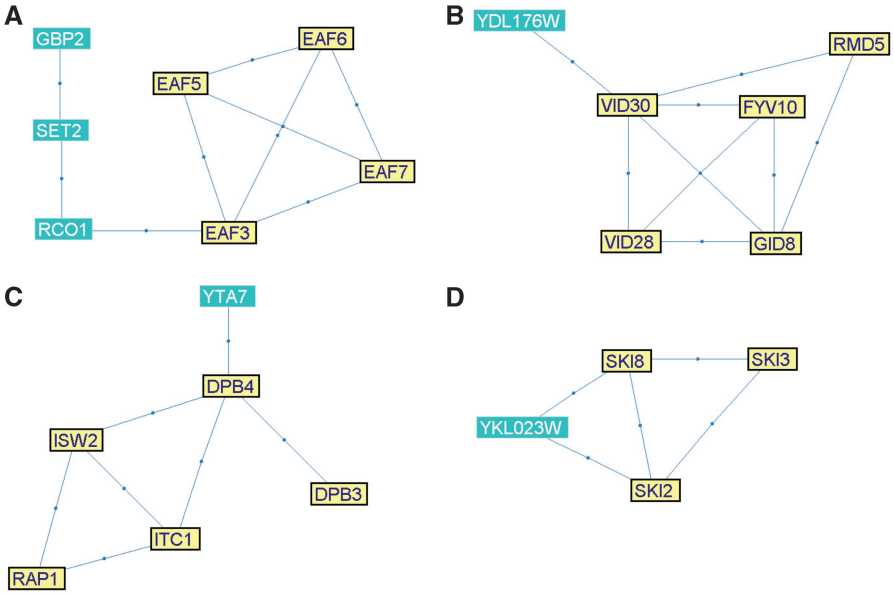


Figure 4 Modules with proposed novel protein functional annotations. Edges correspond to PIs. In each module, genes associated with the main annotation are drawn in yellow and with a thick border. (**A**) Module 14. The highlighted (yellow) genes belong to the NuA4 histone acetyltransferase complex. (**B**) Module 17. Genes associated with gluconeogenesis are highlighted. (**C**) Module 25. Genes associated with chromatin silencing at the telomere are highlighted. (**D**) Module 27. SKI complex genes are highlighted.

et al, 2004). In addition, *YTA7* was recently shown to be required for preventing the spreading of silencing beyond the heterochromatic *HMR* locus (Jambunathan *et al*, 2005). A better characterization of its role will require genomic location studies to characterize its genomic distribution (Ren *et al*, 2000).

- Module 27 contains YKL023W, a protein of unknown function, together with three known members of the SKI complex (Ski2, Ski3 and Ski8; Figure 4D). The SKI complex is involved in exosome-mediated 3′–5′ mRNA degradation and the inhibition of translation of non-poly(A) mRNAs. *YKL023W* was shown to physically interact with a fragment of Nmd2, involved in nonsense-mediated mRNA decay (He *et al*, 1997). We thus suggest that *YKL023W* is involved in mRNA degradation. Further insights into this role will require characterization of some RNA forms processed by the exosome, such as U4 snRNA (van Hoof *et al*, 2000), in a strain deleted for *YKL023W*.

Phenotype analysis

Our algorithm partitions the genes into modules based on GIs and PIs, both of which are usually measured in rich medium. We tested the similarity between the phenotypes exhibited by mutants of genes in the same module in other growth conditions. To this end, we used data from the high-throughput phenotype profiling performed by Brown *et al* (2006). We defined *phenotypic similarity* as the Pearson correlation between the phenotypic profiles of the mutants. We found that genes within the same module tended to exhibit phenotypic similarity far greater than expected at random (average r=0.424, P<0.01). Examples of highly coherent modules include the modules 50 ('Postreplication DNA repair', the genes are required for survival following treatment with DNA-damaging factors such as UV, IR, cisplastin and oxaloplatin), 20 ('HIR', a strong phenotype in environments with a high or low pH and high salt) and 14 ('Elongator', a strong phenotype after treatments with antimycin, benomyl, idarubicin and in elevated pH and salinity). The full list appears in Supplementary information.

We also examined the phenotypic similarity in CMPs. The average phenotypic similarity between genes in different modules that constitute a CMP was 0.156, as opposed to 0.087 between non-complementary module pairs (P<0.001). Interestingly, we also observed several CMPs with very dissimilar phenotypic profiles. The most dissimilar pair (r=−0.25) was formed by modules 49 and 18 ('SAGA'; Supplementary Figure 5). Both modules contain deubiquitination complexes, and in particular the ubiquitin-specific proteases Ubp3 and Ubp8. In this case, the negative correlation probably results from the combination of largely different specificity of the proteases (Zhang, 2003), and partial functional buffering, reflected in the aggravating GIs between the modules.

A map of modules and their relations

One of the merits of our approach is its ability to identify, on top of the modular decomposition, complementarity between modules. We identified 153 CMPs in the ChromBio E-MAP. A map of the modules we identified in the ChromBio E-MAP and their relationships is shown in Figure 5. We used the various annotations and, where possible, manually assigned module names, which are used below (listed in Supplementary information). Coarse-grained annotation of the module map into main cellular processes (Figure 5) reveals a complex picture of interplay between modules, indicating the pleio-

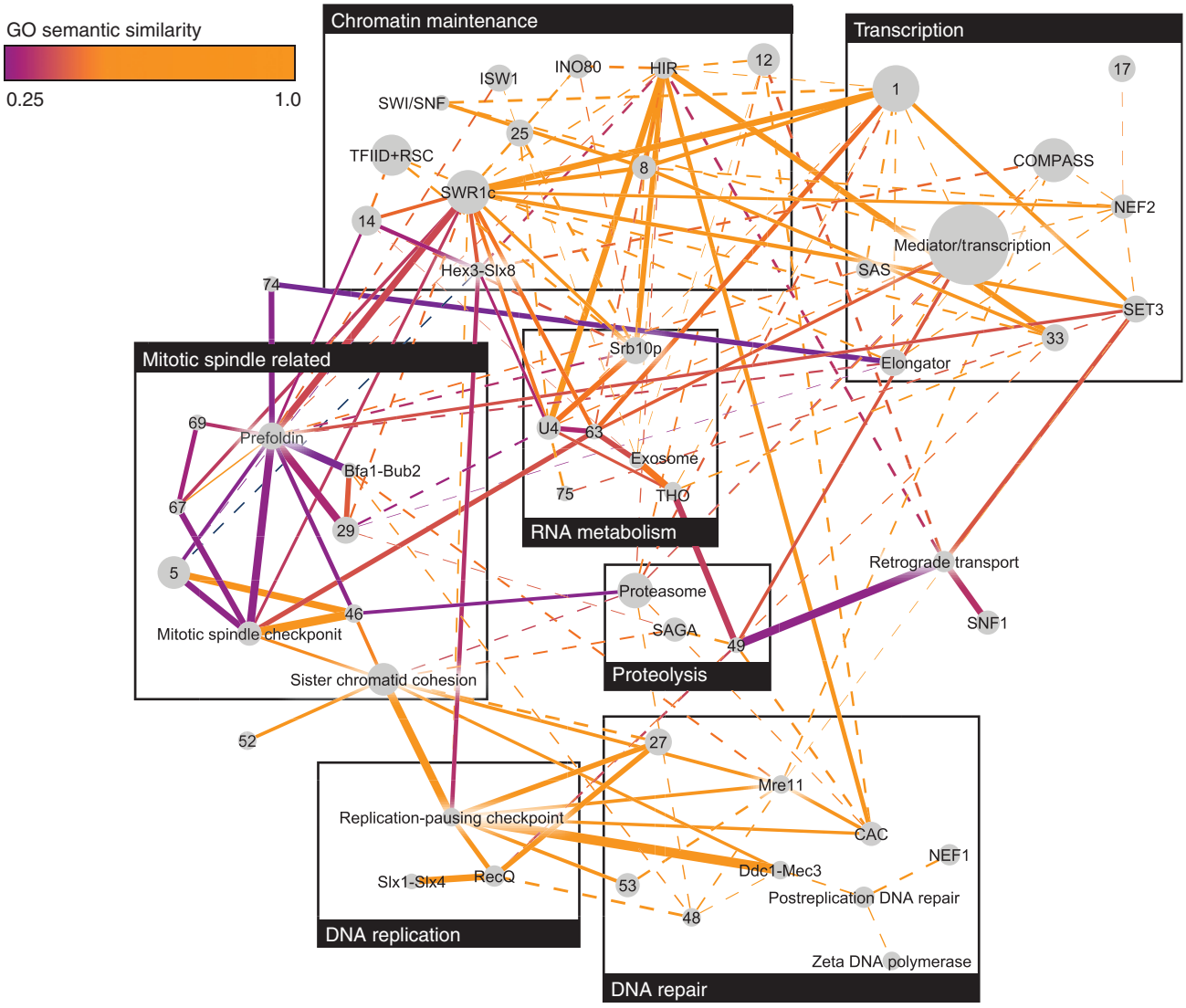


Figure 5 Modules identified in the ChromBio E-MAP and relationships among them. Every node in the network represents a module. Node radius is proportional to the module's size. Node labels are the module number or its primary annotation. Edges connect pairs of modules that form a CMP. The edge width is inversely proportional to the average *S*-score between the two modules in the CMP: thicker edges correspond to stronger aggravating GIs, dashed edges correspond to weak aggravating GIs ($-3 \leqslant$ *S*-score $\leqslant 0$). Edge color is proportional to the GO semantic similarity (Lord *et al*, 2003) between cousins in the CMP. Figure was produced using Cytoscape (Shannon *et al*, 2003).

tropy of the genes involved in chromosome biology. Evidently, most CMPs are formed by modules annotated by similar biological processes (Figure 5). In addition, a large number of CMPs link transcription with chromatin modification and DNA repair with DNA replication. Using GO semantic similarity (Lord *et al*, 2003), we found a significant negative correlation between the average *S*-scores and the functional similarity over all module pairs (Spearman correlation $\rho = -0.105$, $P = 7.38 \times 10^{-6}$). Importantly, this correlation was much higher than the correlation between functional similarity and *S*-scores for individual gene pairs ($\rho = -0.023$). This suggests that redundancy is manifested more strongly at the level of the functional unit, i.e. the module, than on the level of individual genes. We provide several examples of how CMPs formed by seemingly functionally unrelated modules can lead to biological insight. Note that these relationships could not be identified by methods using solely *S*-score profile similarity, as in all cases the similarity between the *S*-score profiles of genes from different modules was close to 0 (Figure 6).

The role of nuclear pore in the mitotic spindle checkpoint

An interesting CMP linking seemingly unrelated processes consists of modules 21 ('mitotic spindle checkpoint') and 63 (Figure 6A). Module 63 contains two genes: *SAC3* and *THP1*, both associated with the nuclear pore, with roles in transcription regulation and mRNA export. Some evidence of a relationship between the nuclear pore and the mitotic spindle checkpoint can be found in the literature. The spindle

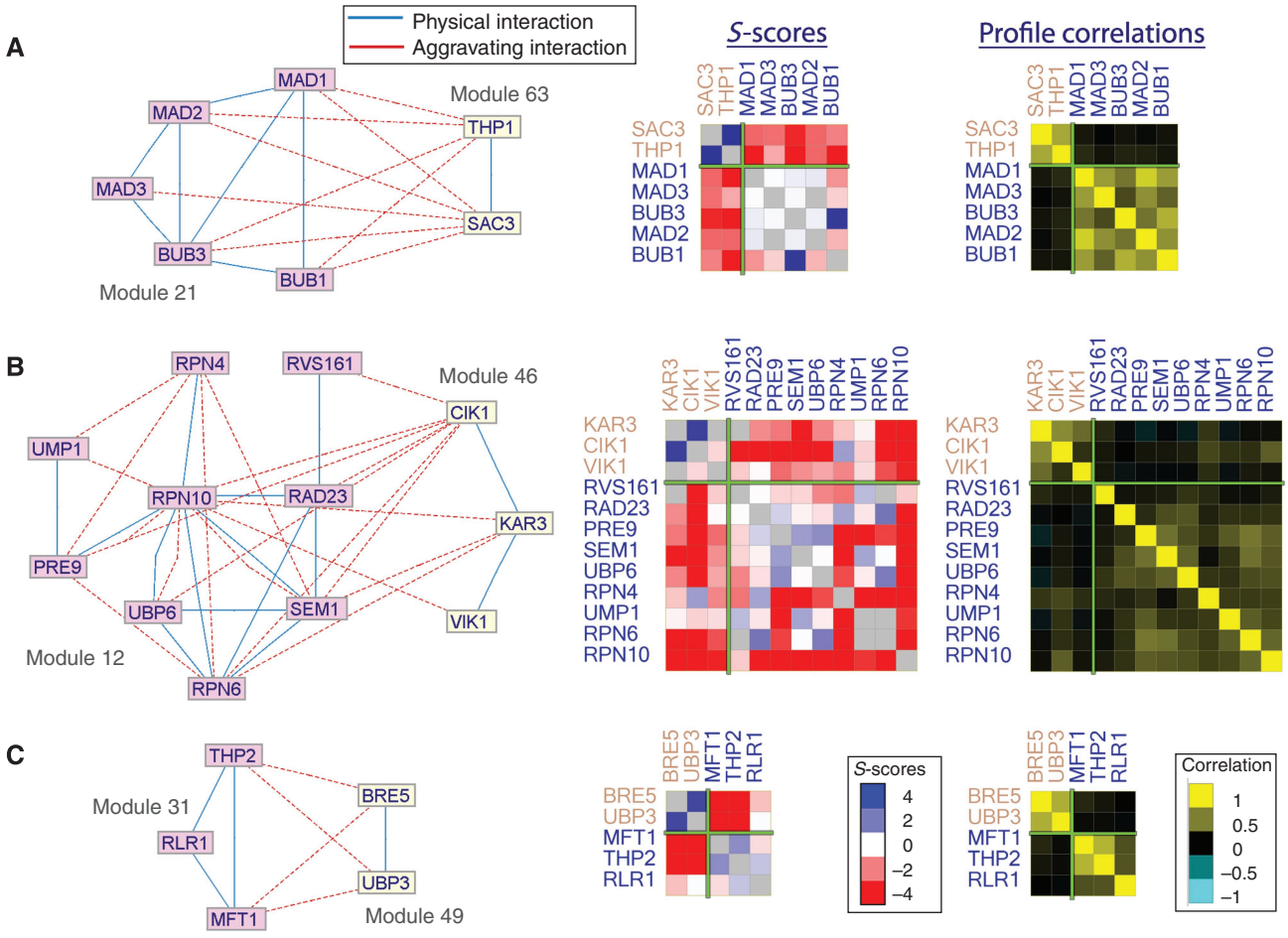


Figure 6 CMP examples. In each example, on the left the two subnetworks forming the pair are shown in different colors. In the middle, the *S*-scores between the genes in the CMP are color-coded. Blue rectangles correspond to alleviating GIs and red rectangles correspond to aggravating GIs. On the right, the correlations between the *S*-score profiles of genes in the CMP are color-coded. Green lines separate the modules.

checkpoint proteins Mad1 and Mad2 (both part of the module 21) were shown to reside predominantly at the nuclear pore throughout the cell cycle (Iouk *et al*, 2002). Several components of the nuclear pore complex (such as Nup170) are specifically associated with chromosome segregation (Iouk *et al*, 2002; Scott *et al*, 2005). Furthermore, Mad1 has a role in transport of specific proteins, such as Pho4, through the nuclear pore (Iouk *et al*, 2002). A role for nuclear pore complexes in the spindle assembly was also shown in higher eukaryotes (Orjalo *et al*, 2006). However, we found no reports of this novel relationship between the Sac3-Thp1 complex and the mitotic spindle checkpoint proteins. *sac3* deletion mutants accumulate in mitosis as large budded cells with extended microtubules, and have an increased rate of chromosome loss compared to wild-type strains (Bauer and Kolling, 1996). As evident in Figure 5, the genes in both modules exhibit GIs with several other modules, and thus the specific elucidation of the connection between Sac3-Thp1 and the mitotic spindle checkpoint would have been very difficult without a focused module map such as the one presented here. Moreover, this connection could not be picked up using *S*-score correlations alone, as the smallest hierarchical clustering subtree that contained the genes in modules 21 and 63 consisted of 231 genes.

The role of the proteasome in mitosis

Another CMP that crosses process boundaries and connects seemingly unrelated modules links module 12 ('Proteasome') with module 46 (Figure 6B). Module 46 contains three proteins (Kar3, Cik1 and Vik1) involved in microtubule-related processes in mitosis and meiosis. Kar3 is a kinesin-14 protein that forms heterodimers with both Cik1 and Vik3 and acts as a motor to pull chromosomes apart. The proteasome (the complex in charge of most protein degradation in the cell) is known to affect progression through cell cycle (Gordon and Roof, 2001; May and Hardwick, 2006). Inspection of single-deletion phenotypes reveals that mutants of genes from module 12 (in particular Rpn10, Sem1 and Ump1) show relative benomyl resistance (Brown *et al*, 2006). Benomyl is an antimitotic drug that destabilizes microtubules and inhibits microtubule-mediated processes, including nuclear division, nuclear migration and nuclear fusion (Hampsey, 1997). The fact that we observe particularly strong aggravating GIs

between the proteasome and the three members of module 46 suggests another link between proteolysis and the mitotic spindle, involving the Kar3 kinesin. One possible explanation for this relation is that alternative kinesin motors are prevented from functioning by a protein(s) that is a substrate for proteasomal degradation. Thus, lack of proteasome activity is genetically equivalent to lack of the alternative motor, exhibiting strong aggravating GIs. A similar parallel pathway is the one that restricts the activity of the alternative kinesin motors Cin8 and Kip1 by CDK-mediated proteasomal degradation (Crasta *et al*, 2006).

Deubiqutination and the THO complex

Module 49 contains Bre5 and Ubp3, which together form a deubiquitination complex with known roles in regulating vesicle traffic (Cohen *et al*, 2003), transcriptional regulation through TFIID (Auty *et al*, 2004) and DNA damage (Bilsland *et al*, 2007). These roles closely correspond to the CMPs that include module 49 (Figure 5). Our map shows a strong GI between this module and module 31, which contains three proteins from the THO complex, involved in transcription elongation and its coupling to mRNA export (Figure 6C). Our analysis thus uncovers a coordinated activity of the Bre5-Ubp3 deubiquitination and the THO complexes, most likely during transcription elongation. Such coordination might be required to prevent DNA damage from occurring during transcription; indeed, mutations in members of either complex result in increased sensitivity to DNA-damaging agents and hyper-recombination (Bilsland *et al*, 2007; Garcia-Rubio *et al*, 2008). In addition, recent experiments demonstrate a new role for the THO complex in transcription-coupled DNA damage repair (Gaillard *et al*, 2007). A connection was found between THO complex activity during transcription, and an alternative DNA repair pathway involving ubiquitin-mediated inactivation of RNA polymerase II (Somesh *et al*, 2005). On the basis of our results, we propose that under specific circumstances, deubiquitination of RNA polymerase II by the Bre5-Ubp3 complex may allow resumption of transcription.

Discussion

Analysis of GI data is an important challenge in computational biology. It was previously demonstrated that integrated analysis of GIs and PIs is a powerful approach for outlining pathways and for identifying pairs of complementing pathways (Kelley and Ideker, 2005; Ulitsky and Shamir, 2007b). Here, we have shown how this integration can be extended in two important directions. First, we handle a richer source of GI data, provided by the E-MAP technology. Second, we describe an algorithmic approach that is capable of extracting a comprehensive map of multiple modules along with their relationships, rather than focusing on a single module or on a module pair. This approach is capable of identifying significant modules that exhibit weak but consistent GIs.

As our formulation of the module-finding problem is computationally hard, we use an efficient greedy heuristic for finding high-scoring partitions. As a very large percentage of the modules we identify correspond to known complexes or pathways, it is evident that this heuristic performs quite well in detecting functional modules. However, as a local search algorithm, our algorithm may converge to a local minimum. More precise algorithms for the problem could further improve the results. Addition of an ability to assign confidence to individual predictions is also expected to boost the applicability of our method. In the PPI network used in this study, we chose to exclude yeast two-hybrid interactions as we found that this improved the results. However, this exclusion may bias our current results toward detection of protein complexes. PI confidence schemes (Qi *et al*, 2006; Suthram *et al*, 2006) should be helpful for a better incorporation of all possible interaction evidence into our framework.

The terminology of a 'module' is frequently used in different settings in systems biology (Hartwell *et al*, 1999). On some level, the entire collection of genes tested in the ChromBio E-MAP can be considered a module, as they were all selected based on their role in chromosome biology. Some methods for analysis of GI data (e.g. Segre *et al*, 2005; Collins *et al*, 2007b) produce a hierarchical collection of modules. This approach has some advantages as description of biological processes is inherently hierarchical (e.g., different chromatin remodeling complexes form a 'chromatin remodeling' module). However, systematic prediction of gene function and module function is more difficult in this setting. A hierarchical tree for the ChromBio E-MAP encompasses hundreds of highly overlapping modules. Here, we use PI data in an attempt to identify distinct modules of genes acting cooperatively in the cell, which can be used for systematic prediction.

We compared two methods for scoring gene similarity: one based on alleviating interactions and another based on similarity of GI profiles across the entire E-MAP. Our results indicate that the use of profile similarity is generally superior when analyzing the ChromBio E-MAP. A recent study by Bandyopadhyay *et al* (2008), which was published while this article was in revision, used a combination of PIs and GIs, and found that modules enriched with aggravating interactions are also of interest, as they frequently correspond to essential complexes. It was also suggested that pairs of pathways could exhibit multiple alleviating interactions between them in some cases (Segre *et al*, 2005). Therefore, further research on alternative scoring schemes may reveal other types of interactions within functional modules.

The main contribution of our approach to the analysis of E-MAP data is in our ability to identify not only the modules in the data but also the relationships among them. As we illustrate above, analysis of the data in light of the CMP relationship is a powerful tool for improving our understanding of the roles played by the modules.

Materials and methods

Problem formulation and the probabilistic model

We are given a PI network $G=(V, E)$ and a matrix of GI scores S (which we denote S-scores as in Collins *et al*, 2006). We are interested in obtaining a partition of the network nodes into subsets $M=\{M_1, \ldots, M_m, R\}$, in which each module M_i corresponds to a cohesive biological unit and R is a set of singleton genes that do not belong to modules. We distinguish between two types of module pairs: (a) module pairs exhibiting a large number of aggravating GIs, which we call *CMPs* and

(b) pairs of unrelated modules, which we call *neutral module pairs* (NMPs). We refer to a pair of genes as: (a) *siblings* if both genes are assigned to the same module; (b) *cousins* if they are assigned to two different modules that together form a CMP and (c) *strangers* otherwise (see toy examples in Figure 1). The modular decomposition we seek to score consists of the partition M alongside the set of CMPs $C=\{(M_i, M_j)\}$.

We tested four different problem formulations; the formulations differ in the way they treat within-module similarity and connectivity of a module. We denote the different formulations *Alleviating*, *AlleviatingConnected*, *Correlated* and *CorrelatedConnected*. In all formulations, we modeled the set of S-scores as coming from a mixture of three Gaussian distributions: G_m for pairs of genes with exceptionally high scores (corresponding to alleviating GIs); G_f for pairs of genes with exceptionally low scores (corresponding to aggravating GIs) and G_n for pairs with neutral S-scores. These assumptions have a theoretical justification (Sharan and Shamir, 2000), and we verified that they hold on the E-MAP data using quantile plots (see Supplementary Figure 1 and Supplementary information).

The Alleviating model

We first describe the *Alleviating* model formulation. In this variant, we looked for modules with the following properties: (a) siblings exhibit mostly alleviating GIs and (b) cousins exhibit mostly aggravating GIs. We formulate the score of a putative solution as a hypothesis-testing question. Given the partition M and the set of CMPs C, the *null* hypothesis H_0 is: M is a random partition, and the *modular* hypothesis H_1 is: M exhibits a biologically plausible modularity. Formally, in the modular hypothesis: (a) the S-scores between siblings come from G_m with a high probability β_m and from G_n otherwise; (b) the S-scores between cousins come from G_f with a high probability β_f and from G_n otherwise and (c) The S-scores between strangers come from distribution G_m with probability p_m, from G_f with probability p_f, and from G_n otherwise. Thus, the likelihood of an S-score between two genes under the module hypothesis is:

$$P(S_{ij}|H_1)=\begin{pmatrix}\beta_m P_{G_m}(S_{ij})+(1-\beta_m)P_{G_n}(S_{ij}) & \text{if} & i,j \text{ are siblings}\\ \beta_f P_{G_f}(S_{ij})+(1-\beta_f)P_{G_n}(S_{ij}) & \text{if} & i,j \text{ are cousins}\\ p_m P_{G_m}(S_{ij})+p_f P_{G_f}(S_{ij})+(1-p_m-p_f)P_{G_n}(S_{ij}) & \text{if} & i,j \text{ are strangers}\end{pmatrix}$$

Under the null hypothesis, for each gene pair, the probability that its S-score comes from distribution G_x is p_x. The probability under the null model is thus: $P(S_{ij}|H_0)=p_m P_{G_m}(S_{ij})+p_f P_{G_f}(S_{ij})+p_n P_{G_n}(S_{ij})$. By setting the *partition score* to $\log P(S|H_1)/P(S|H_0)$, we get that by maximizing this score we obtain partitions of maximum likelihood ratio. Assuming independence between gene pairs, the partition score can be decomposed over all pairs of nodes:

$$\log\frac{P(S|H_1)}{P(S|H_0)}=\log\left(\prod_{i,j}\frac{P(S_{ij}|H_1)}{P(S_{ij}|H_0)}\right)=\sum_{i,j}\log\frac{P(S_{ij}|H_1)}{P(S_{ij}|H_0)}$$

Note that if we denote

$$W_s(i,j)=\log\frac{\beta_m P_{G_m}(S_{ij})+(1-\beta_m)P_{G_n}(S_{ij})}{p_m P_{G_m}(S_{ij})+p_f P_{G_f}(S_{ij})+p_n P_{G_n}(S_{ij})}$$

and

$$W_c(i,j)=\frac{P(S_{ij}|H_1)}{P(S_{ij}|H_0)}=\frac{\beta_f P_{G_f}(S_{ij})+(1-\beta_f)P_{G_n}(S_{ij})}{p_m P_{G_m}(S_{ij})+p_f P_{G_f}(S_{ij})+p_n P_{G_n}(S_{ij})}$$

the partition score is $W_p=\sum_{i,j\in\text{siblings}}W_s(i,j)+\sum_{i,j\in\text{cousins}}W_c(i,j)$.

The Correlated model

The *Correlated* model formulation scores GIs between cousins as before, but differs in scoring GIs between siblings. Instead of scoring a pair of genes based on the single GI between them, it scores the pair based on their full GI profiles. The same score was used with hierarchical clustering in Collins *et al* (2006). Let C_{ij} denote the correlation between the GI profiles of genes i and j (which we call the C-score). We model the distribution of C-scores as a mixture of two Gaussian distributions, G_m^C for siblings and G_n^C for non-siblings (see Supplementary Figure 1 and Supplementary information). In the model hypothesis, we assume that correlations between the profiles of genes within the same module come from G_m^C with probability β_m^C and from G_m^C otherwise. The likelihood of the C-score under the module hypothesis is thus:

$$P(S_{ij}|H_1)=\begin{pmatrix}\beta_m^C P_{G_m^C}(C_{ij})+(1-\beta_m^C)P_{G_n^C}(C_{ij}) & \text{if} & i,j \text{ are siblings}\\ \beta_f P_{G_f}(S_{ij})+(1-\beta_f)P_{G_n}(S_{ij}) & \text{if} & i,j \text{ are cousins}\\ p_m P_{G_m}(S_{ij})+p_f P_{G_f}(S_{ij})+p_n P_{G_n}(S_{ij}) & \text{if} & i,j \text{ are strangers}\end{pmatrix}$$

Connectivity requirements

We tested two variants for each of the two models described above: one that used solely the E-MAP data and another in which each module was required to induce a connected subnetwork in G. We denote the latter models as *AlleviatingConnected* and *CorrelatedConnected*.

Finding high-scoring partitions

We first established that the problems we are studying are computationally hard by a reduction from the related correlation clustering problem (see Supplementary information). While several approximation algorithms for the latter problem are available (Demaine and Immorlica, 2003; Demaine *et al*, 2006), they cannot be applied directly in our setting. We thus developed a greedy heuristic for detection of high-scoring partitions. Starting from a partition in which each module contains a single node from V, we iteratively apply two update steps. In the first step, the node whose module re-assignment provided the highest score improvement is selected and re-assigned accordingly. When no such node is found, we look for pairs of modules that could be merged to improve the partition score. In the *Connected* formulations, we require that the re-assignments maintain the connectivity of all the modules. In the second step, the set of CMPs is re-computed. For every pair of modules M_i and M_j, we compute the contribution to the score of the solution if (M_i, M_j) is included in the set of CMPs: $\Sigma_{x\in M_i,\ y\in M_j}W_c(x,y)$. The pair is included in the CMP set if this contribution is significantly high (see below).

We found that the above algorithm has difficulties in finding good improving moves when starting from singleton sets. We therefore developed a two-phase approach: we first execute the greedy algorithm until convergence when using only the first step, i.e. keeping C empty. In the second phase, we execute the full algorithm as described above.

Identifying significant CMPs

To assess each candidate CMP (M_1, M_2), we evaluated the significance of the aggravating GIs between the modules given their overall GI profiles. To this end, for every gene $g_i\in M_1$, we compared the values of the W_p weights between g_i and the genes in M_2 to the entire weight profile of g_i using the Wilcoxon rank-sum test. Let us denote the significance by p_i^1. $\{p_i^1\}$ is then transformed into a single significance level using the z-transform (Stouffer's method; Hedges and Olkin, 1985). p^2 is computed in a similar way, evaluating the significance of the weights between M_1 and M_2 given the weight profiles of the genes in M_2. Finally, M_1 and M_2 are declared as CMPs if and only if $\max(p^1, p^2)<0.005$. Note that these P-values are not corrected for multiple testing due to evaluation of a large number of possible CMPs by the algorithm. Therefore, this score is a heuristic, which, as we shall show, is successful as identifying biologically meaningful CMPs.

Parameter estimation

The parameters of the Gaussian distributions (including p_m and p_f) were estimated using a standard expectation-maximization algorithm (Bilmes, 1997). In all the results reported here, we used $\beta_m=\beta_f=0.7$. We validated that the results reported here are robust to the choice of these parameters (see Supplementary information).

Hierarchical clustering analysis

Hierarchical clustering of the E-MAP data was performed using average linkage as in Collins *et al* (2007b). Pearson correlation was

used as a distance measure between pairs of GI profiles. When computing the correlation between profiles X_i and X_j, only positions in which neither profile had missing data were used. For comparison with other methods, modules were constructed using the hierarchical clustering tree, by extracting maximal subtrees in which the average correlation of the GI patterns was above a threshold t.

Assessing the reliability of function prediction

We performed cross-validation to assess the reliability of function prediction using the modular partition. The following process was repeated for each annotated gene in every module. We hid the gene's annotation and predicted it based on the annotations of the rest of the module's genes. We used the GO biological process annotation and predicted a function only if its enrichment in the module had $P<0.001$. A prediction was considered correct if the majority of the predicted biological processes were correct, and wrong otherwise. The reliability was defined as the fraction of correct predictions. All GO biological process categories with at least two genes in the E-MAP were considered. To predict a relatively narrow function, we considered only genes that shared at least one GO category with no more than 30 other genes in the E-MAP. In total, 204 genes were considered.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Roded Sharan, Eytan Ruppin, Trey Ideker and Nevan Krogan for helpful discussions regarding this study. We thank the referees of this study for many helpful comments. IU is a fellow of the Edmond J Safra Bioinformatics program at Tel-Aviv University. Research in the MK lab was supported by grants from the Israel Science Fund and the Israel Ministry of Science and Technology. RS was supported in part by the Wolfson foundation and by the Raymond and Beverly Sackler Chair for Bioinformatics at Tel Aviv University.

References

Askree SH, Yehuda T, Smolikov S, Gurevich R, Hawk J, Coker C, Krauskopf A, Kupiec M, McEachern MJ (2004) A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc Natl Acad Sci USA* **101:** 8658–8663

Auty R, Steen H, Myers LC, Persinger J, Bartholomew B, Gygi SP, Buratowski S (2004) Purification of active TFIID from *Saccharomyces cerevisiae*. Extensive promoter contacts and co-activator function. *J Biol Chem* **279:** 49973–49981

Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T (2008) Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* **4:** e1000065

Bauer A, Kolling R (1996) The SAC3 gene encodes a nuclear protein required for normal progression of mitosis. *J Cell Sci* **109** (Part 6)**:** 1575–1583

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

Bilmes JA (1997) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, Berkeley, CA, Technical Report ICSI-TR-97-021

Bilsland E, Hult M, Bell SD, Sunnerhagen P, Downs JA (2007) The Bre5/Ubp3 ubiquitin protease complex from budding yeast contributes to the cellular response to DNA damage. *DNA Repair (Amst)* **6:** 1471–1484

Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinformatics* **7:** 488

Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, Wu HI, McCann KE, Troyanskaya OG, Brown JM (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol Syst Biol* **2:** 0001

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res* **26:** 73–79

Cohen M, Stutz F, Belgareh N, Haguenauer-Tsapis R, Dargemont C (2003) Ubp3 requires a cofactor, Bre5, to specifically de-ubiquitinate the COPII protein, Sec23. *Nat Cell Biol* **5:** 661–667

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ (2007a) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteomics* **6:** 439–450

Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z *et al* (2007b) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446:** 806–810

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

Crasta K, Huang P, Morgan G, Winey M, Surana U (2006) Cdk1 regulates centrosome separation by restraining proteolysis of microtubule-associated proteins. *EMBO J* **25:** 2551–2563

Demaine ED, Emanuel D, Fiat A, Immorlica N (2006) Correlation clustering in general weighted graphs. *Theor Comput Sci* **361:** 172–187

Demaine ED, Immorlica N (2003) Correlation clustering with partial information. *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques: 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2003, and 7th International Workshop on Randomization and Approximation Techniques in Computer Science, RANDOM 2003*. Princeton, NJ, USA, 24–26 August 2003: Proceedings

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30:** 1575–1584

Ferre S, King RD (2006) Finding motifs in protein secondary structure for use in function prediction. *J Comput Biol* **13:** 719–731

Gaillard H, Wellinger RE, Aguilera A (2007) A new connection of mRNP biogenesis and export with transcription-coupled repair. *Nucleic Acids Res* **35:** 3893–3906

Garcia-Rubio M, Chavez S, Huertas P, Tous C, Jimeno S, Luna R, Aguilera A (2008) Different physiological relevance of yeast THO/TREX subunits in gene expression and genome integrity. *Mol Genet Genomics* **279:** 123–132

Gordon DM, Roof DM (2001) Degradation of the kinesin Kip1p at anaphase onset is mediated by the anaphase-promoting complex and Cdc20p. *Proc Natl Acad Sci USA* **98:** 12515–12520

Hampsey M (1997) A review of phenotypes in *Saccharomyces cerevisiae*. *Yeast* **13:** 1099–1133

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* **402:** C47–C52

He F, Brown AH, Jacobson A (1997) Upf1p, Nmd2p, and Upf3p are interacting components of the yeast nonsense-mediated mRNA decay pathway. *Mol Cell Biol* **17:** 1580–1594

Hedges LV, Olkin I (1985) *Statistical Methods for Meta-Analysis*. Orlando: Academic Press

Hurt E, Luo MJ, Rother S, Reed R, Strasser K (2004) Cotranscriptional recruitment of the serine–arginine-rich (SR)-like proteins Gbp2 and Hrb1 to nascent mRNA via the TREX complex. *Proc Natl Acad Sci USA* **101:** 1858–1862

Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Mol Syst Biol* **3:** 86

Iouk T, Kerscher O, Scott RJ, Basrai MA, Wozniak RW (2002) The yeast nuclear pore complex functionally interacts with components of the spindle assembly checkpoint. *J Cell Biol* **159:** 807–819

Jambunathan N, Martinez AW, Robert EC, Agochukwu NB, Ibos ME, Dugas SL, Donze D (2005) Multiple bromodomain genes are involved in restricting the spread of heterochromatic silencing at the *Saccharomyces cerevisiae* HMR-tRNA boundary. *Genetics* **171:** 913–922

Keene JD (2007) RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8:** 533–543

Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23:** 561–566

Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* **19:** 1275–1283

May KM, Hardwick KG (2006) The spindle checkpoint. *J Cell Sci* **119:** 4139–4142

Orjalo AV, Arnaoutov A, Shen Z, Boyarchuk Y, Zeitlin SG, Fontoura B, Briggs S, Dasso M, Forbes DJ (2006) The Nup107-160 nucleoporin complex is required for correct bipolar spindle assembly. *Mol Biol Cell* **17:** 3806–3818

Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124:** 1069–1081

Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* **63:** 490–500

Regelmann J, Schule T, Josupeit FS, Horak J, Rose M, Entian KD, Thumm M, Wolf DH (2003) Catabolite degradation of fructose-1,6-bisphosphatase in the yeast *Saccharomyces cerevisiae*: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways. *Mol Biol Cell* **14:** 1652–1663

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309

Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123:** 507–519

Scott RJ, Lusk CP, Dilworth DJ, Aitchison JD, Wozniak RW (2005) Interactions between Mad1p and the nuclear transport machinery in the yeast *Saccharomyces cerevisiae*. *Mol Biol Cell* **16:** 4362–4374

Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* **37:** 77–83

Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R (2005) EXPANDER––an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6:** 232

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504

Sharan R, Shamir R (2000) CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* **8:** 307–316

Somesh BP, Reid J, Liu WF, Sogaard TM, Erdjument-Bromage H, Tempst P, Svejstrup JQ (2005) Multiple mechanisms confining RNA polymerase II ubiquitylation to polymerases undergoing transcriptional arrest. *Cell* **121:** 913–923

St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39:** 199–206

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34:** D535–D539

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

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L *et al* (2004) Global mapping of the yeast genetic interaction network. *Science* **303:** 808–813

Tucker CL, Fields S (2003) Lethal combinations. *Nat Genet* **35:** 204–205

Ulitsky I, Shamir R (2007a) Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol* **1:** 8

Ulitsky I, Shamir R (2007b) Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks. *Mol Syst Biol* **3:** 104

van Hoof A, Lennertz P, Parker R (2000) Yeast exosome mutants accumulate 3′-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. *Mol Cell Biol* **20:** 441–452

Van Rijsbergen CJ (1979) *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann

Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* **4:** 6

Zhang Y (2003) Transcriptional regulation by histone ubiquitination and deubiquitination. *Genes Dev* **17:** 2733–2740

9. Discussion

In this thesis we described our study on molecular networks and their integration with diverse genomic data. We specifically focused on algorithms for identifying modules through joint analysis of PPI networks and either gene expression profiles or GIs. The research in this thesis integrates concepts from biology, computer science, and statistics. We approached the problems from the computer science perspective, and then analyzed real biological data to demonstrate the biological implications of our methods. Moreover, we demonstrated the advantages of our methodology over extant methods.

We repeatedly validated and developed our methods in two channels. First, we utilized publically available datasets of gene networks and gene expression. Second, we established collaborations with leading biological laboratories in Israel, United States, Germany, France, Austria and the United Kingdom, and conducted joint research that combines our computational methods and their experimental data. Of particular note is our extensive collaboration with the Loring lab at the Scripps Institute in San Diego. Two studies that were performed as part of this collaboration have been published [2, 145], two others are submitted for publication, and several others are still under way. Except for the study described in Chapter 4, these works are not included in this thesis. Our collaborations have served as the stimulus for the development of several methods, and in particular those described in Chapters 5 and 5.

In this chapter we first summarize the methods introduced in this thesis and discuss how they can be used to explore biological questions, and how their performance can be assessed. Finally, we discuss potential future directions that can stem from the results in this thesis.

9.1 Exploiting modularity in biological systems

The idea that many crucial biological processes are carried out by functional modules, rather than by individual molecules, became widely accepted in the last decade [10, 14, 120, 146]. Such modules can be predicted through computational analysis of diverse data types, and in particular by combining networks of physical interactions, such as the PPI networks that we mainly exploit here, with complementary information, such as gene expression and genetic interactions. The basic ingredient of the methods we describe here is a requirement that each module forms a connected component in the PPI network. On top of this requirement, we developed various scores based on gene expression or genetic

interactions. Optimization of these scores results in a collection of functional modules.

9.1.1 Identifying modules using PPI networks and gene expression data

In order to identify modules using PPI networks and gene expression data, we first developed the MATISSE algorithm (Chapter 2) which detects PPI subnetworks that exhibit coherent expression patterns across the entire dataset. MATISSE overcomes several drawbacks of previously suggested approaches, such as bias towards dense pathways [110] or discovery of modules that do not form connected subnetworks [107]. The MATISSE approach was further improved in Chapter 6, which introduced CEZANNE, in which the basic connectivity criterion is replaced with a requirement for confident connectivity, based on confidence values for individual interactions in the PPI network. This improvement addresses the noisy nature of current PPI networks [20, 46].

Although using Pearson correlation as a criterion for expression similarity typically results in biologically relevant modules (Chapters 2 and 6), other criteria could be more appropriate when the samples are labeled with clinical or other parameters. In this case it is frequently desirable to identify subnetworks whose expression is correlated with one of the parameters. For example, when samples represent different cell line groups (as in Chapter 3), it is desirable to extract subnetworks that are differentially expressed in specific groups. In addition to the levels of differential expression, information about co-expression can also be useful in this context, as it is possible that, e.g., several distinct pathways are up-regulated in a specific sample group. In this case, the members of each cluster are expected to be tightly co-expressed both within the sample group and outside it. Thus, our approaches described in Chapters 3 and 5 use an expression score that combines differential expression and co-expression.

Most methods for quantifying differential expression of pathways, including those we proposed in Chapters 3 and 5, share an underlying assumption that a differentially expressed module consists of genes all of which are significantly differentially expressed. This assumption can be too restrictive in human disease studies, as mounting evidence suggests that, at least in cancer, different diseased individuals harbor distinct sets of genetic and transcriptional dysregulations, which tend to occur in specific disease-relevant pathways [147-149]. In such cases, the differential expression of most pathway genes may not be significant when analyzed across the entire case-control dataset, but each individual case will carry some alternations in pathway genes. We addressed this problem in

Chapter 4, and designed a computational framework that is capable of uncovering disease-related pathways.

9.1.2 Combining physical and genetic interactions

Transcriptional and genetic evidence tends to implicate distinct genes and pathways in a specific phenotype [150]. Thus, using genetic evidence for module finding is expected to reveal insights complementary to those identified using expression data. In Chapters 7 and 8 we described two approaches for identifying functional modules by combining GI and PPI networks. In Chapter 7 we described an approach that is capable of identifying pairs of partially redundant pathways using PPIs and qualitative negative GIs. In Chapter 8 we extended this approach to handle both positive and negative quantitative GIs, and to identify simultaneously a collection of modules. In both cases we showed that in addition to facilitating module identification, the PPI and the GI data can be used for organizing modules into high-order structures, through identification of *pivot proteins* connected to pairs of partially redundant pathways (Chapter 7) and construction of *module maps* where complimentary modules are connected by multiple negative GIs (Chapter 8).

9.2 Biological questions that can be addressed by a module-based approach

Assignment of genes into modules can be useful for diverse tasks, some of which were exemplified in this thesis:

- **Discovery of novel pathways**: If most of the discovered modules correspond to known pathways, it is likely that some of the other modules correspond to novel pathways. In addition, if the modules are based on a specific gene expression dataset, it is likely that the novel pathways are related to the specific conditions in which these genes are differentially expressed. We exemplified this approach by identifying a putative new cytoskeleton-related pathway involved in DNA damage response (Chapter 6).
- **Discovery of relationships between modules:** By comprehensive modeling of high-throughput data it is possible not only to identify modules, but also to construct a map of inter-modular relations. In a module map we constructed using a PPI network and quantitative GIs (Chapter 8), pairs of modules are connected if they are linked by many negative GIs. It is thus likely that these module pairs are functionally related. Indeed, we find that such module pairs tend to share common

functions. In four cases where no common function was previously reported, we predicted novel functional connections.

- **Prediction of protein function**: When a significant fraction of the genes in a module are annotated with a common function, the rest of the genes in the same module can be predicted to share this function. This *guilt by association* principle has been previously used in studies utilizing PPI networks (reviewed [37], a review that is not part of this thesis), and in studies combining together heterogeneous data [85]. We used this method to predict novel functions using modules based on PPIs and expression profiles (Chapter 6), or PPIs and GIs (Chapter 8).
- **Prediction of functionally important interactions:** One of the advantages of defining modules as subnetworks rather than as gene sets is the fact that the edges in the subnetwork can provide further biological insight. For example, PluriNet, a module we identified as up-regulated in human pluripotent stem cells (Chapter 3), consisted of at least two distinct sub-modules, one related to the transcription factor NANOG and another to cell cycle progression. As suggested in Chapter 3, it is likely that several edges connecting members of these sub-modules are crucial for the regulation of cell cycle progression by regulators of pluripotency.
- **Prediction of candidate drug targets:** If gene expression data from disease studies are used for module finding, and some modules are identified as up-regulated in diseased individuals, the members of this module constitute possible targets for therapeutic intervention. For example, we identified a focused 14-gene subnetwork as up-regulated in Huntington's disease (HD, Chapter 4). One of the genes in this subnetwork, HDAC1, is a target of HDAC inhibitors, currently tested as promising drugs for treatment of HD [151-153]. Another gene in the same module, MSH6, was recently suggested as a potential target for therapy that can decrease the length of the CAG repeat that causes HD [154].
- **Discovery of novel biological phenomena:** A collection of functional modules can be used to deduce global organization principles of specific biological systems. Such discoveries were presented in Chapters 7 and 8. For example, we find that genes connected to at least two modules that exhibit genetic buffering between them (which we call *pivot* proteins) tend to be essential and conserved in evolution.

9.3 Evaluating performance

We utilized several methods for the evaluation of the performance of our module finding methods, and their comparison with other methods. The most common method for evaluating the quality of a collection of putative functional modules is by testing the enrichment of each module for sets of genes that are known to be functionally-related. In the studies described in this thesis we utilized for these tests Gene Ontology (GO) annotations [155], and occasionally gene sets from other databases, such as KEGG [28], MIPS [156], MSigDB [70] and SGD [157]. Once we established which modules are significantly enriched for specific functional annotations, we measured specificity (fraction of modules enriched for at least one annotation) and sensitivity (fraction of annotations enriched in at least one module).

An alternative way of evaluating performance is to test the method for robustness to parameter choices (c.f., Chapter 8) and robustness to noise in the input data (c.f., Chapter 6). Finally, when we focused on detection of subnetworks pertinent to a particular disease, we also tested whether the subnetworks contained genes mutations in which increase the susceptibility to the disease (Chapter 4).

9.4 Access to the tools described in this thesis

Importantly, we developed a graphical user interface for most of the methods described here. The MATISSE software package (http://acgt.cs.tau.ac.il/matisse) contains an implementation of the methods described in Chapters 2, 4 and 6. Since its release in 2007, it was downloaded by over 200 different users and has been used with our support for research in the Loring lab at UCSD [2], Shiloh lab at TAU, Yarden lab at the Wezimann Institute, Department of Oncology in University of Cambridge and at Stanford University. In addition, we added an implementation of the MATISSE algorithm (Chapter 2) to version 5.0 of the Expander microarray analysis suite (http://acgt.cs.tau.ac.il/expander), which is used in hundreds of labs worldwide. Finally, several components of our computational infrastructure for network analysis were used in the development of two tools for analysis of biological networks, MetaReg [158] and SPIKE [33], which are not included in this thesis.

9.5 Future research

The studies described in this thesis can serve as a basis for several research directions.

- **Performance improvements.** All the major computational problems that we formulated and addressed in this thesis are NP-Hard. The basic problem underlying the formulations described in Chapters 2, 3, 5 and 6 is finding the heaviest subgraph in a graph that contains edges with both positive and negative weights. We have proven the hardness of this problem using a simple reduction from Max-Clique in Chapter 3. The problem we described in Chapter 4 is NP-Hard by reduction from Set-Cover. Finally, the problems we described in Chapters 7 and 8 are NP-Hard by reductions from Maximum Weight Biclique and from Correlation Clustering, respectively. Therefore, in most cases, we proposed heuristic algorithms. Exceptions are our algorithms for detection of dysregulated pathways (Chapter 4), for three of which we proved an approximation bound. Development of novel algorithms with provable approximation bounds for the problems described in this thesis could significantly improve the accuracy and utility of the identified modules. Another frontier is development of efficient exact algorithms for these probes, such as those based on Integer Linear Programming, which has been successfully used in related problems [114, 159].
- **Detection of subnetworks differentially expressed in a subset of the conditions.** The approaches based on gene expression data that are described in Chapters 2, 3, 5 and 6 are all based on expression similarity across all the studied conditions. They are therefore similar to clustering of gene expression patterns. It should be possible to extend the MATISSE model to a biclustering method, which will extract subnetworks whose genes exhibit high similarity across a *subset* of the expression profiles. This extension can be based on the SAMBA framework [160], where the expression data are represented as a bipartite graph $G=(U,V)$, in which U is a set of conditions, V is the set of genes, and an edge (u,v) connects v to u if v is differentially expressed in u. The goal is to detect heavy subgraphs (U',V'). By adding connectivity constraints to U' we get a problem similar both to the problem described in Chapter 2 and to the one in Chapter 7. A combination of our approaches described in those chapters has a potential to be successful in detecting such bicluster-like subnetworks.

- **Detection of pathways targeted by small molecules.** The biclustering-like extension described above can be further adjusted to accommodate data on cell line drug sensitivity. Kutalik et al. [161] have recently described an approach that uses a combination of gene expression profiles of untreated cell lines with data on cell line drug sensitivity to identify co-modules (G,C,D), which consist of G – a group of genes, C- a group of cell lines and D – a set drugs, such the genes in G are differentially expressed in C, and the cell lines in C are sensitive to the drugs in D. This approach was shown to accurately predict drug-gene associations (based on the co-appearance in co-modules). Introduction of network-connectivity constraints to G is likely to increase the likelihood that co-modules represent actual cellular pathways. Computationally, this problem can be modeled as a heaviest subgraph problem in a tri-partite graph of genes vs. conditions vs. drugs. Several ideas described in this thesis (for example, the biclique identification algorithm in Chapter 7) should be useful for designing efficient heuristics for this problem.
- **Using genotypes and gene expression to identify pathways dysregulated in human disease.** High-throughput DNA sequencing makes it possible to study the genome-wide genetic landscape of complex diseases on an unprecedented scale. Recent studies of several cancer types using next-generation sequencing techniques revealed that even though the relatively common mutations are restricted to a small percentage of the affected individuals, in several key pathways at least one protein is mutated in most sick individuals [147-149]. So far, these pathways were identified manually, using prior knowledge on disease biology. By extending our method described in Chapter 4 to handle genetic information on mutations in protein-coding genes it may be possible to detect disease-relevant pathways at higher accuracy.

In this thesis, we described several novel methods for identifying functional modules using diverse genomic data. Application of these methods to large scale genomics data enables identification of novel pathways alongside the cellular context in which they are active. In addition, as we have shown, our module-based approach is powerful for elucidating questions about individual genes and relationships between pathways. Such modular approaches are expected to play a key role in the upcoming genomic challenges, as scientists will use datasets of increasing dimensions to improve our understanding of cellular biology and to identify the pathways relevant for diagnosis and treatment of human disease.

Appendix

Acronyms

ChIP – Chromatin Immunopercipitatioin

E-MAP – Epistatic Mini-Array Profiles

GEO – Gene Expression Onmibus

GI – Genetic Interaction

GO – Gene Ontology

HD – Huntington's Disease

ILP – Integer Linear Programming

miRNA - MicroRNA

PBM – Protein Binding Microarray

PDI – Protein-DNA Interaction

PPI – Protein Protein Interaction

TF – Transcription Factor

Bibliography

1. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data**. *BMC Syst Biol* 2007, **1**:8.
2. Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, Park IH, Rao MS, Shamir R, Schwartz PH *et al*: **Regulatory networks define phenotypic classes of human stem cell lines**. *Nature* 2008.
3. Ulitsky I, Karp RM, Shamir R: **Detecting Disease-Specific Dysregulated Pathways Via Analysis of Clinical Expression Profiles**. In: *Research in Computational Molecular Biology (RECOMB) 2008: 2008*: Springer; 2008: 347.
4. Ulitsky I, Shamir R: **Detecting Pathways Transcriptionally Correlated with Clinical Parameters**. In: *Computational Systems Bioinformatics (CSB) 2008*: Imperial College Press: 249-260.
5. Ulitsky I, Shamir R: **Identifying functional modules using expression profiles and confidence-scored protein interactions**. *Bioinformatics* 2009, **25**(9):1158-1164.
6. Ulitsky I, Shamir R: **Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks**. *Mol Syst Biol* 2007, **3**:104.
7. Ulitsky I, Shlomi T, Kupiec M, Shamir R: **From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions**. *Mol Syst Biol* 2008, **4**:209.
8. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annu Rev Genomics Hum Genet* 2001, **2**:343-372.
9. Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**(5560):1662-1664.
10. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**(6761 Suppl):C47-52.
11. Suter B, Auerbach D, Stagljar I: **Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond**. *Biotechniques* 2006, **40**(5):625-644.
12. Ito T, Chiba T, Yoshida M: **Exploring the protein interactome using comprehensive two-hybrid projects**. *Trends Biotechnol* 2001, **19**(10 Suppl):S23-27.
13. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP *et al*: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae**. *Nature* 2006, **440**(7084):637-643.
14. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al*: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440**(7084):631-636.
15. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S *et al*: **A human protein-protein interaction network: a resource for annotating the proteome**. *Cell* 2005, **122**(6):957-968.
16. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N *et al*: **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature* 2005, **437**(7062):1173-1178.
17. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M *et al*: **Large-scale mapping of human protein-protein interactions by mass spectrometry**. *Mol Syst Biol* 2007, **3**:89.
18. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A *et al*: **Comprehensive curation and analysis of**

global interaction networks in Saccharomyces cerevisiae. *J Biol* 2006, **5**(4):11.
19. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B *et al*: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets**. *Nat Genet* 2006, **38**(3):285-293.
20. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.
21. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120.
22. Bulyk ML: **DNA microarray technologies for measuring protein-DNA interactions**. *Curr Opin Biotechnol* 2006, **17**(4):422-430.
23. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.
24. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, **290**(5500):2306-2309.
25. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A *et al*: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nat Methods* 2007, **4**(8):651-657.
26. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M *et al*: **High-resolution DNA-binding specificity analysis of yeast transcription factors**. *Genome Res* 2009, **19**(4):556-566.
27. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells**. *Genome Res* 2003, **13**(5):773-780.
28. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
29. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R *et al*: **Global analysis of protein phosphorylation in yeast**. *Nature* 2005, **438**(7068):679-684.
30. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K *et al*: **Systematic discovery of in vivo phosphorylation networks**. *Cell* 2007, **129**(7):1415-1426.
31. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs**. *Nature* 2005, **433**(7027):769-773.
32. Shalgi R, Lieber D, Oren M, Pilpel Y: **Global and local architecture of the mammalian microRNA-transcription factor regulatory network**. *PLoS Comput Biol* 2007, **3**(7):e131.
33. Elkon R, Vesterman R, Amit N, Ulitsky I, Zohar I, Weisz M, Mass G, Orlev N, Sternberg G, Blekhman R *et al*: **SPIKE - a database, visualization and analysis tool of cellular signaling pathways**. *BMC Bioinformatics* 2008, **9**:110.
34. Balazsi G, Barabasi AL, Oltvai ZN: **Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli**. *Proc Natl Acad Sci U S A* 2005, **102**(22):7841-7846.
35. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli**. *Nat Genet* 2002, **31**(1):64-68.
36. Eisenberg E, Levanon EY: **Preferential attachment in the protein network evolution**. *Phys Rev Lett* 2003, **91**(13):138701.

37. Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function**. *Mol Syst Biol* 2007, **3**:88.
38. Steffen M, Petti A, Aach J, D'Haeseleer P, Church G: **Automated modelling of signal transduction networks**. *BMC Bioinformatics* 2002, **3**:34.
39. Scott J, Ideker T, Karp RM, Sharan R: **Efficient algorithms for detecting signaling pathways in protein interaction networks**. *J Comput Biol* 2006, **13**(2):133-144.
40. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B *et al*: **Network modeling links breast cancer susceptibility and centrosome dysfunction**. *Nat Genet* 2007, **39**(11):1338-1349.
41. Pan W: **Network-based model weighting to detect multiple loci influencing complex diseases**. *Hum Genet* 2008, **124**(3):225-234.
42. Begley TJ, Rosenbach AS, Ideker T, Samson LD: **Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping**. *Mol Cancer Res* 2002, **1**(2):103-112.
43. Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, Van Houten B: **Integrating phenotypic and expression profiles to map arsenic-response networks**. *Genome Biol* 2004, **5**(12):R95.
44. Ideker T, Sharan R: **Protein networks in disease**. *Genome Res* 2008, **18**(4):644-652.
45. Xu J, Li Y: **Discovering disease-genes by topological features in human protein-protein interaction network**. *Bioinformatics* 2006, **22**(22):2800-2805.
46. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T: **A direct comparison of protein interaction confidence assignment schemes**. *BMC Bioinformatics* 2006, **7**:360.
47. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network**. *Nat Biotechnol* 2005, **23**(8):951-959.
48. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: **Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2007, **6**(3):439-450.
49. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D358-362.
50. Li D, Liu W, Liu Z, Wang J, Liu Q, Zhu Y, He F: **PRINCESS, a protein interaction confidence evaluation system with multiple data sources**. *Mol Cell Proteomics* 2008, **7**(6):1043-1052.
51. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.
52. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL: **Multiplexed biochemical assays with biological chips**. *Nature* 1993, **364**(6437):555-556.
53. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA *et al*: **NCBI GEO: archive for high-throughput functional genomic data**. *Nucleic Acids Res* 2009, **37**(Database issue):D885-890.
54. 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* 1998, **9**(12):3273-3297.

55. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO *et al*: **Identification of genes periodically expressed in the human cell cycle and their expression in tumors**. *Mol Biol Cell* 2002, **13**(6):1977-2000.
56. 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* 2000, **11**(12):4241-4257.
57. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p**. *Mol Biol Cell* 2001, **12**(10):2987-3003.
58. O'Rourke SM, Herskowitz I: **Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis**. *Mol Biol Cell* 2004, **15**(2):532-542.
59. Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO, Botstein D: **Diverse and specific gene expression responses to stresses in cultured human cells**. *Mol Biol Cell* 2004, **15**(5):2361-2374.
60. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6062-6067.
61. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO: **Individuality and variation in gene expression patterns in human blood**. *Proc Natl Acad Sci U S A* 2003, **100**(4):1896-1901.
62. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531-537.
63. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al*: **A gene-expression signature as a predictor of survival in breast cancer**. *N Engl J Med* 2002, **347**(25):1999-2009.
64. Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P, Bouzou B, Jensen RV *et al*: **Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease**. *Proc Natl Acad Sci U S A* 2005, **102**(31):11023-11028.
65. Burczynski ME, Peterson RL, Twine NC, Zuberek KA, Brodeur BJ, Casciotti L, Maganti V, Reddy PS, Strahs A, Immermann F *et al*: **Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells**. *J Mol Diagn* 2006, **8**(1):51-61.
66. Butte A: **The use and analysis of microarray data**. *Nat Rev Drug Discov* 2002, **1**(12):951-960.
67. Sharan R, Elkon R, Shamir R: **Cluster analysis and its applications to gene expression data**. *Ernst Schering Res Found Workshop* 2002(38):83-108.
68. Madeira SC, Oliveira AL: **Biclustering algorithms for biological data analysis: a survey**. *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1**(1):24-45.
69. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.
70. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
71. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment**. *BMC Bioinformatics* 2005, **6**:144.

72. Jiang Z, Gentleman R: **Extensions to gene set enrichment**. *Bioinformatics* 2007, **23**(3):306-313.
73. Kurhekar MP, Adak S, Jhunjhunwala S, Raghupathy K: **Genome-wide pathway analysis and visualization using gene expression data**. *Pac Symp Biocomput* 2002:462-473.
74. Zien A, Kuffner R, Zimmer R, Lengauer T: **Analysis of gene expression data with pathway scores**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:407-417.
75. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **12**(1):37-46.
76. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network**. *Nature* 2004, **430**(6995):88-93.
77. Vert JP, Kanehisa M: **Extracting active pathways from gene expression data**. *Bioinformatics* 2003, **19 Suppl 2**:ii238-244.
78. Segal E, Friedman N, Kaminski N, Regev A, Koller D: **From signatures to models: understanding cancer using microarrays**. *Nat Genet* 2005, **37 Suppl**:S38-45.
79. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, **16**(10):906-914.
80. Finocchiaro G, Mancuso F, Muller H: **Mining published lists of cancer related microarray experiments: identification of a gene expression signature having a critical role in cell-cycle control**. *BMC Bioinformatics* 2005, **6 Suppl 4**:S14.
81. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer**. *Cancer Res* 2002, **62**(15):4427-4433.
82. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM: **Mining for regulatory programs in the cancer transcriptome**. *Nat Genet* 2005, **37**(6):579-583.
83. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer**. *Nat Genet* 2004, **36**(10):1090-1098.
84. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ *et al*: **Integrative molecular concept modeling of prostate cancer progression**. *Nat Genet* 2007, **39**(1):41-51.
85. Tanay A, Steinfeld I, Kupiec M, Shamir R: **Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium**. *Mol Syst Biol* 2005, **1**:2005 0002.
86. Huttenhower C, Hibbs M, Myers C, Troyanskaya OG: **A scalable method for integration and functional analysis of multiple microarray datasets**. *Bioinformatics* 2006, **22**(23):2890-2897.
87. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166-176.
88. Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH: **Functional annotation and network reconstruction through cross-platform integration of microarray data**. *Nat Biotechnol* 2005, **23**(2):238-243.
89. Hu H, Yan X, Huang Y, Han J, Zhou XJ: **Mining coherent dense subgraphs across massive biological networks for functional discovery**. *Bioinformatics* 2005, **21 Suppl 1**:i213-221.

90. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae**. *Nat Genet* 2001, **29**(4):482-486.
91. Hahn A, Rahnenfuhrer J, Talwar P, Lengauer T: **Confirmation of human protein interaction data by human expression data**. *BMC Bioinformatics* 2005, **6**:112.
92. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**(5644):449-453.
93. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis**. *Mol Syst Biol* 2007, **3**:140.
94. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP: **Classification of microarray data using gene networks**. *BMC Bioinformatics* 2007, **8**:35.
95. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL: **Dynamic modularity in protein interaction networks predicts breast cancer outcome**. *Nat Biotechnol* 2009, **27**(2):199-204.
96. Ma X, Lee H, Wang L, Sun F: **CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data**. *Bioinformatics* 2007, **23**(2):215-221.
97. Wei P, Pan W: **Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model**. *Bioinformatics* 2008, **24**(3):404-411.
98. Li C, Li H: **Network-constrained regularization and variable selection for analysis of genomic data**. *Bioinformatics* 2008, **24**(9):1175-1182.
99. de Lichtenberg U, Jensen LJ, Brunak S, Bork P: **Dynamic complex formation during the yeast cell cycle**. *Science* 2005, **307**(5710):724-727.
100. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 2004, **431**(7006):308-312.
101. Wachi S, Yoneda K, Wu R: **Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues**. *Bioinformatics* 2005, **21**(23):4205-4208.
102. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks**. *Bioinformatics* 2002, **18 Suppl 1**:S233-240.
103. Cabusora L, Sutton E, Fulmer A, Forst CV: **Differential network expression during drug and stress response**. *Bioinformatics* 2005, **21**(12):2898-2905.
104. Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S: **Network-based analysis of affected biological processes in type 2 diabetes models**. *PLoS Genet* 2007, **3**(6):e96.
105. Rajagopalan D, Agarwal P: **Inferring pathways from gene lists using a literature-derived network of biological relationships**. *Bioinformatics* 2005, **21**(6):788-793.
106. Bandyopadhyay S, Kelley R, Ideker T: **Discovering regulated networks during HIV-1 latency and reactivation**. *Pac Symp Biocomput* 2006:354-366.
107. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data**. *Bioinformatics* 2002, **18 Suppl 1**:S145-154.
108. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
109. Guo Z, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D *et al*: **Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network**. *Bioinformatics* 2007, **23**(16):2121-2128.

110. Segal E, Wang H, Koller D: **Discovering molecular pathways from protein interaction and gene expression data**. *Bioinformatics* 2003, **19 Suppl 1**:i264-271.
111. Ge R, Ester M, Gao B, Hu Z, Bhattacharya B, Ben-Moshe B: **Joint cluster analysis of attribute data and relationship data: The connected k-center problem, algorithms and applications**. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2008, **2**(2).
112. Breitling R, Amtmann A, Herzyk P: **Graph-based iterative Group Analysis enhances microarray interpretation**. *BMC Bioinformatics* 2004, **5**:100.
113. Nacu S, Critchley-Thorne R, Lee P, Holmes S: **Gene expression network analysis and applications to immunology**. *Bioinformatics* 2007, **23**(7):850-858.
114. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach**. *Bioinformatics* 2008, **24**(13):i223-231.
115. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al*: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418**(6896):387-391.
116. Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, Wu HI, McCann KE, Troyanskaya OG, Brown JM: **Global analysis of gene function in yeast by quantitative phenotypic profiling**. *Mol Syst Biol* 2006, **2**:2006 0001.
117. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH *et al*: **Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast**. *Cell* 2006, **126**(3):611-625.
118. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D *et al*: **The chemical genomic portrait of yeast: uncovering a phenotype for all genes**. *Science* 2008, **320**(5874):362-365.
119. Hartman JLt, Garvik B, Hartwell L: **Principles for the buffering of genetic variation**. *Science* 2001, **291**(5506):1001-1004.
120. Segre D, Deluna A, Church GM, Kishony R: **Modular epistasis in yeast metabolism**. *Nat Genet* 2005, **37**(1):77-83.
121. Beyer A, Bandyopadhyay S, Ideker T: **Integrating physical and genetic maps: from genomes to interaction networks**. *Nat Rev Genet* 2007, **8**(9):699-710.
122. Mani R, St Onge RP, Hartman JLt, Giaever G, Roth FP: **Defining genetic interaction**. *Proc Natl Acad Sci U S A* 2008, **105**(9):3461-3466.
123. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF *et al*: **Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile**. *Cell* 2005, **123**(3):507-519.
124. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H *et al*: **Systematic genetic analysis with ordered arrays of yeast deletion mutants**. *Science* 2001, **294**(5550):2364-2368.
125. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M *et al*: **Global mapping of the yeast genetic interaction network**. *Science* 2004, **303**(5659):808-813.
126. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD: **A DNA integrity network in the yeast Saccharomyces cerevisiae**. *Cell* 2006, **124**(5):1069-1081.
127. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M *et al*: **Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map**. *Nature* 2007, **446**(7137):806-810.
128. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park HO, Hayles J *et al*: **Conservation and rewiring of functional**

modules revealed by an epistasis map in fission yeast. *Science* 2008, **322**(5900):405-410.
129. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G: **Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions**. *Nat Genet* 2007, **39**(2):199-206.
130. Ozier O, Amin N, Ideker T: **Global architecture of genetic interactions on the protein network**. *Nat Biotechnol* 2003, **21**(5):490-491.
131. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks**. *Nat Biotechnol* 2005, **23**(5):561-566.
132. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS: **Gene function prediction from congruent synthetic lethal interactions in yeast**. *Mol Syst Biol* 2005, **1**:2005 0026.
133. Qi Y, Ye P, Bader JS: **Genetic Interaction Motif Finding by expectation maximization--a novel statistical model for inferring gene modules from synthetic lethality**. *BMC Bioinformatics* 2005, **6**:288.
134. Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP: **Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network**. *J Biol* 2005, **4**(2):6.
135. Le Meur N, Gentleman R: **Modeling synthetic lethality**. *Genome Biol* 2008, **9**(9):R135.
136. Collins SR, Schuldiner M, Krogan NJ, Weissman JS: **A strategy for extracting and analyzing large-scale quantitative epistatic interaction data**. *Genome Biol* 2006, **7**(7):R63.
137. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T: **Functional maps of protein complexes from quantitative genetic interaction data**. *PLoS Comput Biol* 2008, **4**(4):e1000065.
138. Pu S, Ronen K, Vlasblom J, Greenblatt J, Wodak SJ: **Local coherence in genetic interaction patterns reveals prevalent functional versatility**. *Bioinformatics* 2008, **24**(20):2376-2383.
139. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H *et al*: **Combining biological networks to predict genetic interactions**. *Proc Natl Acad Sci U S A* 2004, **101**(44):15682-15687.
140. Qi Y, Suhail Y, Lin YY, Boeke JD, Bader JS: **Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions**. *Genome Res* 2008.
141. Paladugu SR, Zhao S, Ray A, Raval A: **Mining protein networks for synthetic genetic interactions**. *BMC Bioinformatics* 2008, **9**(1):426.
142. Zhong W, Sternberg PW: **Genome-wide prediction of C. elegans genetic interactions**. *Science* 2006, **311**(5766):1481-1484.
143. Jarvinen AP, Hiissa J, Elo LL, Aittokallio T: **Predicting quantitative genetic interactions by means of sequential matrix approximation**. *PLoS ONE* 2008, **3**(9):e3284.
144. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR, Boone C: **Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways**. *Nat Biotechnol* 2004, **22**(1):62-69.
145. Laurent LC, Chen J, Ulitsky I, Mueller FJ, Lu C, Shamir R, Fan JB, Loring JF: **Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence**. *Stem Cells* 2008, **26**(6):1506-1516.
146. Rives AW, Galitski T: **Modular organization of cellular networks**. *Proc Natl Acad Sci U S A* 2003, **100**(3):1128-1133.

147. McLendon R, Friedman A, Bigner D, Van Meir E, Brat D, Mastrogianakis G, Olson J, Mikkelsen T, Lehman N, Aldape K: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061-1068.
148. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB *et al*: **Somatic mutations affect key pathways in lung adenocarcinoma**. *Nature* 2008, **455**(7216):1069-1075.
149. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB *et al*: **Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia**. *Nature* 2007, **446**(7137):758-764.
150. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, Auluck PK, Geddie ML, Valastyan JS, Karger DR *et al*: **Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity**. *Nat Genet* 2009, **41**(3):316-323.
151. Thomas EA, Coppola G, Desplats PA, Tang B, Soragni E, Burnett R, Gao F, Fitzgerald KM, Borok JF, Herman D *et al*: **The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice**. *Proc Natl Acad Sci U S A* 2008, **105**(40):15564-15569.
152. Kazantsev AG, Thompson LM: **Therapeutic application of histone deacetylase inhibitors for central nervous system disorders**. *Nat Rev Drug Discov* 2008, **7**(10):854-868.
153. Dompierre JP, Godin JD, Charrin BC, Cordelieres FP, King SJ, Humbert S, Saudou F: **Histone deacetylase 6 inhibition compensates for the transport deficit in Huntington's disease by increasing tubulin acetylation**. *J Neurosci* 2007, **27**(13):3571-3583.
154. Dragileva E, Hendricks A, Teed A, Gillis T, Lopez ET, Friedberg EC, Kucherlapati R, Edelmann W, Lunetta KL, MacDonald ME *et al*: **Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes**. *Neurobiol Dis* 2009, **33**(1):37-47.
155. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
156. Mewes HW, Hani J, Pfeiffer F, Frishman D: **MIPS: a database for protein sequences and complete genomes**. *Nucleic Acids Research* 1998, **26**(1):33-37.
157. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res* 1998, **26**(1):73-79.
158. Ulitsky I, Gat-Viks I, Shamir R: **MetaReg: a platform for modeling, analysis and visualization of biological systems using large-scale experimental data**. *Genome Biol* 2008, **9**(1):R1.
159. Bruckner S, Hüffner F, Karp RM, Shamir R, Sharan R: **Topology-Free Querying of Protein Interaction Networks**. In: *Research in Computational Molecular Biology (RECOMB) 2009; Tucson, AZ*: Springer: 74-89.
160. Tanay A, Sharan R, Shamir R: **Discovering statistically significant biclusters in gene expression data**. *Bioinformatics* 2002, **18 Suppl 1**:S136-144.
161. Kutalik Z, Beckmann JS, Bergmann S: **A modular approach for integrative analysis of large-scale gene-expression and drug-response data**. *Nat Biotechnol* 2008, **26**(5):531-539.
162. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.
163. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data**. *Bioinformatics* 2002, **18**(11):1454-1461.

164. Dudoit S, Yang Y, Callow M, Speed T: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments**. *Statistica Sinica* 2002, **12**(1):111-140.

אוניברסיטת תל-אביב TEL AVIV UNIVERSITY

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלבטניק

אלגוריתמים מבוססי-רשת לניתוח מידע ביו-רפואי רבגוני

חיבור לשם קבלת תואר "דוקטור לפילוסופיה"

מאת **איגור אוליצקי**

בהנחייתו של פרופ' **רון שמיר**

הוגש לסנאט של אוניברסיטת ת"א

יולי 2009

תמצית

פריצות דרך בביוטכנולוגיה בעשור האחרון מאפשרות איסוף מידע ביו-רפואי במימדים חסרי תקדים. מגוון שיטות חדשניות מאפשרות מיפוי רחב היקף של קשרים מסוגים שונים בין גנים או בין תוצריהם החלבוניים. אוסף הקשרים מסוג מסוים ניתן לייצוג כרשת כלל-גנומית. שיטות אחרות מאפשרות למדוד את הכמות והפעילות של גנים בתנאים שונים. המידע שנאסף באמצעות רוב השיטות הקיימות הוא רועש וקשה לפירוש, ולכן שילוב נתונים ממקורות שונים יכול לסייע רבות ביכולת ניתוחם. בנוסף, רוב הפעילויות התאיות מתבצעות על ידי פעילות מתואמת של תוצרי כמה גנים, הנקראים יחד מודול פונקציונאלי. אתגר מרכזי של הביולוגיה החישובית הוא לזהות מודולים אלה ולגלות בעזרתם תגליות ביולוגיות.

בתזה זו פיתחנו מספר שיטות חישוביות למציאת מודולים פונקציונאליים באמצעות מידע מגוון (הטרוגני). התמקדנו בניתוח של רשתות קשרי חלבון-חלבון, רשתות קשרים גנטיים, ומידע על ביטוי גנים, אך הכלים החישוביים שפיתחנו מתאימים גם לסוגי נתונים נוספים. הדגמנו את היעילות של השיטות הללו באמצעות נתונים שנאספו במגוון מערכות ביולוגיות בשמר האפייה ובאדם. שיטותינו מאפשרות חיזוי תפקידיהם של גנים בודדים ושל קבוצות גנים, מציאת מסלולים מולקולאריים וקומפלקסים חדשים והקשרים ביניהם, וחיזוי של קשרים בעלי חשיבות פונקציונאלית. בהקשר הקליני, המודולים שזיהינו יכולים לשמש כסמן דיאגנוסטי למספר מחלות, מצביעים על המסלולים המולקולאריים שנפגעים במחלות אלה, ומציעים מטרות אפשריות לפיתוח של טיפול תרופתי.

<u>תקציר</u>

רקע כללי

בעשור האחרון, שיפורים עצומים בטכנולוגיות ריצוף דנ"א אפשרו את קביעת הרצף של גנומים שלמים ואת זיהוי רוב אבני הבניין של המערכת התאית. במקביל, התפתחויות בתחומים אחרים של הביוטכנולוגיה אפשרו מיפוי רחב היקף של כמות הרכיבים, מיקומם בתוך התא ומגוון הקשרים ביניהם. התפתחויות אלה הובילו לשינוי דרמטי במחקר הביולוגי ולהתפתחות מואצת של ה**ביולוגיה של מערכות** כמדע בין-תחומי חדש. ביולוגיה של מערכות בוחנת את המערכת הביולוגית בראייה כוללת, ומנסה לפענח את העקרונות הגלובליים שלה על מנת להבין ולחזות את התנהגותה.

אחת התכונות המרתקות של המערכות ביולוגיות היא ה**מודולאריות** שלהן. רוב התהליכים בתא מסתמכים על פעילות מתואמת של קבוצות תוצרים של גנים רבים, הנקראות **מדולים פונקציונאליים** [10]. ניתן לראות את התפקוד הפיזיולוגי של תאים ואורגניזמים כפעילות מתואמת ומשולבת של מודולים אלה. אחת מהמטרות המרכזיות של הביולוגיה החישובית היא מציאת המודולים הללו ושיפור ההבנה הביולוגית באמצעות הניתוח שלהם.

רשתות ביולוגיות

מידע ביולוגי מגוון יכול להיות מיוצג על ידי רשתות של קשרים בין גנים או בין התוצרים החלבוניים שלהם. בעקרון, אם הידע שלנו לגבי הביולוגיה התאית היה מלא, יכולנו להרכיב ולהשתמש ברשת בקרה כלל-גנומית שבאמצעותה ניתן היה לחזות בצורה מהימנה את הכמות, המיקום ורמת הפעילות של כל פרודה בתא בכל זמן נתון. למרבה הצער, מיפוי של רשת בעלת היקף ורזולוציה כאלה לא הושג עד היום באף אורגניזם, ולא צפוי בשנים הקרובות. למרות זאת, מספר שיטות ביוטכנולוגיות מאפשרות פענוח רשתות המייצגות חלקים מסוימים של הרשת הבקרה המלאה. קשרים בין גנים ניתנים למדידה הן באמצעות שיטות קלאסיות שמתמקדות בגנים בודדים והן באמצעות מספר שיטות כלל-גנומיות חדשניות, אולם בשני המקרים המדידות מכילות כמויות ניכרות של רעש. שילוב מידע משני סוגי הניסויים יצר את הרשתות המקיפות ביותר שזמינות היום למחקר. הקשרים ברשתות הללו כוללים: (1) **קשרי חלבון-חלבון** המעידים על מגע פיזי בין שני החלבונים; (2) **קשרי גורם שעתוק-דנ"א** המעידים על מגע פיזי בין חלבון המשמש כגורם שעתוק לבין רצף דנ"א שנמצא לרוב בפרומוטר של גן אחר; (3) **קשרי בקרה** שבהם גן אחד מבקר את הפעילות של גן אחר. קשרי גורם שעתוק-דנ"א הם סוג אחד של קשר בקרה. סוגים נוספים הם קשר בין חלבון מזרחן לסובסטרט שלו ובין מיקרו-רנ"א לגן

המטרה שלו. (4) **קשרים מטבוליים** המעידים על כך שתוצר של אנזים אחד משמש כסובסטרט של אנזים אחר. (5) **קשרים גנטיים** המעידים על כך שהפנוטיפ המתקבל במצב של חוסר של שני גנים שונה משמעותית מזה הצפוי לפי הפנוטיפים הנגרמים מחוסר של כל אחד מהגנים בנפרד.

ניתוח חישובי של רשתות ביולוגיות

התכונות המרתקות והפוטנציאל העצום של הרשתות הביולוגיות הובילו לפיתוח מספר רב של שיטות חישוביות חדשות. המחקרים הראשוניים התמקדו בתכונות בסיסיות של הרשתות וחקרו מדדים טופולוגיים שונים [162], שכיחות של תתי-גרפים קטנים [35], ואת האבולוציה של הרשת [36]. עשרות שיטות חישוביות פותחו לזיהוי תתי-רשתות צפופות, על מנת לזהות קומפלקסים חלבוניים חדשים ולחזות תפקידים חדשים לגנים (ראה סקירה ב-[37]). קבוצה נוספת של שיטות מגלה מסלולים ארוכים ברשתות קשרי חלבונים שעשויים להתאים למסלולים ליניאריים להעברת סיגנל [38, 39].

עבודות רבות הראו ששימוש ברשתות קשרי חלבון-חלבון וגורם שעתוק-דנ"א מסייע בפירוש מידע גנומי ממקורות אחרים. השימוש ברשתות לפירוש מידע על ביטוי גנים ועל קשרים גנטיים מתואר להלן. בנוסף, ניתוח משופר תוך שימוש ברשתות תואר עבור גנוטיפים [40, 41], פנוטיפים של חוסרים בגנים [42, 43], ומחלות [44, 45].

מדידות רחבות היקף של רמות הביטוי הגנטי

אחת ההתפתחויות המשמעותיים ביותר בביולוגיה מולקולארית ב-15 השנה האחרונות היא הפיתוח של הטכנולוגיה של שבבי דנ"א [51, 52]. שבב דנ"א מכיל מערך של אוליגונוקליאוטידים שמשמשים כגלאים לכמויות של חומצות גרעין. בניסוי אחד המשתמש בשבבי דנ"א ניתן למדוד את רמות הביטוי של אלפי גנים בתנאי מסוים. עד היום, שבבי דנ"א שימשו בעיקר למדידה רחבת היקף של כמויות רנ"א שליח. נכון לאפריל 2009, כמעט 300,000 פרופילים כאלו זמינים במאגרים הציבוריים [53].

מדידה רחבת היקף של כמויות רנ"א שליח יכולה לעזור לפתרון מגוון שאלות ביולוגיות. בסוג אחד של ניסויים התאים נחשפים לטיפולים שונים ורמות הרנ"א שליח נמדדות במספר נקודות זמן אחרי הטיפול. ניסויים קלאסיים מסוג זה חקרו את התקדמות מחזור התא [54, 55] והתגובה לעקות שונות בשמרים ובבני אדם [56-59]. מגוון נוסף של ניסויים משווה בין פרופילי ביטוי של תאים הלקוחים מרקמות או מאוכלוסיות שונות [60,

[61]. ניסויים קליניים המשתמשים בשבבי דנ"א משווים לרוב רקמות הלקוחות מאנשים חולים לאלו של אנשים בריאים [62-65].

ניתוח חישובי של תוצאות שבבי דנ"א

ההיקף הגדול של ניסויים בשבבי דנ"א (מחקר טיפוסי בתאים אנושיים מודד רמות ביטוי של כ-20,000 גנים בעשרות ואף מאות דגימות) והרעש הטכני והביולוגי דורשים פיתוח כלים חישוביים ייעודיים לניתוח המידע. מגוון שיטות לביצוע שלבים שונים בניתוח מידע משבבי דנ"א פותחו עד היום [66]. שיטות אלה כוללות שיטות קיבוץ למציאת צבירים של גנים ותנאים בעלי תבניות ביטוי דומות [67, 68], ולסיווג דגימות באמצעות מידע על ביטוי הגנים [69]. מספר שיטות סטטיסטיות פותחו למציאת גנים שהביטוי שלהם שונה בצורה מובהקת בין שני סוגים של דגימות או נמצא במתאם חזק לפרמטר קליני רציף [163, 164]. למרות שלרוב ניתן להציע היפותזות חדשות על סמך ביטוי משתנה של גנים בודדים, לעתים קרובות, זיהוי של ביטוי משתנה של קבוצת גנים השייכים לתהליך תאי מסוים יכול להיות מועיל אף יותר. גישה זו עשויה לגלות אותות חלשים, שאינם נחשפים כאשר כל גן מנותח בנפרד. גישה אחת לבעיה זו משתמש בקבוצות מוגדרות של גנים שתויגו כבעלי פונקציה משותפת [70-72], כמשתייכים לאותו מסלול מטאבולי [73, 74] או לאותו קומפלקס חלבוני [75]. החסרונות של שיטות אלה הם הקושי שבקביעה אילו גנים משתייכים לקבוצות אלה, ושבמקרים רבים, רק חלק מהמסלול המולקולארי משתנה ברמת הביטוי שלו. השימוש ברשתות גנים, ובפרט ברשתות קשרי חלבונים, יכול להועיל בפתרון בעיה זו.

השיטות האלגוריתמיות לשילוב רשתות ומידע על ביטוי גנים יכולות להבליט את שינויי הביטוי בעלי המשמעות הביולוגית. גנים המחוברים ברשת קשרי חלבונים נוטים להתבטא בצורה דומה [90, 91], ושימוש ברשתות קשרי חלבונים הועיל לשיפור סיווג דגימות באמצעות רמות ביטוי גנים [93-95] ובגילוי גנים בעלי ביטוי שונה בדגימות מסוגים שונים [96-98]. אחד מתחומי המחקר העיקריים בשילוב מידע מרשתות וביטוי גנים הוא זיהוי תתי-רשתות בעלות דפוס ביטוי מעניין. ניתן לסווג את הגישות הללו לארבע קבוצות עיקריות: (1) זיהוי תתי-רשתות הפעילות בדגימה ספציפית [34, 99-101]; (2) זיהוי תתי רשתות הפעילות בתת קבוצה של התנאים שנבדקו [102-106]; (3) גילוי תתי-רשתות שהגנים שבהן מראים מתאם הדדי גדול לאורך כל התנאים [107, 109,

110]; (4) זיהוי תתי רשתות שהביטוי שלהן מבדיל בין שתי קבוצות תנאים [112-114]. שימוש בשיטות אלה אפשר זיהוי של מודולים פונקציונאליים רבים הקשורים לתנאים ביולוגיים מגוונים ולמחלות שונות.

רשתות קשרים גנטיים

לאחרונה, שימוש בשבבי דנ"א ופיתוחים ביוטכנולוגיים אחרים, הובילו למיפוי רחב היקף של הפנוטיפים של חוסרים של גנים בודדים, בעיקר בשמר S. cerevisiae [115-118]. מחקרים אלה הראו שרק כ-18% מהגנים באורגניזם זה חיוניים לגידול על מצע עשיר [115]. תוצאה זו הובילה לקביעה שגיבוי (buffering) גנטי נפוץ מאוד בשמרים ובאורגניזמים אחרים [119]. כאשר הפנוטיפ של מוטנט החסר שני גנים שונה מהצפוי לפי הפנוטיפים של המוטנטים החסרים כל גן בנפרד, נקבע כי בין הגנים הללו קיים קשר. ניתן לסווג את הקשרים הגנטיים לקשרים חיוביים, נייטרליים ושליליים [120, 121]. **קשר גנטי שלילי** נקבע כאשר השרידות של מוטנט החסר את שני הגנים נמוכה מהצפוי ו**קשר גנטי חיובי** נקבע כאשר השרידות גבוהה מהצפוי. השרידות הצפויה נקבעת לרוב על-ידי מכפלת השרידויות של המוטנטים החסרים גן בודד [120, 122, 123]. מספר שיטות מולקולאריות חדשות מאפשרות מיפוי רחב היקף של קשרים גנטיים [123-129].

מחקרים מוקדמים על רשת הקשרים הגנטיים הראו שגנים בעלי קשרים גנטיים נוטים לחלוק גם קשר פיזי (קשרי חלבון-חלבון וגורם שעתוק-דנ"א) [124, 125], ושגנים בעלי קשרים גנטיים רבים נוטים לקיים גם קשרים פיזיים רבים [130]. מחקרים אלה הובילו למסקנה ששילוב של רשת הקשרים הגנטיים עם רשת של קשרים פיזיים עשוי להוביל למסקנות ביולוגיות חדשניות. מחקר המשך של Kelley ו-Ideker [131] הראה שאם משלבים יחד את שתי הרשתות מגלים שרוב הקשרים הגנטיים השליליים מחברים מסלולים מקבילים ברשת הפיזית. מחקרים מאוחרים יותר אשררו מסקנה זו [132-135]. לעומתם, הקשרים הגנטיים החיוביים מתרחשים לרוב בין גנים השייכים לאותו תהליך [127], ונובעים מאפקט חזק של חוסר באחד הגנים על פעילות התהליך, אפקט הגורם לכך שחוסר בגן נוסף מאותו תהליך הוא חסר השפעה.

ניתוח קשרים גנטיים כמותיים נעשה לרוב באמצעות קיבוץ היררכי [123, 127, 136]. לאחרונה, הוצעו שיטות חדשות לניתוח רשתות אלה המשלבות קיבוץ היררכי עם נתונים על קשרים פיזיים [137] או דו-קיבוץ (biclustering) המאפשר לגלות צבירים חופפים של

גנים [138]. שיטות נוספות משתמשות ברשת הקשרים הגנטיים לחיזוי קשרים גנטיים חדשים [139-143], לחיזוי תפקידי גנים חדשים [132] ולחיזוי גנים המעוכבים על ידי תרכובות כימיות [144].

תקציר המאמרים הכלולים בתזה

עבודה זו מסתמכת על שבעה מאמרים, אשר התפרסמו בכתבי עת מדעיים או הוצגו בכנסים מדעיים. להלן פירוט תקצירי המאמרים:

1. **Identification of Functional Modules using Network Topology and High-Throughput Data**

Igor Ulitsky and Ron Shamir

Published in *BMC Systems Biology* [1].

ניתוח רשתות קשרי חלבונים ונתונים רחבי היקף על ביטוי גנים מתבצע בדרך כלל בנפרד ובאמצעות שיטות שונות. חקר משולב של שני סוגי המידע עשוי לשפר את איכות הניתוח על-ידי שילוב התכונות הטופולוגיות של הרשת עם תכונות הנגזרות מתבניות ביטוי הגנים. בעבודה זו אנו מתארים מסגרת אלגוריתמית חדשה לבעיה זו. בשלב ראשון, אנו מחשבים את הדמיון בין תבניות הביטוי של כלל הגנים. בשלב השני, בהינתן רשת קשרי חלבונים וערכי הדמיון בין כלל זוגות הגנים, אנו מחפשים תתי-רשתות קשירות (או מודולים) שהגנים שבהן דומים זה לזה. פיתחנו אלגוריתמים לבעיה זו, והערכנו את הביצועים שלהם באמצעות נתונים על תגובת השמר *S. cerevisiae* למליחות גבוהה ונתונים על ביטוי גנים במהלך מחזור התא האנושי. התוצאות שלנו מראות שבאמצעות השיטה המוצעת ניתן לקבל מודולים פונקציונאליים ממוקדים ובעלי משמעות ביולוגית.

2. **Regulatory Networks Define Phenotypic Classes of Human Stem Cell Lines**

Franz-Josef Müller, Louise C. Laurent, Denis Kostka, Igor Ulitsky, Ron Williams, Cristina Lu, Mahendra S. Rao, Ron Shamir, Philip H. Schwartz, Nils O. Schmidt, Jeanne F. Loring

Published in *Nature* [2].

תאי גזע מוגדרים כאוכלוסייה מתחדשת של תאים בעלי יכולת להתמיין למגוון סוגים רחב. המונח 'תאי גזע' מתייחס למגוון תאים רחב, החל מתאים פלוריפוטנטיים, כדוגמת תאי גזע עובריים, אשר יכולים להתחלק ולהתמיין כמעט ללא מגבלות, ועד תאי גזע בוגרים, בעלי יכולת התמיינות הרבה יותר מוגבלת. ריבוי הדיווחים על מקורות חדשים של תאי גזע והציפייה שהם ימלא תפקיד מרכזי ברפואה מחדשת (Regenerative Medicine) הצריך שיטה כללית ומהימנה לסיווג התאים הללו.

בעבודה זו, יצרנו וניתחנו מסד נתונים של תבניות כלל-גנומיות של ביטוי גנים (stem' cell matrix') שמאפשר סיווג של תאי גזע אנושיים באמצעות מידע על מגוון רחב של תאי גזע ותאים ממוינים. באמצעות שיטת קיבוץ סיווגנו כ-150 דגימות תאים ל-12 צבירים וגילינו שתאי גזע פלוריפוטנטיים מתקבצים יחד לצביר אחד, בעוד שתאים מסוגים אחרים, כגון תאי גזע של מערכת העצבים, הם הרבה יותר מגוונים. באמצעות ניתוח חישובי נוסף גילינו תת-רשת של קשרי חלבונים שהביטוי שלה מייחד את התאים הפלוריפוטניטים. ניתוח מידע ממחקרים אחרים הראה שהגנים בתת רשת זו מתבטאים בתאים פלוריפוטנטים נוספים, כגון תאי גזע עובריים בעכברים ותאי ביצית אנושיים. עבודה זו מציעה אסטרטגיה חדשה לסיווג תאי גזע ותומכת ברעיון שפלוריפוטנטיות והתחדשות מבוקרות על ידי רשתות בקרה מורכבות.

3. **Detecting Disease-Specific Dysregulated Pathways via Analysis of Clinical Expression Profiles**

Igor Ulitsky, Richard M. Karp and Ron Shamir

Published in *Proceedings of the Eleventh Annual International Conference on Research in Computational Molecular Biology* (*RECOMB 2008*) [3].

במאמר זה אנו מציעים שיטה לזיהוי תתי-רשתות קשירות המועשרות בצורה מובהקת בגנים אשר הביטוי שלהם משתבש במצב של מחלה. ניתן להשתמש בתתי הרשתות הללו לאבחון המחלה, לזיהוי מסלולים מולקולאריים המשפיעים עליה, ולמציאת מטרות אפשריות לטיפול תרופתי. בשלב ראשון, השיטה שלנו משתמשת במידע על ביטוי גנים בחולים בהשוואה לאוכלוסיות בריאות, ומזהה קבוצות של גנים שביטויים השתבש בכל אחד מהחולים. בשלב שני, אנו מחפשים תתי-רשתות קשירות שבהן התרחשו שיבושי ביטוי רבים ברוב החולים. האלגוריתם המרכזי שלנו מחפש תתי רשתות קשירות מינימאליות שבהן מספר הגנים המשובשים בכל חולה הוא מעל סף מסוים. באמצעות שיטה זו חקרנו את ביטוי הגנים באזור גרעין זנבי (caudate nucleus) של המח באנשים בריאים ובחולים במחלת הנטינגטון וביצענו ניתוח על (meta-analysis) של שישה מחקרים על סרטן השד. בשני המקרים קיבלנו תתי-רשתות בעלות מובהקות סטטיסטית אשר סיפקו הסברים ספציפיים לגורמים המולקולאריים של מחלות אלה.

4. **Detecting Pathways Transcriptionally Correlated with Clinical Parameters**

Igor Ulitsky and R. Shamir

Published in *Proceedings of Computational Systems Bioinformatics (CSB) 2008* [4].

בעבודה זו אנו מתארים שיטה חדשה למציאת תתי רשתות קשירות עם תבניות ביטוי קוהרנטיות שנמצאות במתאם עם פרמטר כלשהו של הדגימות שנבדקו. השיטה המוצעת מתאימה גם לפרמטרים רציפים, כמו גיל או גודל גידול, וגם לפרמטרים קטגוריים, כגון מין או גנוטיפ. באמצעות השיטה אנו מנתחים נתונים מחולות בסרטן שד, ומגלים מודולים הקשורים לתשעה פרמטרים קליניים שונים, כולל גיל החולה, גודל הגידול, ומשך השרידות ללא גרורות. השיטה שלנו מסוגלת לגלות מסלולים מולקולאריים הקשורים למחלה אשר אינם מזוהים על ידי שיטות אחרות. התוצאות של העבודה תומכות במספר היפותזות לגבי המסלולים המולקולאריים התורמים לשונות בין הגידולים, ומציעות מספר היפותזות חדשות.

5. **Identifying Functional Modules Using Expression Profiles and Confidence-Scored Protein Interactions**

Igor Ulitsky and Ron Shamir

Published in *Bioinformatics* [5].

השיטות השונות לשילוב נתונים של ביטוי גנים עם רשתות קשרי חלבונים, מושפעות לרעה על ידי השונות המשמעותית באמינותם של הקשרים ברשת. בעבודה זו, אנו מציגים את CEZANNE, שיטה המשתמשת במדדי אמינות של קשרי חלבון-חלבון למציאת קבוצות של גנים הקשורים זה לזה מבחינה פונקציונאלית וגם מבוטאים יחדיו. אנו מציעים מודל הסתברותי חדש המשמש כבסיס לקשתות של הרשת, כך שהסבירות שקבוצת הגנים יוצרת רכיב קשירות ברשת מיוצגת על ידי משקל החתך המינימאלי בגרף. אנו משתמשים ב-CEZANNE לניתוח נתונים על ביטוי גנים אחרי חשיפה של השמר *S. cerevisiae* לנזק דנ"א, ומזהים מגוון מודולים ידועים יחד עם מספר מודולים חדשים. בנוסף, השימוש בשיטה החדשה מאפשר לנו לזהות מספר תפקידים חדשים לגנים ספציפיים. לבסוף, אנו מראים שהביצועים של CEZANNE עולים על הביצועים של שיטות אחרות לניתוח מידע על ביטוי גנים ורשתות.

6. **Redundancy and Protein Essentiality Revealed in the *S. cerevisiae* Interaction Networks**

Igor Ulitsky and Ron Shamir

Published in *Molecular Systems Biology* [6].

בעבודה זו פיתחנו כלים אנליטיים חדשים לניתוח קשרים גנטיים בהקשר של קשרים פיזיים, תוך הרחבה של מודל שהוצע בספרות על ידי Kelley ו-Ideker. המודל שלנו מפרש רבים מהקשרים הגנטיים כקשרים המחברים בין מסלולים מולקולאריים מקבילים ברשת הפיזית. אנו משתמשים במודל הזה לניתוח רשתות בשמר *S. cerevisiae* ומגלים 140 מודלים של מסלולים מקבילים המסבירים יחדיו 3,765 קשרים גנטיים, מספר כמעט כפול מזה שדווח לפני כן. הגנים שנמצאים במודלים אלה נוטים להיות בעלי זמן מחצית חיים קצר יותר ולהכיל יותר אתרי זרחון, תכונות המצביעות על כך שגנים אלא עוברים רגולציה הדוקה בשל פעילותם במסלולים יתירים. בנוסף, אנו מזהים 'חלבוני ציר' בעלי קשרים פיזיים רבים לשני המסלולים המרכיבים מודל, ומראים שחלבונים אלה נוטים להיות חיוניים ושמורים. הניתוח שלנו שופך אור חדש על הארגון של מנגנוני התא, וחושף מספר תפקידים חדשים עבור גנים מסוימים.

7. From E-MAPs to Module Maps: Dissecting Quantitative Genetic Interactions Using Physical Interactions

Igor Ulitsky, Tomer Shlomi, Martin Kupiec and Ron Shamir

Published in *Molecular Systems Biology* [7].

בעבודה זו אנו מרחיבים את המודל שהצענו בעבודה הקודמת בשני היבטים. בעוד שהשיטה הקודמת יכלה לזהות רק זוגות של מודולים פונקציונאליים, השיטה החדשה מאפשרת למצוא סימולטנית אוסף כלשהו של מודולים. בנוסף, המודל החדש מתאים למידע כמותי על קשרים גנטיים, כולל קשרים חיוביים ושליליים. השתמשנו בשיטה זו למיפוי מודולים של המנגנונים המולקולאריים השונים הקשורים לכרומוזומים בשמרים, והראנו שהמפה מספקת רמזים חשובים לגילוי תפקידים של גנים בודדים ושל מודולים שלמים.