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 $V_{\mathrm{sim}}\subseteq V$ and a symmetric matrix S where S_{ij} is the similarity between v_i and v_j, where $v_i, v_j \in V_{\mathrm{sim}}$. The goal is to find disjoint subsets $U_1, U_2, \dots, U_m$, 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 $V \setminus V_{\mathrm{sim}}$ *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 $U \setminus U'$ 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 $U \setminus W$. 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

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

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

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

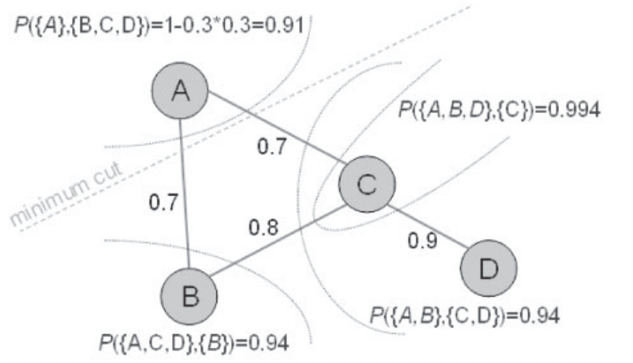

Fig. 1. A q-connected module for $q=0.9$. The numbers of the edges indicate edge probabilities. The probability of missing edges is 0. For every possible partition of the nodes into two sets, the probability that at least one true interaction connects the two sets exceeds 0.9. Four such partitions are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

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

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,

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

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	$1.2\cdot10^{-117}$	Ribosome	$5.9\cdot10^{-91}$
	Translation	$1.0\cdot10^{-85}$	Eukaryotic 43S preinitiation complex	$3.8\cdot10^{-49}$
	rRNA processing	$7.5\cdot10^{-79}$	Small nucleolar ribonucleoprotein complex	$1.5\cdot10^{-41}$
	35S primary transcript processing	$4.6\cdot10^{-44}$	DNA-directed RNA polymerase III complex	$3.1\cdot10^{-17}$
	Ribosome assembly	$4.3\cdot10^{-39}$	Exosome (RNase complex)	$4.4\cdot10^{-15}$
	Ribosomal large subunit biogenesis	$9.2\cdot10^{-14}$	DNA-directed RNA polymerase I complex	$5.7\cdot10^{-14}$
	rRNA modification	$4.4\cdot10^{-12}$	Noc complex	$3.2\cdot10^{-6}$
2 (38)	Protein catabolism	$1.8\cdot10^{-46}$	Proteasome complex (sensu Eukaryota)	$5.7\cdot10^{-71}$
	Proteolysis	$9.0\cdot10^{-44}$	Proteasome core complex (sensu Eukaryota)	$9.4\cdot10^{-32}$
	Ubiquitin cycle	$1.1\cdot10^{-42}$		
3 (12)	Histone acetylation	$3.6\cdot10^{-13}$	Histone acetyltransferase complex	$2.1\cdot10^{-12}$
	Chromatin modification	$5.9\cdot10^{-11}$		
	Transcription from RNA polymerase II promoter	$1.4\cdot10^{-6}$		
4 (12)	Translation	$1.1\cdot10^{-14}$	Ribosome	$1.4\cdot10^{-15}$
5 (12)	Nuclear mRNA splicing, via spliceosome	$3.5\cdot10^{-21}$	Spliceosome complex	$3.5\cdot10^{-17}$
			Small nuclear ribonucleoprotein complex	$2.5\cdot10^{-15}$
6 (10)	Barbed-end actin filament capping	$4.8\cdot10^{-6}$	F-actin capping protein complex	$4.8\cdot10^{-6}$
	Endocytosis	$1.1\cdot10^{-5}$		
	Cytoskeleton organization and biogenesis	$2.8\cdot10^{-5}$		
7 (8)	Establishment and/or maintenance of chromatin architecture	$1.1\cdot10^{-5}$	Chromatin remodeling complex	$4.6\cdot10^{-6}$
8 (7)	Glycogen metabolism	$3.0\cdot10^{-8}$	Protein phosphatase type 1 complex	$3.3\cdot10^{-5}$
	Sporulation (sensu Fungi)	$2.0\cdot10^{-6}$		
9 (6)	Translation	$1.1\cdot10^{-7}$	Ribosome	$4.0\cdot10^{-8}$
10 (6)	tRNA processing	$2.5\cdot10^{-14}$	Ribonuclease P complex	$9.2\cdot10^{-8}$
	rRNA processing	$2.2\cdot10^{-9}$		
11 (4)	Trehalose biosynthesis	$6.8\cdot10^{-14}$	Alpha, alpha-trehalose-phosphate synthase complex (UDP-forming)	$6.8\cdot10^{-14}$
12 (4)	Ubiquitin-dependent protein catabolism	$5.2\cdot10^{-7}$		
13 (3)	Pseudohyphal growth	$9.8\cdot10^{-7}$	cAMP-dependent protein kinase complex	$9.6\cdot10^{-7}$
14 (3)	Proteasome assembly	$3.2\cdot10^{-6}$		
	Protein folding	$3.9\cdot10^{-6}$		

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.

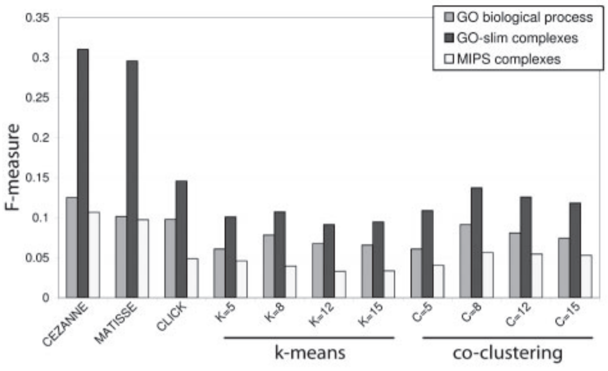

Fig. 2. Performance of several module finding methods. All GO annotations were used in the comparison. The *F*-measure evaluates sensitivity and specificity (see text).

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

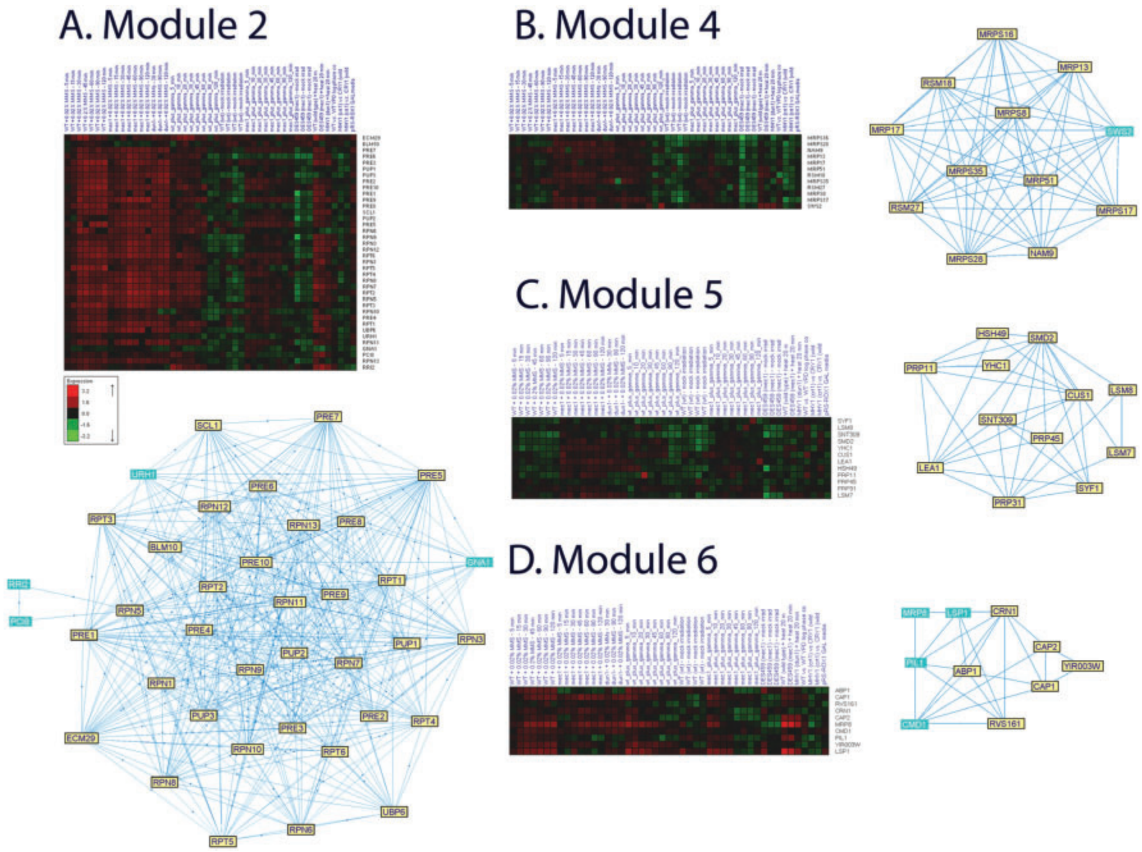

Fig. 3. Modules identified in *S. cerevisiae* response to DNA damage. For each module, the expression heat-map is presented together with the interaction network. In each subnetwork, the genes belonging to the dominant annotation are highlighted. (**A**) Members of the proteasome are in yellow. (**B**) 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) ($P = 4.8 \bullet 10^{-6}$), 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.