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

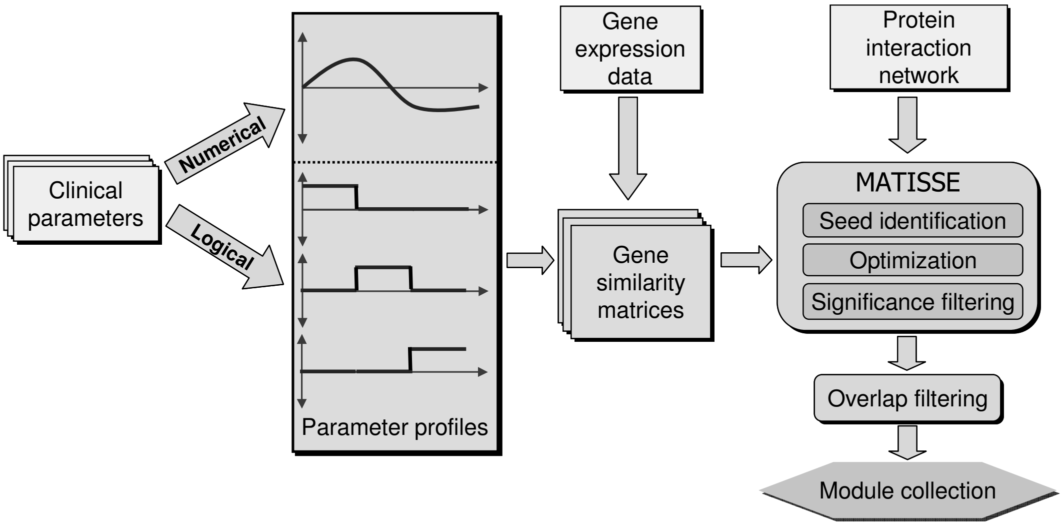

Fig. 1. Study outline. Clinical parameters are used to generate a collection of parameter profiles. The parameter profiles are used, together with gene expression data, to generate gene similarity scores. These scores, together with a protein interaction network serve as an input to MATISSE, which identifies 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.

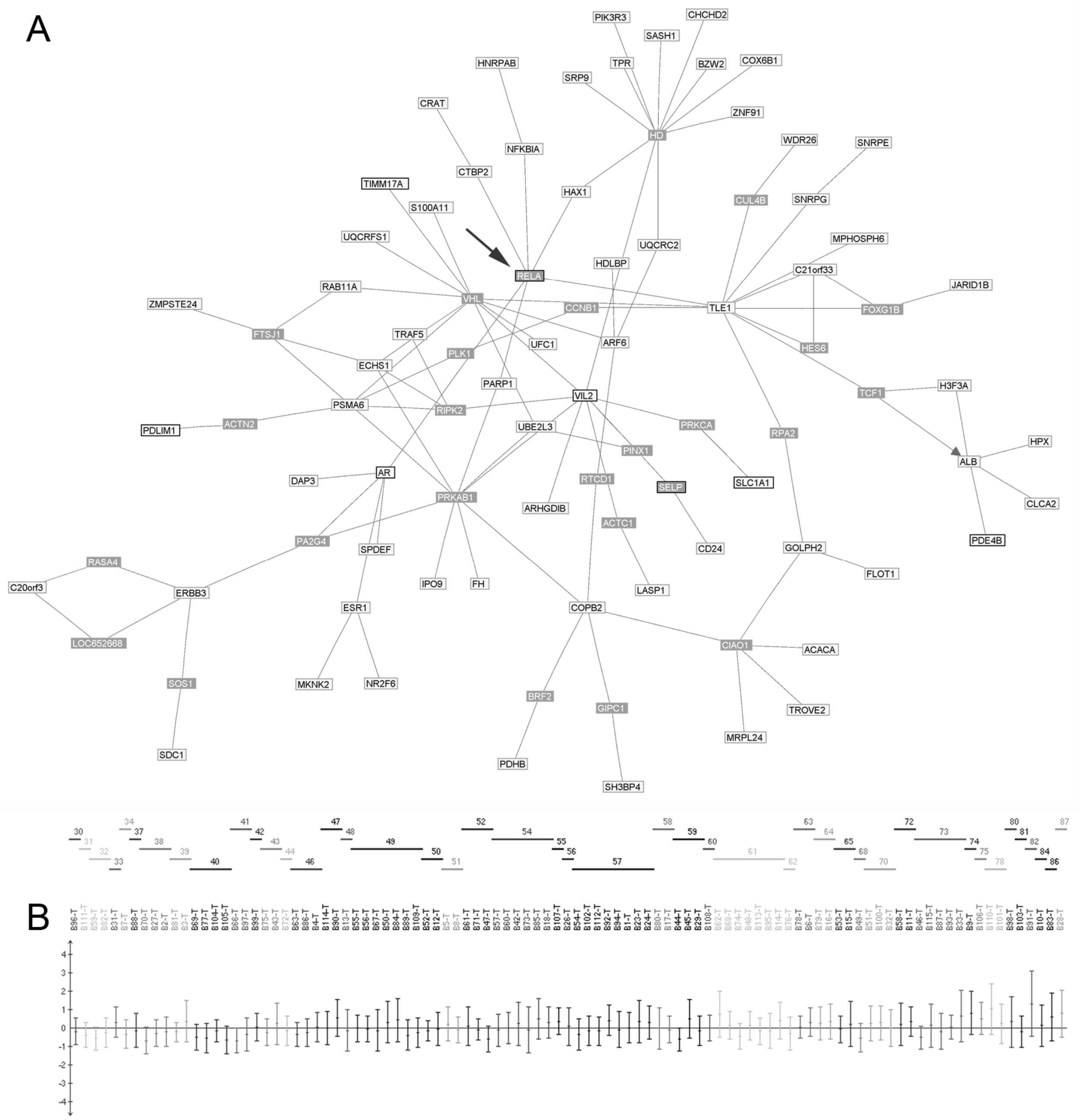

Fig. 2. BC-1 module related to age at diagnosis. (A) The subnetwork view of the module. Front nodes have a brighter background color. Gene overlapping the MSigDB RHESUS_AGING_UP category have thicker border. The arrow points at the RELA transcription factor. (B) Average expression levels of BC-1. Numbers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