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

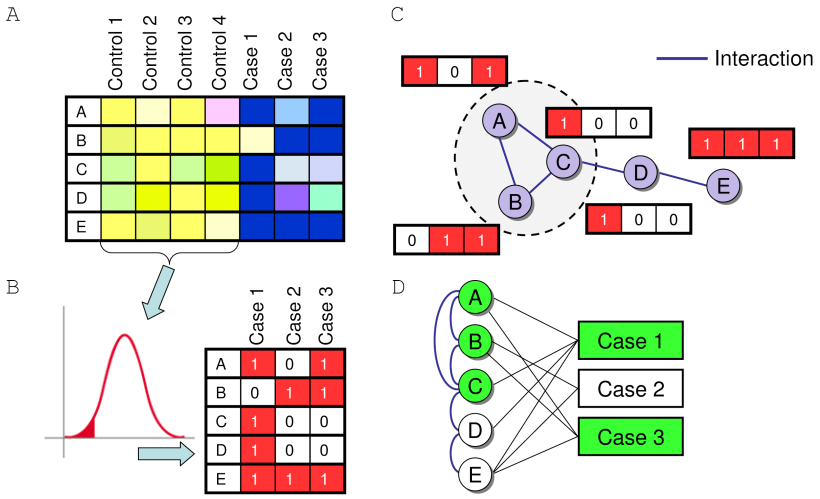

Fig. 1. From case-control profiles to dysregulated pathways. (A) The first input to our method is the gene expression matrix where the columns correspond to samples taken from case/control subjects and rows correspond to genes. (B) In a preprocessing step, differential expression is analyzed and, for 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

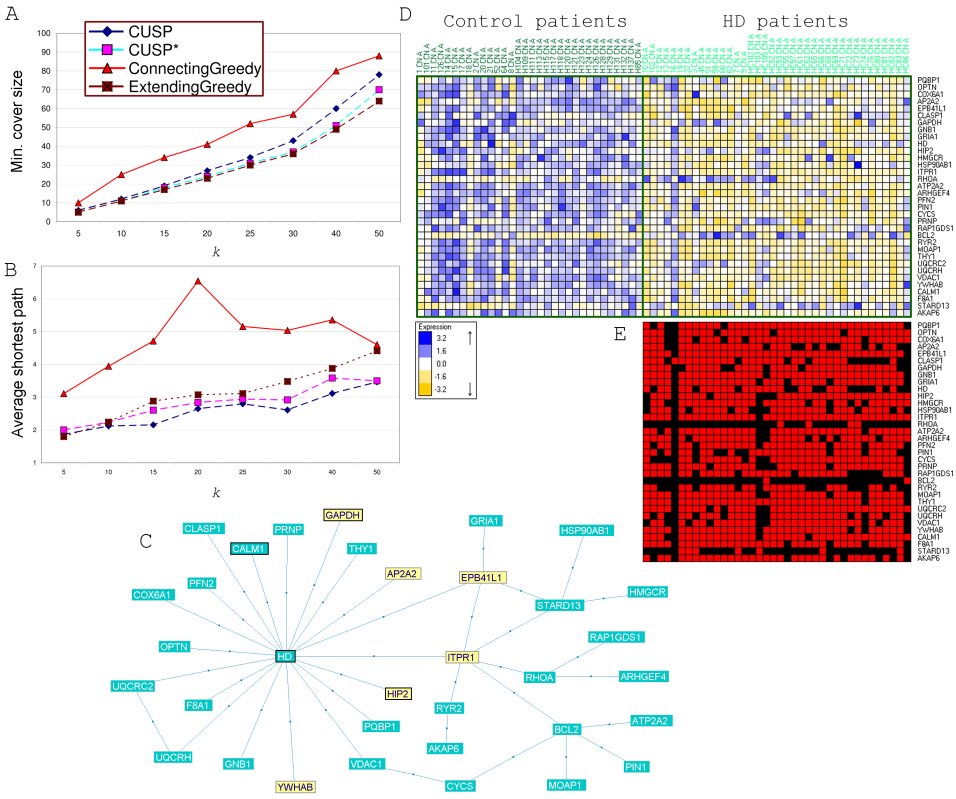

Fig. 2. Subnetwork identified by the CUSP algorithm as down-regulated in the caudate nucleus of HD patients. (A) Comparison of the minimal cover size obtained by the greedy and the CUSP algorithms. (B) Comparison of the average shortest path length between nodes in the minimal cover obtained by the greedy 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

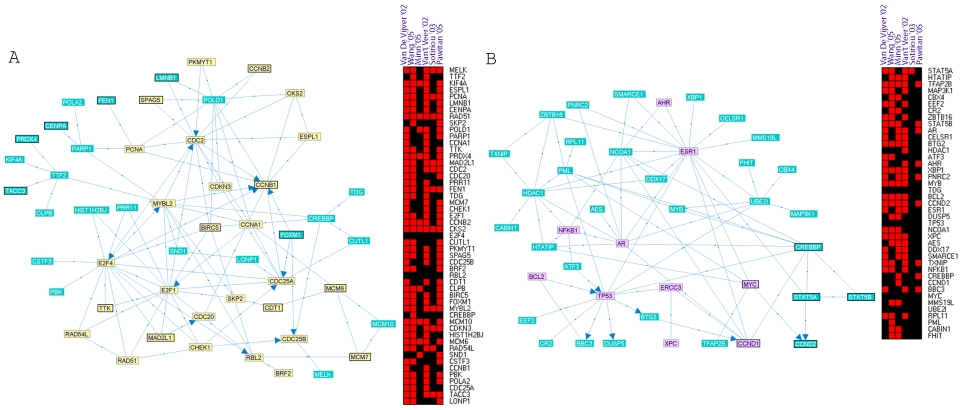

Fig. 3. **DPs identified in breast cancer meta-analysis**. In the differential expression maps (right) red cells correspond to differentially expressed genes. (A) a DP up-regulated in poor prognosis breast cancers ($k = 40, p < 0.005$). Cell cycle genes (from GO) are in yellow. YY1 regulated genes are 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)