Journal of Bioinformatics and Computational Biology
Vol. 10, No. 1 (2012) 1240002 (15 pages)
© Imperial College Press
DOI: 10.1142/S0219720012400021

Imperial College Press
www.icpress.co.uk

RANKING DIFFERENTIAL HUBS IN GENE CO-EXPRESSION NETWORKS

OMAR ODIBAT* and CHANDAN K. REDDY†

Department of Computer Science
Wayne State University
Detroit, MI 48228, USA
**odibat@wayne.edu*
†reddy@cs.wayne.edu

Accepted 23 November 2011
Published 12 January 2012

Identifying the genes that change their expressions between two conditions (such as normal versus cancer) is a crucial task that can help in understanding the causes of diseases. Differential networking has emerged as a powerful approach to detect the changes in network structures and to identify the differentially connected genes among two networks. However, existing differential network-based methods primarily depend on pairwise comparisons of the genes based on their connectivity. Therefore, these methods cannot capture the essential topological changes in the network structures. In this paper, we propose a novel algorithm, *DiffRank*, which ranks the genes based on their contribution to the differences between the two networks. To achieve this goal, we define two novel structural scoring measures: a local structure measure (*differential connectivity*) and a global structure measure (*differential betweenness centrality*). These measures are optimized by propagating the scores through the network structure and then ranking the genes based on these propagated scores. We demonstrate the effectiveness of *DiffRank* on synthetic and real datasets. For the synthetic datasets, we developed a simulator for generating synthetic differential scale-free networks, and we compared our method with existing methods. The comparisons show that our algorithm outperforms these existing methods. For the real datasets, we apply the proposed algorithm on several gene expression datasets and demonstrate that the proposed method provides biologically interesting results.

Keywords: Differential analysis; differential networking; connectivity; centrality; shortest paths; co-expression networks; differential hubs.

1. Introduction

Microarray studies are used to measure the expression level of thousands of genes under different conditions. These conditions could be different tissue types (normal versus cancerous),[1,2] different stages of cancer (early stage versus developed stage)[3] or different time points.[4] Differential analysis of networks has led to important results in studying the phenotypic differences across different conditions.[5] The set of genes that causes network topological changes may serve as biomarkers[6] and it can provide insights into disease-specific alterations.[7]

The goal of differential network analysis is to identify the differentially connected genes (or differential hubs). Although this type of analysis focuses on identifying single genes as differential hubs, the correlation between each gene and with the other genes is considered rather than testing each gene individually as in the differential expression (DE)[8] and the differential variability (DV)[9] methods. Both DE and DV methods depend on statistically testing each gene individually using the t-test and the F-test, respectively. Therefore, these methods do not capture the relationships between the genes. To overcome these problems, networks have been successfully used to model the gene activities and their interactions. These networks consist of genes as the nodes and the interactions between them as the edges. Studying the topology and functionality of these networks can provide valuable knowledge for understanding the roles of genes in several diseases.[7]

The main challenge in the differential network analysis is to identify the important differences between two networks. A naive solution is to transfer this problem to solving the subgraph isomorphism problem. Unfortunately, it was shown that solving the subgraph isomorphism problem is an NP-complete problem.[10]

To compare the genes between two gene networks, several differential measures such as differential connectivity have been defined in Refs. 5, 11 and 12, some methods are based on performing permutations and statistical tests such as the MDA test.[4] However, most of these methods depend on pair-wise comparisons of the genes based on their degrees. Therefore, we propose an efficient algorithm to capture all the local and global changes between two networks.

In this paper, we propose a new differential network analysis algorithm (*DiffRank*) that can overcome these drawbacks. The proposed method captures the changes in the edges (local changes) and the change in the centrality of each gene (global changes). As an example, two networks are shown in Fig. 1. In this example, it can be seen that gene 4 should be identified as the differential gene when comparing network A and network B. However, this gene has the same degree (which is 3) in both networks. Therefore, depending only on the comparison of the degree of each gene cannot capture all the differences between two gene

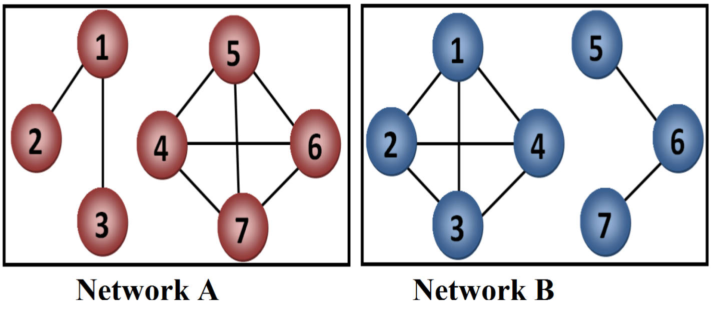


Fig. 1. A simple illustration of differential hubs.

networks. Using the proposed method, gene 4 will be the top-ranked differential gene in Fig. 1.

In this paper, we propose *DiffRank* as an efficient and approximate solution to rank the genes based on their contribution in the differences between two gene networks. We propose two new measures for each node: *differential connectivity* and *differential centrality.* These measures are propagated through the network and are optimized to capture the topological changes between two networks. We show the performance of the proposed algorithm on some synthetic examples, and we develop a simulator for generating synthetic differential scale-free networks to evaluate the proposed algorithm and to compare it with other methods. For the real-world datasets, we use four cancer datasets and show the functional enrichment analysis on all the four datasets. We also illustrate the significance of the proposed algorithm by showing the overlap between our results and some results published in the literature.

2. Method

2.1. *Preliminaries*

Given two gene networks, represented by graphs $G^A(V, E^A)$ and $G^B(V, E^B)$, where V is the set of N nodes and E^c is the set of edges in G^c, $c \in \{A, B\}$. An edge between two genes u and v, with a weight $w^c(u, v)$ in G^c, determines the strength of the interaction between the genes. The weight of each edge must be a non-negative value, 0 if the nodes are not connected to each other, or 1 in unweighted graphs. We denote the degree of gene v in network c as k_v^c. The proposed algorithm can be applied on both directed and undirected networks. In this work, we focus our discussion on undirected networks.

Given two networks, G^A *and* G^B*, the goal is to find the top differential genes that best explain the differences between the networks. The output is a vector* $\Pi = \langle \pi_1, \pi_2, \ldots, \pi_N \rangle$*, where* π_v *denotes the rank of the differential gene* v*.*

2.2. *Differential measures*

The proposed model is composed of two measures: *differential connectivity* and *differential betweenness centrality.* These measures are optimized to capture the changes in the local structure and the changes in the global structure between two the networks, respectively.

2.2.1. *Differential connectivity*

Genes with the highest number of edges, known as hubs, play central roles in the analysis of networks. Differential connectivity measures the local differences between two networks, G^A and G^B, by considering the actual weights of all the edges, and it is defined as follows:

$$\Delta C^i(v) = \sum_{u=1}^{N} \frac{|w^A(u, v) - w^B(u, v)| \cdot \pi_u^i}{\sum_{z=1}^{N} |w^A(u, z) - w^B(u, z)|}, \tag{1}$$

where π_v^i is the differential scores (or rank) of node v at the ith iteration. It is initialized to $\frac{1}{N}$ and will be updated in each iteration (it can also be used to incorporate prior knowledge). If a given gene has the same set of edges in both networks with the same weights, then the differential connectivity of that node will be 0. On the other hand, when a node has different sets of edges (such as gene 4 in Fig. 1), it will get a high value for the differential connectivity. In addition to the number of edges and their weights, the differential connectivity of each gene also depends on the differential scores of the neighbors it is connected to. A gene will be assigned a higher score if it is connected to many differential genes. Given two genes, u and v, the propagation of the differential score from u to v depends on three factors:

(1) The weight of the edge (u, v) in both networks, denoted by $|w^A(u,v) - w^B(u,v)|$.
(2) The current score of the gene u, denoted by π_u^i.
(3) The weights of all the edges connected to u, denoted by $\sum_{z=1}^{N} |w^A(u,z) - w^B(u,z)|$.

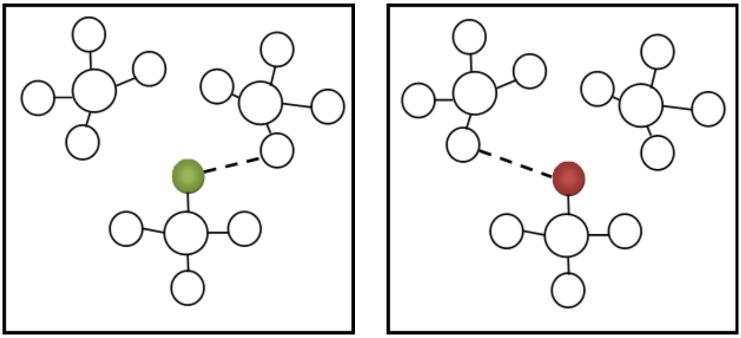
2.2.2. *Differential centrality*

Centrality is an important measure in understanding biological networks because it is difficult to detect the changes in the expression level of the central genes by single gene analysis. However, these changes could significantly alter the topology of the network.[13] Hence, we integrate the notion of gene centrality into the proposed algorithm.

Betweenness Centrality (BC) can be used to measure the centrality of each node, which is proportional to the sum of the shortest paths passing through it.[14] If P_{st} is the number of the shortest paths from node s to node t, where $s \neq t$, and $P_{st}(v)$ is the number of the shortest paths from s to t that pass through a node v, where $s \neq v$ and $t \neq v$, then the BC of the node v can be computed as $BC(v) = \sum_{s \neq t} \frac{P_{st}(v)}{P_{st}}$.[13] In gene co-expression networks, the weights of the edges represent the correlation between the genes. Therefore, distance values should be calculated from the correlation values in order to calculate the shortest paths. For example, if $w(u, v)$ is the correlation between two genes, then the distance between the two genes could be computed as $1 - w(u, v)$.

Comparing the values of BC may not detect the topological changes. For example, the shaded gene in Fig. 2 has the same value of BC (which is 6) in both networks. However, the shortest paths that pass through that gene are different. Therefore, we propose to consider the shortest paths in our method. Let SP_v^c be a binary $N \times N$ matrix, such that $SP_v^c(s,t) = 1$ if one of the shortest paths from s to t passes through the node v in network $c = \{A, B\}$, where $s \neq t$, and it is 0 otherwise. We define differential betweenness centrality of a node v as follows:

$$\Delta BC(v) = \sum_{s=1}^{N} \sum_{t=1}^{N} |SP_v^A(s,t) - SP_v^B(s,t)|. \tag{2}$$



(a) Network A (b) Network B

Fig. 2. A simple illustration for differential betweenness centrality.

2.3. *The DiffRank algorithm*

We propose *DiffRank* algorithm, which iteratively optimizes an objective function that is a linear combination of differential connectivity and differential betweenness centrality (parameterized by λ) within a PageRank-style framework,[15] such that the rank of each node v is computed as follows:

$$\pi_v^i = (1-\lambda) \cdot \frac{\Delta BC(v)}{\sum_{u=1}^{N} \Delta BC(u)} + \lambda \cdot \Delta C^i(v). \tag{3}$$

The parameter λ controls the trade-off between differential connectivity and differential betweenness centrality. It can be assigned any value in the range $[0, 1]$. When $\lambda = 0$, the ranking depends only on the differential betweenness centrality, and when $\lambda = 1$, the ranking depends only on the differential connectivity. Any other value of λ combines both terms in the ranking. In this paper, we set λ to 0.75 based on some of the preliminary experiments we performed. The integration of the ΔBC term into Equ. (3) adds significant global topological information to the differential analysis of networks.

2.4. *Condition-specific analysis*

It is important to find the genes that are differentially rewired in the cancer cells. For this purpose, we introduce a second version of the proposed algorithm based on the particular network of interest. To find the differential nodes in network B, the differential connectivity (ΔC) for each gene can be redefined as follows:

$$\Delta C'^i(v) = \sum_{u=1}^{N} \frac{\max(w_B(u,v) - w_A(u,v), 0) \cdot \pi_u^i}{\sum_{z=1}^{N} \max(w_B(u,z) - w_A(u,z), 0)}. \tag{4}$$

This new definition excludes any edge in the network of interest if the corresponding edge in the other network has a higher weight. Similarly, the new definition of differential betweenness centrality, ΔBC, includes the unique shortest paths that are in the network of interest and excludes the unique shortest paths in the other network.

$$\Delta BC'(v) = \sum_{s=1}^{N} \sum_{t=1}^{N} \max(SP_B^v(s,t) - SP_A^v(s,t), 0). \tag{5}$$

The second version of *DiffRank* is modified as follows:

$$\pi_v^i = (1-\lambda) \cdot \frac{\Delta BC'(v)}{\sum_{u=1}^{N} \Delta BC'(u)} + \lambda \cdot \Delta C'^i(v). \tag{6}$$

These two versions of *DiffRank* can solve the following problems:

(1) Find the top differential genes; this can be solved by the first version of *DiffRank*. In this version, we solve the phenotypic distinction problem.
(2) Find condition-specific differential genes; this can be solved by the second version of *DiffRank*. In this type of analysis, we focus on the set of genes that are active in the cancer networks (identifying disease-causing genes).

2.5. *Preservation and convergence*

To begin with, all the nodes are initialized to $\frac{1}{N}$ (uniform distribution), so that the sum of the rankings is 1 i.e. $\sum_{v=1}^{N} \pi_v^i = 1$. The rankings will be updated in each iteration. There is no need to normalize after each step since the sum of the rankings is preserved to unity.

Lemma 1. *The sum of the node ranks Π_Δ obtained by DiffRank is preserved to unity.*

Proof. Let us assume that the algorithm is at the iteration i and $\sum_{v=1}^{N} \pi_v^i = 1$. Now, we will show that the sum of the rankings is preserved for the next iteration $(i+1)$:

$$\begin{aligned}
\sum_{v=1}^{N} \pi_v^{i+1} &= \sum_{v=1}^{N} \left(\frac{(1-\lambda).\Delta BC(v)}{\sum_{u=1}^{N} \Delta BC(u)} + \lambda \cdot \sum_{u=1}^{N} \Delta DC^i(v) \right) \\
&= (1-\lambda) \cdot \left(\frac{\sum_{v=1}^{N} \Delta BC(v)}{\sum_{u=1}^{N} \Delta BC(u)} \right) \\
&\quad + \lambda \cdot \left(\sum_{v=1}^{N} \sum_{u=1}^{N} \frac{|w^A(u,v) - w^B(u,v)|.\pi_u^i}{\sum_{z=1}^{N} |w^A(u,z) - w^B(u,z)|} \right) \\
&= (1-\lambda) + \lambda \cdot \left(\sum_{u=1}^{N} \pi_u^i \frac{\sum_{v=1}^{N} |w^A(u,v) - w^B(u,v)|}{\sum_{z=1}^{N} |w^A(u,z) - w^B(u,z)|} \right)
\end{aligned}$$

$$= (1-\lambda) + \lambda \cdot \sum_{u=1}^{N} \pi_u^i$$
$$= (1-\lambda) + \lambda = 1.$$

□

One issue that needs to be resolved is handling the sinks (or isolated nodes). These nodes will be assigned uniform weighted edges to each other node in the network in order to ensure the convergence of the *DiffRank* algorithm.[16]

Theorem 1. *The result from the DiffRank model converges to a unique rank vector.*

Proof. Let us define $M^{N\times N}$ as a square matrix, such that

$$M_{uv} = \frac{|w^A(u,v) - w^B(u,v)|}{\sum_{z=1}^{N} |w^A(u,z) - w^B(u,z)|}.$$

We replace all rows with zeros by $\frac{1}{N}$. Now, M is considered to be a stochastic matrix in which the sum of each row is 1: $\sum_{v=1}^{N} M_{uv} = 1$, $1 \le u \le N$. Let P denote a vector of length N, such that

$$P_v = \frac{\Delta BC(v)}{\sum_{u=1}^{N} \Delta BC(u)};$$

then we will have $\sum_{v=1}^{N} P_v = 1$. Finally, we define a new matrix M' as follows:

$$M' = \lambda \cdot M + (1-\lambda) \cdot P^T.$$

The combination of the stochastic matrix M, and the vector P reduces the effect of the isolated nodes $\lambda \in [0,1]$. Now, the rank vector Π_Δ can be computed by solving the following eigenvector problem:

$${\Pi_\Delta}^T M' = {\Pi_\Delta}^T.$$

Since M' is a stochastic matrix, the *DiffRank* model is reduced to a personalized PageRank model for which a unique solution is guaranteed.[15,16] □

2.6. *Scalability*

While the differential connectivity is computed in a linear time, computing the differential centrality is time-consuming because it requires finding the shortest paths between the genes. Using the traditional Dijkstra's algorithm, computing the shortest paths between two nodes requires $O(m + n\log(n))$, where m is the number of links and n is the number of nodes in the graph and solving all-pairs shortest paths requires $O(nm + n^2 \log n)$ time and $O(n^2)$ space.[17] However, some recent methods have been proposed to reduce the computational overhead by using approximation methods,[17] which can potentially help in efficiently applying *DiffRank* on large-scale networks. In our previous work, we applied the *DiffRank* algorithm in other domains such as the co-authorship networks.[18]

3. Experiments on Synthetic Datasets

Given the ith gene, $k^A(i)$ and $k^B(i)$ are the connectivity of the ith gene in networks A and B, respectively;

(1) (ΔPR): As a baseline method, we used the difference between the scores computed by the PageRank algorithm[19] in the two networks and is defined as follows:

$$\Delta PR(v) = |PR^A(v) - PR^B(v)|, \tag{7}$$

where $PR^K(v)$ is the score for the gene v obtained by applying PageRank on network K.

(2) (DH): Differential Hubbing was defined based on the degrees of each gene as follows[12]:

$$DH(v) = K_i^A - K_i^B. \tag{8}$$

(3) (DC): Differential Connectivity was defined based on the degrees of each gene as follows[11]:

$$DC(v) = \log_{10}\left(\frac{K_i^A}{K_i^B}\right). \tag{9}$$

(4) ($DiffK$) is defined as follows[5]:

$$DiffK(v) = |K^A(v) - K^B(v)|, \tag{10}$$

where $K^A(v) = \frac{k^A(v)}{\max(k^A)}$ and $K^B(v) = \frac{k^B(v)}{\max(k^B)}$.

3.1. *Synthetic differential scale-free networks*

We developed a simulator to generate synthetic differential scale-free networks. Initially, we started with a small network as a seed and then followed the preferential attachment rule[20] in adding new nodes. This rule assumes the probability of receiving new edges increases with the increase in node degree. To generate two differential networks of size n, we start with the same seed for each network of size m; then we generate the remaining $n - m$ nodes for each network separately.

3.2. *Evaluation measures*

Since there is no standard measure for comparing two networks, we developed two evaluation measures, and we used the *Kendall's Tau* statistic[21] to measure the correlation between the evaluation measures and the ranking algorithms.

Local structure measure (M_L): This measure depends on comparing the edges of each node to find the differential genes. It is a local measure which is defined as follows:

$$M_L(v) = \sum_{u=1}^{N} [w^A(u,v) - w^B(u,v)]^2. \tag{11}$$

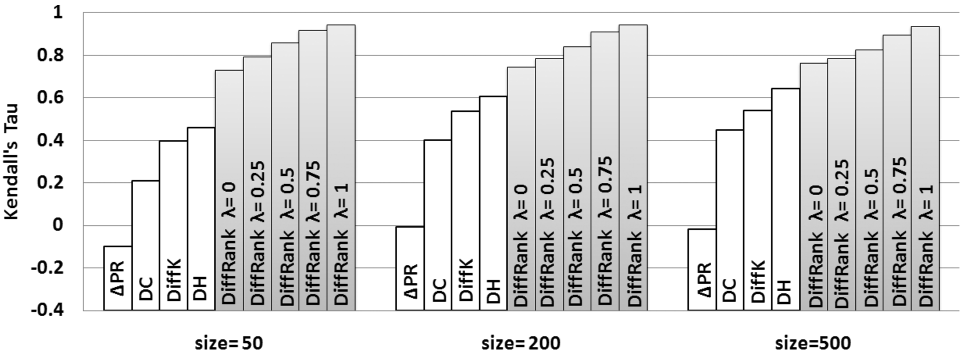


Fig. 3. Results on simulated networks evaluated based on the local measure (M_L).

Global structure measure (M_G): This measure captures the global changes in the gene networks and it uses the shortest paths in the computation as follows: Let us define $\text{dist}(u, v, G^c)$ to be the distance between the nodes u and v in graph G^c computed through the shortest path between them, and let $G_z^{c'}$ be the same as G^c except that all the edges for node z are removed. Then, we define $\Delta_z\text{dist}(u, v, G^c) = [\text{dist}(u, v, G^c) - \text{dist}(u, v, G_z^{c'})]^2$. Finally, M_G is defined as follows:

$$M_G(z) = \sum_{u=1}^{N} \sum_{v=1}^{N} [\Delta\text{dist}(u, v, G^A) - \Delta\text{dist}(u, v, G^B)]^2. \tag{12}$$

M_G measures the importance of each node to all other nodes in the network. It captures the contribution of each gene in the global structure of the network by considering the changes in the shortest paths between each pair of genes.

3.3. *Results on the simulated network datasets*

Figure 3 shows the results on the simulated data for different network sizes: 50, 200 and 500 evaluated using M_L. These results are the average of 10 runs. As shown in Fig. 3, it is obvious that as the value of λ increases from 0 to 1, better results are obtained. This is because the M_L measure depends only on the connectivity and does not include the centrality component. However, regardless of the value of λ, the *DiffRank* algorithm outperforms the other methods in all of the cases. Figure 4 shows the results of the simulated data for different network sizes: 50, 200 and 500 evaluated using M_G. These results are the average of 10 runs. Again, regardless the value of λ, the *DiffRank* algorithm outperforms the other methods in all the cases.

4. Experiments on Real Datasets

Table 1 shows the four real-world datasets used in our experiments. For each dataset, we built a network for each class; then, we ran the proposed method on the two networks.

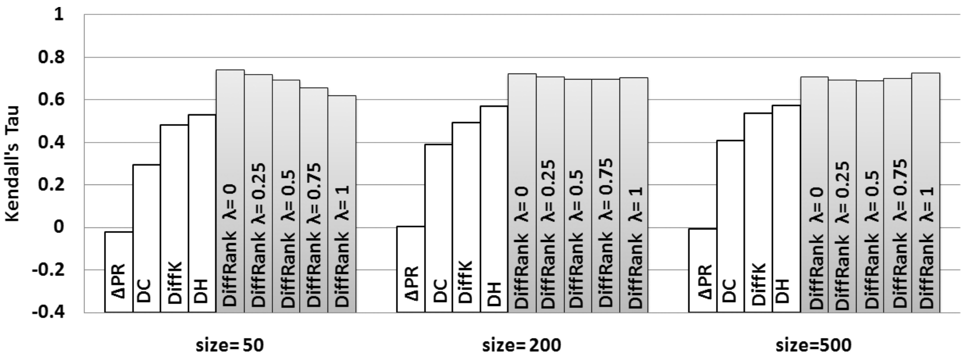


Fig. 4. Results on simulated networks evaluated based on the global measure (M_G).

Table 1. Description of the four gene expression datasets used in our experiments.

		Class A		Class B	
Dataset	Genes	Description	Samples	Description	Samples
Leukemia[22]	3051	AML	11	ALL	27
Medulloblastoma[23]	2059	Metastatic	10	Non-metastatic	13
Lung cancer[2]	1975	Normal	67	Tumor	102
Gastric cancer[1]	7192	Normal	8	Tumor	22

4.1. *Constructing the gene co-expression network*

Mutual Information (MI) can be used to measure the correlation between different genes, and it outperforms Pearson correlation and other linear measurements because it can capture nonlinear dependencies.[24] Therefore, we used MI to construct the gene networks. To find the threshold for the MI values, we followed the rank-based approach that was proposed in Ref. 25. The MI between each gene and all other genes are computed and ranked; then, each gene will be connected to the top d genes that are similar to it. Based on this approach, the minimum degree is d, the mean degree is between d and $2d$ and the maximum degree can be $N - 1$. There are two main advantages of this approach over the other value-based approaches[25]: First, the network will contain only reliable edges. Second, there will be no isolated nodes in the networks. We used $d = 5$, and the resulting networks for each class are given in Table 2. This table shows the minimum, the mean and the maximum of the degrees. However, it is worth mentioning that the proposed algorithm can be applied on any network regardless of the construction method used.

4.2. *Biological evaluation*

To evaluate the results of proposed algorithm, we used the DAVID functional annotation tool[26] to identify enriched biological GO terms and biological pathways of the top 100 ranked genes in each dataset, and we showed the top five biological

Table 2. Degree distribution of the networks built for our experiments.

Dataset	Class	Min	Mean	Max
Leukemia	AML	5	8.7	96
	ALL	5	8.8	120
Medulloblastoma	Metastatic	5	8.5	66
	Non-metastatic	5	9.0	743
Lung cancer	Normal	5	9.9	878
	Tumor	5	9.9	858
Gastric cancer	Normal	5	9.4	288
	Tumor	5	8.5	248

terms ranked based on their corrected p-values. In addition, we compared the top 100 ranked genes with the previously published results in the original papers from which we obtained the datasets.

4.3. *Results on gene expression datasets*

The top three differential genes from each dataset are shown in Table 3. In this table, we present the degrees of each gene in network A, network B and the common edges between the two classes. Table 4 shows the top five enriched biological terms for each dataset using the DAVID tool.[26]

(i) The Leukemia Dataset: The leukemia data contains the expression profiles of 3051 genes in 38 tumor samples. In this dataset, there are 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples.[22] For this dataset, we applied the version 1 of the proposed *DiffRank* algorithm. In addition to the functional enrichment analysis, we compared our results with the previously published results, and we found some differential genes, such as *M80254_at* (*CyP3*) and *M27891_at* (*Cystatin C*), were reported in Ref. 22 among the most highly correlated genes with AML-ALL class distinction.

Table 3. Top 3 differential genes obtained from the gene expression datasets.

Dataset	Rank	Gene Name	Degree in Class A	Degree in Class B	Common Edges
Leukemia	1	M26692_s_at	21	92	1
	2	X03934_at	120	5	1
	3	D87459_at	6	96	0
Medulloblastoma	1	196_s_at	5	743	3
	2	2008_s_at	5	709	2
	3	664_at	25	678	6
Lung cancer	1	MTHFR	15	659	11
	2	BAI1	84	492	52
	3	CSF1	530	851	496
Gastric cancer	1	HG1751HT1768_s_at	22	248	0
	2	M10098_5_at	123	224	7
	3	M11722_at	62	181	2

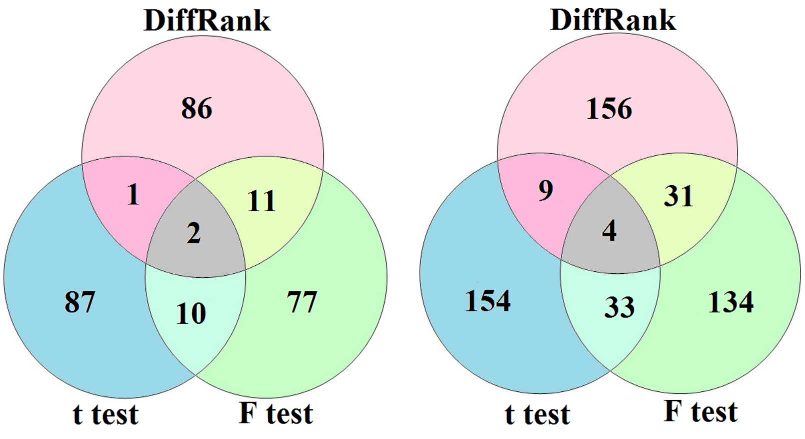
Table 4. Top five enriched biological terms obtained from the gene expression datasets.

Dataset	Term	Fold Enrichment	Corrected p-value
Leukemia	transmembrane protein	4.51	$2.9E-03$
	GO:0005829 cytosol	2.66	$1.1E-02$
	GO:0033273 response to vitamin	15	$1.8E-02$
	GO:0002520 immune system development	5.98	$2.3E-02$
	GO:0048534 lymphoid organ development	6.35	$2.8E-02$
Medulloblastoma	hsa05200:Pathways in cancer	4.83	$1.7E-06$
	kinase	5.47	$4.8E-06$
	ATP	9.75	$1.3E-05$
	domain:Protein kinase	6.64	$1.9E-05$
	nucleotide-binding	3.22	$1.9E-05$
Lung cancer	acetylation	2.73	$2.3E-06$
	Proto-oncogene	10.14	$3.2E-06$
	disease mutation	3.30	$4.1E-06$
	phosphoproteinr	1.71	$4.5E-06$
	nucleus	2.13	$4.9E-06$
Gastric cancer	GO:0005576 extracellular region	2.57	$1.3E-04$
	signal peptide	2.21	$1.3E-03$
	GO:0005615 extracellular space	3.59	$3.1E-03$
	disulfide bond	2.10	$3.5E-03$
	GO:0044459 plasma membrane part	2.0	$4.1E-03$

(ii) The Medulloblastoma Dataset: Medulloblastoma is a common malignant brain tumor of childhood. The medulloblastoma dataset[23] contains gene expression profiles of primary medulloblastomas clinically designated as either metastatic or non-metastatic. For this dataset, we applied the version 1 of the proposed *DiffRank* algorithm and found some statistically significant pathways such as: pathways in cancer, chemokine signaling pathway and mitogen-activated protein kinase (MAPK) signaling pathway, which have p-values $= 1.7E-06, 4.0 E-04$ and $1.0E-02$, respectively. The MAPK signal transduction pathway was reported as an upregulated pathway in the metastatic tumors that is relevant to the study of the metastatic disease.[23] In addition, some of the top differential genes were reported in Ref. 23 among the genes differentiating metastatic from non-metastatic tumors, such as *2042_s_at, 311_s_at* and *1001_at*.

(iii) The Lung Cancer Dataset: This dataset[2] contains the expression profiles of 1975 genes in normal and lung cancer samples. For this dataset, we applied the version 2 of the proposed *DiffRank* algorithm. When compared with the previously published results on the same dataset, we found that some of the top-ranked genes, such as {*CLDN14, PAX7, SDCBP, TADA3L, ITGA2B*}, were also reported in the differential patterns discovered by the subspace differential co-expression analysis proposed in Ref. 2.

(iv) The Gastric Cancer Dataset: The gastric cancer dataset[1] contains the expression profiles of 7192 genes in normal and gastric cancer samples. For this dataset, we applied the version 2 of the proposed *DiffRank* algorithm and found



(a) Top 100 genes.

(b) Top 200 genes.

Fig. 5. The overlap between the results of the *DiffRank* algorithm, the *t*-test and the F-test. The numbers are the averages of the four datasets (a) based on the top 100 genes in each method and (b) based on the top 200 genes in each method.

some of the top ranked genes such as *X51441_s_at* and *Y07755_at* had been reported as highly expressed genes in gastric tumors in Ref. 1.

4.4. *The relationships between DiffRank and other approaches*

The relationships between the top ranked genes from the *DiffRank* algorithm, DE (represented by the *t*-test) and DV methods (represented by the F-test) are shown in Fig. 5. The numbers in this figure are the averages of the rankings from the four datasets. As shown in Fig. 5, most of the genes identified by one approach cannot be identified by the other approaches. This fact explains why we found a few number of genes that were previously published and were top-ranked by our algorithm. Furthermore, some of the top-ranked genes have not been annotated yet. For example the top-ranked gene from the gastric dataset, $HG1751HT1768\_s\_at$, has no annotations according to the NCBI.[a] As shown in Table 3, this gene has 22 edges in the normal network and 248 different edges in the tumor network. From these numbers, one can observe that this gene may be involved in important biological processes relevant to the gastric cancer. Such genes can further be investigated.

5. Conclusion

In this paper, we proposed a novel differential networking algorithm to find the differential genes in gene networks that represent two biological conditions such as normal and cancer. The proposed algorithm, *DiffRank*, can effectively capture the local and the global changes in the topological structures between two given gene

[a] http://www.ncbi.nlm.nih.gov/

networks. The experiments on synthetic datasets show that the proposed algorithm is effective and outperforms various baseline methods, and the results on the gene expression datasets were evaluated using the DAVID functional annotation tool. The proposed method is independent of the network construction procedure and can be applied on both directed and undirected networks. Prior knowledge can be incorporated into our algorithm by assigning high scores to the set of relevant genes[27] rather than using a uniform distribution for the initialization of the ranking vector. We also plan to study *DiffRank* in the context of gene regulatory networks.

References

1. Hippo Y, Taniguchi H, Tsutsumi S *et al.*, Global gene expression analysis of gastric cancer by oligonucleotide microarrays, *Cancer Res* **62**:233−240, 2002.
2. Fang G, Kuang R, Pandey G *et al.*, Subspace differential coexpression analysis: Problem definition and a general approach, *Pac Symp Biocomput* pp. 145−156, 2010.
3. Odibat O, Reddy CK, Giroux CN, Differential biclustering for gene expression analysis, *Proc. ACM Conference on Bioinformatics and Computational Biology (BCB)*, Niagara Falls, NY, USA, pp. 275−284, 2010.
4. Gill R, Datta S, Datta S, A statistical framework for differential network analysis from microarray data, *BMC Bioinformatics* **11**:95, 2010.
5. Fuller TF, Ghazalpour A, Aten JE *et al.*, Weighted gene coexpression network analysis strategies applied to mouse weight, *Mamm Genome* **18**:463−472, 2007.
6. Zhang B, Li H, Riggins RB *et al.*, Differential dependency network analysis to identify condition-specific topological changes in biological networks, *Bioinformatics* **25**:526−532, 2009.
7. de la Fuente A, From 'differential expression' to 'differential networking' identification of dysfunctional regulatory networks in diseases, *Trends Genet* **26**:326−333, 2010.
8. Tusher VG, Tibshirani R, Chu G, Significance analysis of microarrays applied to the ionizing radiation response, *PNAS* **98**:5116−5121, 2001.
9. Ho JW, Stefani M, dos Remedios CG *et al.*, Differential variability analysis of gene expression and its application to human diseases, *Bioinformatics* **24**:i390−398, 2008.
10. Pržulj N, Biological network comparison using graphlet degree distribution, *Bioinformatics* **23**:e177−e183, 2007.
11. Reverter A, Ingham A, Lehnert SA *et al.*, Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer, *Bioinformatics* **22**:2396−2404, 2006.
12. Hudson NJ, Reverter A, Dalrymple BP, A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation, *PLoS Comput Biol* **5**: e1000382, 2009.
13. Francesconi M, Remondini D, Neretti N *et al.*, Reconstructing networks of pathways via significance analysis of their intersections, *BMC Bioinformatics* **9**:S9, 2008.
14. Freeman LC, A set of measures of centrality based on betweenness, *Sociometry* **40**:35−41, 1977.
15. Haveliwala TH, Topic-Sensitive PageRank: A context-sensitive ranking algorithm for web search, *IEEE Trans Knowl Data Eng* **15**:784−796, 2003.
16. Langville AN, Meyer CD, Deeper inside PageRank, *Internet Math* **1**:335−380, 2004.
17. Gubichev A, Bedathur S, Seufert S *et al.*, Fast and accurate estimation of shortest paths in large graphs, *Proc. 19th ACM Int. Conf. Information and Knowledge Management*, Toronto, ON, Canada, pp. 499−508, 2010.

18. Odibat O, Reddy CK, Mining differential hubs in homogenous networks, *Proc. Ninth Workshop on Mining and Learning with Graphs*, San Diego, CA, USA, 2011.
19. Page L, Brin S, Motwani R *et al.*, The PageRank citation ranking: Bringing order to the web, *Technical Report*, Stanford University, 1998.
20. Barabasi AL, Albert R, Emergence of scaling in random networks, *Science* **286**:509−512, 1999.
21. Lapata M, Automatic evaluation of information ordering: Kendall's Tau, *Comput Linguistics* **32**:471−484, 2006.
22. Golub TR, Slonim DK, Tamayo P *et al.*, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286**:531−537, 1999.
23. Macdonald TJ, Brown KM, Lafleur B *et al.*, Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease, *Nat Genet* **29**:143−152, 2001.
24. Torkamani A, Schork NJ, Identification of rare cancer driver mutations by network reconstruction, *Genome Res* **19**:1570−1578, 2009.
25. Ruan J, Dean AK, Zhang W, A general co-expression network-based approach to gene expression analysis: Comparison and applications, *BMC Sys Biol* **4**:8, 2010.
26. Huang da W, Sherman BT, Lempicki RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc* **4**:44−57, 2009.
27. Chen J, Aronow BJ, Jegga AG, Disease candidate gene identification and prioritization using protein interaction networks, *BMC Bioinformatics* **10**:73, 2009.

Omar Odibat received his B.S. in Computer Science from Yarmouk University in 2003 and M.S. from University of Jordan in 2005. He is currently a Ph.D. candidate at the Department of Computer Science at the Wayne State University. His research interests include data mining and bioinformatics. He is a student member of ACM and SIAM.

Chandan Reddy is an Assistant Professor in the Department of Computer Science at Wayne State University. He received his M.S. from Michigan State University and Ph.D. from Cornell University. His current research interests include data mining and machine learning with applications to bioinformatics, healthcare informatics and social networks. He published over 40 peer-reviewed articles in leading conferences and journals. He was a recipient of the best application paper award at the SIGKDD 2010 conference. He is a member of IEEE, ACM and SIAM.