Ranking Differential Genes in Co-expression Networks

Omar Odibat
Dept. of Computer Science
Wayne State University
Detroit, MI 48202
odibat@wayne.edu

Chandan K. Reddy
Dept. of Computer Science
Wayne State University
Detroit, MI 48202
reddy@cs.wayne.edu

ABSTRACT

Identifying the genes that change between two conditions, such as normal versus cancer, is a crucial task in understanding the causes of diseases. Differential networking has emerged as a powerful approach to achieve this task and to detect the changes in the corresponding network structures. The goal of differential networking is to identify the differentially connected genes between two networks. However, the current differential networking methods primarily depend on pair-wise comparisons of the genes based on their degrees in the two networks. Therefore, these methods cannot capture all the topological changes in the network structure. In this paper, we propose a novel differential networking algorithm, *DiffRank*, to rank the genes based on their contribution to the differences between two gene co-expression networks. To achieve this goal, we define two novel scoring measures: a local structure measure, *differential connectivity*, and a global structure measure, *differential betweenness centrality*. These measures are combined within a PageRank-style framework and optimized by propagating them through the network. Finally, the genes are ranked based on the their propagated scores. We demonstrate the effectiveness of *DiffRank* on several gene expression datasets, and we show that our method provides biologically interesting rankings.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining; J.3 [**Life and Medical Sciences**]: Biology and genetics.

General Terms

Algorithms, Design, Bioinformatics

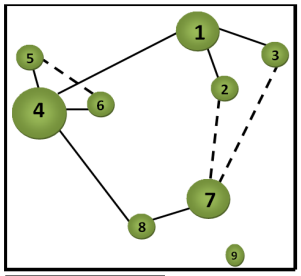
Keywords

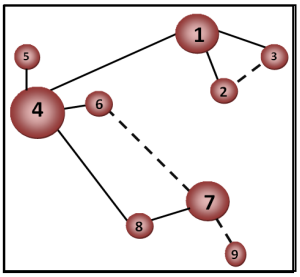
Differential network analysis, ranking, connectivity, centrality, shortest paths, co-expression networks.

ACM-BCB '11, August 1-3, Chicago, IL, USA

1. INTRODUCTION

Microarray studies are used to measure the expression level of thousands of genes under different conditions in different cells. These cells have the same set of genes, but the gene expression and their activities are different. There are several examples of such phenotypic variations: different tissue types: e.g., normal vs cancerous [1, 8], different class types: e.g., acute lymphoblastic leukemia (ALL) vs acute myeloid leukemia samples (AML) [6], different stages of cancer: early stage vs developed stage of prostate cancer [11] or different time points [5]. Differential analysis of networks has shown some promising results in studying the phenotypic differences across different conditions [4]. The set of genes which cause network topological changes may serve as biomarkers [20]. 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 [14]. Hence, we propose *DiffRank* as an efficient and approximate solution to find the differential genes in co-expression networks.





Network A	
gene	degree
4	4
1,7	3
2, 3, 5, 6, 8	2
9	0

Top genes based on comparing the degrees	5, 9
Top genes based on the proposed algorithm	7, 2, 3, 6

Network B	
gene	degree
4	4
1, 7	3
2, 3, 6, 8	2
5,9	1

Figure 1: A simple illustration of differential networks. Network A and network B have the same set of genes but different sets of edges. The solid edges are common in both networks, but the dashed edges exist only in one network. The size of the nodes represents the degree of each gene.

1.1 A Simple Illustration

Figure 1 shows a simple illustration of the concept of dif-

ferential network analysis using two unweighted and undirected networks. The top three hubs from network A based on their degrees are 1, 4 and 7, which are the same top three hubs in network B. However, for the differential analysis purpose, we would like to see the genes 7, 2, 3 and 6 ranked top in the list because they are responsible for the major differences between the networks. These genes are referred to as differentially connected genes or differential genes in short. Considering only the degree of nodes in each network individually is not an accurate measure for identifying the differential genes. As shown in Figure 1, gene 7 has the same degree in both networks, but the edges of gene 7 are different. Therefore, it is crucial to capture the changes in the edges and the changes in the centrality of each gene in differential networking. In this paper, we propose a novel differential network algorithm, DiffRank, to rank the genes based on their contribution in the differences between two gene networks. The proposed algorithm can effectively capture the local and the global changes in the topological structures between the two gene networks

1.2 Related Work

There are some differential networking methods that have been proposed in the literature. In [15], the degree distribution of each network was used to compare two gene networks. Edge-level comparison was used to identify sets of genes whose interactions are impacted by radiation exposure in mice [19]. To compare the genes between two gene networks, several differential measures such as differential connectivity have been defined in [16, 4]. Topological overlap of gene co-expression networks in human and chimpanzee brains was used to identify key drivers of evolutionary change [12]. It is also used by DiffCoEx [18] tool to identify differentially co-expressed modules between two conditions. Differential dependency network (DDN) [20] performs a permutation test to detect local topological changes in gene subnetworks, and Ryan et al. [5] proposed another statistical framework for differential network analysis.

The existing methods depend on pair-wise comparisons of genes in two networks, and they can not capture the global changes in the network structures. In this paper, we propose a new differential network analysis algorithm that can overcome these drawbacks. The proposed method captures the changes in the edges (local changes) and the changes in the centrality of each gene (global changes).

2. THE PROPOSED METHOD

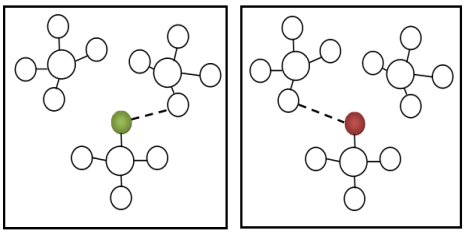
Given two gene networks, represented by graphs $G^A(V, E^A)$ and $G^B(V, E^B)$, where V is the set of N nodes and E^c is the set of edges in G^c, $c \in \{A, B\}$. An edge between two genes u and v, with a weight $w^c(u, v)$ in G^c, determines the strength of the interaction between the genes. We denote the degree of gene v in network c as k_v^c.

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 $\Pi =< \pi_1, \pi_2, ..., \pi_N >$, where π_v denotes the rank of the differential gene v.

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

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

Where π_v^i is the differential scores (or rank) of node v at the i^{th}iteration. It is initialized to $\frac{1}{N}$ and will be updated in each iteration. 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 7 in Figure 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 depends also 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.



(a) Network A (b) Network B

Figure 2: An illustration for differential centrality. The shaded gene has the same betweenness centrality value in both networks, but the paths that pass through that gene are different.

Differential Centrality: Centrality is an important measure in understanding biological networks because it is difficult to detect small changes in the expression level of the central genes. However, these changes could significantly alter the topology of the gene network [3]. Therefore, we integrate gene centrality in 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 [3]. 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 $BC(v) = \sum_{s \neq t} \frac{P_{st}(v)}{P_{st}}$ [3]. Comparing the values of BC may not detect the topological changes. For example, the shaded gene in Figure 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:

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

The proposed DiffRank algorithm is a combination of differential connectivity and differential centrality (parameterized by λ) within a PageRank-style framework [13], such

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

Dataset	Genes	class A		class B		Source
		Description	Samples	Description	Samples	
Leukemia	3051	AML	11	ALL	27	[6]
Medulloblastoma	2059	Metastatic	10	Non-metastatic	13	[10]
Lung cancer	1975	Normal	67	Tumor	102	[2]
Gastric cancer	7192	Normal	8	Tumor	22	[8]

Table 2: Degree distribution of the networks used in the experiments. This table shows the minimum, the mean and the maximum of the degrees.

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

that the rank of each node v is computed as follows:

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

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.

Finding the shortest paths is the most time-consuming computation in the proposed model. Using the traditional Dijkstra's algorithm, computing the shortest paths between two nodes needs $O(m + nlog(n))$ where m is the number of links, and n is the number of nodes in the graph and solving all-pairs shortest paths requires $O(nm + n^2logn)$ time and $O(n^2)$ space [7]. However, Recent methods have been proposed to reduce the computational overhead by using approximation methods [7], which helps in efficiently applying DiffRank on large-scale networks.

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:

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

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.

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

The second version of DiffRank is modified as follows:

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

Using the first version of DiffRank, we can find the top differential genes from two networks to solve the phenotypic distinction problem. The second version of DiffRank can be used to find condition-specific differential genes.

3. EXPERIMENTS ON REAL DATASETS

We used four gene expression datasets as described in Table 1. For each dataset, we built a network for each class; then, we ran the proposed method on the resulting two networks.

3.1 Constructing the Gene Networks

We used Mutual Information (MI) to measure the correlation between different genes in order to construct the gene co-expression networks. To find the threshold for the MI values, we followed the rank-based approach proposed in [17]. 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. First, the network will contain only reliable edges. Second, there will be no isolated nodes in the networks [17]. We used $d = 5$, and the resulting networks for each class are described in Table 2.

3.2 Biological Evaluation

We used the DAVID functional annotation tool [9] to identify enriched biological GO terms and biological pathways of the top 100 ranked genes in each dataset, and we show the top four biological 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.

3.3 Results

(i) Leukemia Dataset: The leukemia data contains the expression profiles of 3051 genes in 38 tumor samples. In this dataset, there are 27 ALL samples and 11 AML samples [6]. For this dataset, we applied the version 1 of the proposed DiffRank algorithm. The top 3 differential genes 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

Table 3: Top 3 differential genes obtained from each dataset.

Dataset	Rank	Gene	k^A	K^B	$k^A \cap k^B$
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: Functional enrichment analysis for the Leukemia cancer dataset.

Term	Count	FE	p-value
transmembrane protein	14	4.51	$2.9E-03$
GO:0005829 cytosol	21	2.66	$1.1E-02$
GO:0033273 response to vitamin	6	15	$1.8E-02$
GO:0002520 immune system development	10	5.98	$2.3E-02$
GO:0048534 lymphoid organ development	10	6.35	$2.8E-02$

two networks. The top 5 enriched biological terms are shown in Table 4 (FE stands for Fold Enrichment). 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 [6] among the most highly correlated genes with AML-ALL class distinction.

(ii) Medulloblastoma Dataset: This dataset [10] 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 DiffRank algorithm. The top 3 differential genes are shown in Table 3, and the top 5 enriched biological terms are shown in Table 5. We also found some significant pathways such as: *Pathways in cancer, Chemokine signaling pathway, MAPK signaling pathway* which have p-values$= 1.7E-06$, $4.0E-04$ and $1.0E-02$, respectively. The mitogen-activated protein kinase **MAPK** signal transduction pathway was reported as an up-regulated pathway in the metastatic tumors that is relevant to the study of the metastatic disease [10]. In addition, some of the top differential genes were reported in [10] among the gene differentiating metastatic from non-metastatic tumors, such as *2042_s_at, 311_s_at* and *1001_at*.

(iii) 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. The top 3 differential genes are shown in Table 3, and the top 5 enriched biological terms are shown in Table 6. More qualitatively, when compared with previous published results on the same dataset, we found that some of the top ranked genes such as *CLDN14, PAX7, SDCBP, TADA3L, ITGA2B* were also

Table 5: Functional enrichment analysis for the Medulloblastoma dataset.

Term	Count	FE	p-value
hsa05200:Pathways in cancer	19	4.83	$1.7E-06$
kinase	18	5.47	$4.8E-06$
ATP	11	9.75	$1.3E-05$
domain:Protein kinase	15	6.64	$1.9E-05$
nucleotide-binding	26	3.22	$1.9E-05$

Table 6: Functional enrichment analysis for the Lung cancer dataset.

Term	Count	FE	p-value
acetylation	37	2.73	$2.3E-06$
Proto-oncogene	12	10.14	$3.2E-06$
disease mutation	27	3.30	$4.1E-06$
phosphoproteinr	64	1.71	$4.5E-06$
nucleus	47	2.13	$4.9E-06$

reported in the differential patterns discovered by the subspace differential co-expression analysis proposed in [2].

(iv) Gastric Cancer Dataset: The Gastric cancer dataset [8] 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. The top 3 differential genes are shown in Table 3, and the top 5 enriched biological terms are shown in Table 7. We also found 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 [8]. Some of the top ranked genes have not been annotated yet. For example the top ranked gene, $HG1751HT1768\_s\_at$, has no annotations according to the NCBI[1]. As shown in Table 3, this gene has 22 edges in the normal network and 248 different edges in the tumor network. Such gene can further be investigated and validated.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel algorithm, DiffRank, to rank the differential genes when analyzing two gene networks that represent two biological conditions. The proposed algo-

[1] *http://www.ncbi.nlm.nih.gov/*

Table 7: Functional enrichment analysis for the Gastric cancer dataset.

Term	Count	FE	p-value
GO:0005576 extracellular region	31	2.57	$1.3E-04$
signal peptide	36	2.21	$1.3E-03$
GO:0005615 extracellular space	15	3.59	$3.1E-03$
disulfide bond	31	2.10	$3.5E-03$
GO:0044459 plasma membrane part	27	2.0	$4.1E-03$

rithm can effectively capture the local and the global changes in the topological structures between two gene networks. The proposed method is independent of the network construction method, and it can be applied on directed and undirected networks. In this paper, we illustrated the performance of the proposed method on the co-expression networks, and in the future we will study DiffRank in the context of gene regulatory networks (GRN). Moreover, the ranking obtained by DiffRank can be integrated into a module detection framework to obtain differential subnetworks.

5. REFERENCES

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–6750, June 1999.

[2] G. Fang, R. Kuang, G. Pandey, M. Steinbach, C. L. Myers, and V. Kumar. Subspace differential coexpression analysis: problem definition and a general approach. *Pacific Symposium on Biocomputing*, pages 145–156, 2010.

[3] M. Francesconi, D. Remondini, N. Neretti, J. Sedivy, L. Cooper, E. Verondini, L. Milanesi, and G. Castellani. Reconstructing networks of pathways via significance analysis of their intersections. *BMC Bioinformatics*, 9(Suppl 4):S9, 2008.

[4] T. Fuller, A. Ghazalpour, J. Aten, T. Drake, A. Lusis, and S. Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18:463–472, 2007.

[5] R. Gill, S. Datta, and S. Datta. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11(1):95, 2010.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct. 1999.

[7] A. Gubichev, S. Bedathur, S. Seufert, and G. Weikum. Fast and accurate estimation of shortest paths in large graphs. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 499–508, New York, NY, USA, 2010.

[8] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J.-M. Chong, M. Fukayama, T. Kodama, and H. Aburatani. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Research*, 62(1):233–240, 2002.

[9] D. W. a. . W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57, Dec. 2009.

[10] T. J. Macdonald, K. M. Brown, B. Lafleur, K. Peterson, C. Lawlor, Y. Chen, R. J. Packer, P. Cogen, and D. A. Stephan. Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nature Genetics*, 29(2):143–152, Oct. 2001.

[11] O. Odibat, C. K. Reddy, and C. N. Giroux. Differential biclustering for gene expression analysis. In *Proceedings of the ACM Conference on Bioinformatics and Computational Biology (BCB)*, pages 275–284. ACM, 2010.

[12] M. C. Oldham, S. Horvath, and D. H. Geschwind. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, 103(47):17973–17978, 2006.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[14] N. Prđulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[15] D. Remondini, B. O'Connell, N. Intrator, J. M. Sedivy, N. Neretti, G. C. Castellani, and L. N. Cooper. Targeting c-Myc-activated genes with a correlation method: Detection of global changes in large gene expression network dynamics. *Proc Natl Acad Sci U S A*, 102(19):6902–6906, 2005.

[16] A. Reverter, A. Ingham, S. A. Lehnert, S.-H. Tan, Y. Wang, A. Ratnakumar, and B. P. Dalrymple. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics*, 22(19):2396–2404, 2006.

[17] J. Ruan, A. Dean, and W. Zhang. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Systems Biology*, 4(1):8, 2010.

[18] B. Tesson, R. Breitling, and R. Jansen. Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11(1):497, 2010.

[19] B. H. Voy, J. A. Scharff, A. D. Perkins, A. M. Saxton, B. Borate, E. J. Chesler, L. K. Branstetter, and M. A. Langston. Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol*, 2(7):e89, 07 2006.

[20] B. Zhang, H. Li, R. B. Riggins, M. Zhan, J. Xuan, Z. Zhang, E. P. Hoffman, R. Clarke, and Y. Wang. Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*, 25(4):526–532, 2009.