Characterizing healthy and disease states by systematically comparing differential correlation networks in lung

Serene Wong, MSc, York University and Ontario Cancer Institute, Toronto

Nick Cercone, PhD, York University, Toronto

Igor Jurisica, PhD, University of Toronto and Ontario Cancer Institute, Toronto

Abstract

Comparing network structures that characterize healthy and disease state is an important problem as it provides insights to the underlying mechanisms and treatments for complex disease. However, it is intractable in general as it requires solving the subgraph isomorphism problem, which is NP-complete. We developed a heuristic algorithm to compare healthy and disease networks using local network properties, and the neighborhoods of their correlation difference networks. The algorithm identifies areas of difference between "healthy" and "disease networks" through their correlation difference networks. Thus, subgraph enumeration is only needed in identified areas. We have validated the proposed algorithm by analyzing three normal lung and lung tumor samples datasets. Results show the ability to identify differences between healthy and lung tumor networks that is consistent across multiple datasets.

Keywords:

Carcinoma, Non-Small-Cell Lung; Gene Expression Networks; Systems Biology

Introduction

Application of network properties to biological networks can bring forth important insights. One area is the development in the relationship between network topology and protein functions, or network topology and the underlining disease mechanism. Some studies, for example, Jeong et al. [1] suggested that most highly connected proteins are those that are most important to survival. Pržulj et al. [2] claimed that lethal proteins are not only highly connected, but they are articulation points. Other studies, for example, Jonsson et al. [3] provided insight of global network properties of cancer proteins, and found that cancer proteins, on average, had twice as many interacting partners as non-cancer proteins.

Comparing network structures that characterize healthy and disease state is an important problem as it provides insights to the underlying mechanisms and treatments for complex disease. However, it is infeasible to compare all aspects of large networks as it requires the solving of subgraph isomorphism problem, which is NP-complete [4]. Thus, heuristics for network comparison have arisen. There are two network comparison classes, global heuristics and local heuristics. Global heuristics use global network properties such as degree distributions and diameters to compare networks. However, global network properties do not contain the detail needed to capture the structural characteristics of biological networks [5]. Thus, more constraining local structure measurements have emerged [5]. Local heuristics use local network properties such as graphlets to measure networks similarities [4]. *Graphlets* are all non-isomorphic connected induced graphs on a certain number of nodes. In Fig. 1, all 3 to 5 node graphlets are shown. Two measures for comparing network similarities based on graphlets have developed, relative graphlet frequency distance (RGF-distance) [6] and graphlet degree distribution agreement (GDD-agreement) [4]. Both RGF-distance and GDD-agreement returns a scalar for the difference between two graphs, but we used graphlets to obtain network structure difference between healthy and disease states.

Although it is computationally expensive to compare all aspects of large networks, however, not all areas are needed to perform comparisons. Instead, methods can be used to locate important areas for comparing healthy graphs and tumor graphs. Thus, comparison of subgraphs can be done only on important areas as

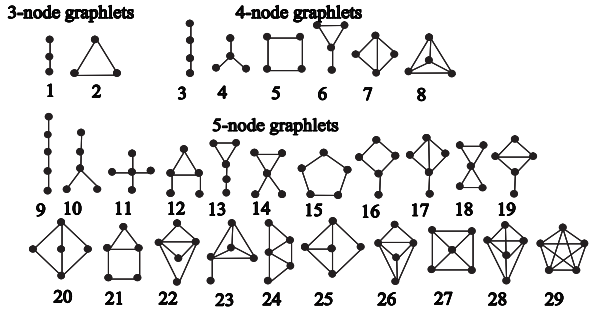

Figure 1- All 3 to 5 node graphlets. Fig. 1 of Modeling interactome: scale-free or geometric [6].

opposed to the entire graph. Since gene expression data provides much information about a disease state [7], we used gene expression datasets to capture healthy and disease states. With the use of an absolute correlation difference network (ACDN) on gene expression datasets, one can identify important areas in the network. The ACDN captures the difference in gene expression correlation values between the healthy and tumor graph, and thus can identify areas with significant difference, which may be biologically meaningful.

In our disease-driven approach, we compared healthy and disease networks based on the neighborhoods of their ACDNs [8]. With a threshold on the absolute correlation values in ACDNs, *x*, guided by both global and local network properties, an ACDN can be used as a backbone to accurately estimate the difference in network structures between healthy and disease networks. There is a trade-off between a large *x* – increased accuracy and increased computational demand, and small *x* – decreased accuracy and small computational demand due to covering fewer nodes. *x* can be set accordingly depending on resource availability and the amount of detail needed.

Methods

Datasets and construction of graphs

Gene expression datasets and gene signatures are the input to the method. Three non-small cell lung cancer (NSCLC) gene expression datasets [9, 10, 11] are used, and are referred to as Hou, Su, and Landi respectively in this paper. Datasets are chosen based on the number of healthy and tumor samples, and the balance between them. Refer to Table 1 for a more detailed characterization of the datasets. To focus on the most relevant genes, for each dataset, we consider genes that intersect with the eighteen validated prognostic NSCLC gene signatures. Two correlation matrices for each dataset, a normal and a tumor matrix, were generated using pairwise Pearson correlations for all gene pairs.

Given a healthy correlation matrix, *H*, and a tumor correlation matrix, *T*, an absolute correlation difference matrix, *D*, is generated. Let H_{ij}, T_{ij} represent the correlation value from the i^{th} row and j^{th} column in *H* and *T* respectively. *D* is defined as the following: $D_{ij} = |H_{ij} - T_{ij}|$. The *ACDN* is constructed by taking the top *x%* of gene pairs in *D*. The choice of *x* is guided by both global and local network properties: shortest path lengths and graphlet distributions. We found that even if *x* is set to be a high threshold, it is sufficient to serve as a backbone to accurately estimate the difference in network structures between healthy and disease graphs. For each vertex, *v*, in the *ACDN*, computes the 5-node graphlets that involves *v*. We use the neighborhood of *ACDN* to estimate the network structure differences between the healthy and tumor graphs.

Table 1 – Descriptions to datasets

Authors	GSE #	Title	Description
J. Hou et al. [9]	GSE19188	Expression data for early stage NSCLC	91 patients, 91 tumor and 65 adjacent normal lung tissue samples
L. J. Su et al. [10]	GSE7670	Expression data from Lung cancer	Pairwise samples from 27 patients
M. T. Landi et al. [11]	GSE10072	Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival	107 lung adenocarcinoma and normal lung samples, 58 tumor and 49 non-tumor tissues

Benchmark for evaluation

Healthy and tumor networks were generated using the top *1%* of the correlated gene pairs. Details of the healthy and tumor networks are found in Table 2. The current method used ACDNs to locate the difference between the healthy and tumor networks. We used Fanmod [12] for graphlet enumeration, and GraphCrunch 2 [13] for graphlet distributions.

Table 2 – Number of nodes and edges in healthy and tumor networks

Dataset	Healthy		Tumor	
	#Nodes	#Edges	#Nodes	#Edges
Hou	463	1956	416	1956
Su	423	1755	433	1755
Landi	321	1755	370	1755

Evaluation of performance was based on 3 categories, see Fig. 2:

1. HEALTHY: graphlets that are in the healthy graph only. The performance of this category is denoted as H_p, the total number of graphlets in the HEALTHY category is denoted as T_{gh}, the number of graphlets in the HEALTHY category obtained from the current method is denoted as C_{gh}.

$$H_p = \frac{C_{gh}}{T_{gh}} \times 100\%$$

2. BOTH: graphlets that are in the healthy and tumor graphs, but with structural differences. The performance of this category is denoted as B_p, the total number of graphlets in the BOTH category is denoted as T_{gb}, the number of graphlets in the BOTH category obtained from the current method is denoted as C_{gb}.

$$B_p = \frac{C_{gb}}{T_{gb}} \times 100\%$$

3. TUMOR: graphlets that are in the tumor graph only. The performance of this category is denoted as T_{up}, the total number of graphlets in the TUMOR category is denoted as T_{gt}, the number of graphlets in the TUMOR category obtained from the current method is denoted as C_{gt}.

$$T_{up} = \frac{C_{gt}}{T_{gt}} \times 100\%$$

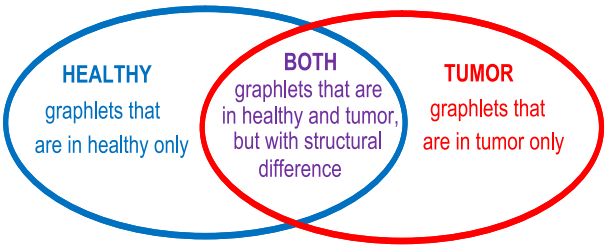

Figure 2- Evaluation of performance was based on 3 categories: HEALTHY, BOTH and TUMOR

Analysis

Three properties are observed in the analysis of ACDNs: 1) the relationship between the number of vertices and edges, 2) the shortest path lengths, and 3) the graphlet distributions.

Number of vertices and edges

Consistently from all three datasets we observe that as the threshold of the ACDN is lowered, the number of vertices increases much slower than the number of edges. Fig. 3 displays this observation across the 3 datasets. Thus, as the threshold of the ACDN is lowered, more edges will be added into the graph than vertices. Therefore, many newly edges interact with existing vertices.

Shortest path lengths

Another observation made among all 3 datasets is that there are many gene pairs with "long" shortest paths between them. As the threshold of the ACDN is lowered, there are fewer "long" shortest paths. For example, at 99.9 percentile, Landi has shortest paths with length 22. At 99.8 percentile, the longest shortest path in Landi is reduced to 14. When the threshold of the ACDN is lowered, many newly-added edges interact with existing vertices; thus, the distance between many gene pairs decrease. Fig. 4 displays this observation. Different bar colors represent different thresholds, for example, landi999 is Landi at 99.9 percentile.

Graphlet distributions

Yet another observation is that there is a peak in graphlet 10, denoted as G10, for all 3 datasets in their graphlet distributions for the 99.9 percentile ACDNs. The "v" shape of G10 is the backbone of many other higher numbered graphlets, for example, G13, G14, G26, G27, G28, G29 etc... If the 5 nodes remain, but

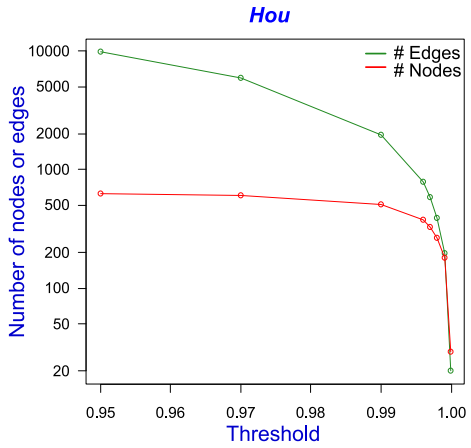

Figure 3a- number of nodes and edges for Hou

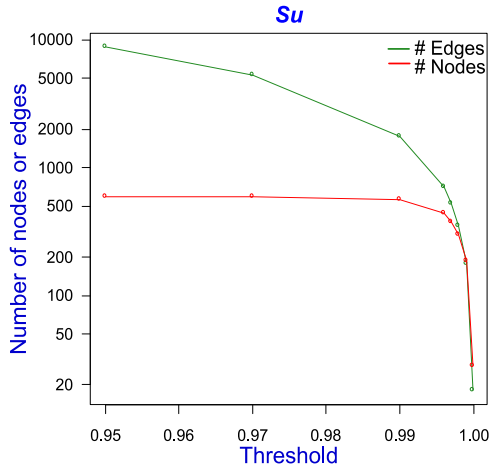

Figure 3b – number of nodes and edges for Su

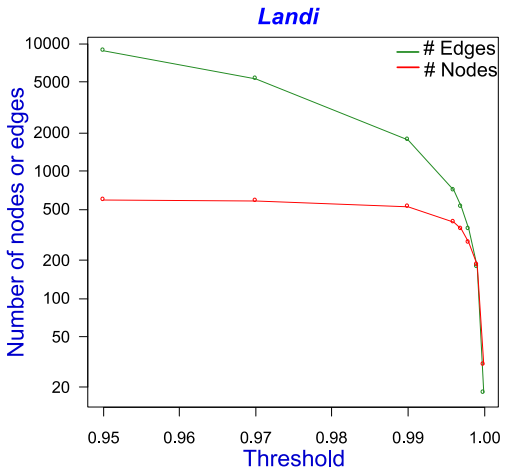

Figure 3c – number of nodes and edges for Landi

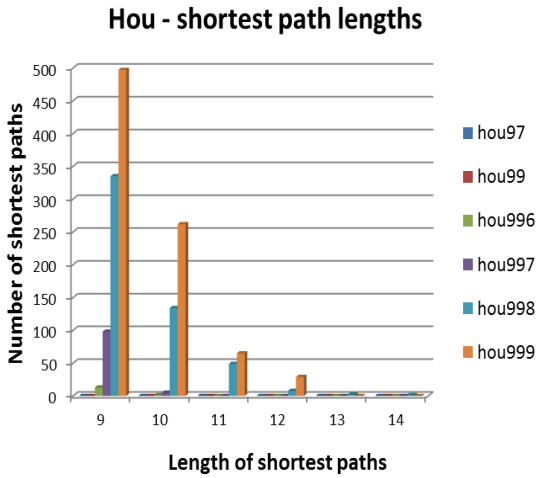

Figure 4a-shortest paths for Hou

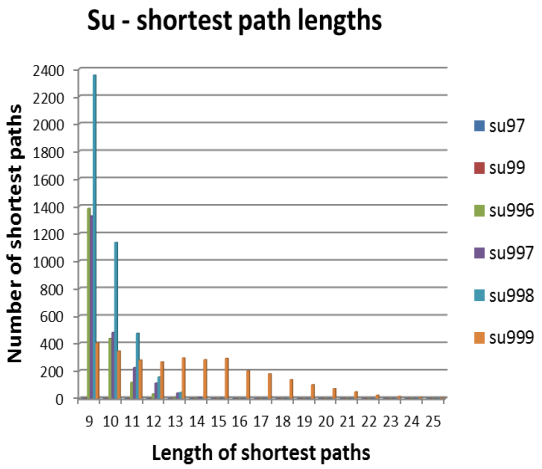

Figure 4b-shortest paths for Su

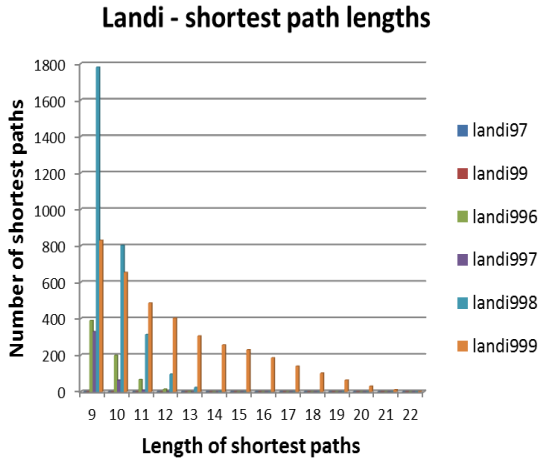

Figure 4c-shortest paths for Landi

with an addition of edges, then G10 will become other shapes. At 99.9 percentile, the ACDNs have many G10s, and "long" shortest paths. The long shortest paths extend a large area that contains difference between healthy and tumor conditions. The "v" shape of G10 branches off to a different local area. As the threshold is lowered, the number of nodes added into the ACDN is less than the number of edges. Many newly-added edges will be formed with existing nodes resulting in the increase number of graphlets above G10, and decrease of shortest path lengths. Therefore, at higher thresholds, the backbone of difference is already formed, and computing the neighborhood around the backbone is sufficient.

Results

Table 3 and 4 are the performance for the ACDN at the 99.9 percentile and 99.8 percentile respectively. At 99.9 percentile, Hou performed very well in all 3 categories of the evaluation performance. However, at 99.9 percentile, Su and Landi did not perform well at the BOTH category. At 99.9 percentile, Landi also did not perform well at the TUMOR category. This is explained by the fact that although they both have long shortest paths at the 99.9 percentile, they have a lot less G10 when compared to Hou. However, both Su and Landi performed well for all 3 categories at 99.8 percentile. At 99.8, both Su and Landi have a lot more G10s when compared with their 99.9 percentile graphlet distributions, refer to Fig. 6. Thus, both the shortest path lengths and graphlet distributions are needed for accurate estimation.

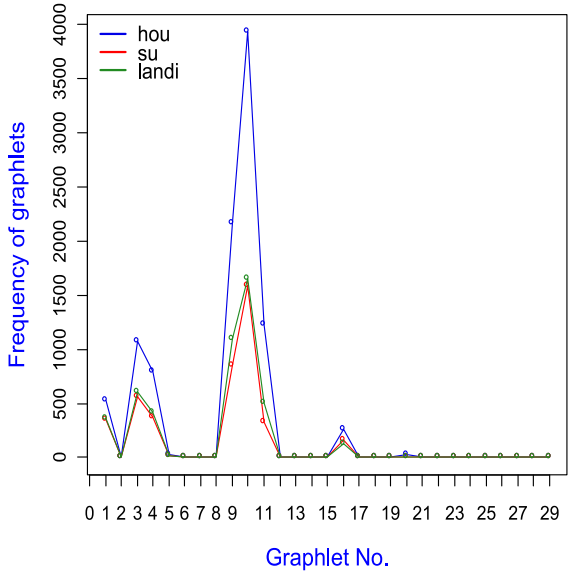

Figure 5- Graphlet distributions for the 99.9 percentile ACDNs for all 3 datasets

Discussions

In our disease-driven approach, we compared healthy and disease graphs based on the neighborhoods of their ACDNs. With a threshold guided by the shortest path lengths and graphlet distributions, an ACDN can be used as a backbone to accurately estimate the difference in network structures between healthy and disease graphs. In the 3 NSCLC datasets, we showed that it is sufficient to achieve accurate estimation in the difference between healthy and tumor states with 99.9 and 99.8 percentile of the ACDNs. This disease-driven approach is not designed to be specific to only NSCLC. Thus, this approach can be applied to other cancers or other diseases.

A future work would be to test the sensitivity of thresholds. Another future work would be to test this approach with larger normal and tumor graph size using the "divide and conquer" strategy. This is because the neighborhoods of vertices in the ACDN can be independently computed in parallel across different CPUs or GPUs. This is especially useful when the size of the normal or tumor graph is too large for subgraph enumeration algorithms to process the entire graph at one time. We anticipate that this is indeed an important property as co-expression graphs can be large, depending on the threshold set for constructing the normal and the tumor graphs. Furthermore, as the number of genes of interest increases, the normal and tumor graphs will grow in size.

Table 3 – Performance for the 99.9 percentile ACDNs for all 3 datasets

Dataset	# Nodes in ACDN	T_{gh}	C_{gh}	H_p	T_{gb}	C_{gb}	B_p	T_{gt}	C_{gt}	T_{up}
Hou	179	4182593	3666267	87.66	8468	8248	97.40	8577395	8520169	99.33
Su	186	6137000	5800794	94.52	26838	19767	73.65	9081990	8404783	92.54
Landi	183	15748654	13801528	87.64	505044	300586	59.52	12180103	6174288	50.69

Table 4 – Performance for the 99.8 percentile ACDNs for all 3 datasets

Dataset	# Nodes in ACDN	T_{gh}	C_{gh}	H_p	T_{gb}	C_{gb}	B_p	T_{gt}	C_{gt}	T_{up}
Hou	266	4182593	4007166	95.81	8468	8325	98.31	8577395	8574195	99.96
Su	301	6137000	6106593	99.50	26838	26633	99.24	9081990	9056966	99.72
Landi	277	15748654	15546347	98.72	505044	472264	93.51	12180103	10164804	83.45

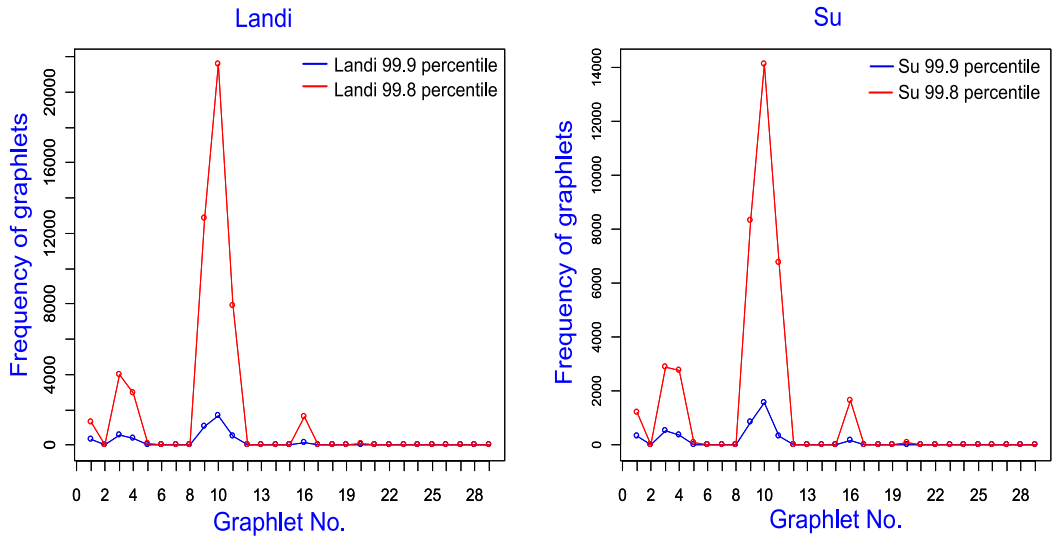

Figure 6- Graphlet distributions for the 99.9 and 99.8 percentile ACDNs for Landi and Su

References

[1] Jeong H, Mason SP, Barabási AL, and Oltvai ZN. Lethality and centrality in protein networks. Nature Brief Communications 2001: 411:41–42.

[2] Pržulj N, Wigle D, and Jurisica I. Functional topology in a network of protein interactions. Bioinformatics 2004: 20(3):340–348.

[3] Jonsson PF and Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics 2006: 22(18):2291–2297.

[4] Pržulj N. Biological network comparison using graphlet degree distribution. Bioinformatics 2007: 23(2):e177–e183.

[5] Pržulj N and Milenković T. Biological data mining, chapter Computational methods for analyzing and modeling biological networks. CRC Press, 2009.

[6] Pržulj N, Corneil DG, and Jurisica I. Modeling interactome: scale-free or geometric? Bioinformatics 2004: 20(18):3508–3515.

[7] Causton HC, Quackenbush J, and Brazma A. Microarray Gene Expression Data Analysis, chapter Introduction. Blackwell Publishing, 2003.

[8] Wong S, Jurisica I, and Cercone N. Systematic, comparative network analysis related to human disease. International Conference on Intelligent Systems for Molecular Biology 2011: poster.

[9] Hou J, Aerts J, den Hamer B, van Ijcken W, den Bakker M, Riegman P, van der Leest C, van der Spek P, Foekens JA, Hoogsteden HC, Grosveld F, and Philipsen S. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010: 5(4).

[10]Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC, Lin CH, Whang-Peng J, Hsu SL, Chen CH, and Huang CY. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. BMC Genomics 2007: 8(140).

[11]Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, Murphy SE, Yang P, Pesatori AC, Consonni D, Bertazzi PA, Wacholder S, Shih JH, Caporaso NE, and Jen J. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS one 2008: 3(2).

[12]Wernicke S and Rasche F. Fanmod: a tool for fast network motif detection. Bioinformatics 2006: 22(9):1152-1153.

[13]Kuchaiev O, Stevanović A, Hayes W and Pržulj N. GraphCrunch 2: Software tool for network modeling, alignment and clustering. BMC Bioinformatics 2011: 12:24.

Address for correspondence

Department of Computer Science and Engineering
York University, CSE 1003, 4700 Keele St.
Toronto, Ontario, Canada, M3J 1P3

and

Ontario Cancer Institute, UHN
101 College St., TMDT 9-305
Toronto, M5G 1L7, Canada