Cancer Biomarker Discovery from Gene Co-expression Networks Using Community Detection Methods

Raihanul Bari Tanvir
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
rtanv003@fiu.edu

Ananda Mohan Mondal
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
amondal@fiu.edu

***Abstract*— Finding the network biomarkers of cancers and the analysis of cancer driving genes that are involved in these biomarkers are essential for understanding the dynamics of cancer. Clusters of genes in co-expression networks are commonly known as functional units. This work is based on the hypothesis that the dense clusters or communities in the gene co-expression networks of cancer patients may represent functional units regarding cancer initiation and progression. In this study, RNA-seq gene expression data of three cancers - Breast Invasive Carcinoma (BRCA), Colorectal Adenocarcinoma (COAD) and Glioblastoma Multiforme (GBM) - from The Cancer Genome Atlas (TCGA) are used to construct gene co-expression networks using Pearson Correlation. Six well-known community detection algorithms are applied on these networks to identify communities with five or more genes. A permutation test is performed to further mine the communities that are conserved in other cancers, thus calling them conserved communities. Then survival analysis is performed on clinical data of three cancers using the conserved community genes as prognostic co-variates. The communities that could distinguish the cancer patients between high- and low-risk groups are considered as cancer biomarkers. In the present study, 16 such network biomarkers are discovered.**

Keywords—Cancer Biomarkers, Community Detection, Gene Co-expression Network, Survival Analysis.

I. INTRODUCTION

Gene expression profiles across samples can be highly correlated, and it is natural to describe their pairwise relationship using graph theoretic techniques or network analyses. Genes that correlate in terms of expression may form complexes, pathways, or participate in regulatory and signaling circuits [1]–[3]. Gene expression patterns, combined with statistical techniques, have been explored in many types of cancer [4]–[9].

There have been some works on the identification of recurrent patterns or biologically significant modules from gene co-expression networks (GCN). In a GCN, nodes represent genes, and edges represent that the pairs of genes connected by edges have significantly similar expression patterns over different samples. In other words, co-expression is a tool that infers edges in a gene network that is based on the "guilt by association" concept. Several methods exist for inferring edges in a GCN. Pearson correlation is the most common co-expression measure used in various studies [3][10]. Another standard method is Mutual Information (MI) [11], which is an information theoretic measure for measuring the nonlinear relationship between genes or other variables. Other notable methods are Spearman Rank Correlation, Euclidean Distance, Angle between a pair of observed expression vectors, and Gaussian graphical models [12]–[13]. A threshold is applied after constructing the co-expression network to retain the most biologically significant correlations between genes. The threshold can be soft [14] or hard [15] and is applied to produce a binary or weighted network.

While the GCN has topological structures reflecting real gene interactions, identifying groups of genes with dense interactions becomes a topic of research interest. These highly connected groups have a higher within-group homogeneity and can be considered as biologically significant modules performing a common task, such as shared regulatory inputs or functional pathways. Different methods are used to find these groups or modules. Lee et al. used hierarchical clustering to find functionally relevant modules [16]. Weighted Gene Correlation Network Analysis (WGCNA) is the most widely used package for finding modules [17] applying hierarchical clustering. It involves soft thresholding during the construction of a GCN. Tang et al. constructed a free scale GCN using WGCNA on gene expression data of breast cancer and found three hub genes [11]. Similarly, other researchers identified critical genes associated with different cancers such as breast, cervical, colon, esophageal, osteosarcoma, and ovarian cancer [18]–[25]. Few of them conducted survival analysis to validate those biomarkers.

To discover network biomarkers, other than GCNs, researchers also used protein-protein interaction (PPI) networks [26]–[28]. Though the PPI and co-expression networks are static in nature, these come with rich information about the dynamic processes such as the behavior of genetic networks in response to DNA damage [29], the prediction of protein subcellular localization [30]–[35], protein function [36], genetic interaction [37], and the process of aging [38].

Graph-theoretic methods are also applied in the analysis of GCN to achieve the same goal. In a recent work, Mondal et al. showed that clique-like and bipartite graphs can be used as

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

building blocks for disease initiation and progression at the protein or gene network level [39]. Using these building blocks, Tanvir et al. discovered network modules related to cancers from GCNs [40]. Shi et al. proposed a new algorithm named Iterative clique enumeration technique (ICE) to discover relatively independent maximal cliques for breast cancer on GEO dataset, and they found some highly correlated modules which may be capable of differentiating different tumor grades [41]. Similarly, Perkins et al. used spectral graph theory on Homo sapiens and Saccharomyces cerevisiae microarray data for clustering genes at various thresholds [42]. Using the centrality analysis method, Zhang et al. discovered the top five hub genes for bladder cancer [43]. Later they found 329 differentially expressed genes (DEG) that are significantly enriched by pathway analysis.

As GCN is a type of biological systems network, network science approach such as community detection can be applied to it as an analysis tool. Community detection in a network can be viewed as identifying the clusters of related nodes. Several studies were conducted on finding communities from biological networks. Feng et al. proposed MiMod, an algorithm for finding communities in GCN [44], which uses a divide-and-conquer strategy through application of biclustering method. Tripathi et al. [45] presented a framework that uses an adapting ensemble method using multiple community detection algorithms to find disease modules in heterogeneous biological networks. Wang et al. proposed Heuristic Graph Clustering Algorithm (HGCA) that selects seed nodes based on various topological characteristics and expands from those seed nodes to form communities from protein-protein interaction networks [46]. Kanter et al. propose a Python-based toolkit that uses the k-Nearest Neighbor to identify communities. Couturier et al. used Louvain, a community detection algorithm to detect subtypes of cancer from scRNA-seq data of five Glioblastoma patients [47].

Genes co-expressed across multiple samples are more likely to correspond to functional groups [16]. Different cancers share common characteristics [48], and conserved functions and genes related to these common characteristics can be found by analyzing GCNs of different types of cancer [49]. So, it is of considerable significance to study the common characteristics of patients with different types of cancer. This rationale motivates our work to find cancer biomarkers using multiple methods from GCNs. In a recent study, Yu et al. [50] used GCNs to mine conserved communities for cancers using Cytoscape plugin MCODE. The present study is the extended version of our previus work [51] and different from Yu et al. [50] in four different aspects: i) in constructing GCN, they used rank-based approach and we used a threshold on Pearson Coeffient, ii) to identify communities, they used Cytoscape plugin MCODE and we used six community detection algorithms, iii) to identify conserved community, they found community from only one cancer and then check the conserved behavior with other cancer but we found communities from each cancer and compared with others, and iv) in survival analysis, they used GGI (gene expression grade index) score to divide the patients into two groups whereas we used K-means clustering. The appraoch using GGI score has reproducibility issue as explained in survival analysis, *Section II-D*.

In this study, RNA-seq gene expression data of three cancers - BRCA, COAD and GBM - from The Cancer Genome Atlas (TCGA) are used to construct GCNs using Pearson Correlation with coefficient value greater than or equal to 0.70 as a threshold. Six well-known community detection algorithms are applied on these networks to identify communities with five or more genes. A permutation test is performed to further mine the communities that are conserved in other cancers, thus calling them conserved communities. Then survival analysis is performed on clinical data of three cancers using the conserved community genes as prognostic co-variates. The communities that could distinguish the cancer patients between high- and low-risk groups are considered as cancer biomarkers. In the present study, 16 such network biomarkers are discovered.

II. Materials and Methods

A. Data Preparation

The RNAseq gene expression and clinical data for three cancers – BRCA, COAD, and GBM - were collected from linkedomics.org [52], which contains well organized collection of multi-omics data of 32 cancers, curated from TCGA project. The number of genes and samples for BRCA, COAD and GBM are given in the Table 1. The gene expression dataset for each of these three cancers contained a large number of zero values which could lead to an erroneous outcome. In order to solve this issue, genes that have average of RPKM value less than equal to 0.3 are removed. This criterion was used in [53]–[55] as a preprocessing step. The number of genes that fulfills this criterion and the remaining number of genes for each cancer are given in the Table I.

Table I: Summary of gene expression data for three cancers

Cancer	# of samples	# of genes	# of genes with avg (RPKM) ≤ 0.3	# of remaining genes
BRCA	1093	20155	1480	18675
COAD	379	19828	1557	18271
GBM	153	19660	1298	18362

B. Gene Co-expression Network Construction

Pearson's Correlation was used to construct the GCN. It computes the correlation coefficient between each pair of genes based on gene expression values. These coefficient values, ranging between -1 to +1, provide information about how two genes are correlated with each other towards causing cancer in terms of their expression values. The pairs of genes are filtered using the criteria of having the absolute value of Pearson Correlation Coefficient (PCC) greater than or equal to 0.70. This threshold is based on the studies by Mondal and Hu [30][32][35], which showed that GCN with PCC $\geq$ 0.7 produces the best results in predicting protein localization. A pair of genes having PCC $\geq$ 0.7 means they are strongly correlated. Table II shows the topology of GCNs derived from three cancers.

It is clear from this table that number of connected components (CC) in each of the co-expression network is very high, most of which are pairs and triples as shown in Figure 1.

Table II: Network Topology of GCNs for three cancers. CC represents connected component.

Cancer	BRCA	COAD	GBM
# of Nodes	6613	6387	8655
# of Edges	99189	269027	194165
# of CCs	539	472	345
# of CCs > 4	37	35	23

But the number of components with more than four edges is much lower, which are used for finding the biomarker communities.

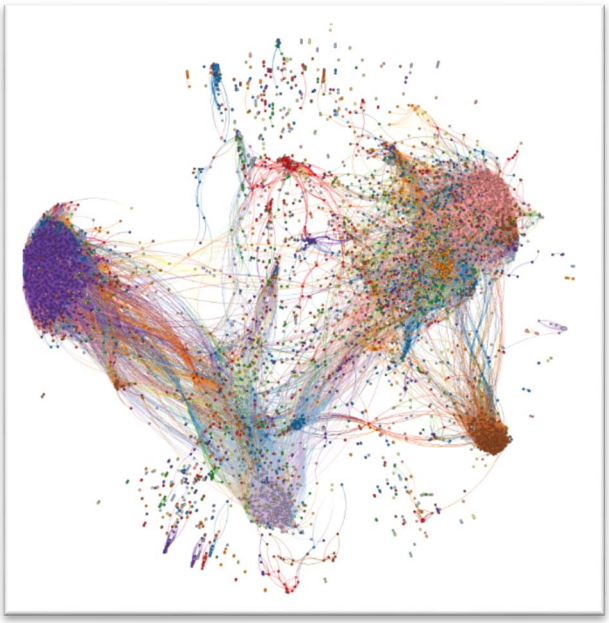

Fig. 1. Gene co-expression network of BRCA. The image was generated using PyGraphistry [56].

After construction of the GCN, six community detection algorithms were used to find subnetworks of interest. The six algorithms are – Fastgreedy [57], Infomap [58], Label Propagation [59], Leading Eigenvector [60], Multilevel [61] and Walktrap [62]. Python package NetworkX [63] and igraph [64] were used for implementation.

C. Permutation Test

Permutation test, following the procedure enumerated by Yu et al. [50], was done on each of the communities to find out which were conserved in all the GCNs. The detailed workflow for the permutation test is described below:

- Let, a community C extracted from GCN of cancer A has n nodes.
- Now a subnetwork is formed using the same n nodes of C from GCN of cancer B. let the number of edges in that subnetwork is e_1
- A random subnetwork of n nodes from GCN of cancer B is formed. let, this subnetwork has e_2 edges.
- Previous step is repeated 1000 times and every time e_2 is greater than e_1 is counted. Let this count be X.
- Now, the P-value for the community, which denotes the significance value for the community in consideration, is calculated by the following equation.

$$P = \frac{X}{1000}$$

In each iteration, a P-value for a community from one cancer with respect to another cancer is obtained. The evaluation of P-value is done for each community with respect to all cancers except the one this community is mined from. The communities with P-value ≤ 0.05 are considered as the conserved communities.

D. Survival Analysis

The gene sets in conserved communities were used in survival analysis to validate them as biomarkers in terms of prognostic risk assessment of cancer patients. For each conserved community, the member genes were taken as co-variates and K-means clustering algorithm was used on cancer samples to create two distinguishing groups assuming that the gene community is capable of differentiating cancer patients in high-risk and low-risk groups. Then Log-rank test and Kaplan-Meier test were done using these two groups to check whether the gene community is really capable of differentiating between high-risk and low-risk groups.

Although Cox Proportional-Hazards Regression Model (CPH) [65] is mostly used to do survival analysis, we chose K-means clustering over CPH. The reason is that CPH requires training samples and after training is done, a scoring method based on the coefficients of CPH is normally used to divide the test samples as two groups, referred to as high-risk and low-risk groups. Then Log-rank and Kaplan-Meier test is done to check the difference between the two groups. In CPH approach, test data is half the size of the total clinical data and the remaining half is used for training. The division of test and training samples is often done randomly, which may result in different output with different division, raising the question of reproducibility. With K-means unlike CPH, one can use all the samples to do survival analysis by creating two groups and the result is reproducible. High-risk and low-risk groups are identified after Kaplan-Meier test.

III. Results and Discussion

A. Communities

In this study, six different community detection algorithms were used on GCNs derived from three cancers - BRCA, COAD and GBM – to discover the communities as possible network modules or biomarkers. Each algorithm resulted in different numbers of communities from different networks, as shown in

Figure 2. Infomap produced the highest numbers of communities - 187 and 162 respectively from GCNs of BRCA and COAD. For GBM, Walktrap produced 112 communities, which is the highest for the same. Leading Eigenvector produced lowest number of communities for all three GCNs - 51 from BRCA, 38 from COAD and 32 from GBM. Total numbers of unique communities produced by six algorithms from BRCA, COAD and GBM are 387, 267 and 247 respectively. The combined number of total unique communities are 899 while only 2 are in common.

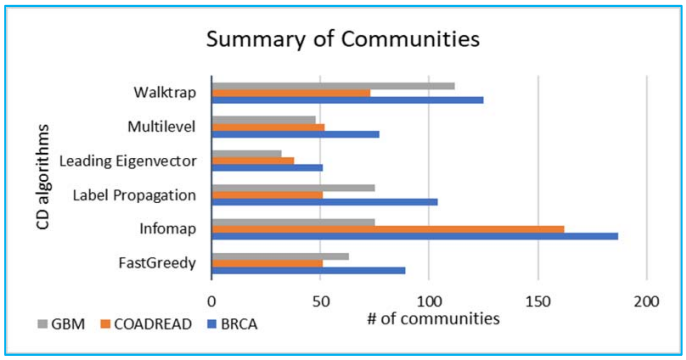

Fig. 2. Number of communities discovered by different algorithms for different GCNs. The communities with number of nodes greater than four are accounted for.

B. Conserved Communities.

Based on permutation test, 64 communities out of 899 are conserved - 27 from BRCA, 23 from COAD and 14 from GBM. Figure 3 presents each of these 64 communities in terms of a) number of genes in each community, b) the GCN network each community is derived from, and c) the algorithm(s) each community is discovered by. Fourteen out of 64 conserved communities are discovered using multiple community detection algorithms. For instance, community #1 is found by three algorithms, Fastgreedy, Infomap and Label Propagation; and community #5 was found by all six algorithms.

C. Survival Analysis

Survival analysis was performed on each of the 64 conserved communities to see if they could significantly distinguish the risk between two groups, high-risk and low-risk, of cancer patients. These two groups are created using K-mean clustering employing the expression values of genes from the community of interest. The set of genes involved in each conserved community were used as prognostic variables to predict the risks of cancer patients in all three cancers. The communities that could distinguish the prognostic risks of patients in multiple cancers are considered as network biomarkers. Finally, the Log-rank test and Kaplan-Meier Estimation were performed to validate their prognostic capabilities.

Based on survival analysis, 16 out of 64 communities could differentiate between high-risk and low-risk groups of patients in one or two cancers. Table III presents the summary of survival analyses of these 16 conserved communities. The triple (Cancer, P-Value, HR) in third column represents – i) "Cancer" represents the type of cancer patients used for survival analysis, ii) "P-value" represents Log-rank test P-value, and iii) "HR" represents the hazard ratio. It is clear from this table that five communities (#5, #19, #27, #59, #63) were able to differentiate prognostically significant patients' group of two cancers and others are capable of differentiating patients' group in one cancer.

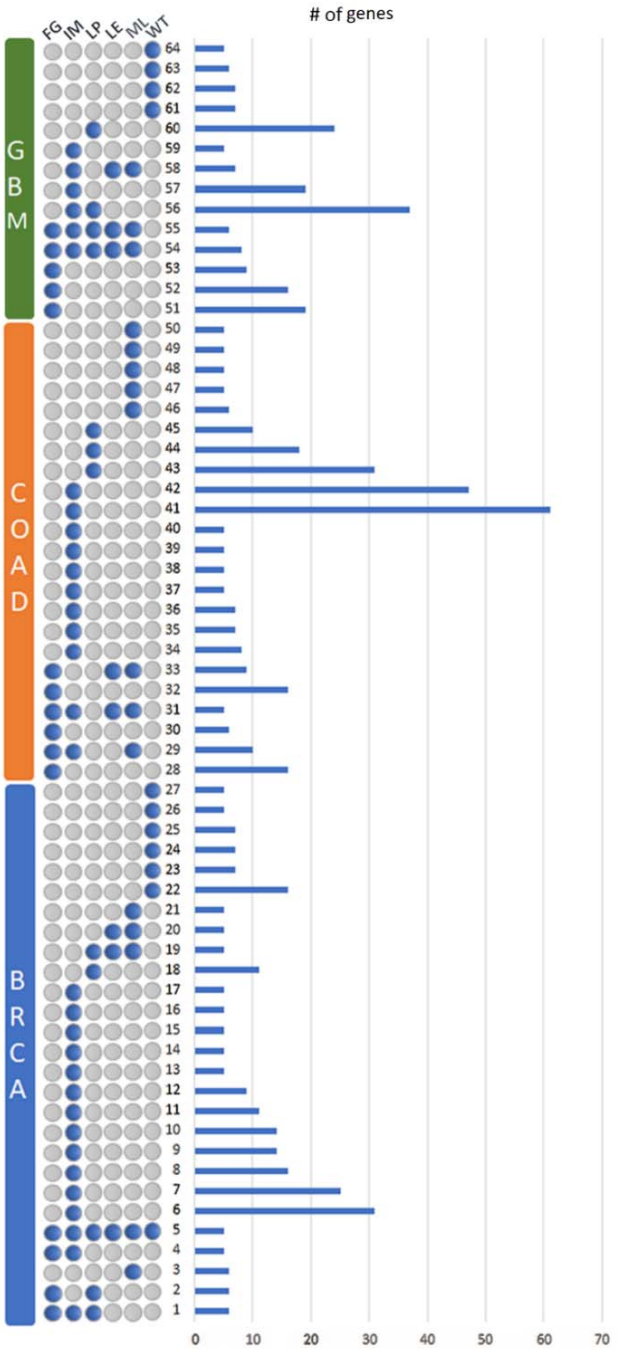

Fig. 3. The summary of 64 conserved communities. The vertical boxes - blue, orange and green - indicate the GCNs from where conserved communities are mined. The blue circle(s) in each row denote(s) the algorithm(s) that discovered the corresponding community. The horizontal bar chart shows the sizes of the communities in terms of number of nodes (genes). X-axis is the number of genes and Y-axis is the communities, indexed from 1 to 64. The names of the algorithms are abbreviated as: FG: Fastgreedy, IM: Infomap, LP: Label Propagation, LE: Leading Eigenvector, ML: Multilevel and WT: Walktrap.

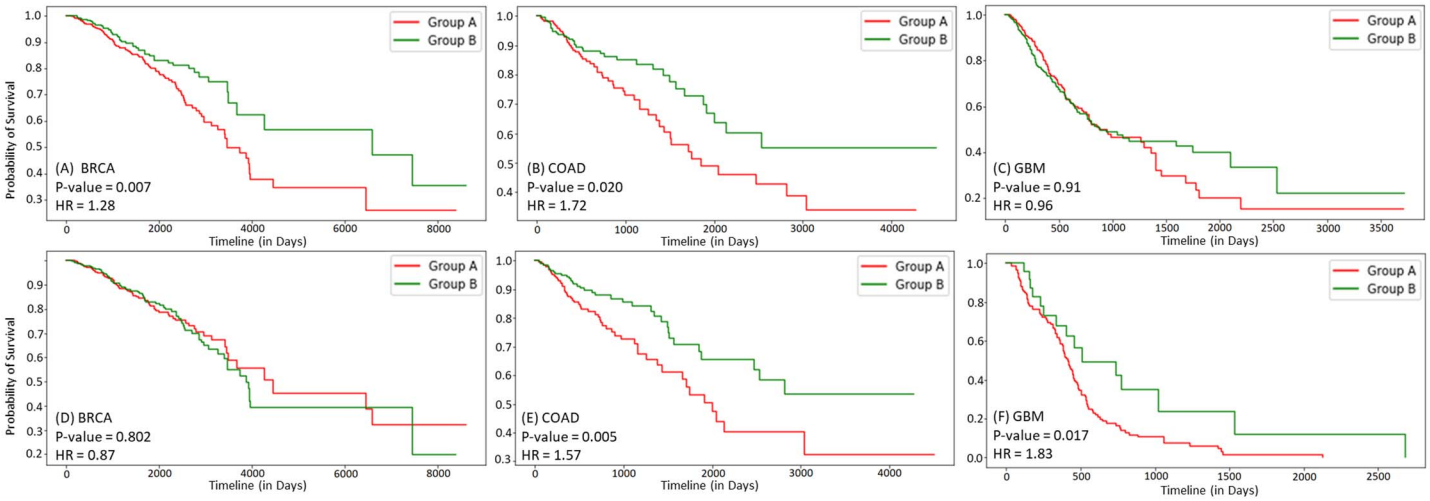

Fig. 4. Kaplan-Meier Curve of cancer patients of (A) BRCA, (B) COAD, (C) GBM using community #27, and (D) BRCA (E) COAD (F) GBM using community #63.

Figure 4 shows the Kaplan-Meier curve employing the gene sets in communities #27 and #63 on BRCA, COAD and GBM cancer patients. Community #27 was able to differentiate two cancer patients' groups in terms of prognostic risk significantly (P- value <= 0.05), namely BRCA (Fig. 4-A) and COAD (Fig 4-B).

Table III: Summary of Survival Analysis of conserved communities. The rows with bold community indexes indicate they were found by multiple community detection algorithms.

Community Index	Corresponding GCN	Cancer Patients, Log-rank test P-value and Hazard Ratio (Cancer, P-value, HR)
5	BRCA	(COAD, 0.0425, 1.43), (GBM, 0.0067, 1.29)
8	BRCA	(COAD, 0.0518, 1.62)
14	BRCA	(BRCA, 0.0143, 1.63)
17	BRCA	(BRCA, 0.0.0066, 1.43)
19	BRCA	(BRCA, 0.0193, 1.31), (GBM, 0.0218. 1.37)
23	BRCA	(COAD, 0.0666,1.21)
27	COAD	(BRCA, 0.0071, 1.28), (COAD, 0.0201, 1.72)
34	COAD	(COAD, 0.0086, 1.82)
36	COAD	(COAD, 0.0367, 1.25)
41	COAD	(COAD,0.0430,1.54)
53	GBM	(GBM, 0.0329, 1.57)
56	GBM	(COAD, 0.0243, 1.45)
57	GBM	(GBM, 0.0226, 1.17)
59	GBM	(BRCA, 0.046, 1.19), (GBM, 0.0492, 1.148)
62	GBM	(GBM, 0.0526, 1.52)
63	GBM	(COAD, 0.0047, 1.57), (GBM, 0.0175,1.83)

But it could not perform the same way for GBM (Fig. 4-C), which is evidenced by a high P-value of 0.91 and a hazard ratio close to 1. Similarly, community #63 was able to differentiate significantly in terms of prognostic risk for COAD (Fig. 4-E), and GBM (Fig. 4-F), but could not do well for BRCA (Fig. 4-D).

IV. Conclusion

Six well-known community detection algorithms are applied to mine probable network biomarkers from gene expression networks of three cancers - BRCA, COAD and GBM. A permutation test is performed to further mine the communities that are conserved in other cancers, thus calling them conserved communities. Then survival analysis is performed on clinical data of three cancers using the conserved community genes as prognostic covariates. The communities that could distinguish the cancer patients between high- and low-risk groups are considered as cancer biomarkers. In the present study, 16 such network biomarkers are discovered. The list of genes from these network biomarkers can be used to discover the trajectory of cancer development by inferring pseudotime from static expression profiles [66].

In the extended version of this paper, functional analysis of the conserved communities will be conducted to provide better insights about the discovered network communities. Another level of validation will be conducted using gene expression and clinical data of different cancers which are not used to derive the conserved communities. In this study, network biomarkers or biomarker communities are mined using mRNA expression data only. It would be worthwhile to study other omics data like miRNA expression, lncRNA expression and DNA methylation to identify network modules for cancer. Finding communities from networks constructed using integration of multi-omics data will provide holistic picture of network modules related to cancer. Graph-based deep learning can be applied to these biological networks to find submodules that might be of biological interest as well.

ACKNOWLEDGMENT

This research is partially funded by NSF CAREER award #1651917 (transferred to #1901628) to AMM.

REFERENCES

[1] Y. Huang et al., "Systematic discovery of functional modules and context-specific functional annotation of human genome," Bioinformatics, vol. 23, no. 13, pp. i222–i229, 2007.

[2] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," Bioinformatics, vol. 18, no. suppl_1, pp. S233–S240, 2002.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci., vol. 95, no. 25, pp. 14863–14868, 1998.

[4] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, vol. 22, no. 15, pp. 1855–1862, 2006.

[5] C. Zheng, D. Huang, L. Zhang, and X. Kong, "Tumor Clustering Using Nonnegative Matrix Factorization With Gene Selection," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 4, pp. 599–607, Jul. 2009.

[6] C. Zheng, L. Zhang, T. Ng, C. K. Shiu, and D. Huang, "Metasample-Based Sparse Representation for Tumor Classification," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 8, no. 5, pp. 1273–1282, 2011.

[7] C. Zheng, L. Zhang, V. T. Ng, C. K. Shiu, and D.-. Huang, "Molecular Pattern Discovery Based on Penalized Matrix Decomposition," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 8, no. 6, pp. 1592–1603, Nov. 2011.

[8] S. Wang, Y. Zhu, W. Jia, and D. Huang, "Robust Classification Method of Tumor Subtype by Using Correlation Filters," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 9, no. 2, pp. 580–591, Mar. 2012.

[9] D. Huang and H. Yu, "Normalized Feature Vectors: A Novel Alignment-Free Sequence Comparison Method Based on the Numbers of Adjacent Amino Acids," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 10, no. 2, pp. 457–467, Mar. 2013.

[10] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks," BMC Bioinformatics, vol. 6, no. 1, p. 227, 2005.

[11] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.," Pac. Symp. Biocomput., pp. 418–429, 1999.

[12] X. Wen et al., "Large-scale temporal gene expression mapping of central nervous system development," Proc. Natl. Acad. Sci., vol. 95, no. 1, pp. 334–339, 1998.

[13] H. Li and J. Gui, "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," Biostatistics, vol. 7, no. 2, pp. 302–317, 2005.

[14] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond, "Gene co-expression network topology provides a framework for molecular characterization of cellular state," Bioinformatics, vol. 20, no. 14, pp. 2242–2250, 2004.

[15] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," Bioinformatics, vol. 24, no. 5, pp. 719–720, 2007.

[16] H. K. Lee, A. K.-H. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets.," Genome Res., vol. 14 6, pp. 1085–1094, 2004.

[17] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis.," Stat. Appl. Genet. Mol. Biol., vol. 4, p. Article17, 2005.

[18] J. Tang et al., "Overexpression of ASPM, CDC20, and TTK Confer a Poorer Prognosis in Breast Cancer Identified by Gene Co-expression Network Analysis," in Front. Oncol., 2019.

[19] H. Lalremmawia and B. K. Tiwary, "Identification of Molecular Biomarkers for Ovarian Cancer using Computational Approaches.," Carcinogenesis, 2019.

[20] A. M. Maertens, V. Tran, A. Kleensang, and T. Hartung, "Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response," in Front. Genet., 2018.

[21] H. Shi, L. Zhang, Y. Qu, L. Hou, L. Wang, and M. Zheng, "Prognostic genes of breast cancer revealed by gene co-expression network analysis," in Oncology letters, 2017.

[22] X. Liu, A.-X. Hu, J.-L. Zhao, and F. Chen, "Identification of Key Gene Modules in Human Osteosarcoma by Co-Expression Analysis Weighted Gene Co-Expression Network Analysis (WGCNA).," J. Cell. Biochem., vol. 118 11, pp. 3953–3959, 2017.

[23] 丛 CongZhang张 and 茜 QianSun孙, "Weighted gene co-expression network analysis of gene modules for the prognosis of esophageal cancer," J. Huazhong Univ. Sci. Technol. [Medical Sci., vol. 37, pp. 319–325, 2017.

[24] R. Liu, W. Zhang, Z. Liu, and H. Zhou, "Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis," in BMC Genomics, 2017.

[25] S. Deng, L. Zhu, and D. Huang, "Predicting Hub Genes Associated with Cervical Cancer through Gene Co-Expression Networks," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 13, pp. 27–35, 2016.

[26] P. Timalsina, K. Charles, and A. M. Mondal, "STRING PPI Score to Characterize Protein Subnetwork Biomarkers for Human Diseases and Pathways," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, 2014, pp. 251–256.

[27] C. Kevin, A. Andrews, and A. Mondal, "Protein Subnetwork Biomarkers for Yeast Using Brute Force Method," in Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP) , 2013, pp. 218–223.

[28] D. . Bett and A. Mondal, "Diffusion Kernel to Identify Missing PPIs in Protein Network Biomarker," in 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 1614–1619.

[29] S. Bandyopadhyay et al., "Rewiring of genetic networks in response to DNA damage.," Science, vol. 330, no. 6009, pp. 1385–1389, Dec. 2010.

[30] A. M. Mondal and J. Hu, "NetLoc: Network based protein localization prediction using protein-protein interaction and co-expression networks," in 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2010, pp. 142–148.

[31] A. M. Mondal, J. Lin, and J. Hu, "Network based subcellular localization prediction for multi-label proteins," in 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW),

2011, pp. 473–480.

[32] A. M. Mondal and J. Hu, “Protein Localization by Integrating Multiple Protein Correlation Networks,” in The 2012 International Conference on Bioinformatics & Computational Biology, 2012, p. 7.

[33] J.-R. Lin, A. M. Mondal, R. Liu, and J. Hu, “Minimalist ensemble algorithms for genome-wide protein localization prediction,” BMC Bioinformatics, vol. 13, no. 1, p. 157, 2012.

[34] A. M. Mondal and J. Hu, “Scored Protein-Protein Interaction to Predict Subcellular Localizations for Yeast Using Diffusion Kernel,” Lect. Notes Comput. Sci., vol. 8251, pp. 647–655, 2013.

[35] A. Mondal and J. Hu, “Network based prediction of protein localisation using diffusion kernel.,” Int. J. Data Min. Bioinform., vol. 9, no. 4, pp. 386–400, 2014.

[36] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen, “Diffusion kernel-based logistic regression models for protein function prediction.,” OMICS, vol. 10, no. 1, pp. 40–55, 2006.

[37] Y. Qi, Y. Suhail, Y. Lin, J. D. Boeke, and J. S. Bader, “Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions.,” Genome Res., vol. 18, no. 12, pp. 1991–2004, Dec. 2008.

[38] F. E. Faisal and T. Milenkovic, “Dynamic networks reveal key players in aging.,” Bioinformatics, vol. 30, no. 12, pp. 1721–1729, Jun. 2014.

[39] A. M. Mondal, C. A. Schultz, M. Sheppard, J. Carson, R. B. Tanvir, and T. Aqila, “Graph Theoretic Concepts as the Building Blocks for Disease Initiation and Progression at Protein Network Level: Identification and Challenges,” in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2713–2719.

[40] R. B. Tanvir, T. Aqila, M. Maharjan, A. Al Mamun, and A. M. Mondal, “Graph Theoretic and Pearson Correlation-Based Discovery of Network Biomarkers for Cancer,” Data, vol. 4, no. 2, p. 81, Jun. 2019.

[41] Z. Shi, C. K. Derow, and B. Zhang, “Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression,” in BMC Systems Biology, 2010.

[42] A. D. Perkins and M. A. Langston, “Threshold selection in gene co-expression networks using spectral graph theory techniques,” BMC Bioinformatics, vol. 10, no. 11, p. S4, 2009.

[43] D.-Q. Zhang, C. Zhou, S.-Z. Chen, Y. Yang, and B. Shi, “Identification of hub genes and pathways associated with bladder cancer based on co-expression network analysis.,” Oncol. Lett., vol. 14 1, pp. 1115–1122, 2017.

[44] Y. Li, B. Liu, J. Li, and G. Li, “MiMod: A New Algorithm for Mining Biological Network Modules,” IEEE Access, vol. 7, pp. 49492–49503, 2019.

[45] B. Tripathi, S. Parthasarathy, H. Sinha, K. Raman, and B. Ravindran, “Adapting Community Detection Algorithms for Disease Module Identification in Heterogeneous Biological Networks,” Front. Genet., vol. 10, p. 164, 2019.

[46] J. Wang, J. Liang, W. Zheng, X. Zhao, and J. Mu, “Protein complex detection algorithm based on multiple topological characteristics in PPI networks,” Inf. Sci. (Ny)., vol. 489, pp. 78–92, 2019.

[47] C. P. Couturier et al., “Single-cell RNA-seq reveals that glioblastoma recapitulates normal brain development,” bioRxiv, 2018.

[48] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” Cell, vol. 144, no. 5, pp. 646–674, 2011.

[49] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, “Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types,” Nat. Commun., vol. 5, p. 3231, 2014.

[50] L.-H. Yu, Q.-W. Huang, and X.-H. Zhou, “Identification of Cancer Hallmarks Based on the Gene Co-expression Networks of Seven Cancers,” Front. Genet., vol. 10, p. 99, 2019.

[51] R. B. Tanvir, M. Maharjan, and A. M. Mondal, “Community Based Cancer Biomarker Identification from Gene Co-expression Network,” in 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB’19), 2019, pp. 545–545.

[52] S. V Vasaikar, P. Straub, J. Wang, and B. Zhang, “LinkedOmics: analyzing multi-omics data within and across 32 cancer types,” Nucleic Acids Res., vol. 46, no. D1, pp. D956–D963, Nov. 2017.

[53] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg, “An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data,” PLOS Comput. Biol., vol. 5, no. 12, pp. 1–11, 2009.

[54] J. W. Rowley et al., “Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes,” Blood, vol. 118, no. 14, pp. e101--e111, 2011.

[55] L. Han et al., “The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes,” Nat. Commun., vol. 5, p. 3963, Jul. 2014.

[56] Graphistry, “PyGraphistry.” .

[57] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top., vol. 70, no. 6, p. 6, Dec. 2004.

[58] M. Rosvall and C. T. Bergstrom, “An information-theoretic framework for resolving community structure in complex networks,” Proc. Natl. Acad. Sci., vol. 104, no. 18, pp. 7327–7331, 2007.

[59] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” Phys. Rev. E, vol. 76, no. 3, p. 36106, Sep. 2007.

[60] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” Phys. Rev. E, vol. 74, no. 3, p. 36104, Sep. 2006.

[61] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech. Theory Exp., vol. 2008, no. 10, p. P10008, Oct. 2008.

[62] P. Pons and M. Latapy, “Computing communities in large networks using random walks (long version).” 2005.

[63] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX,” in Proceedings of the 7th Python in Science Conference (SciPy), 2008, pp. 11–15.

[64] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” InterJournal, vol. Complex Sy, p. 1695, 2006.

[65] E. A. Mauger, R. A. Wolfe, and F. K. Port, “Transient effects in the cox proportional hazards regression model,” Stat. Med., vol. 14, no. 14, pp.

1553–1565, 1995.

[66] T. Aqila, A. Al Mamun, and A. M. Mondal, “Pseudotime Based Discovery of Breast Cancer Heterogeneity,” in 2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2019), 2019.