Comparison of tissue/disease specific integrated networks using directed graphlet signatures

Arzu Burçak Sonmez
Department of Medical Informatics
Informatics Institute
Middle East Technical University
Ankara, Turkey
aburcak@gmail.com

Tolga Can
Department of Computer Engineering
Middle East Technical University
Ankara, Turkey
tcan@metu.edu.tr

ABSTRACT
We present a novel framework for counting small sub-graph patterns in integrated genome-scale networks. An integrated network was built using the physical, regulatory, and metabolic interactions between H. sapiens proteins from the Pathway Commons database. The network was filtered for tissue/disease specific proteins by using a large-scale human transcriptional profiling study, resulting in several tissue and disease-specific sub-networks. In this study, we apply and extend the idea of graphlet counting in undirected protein-protein interaction (PPI) networks to directed multi-labeled networks and represent each network as a vector of graphlet counts. Graphlet counts are assessed for statistical significance by comparison against a set of randomized networks. We present our results on analysis of differential graphlets between different conditions. Our results show that graphlets can be used for identification of systems level differences between disease states.

Categories and Subject Descriptors
J.3 [Computer Applications]: Biology and Genetics; G.2.2 [Mathematics of Computing]: Discrete Mathematics—Graph Theory

General Terms
Algorithms

Keywords
integrated networks, network comparison, directed graphlets

1. INTRODUCTION
With the accumulation of high-throughput omics data in public databases, integrative studies on heterogenous and dynamic biological networks have become possible. Repositories, such as Pathway Commons [1], BioGRID [8], and the Human Protein Reference Database (HPRD) [6], collect and curate associations between genes, proteins, and chemical compounds from various high and low throughput data sources. In addition, there are efforts, such as BioPAX [3], towards a standardized representation and exchange of different types of networks between databases and applications. Although the data for various types of interactions such as regulatory, metabolic, and physical interactions are available in these repositories, joint analysis of these data in a single integrated network remains a challenge. The software suite Paxtools [2] is a rich collection of methods for querying, visualizing, and converting integrated BioPAX networks; however, advanced algorithms, such as graphlet counting, are yet to be added to the expanding repository of this open source project.

Graphlets are small sub-graphs that provide more detailed topological statistics for a graph. As an extension of single node statistics, such as average degree and degree distribution, graphlets give a broader view around a node. Introduced by Pržulj in 2007 [7] graphlets have been shown to be effective in analysis and comparison of biological networks [5]. Due to the combinatorial expansion of different types of graphlets, computationally efficient counting of graphlets is a challenging problem. The combinatorial expansion of different types of graphlets is more dramatic when directed and multi-label edges are considered. In this paper, without tackling computational efficiency, we propose a straightforward method for counting directed multi-label graphlets of size 2-3 and assess the utility of these graphlets in tissue specific networks. To the best of our knowledge, this is the first study to address directed multi-label graphlets in integrated networks. By counting graphlets in several different tissue specific networks, we have identified many statistically significant graphlets.

2. MATERIALS AND METHODS
We have used two main resources to acquire the data used in this study: 1) Pathway Commons [1] and 2) a human body transcriptional profiling study accessible at the NCBI GEO database [4] with accession number GSE7307. From the Pathway Commons database, we obtained an integrated human network of 19,537 proteins with 523,498 undirected edges and 337,117 directed edges. The network contains physical, regulatory, and metabolic relations. By consolidating multiple-edges into single multi-labeled directed edges between protein pairs, the final network contains 1,521,508 directed edges between 760,754 protein pairs.
The human body transcriptional profiling study by Roth et al. (GSE7307) profiled 90 distinct tissue types with several samples per tissue type (677 samples in total) using the Affymetrix U133 plus 2.0 array. We identified active and inactive genes using a normalized expression threshold of 10.0. The subnetwork induced by the active genes in a sample comprised the tissue specific network for that sample.

We used a hashing based strategy to count all the graphlets in a network using a binary encoding of multi-label edges. We used the encoding of edges as keys in a hashable of graphlet counts. We counted up to 3-node graphlets; since, this brute-force counting method is prohibitively expensive for counting graphlets of larger sizes. Nevertheless, the number of distinct 3-node graphlets are on the order of thousands and 3-node graphlets are effective in identifying differences between tissues and diseases.

We compute the z-score of a graphlet count by comparison against the count of that graphlet in an ensemble of randomized networks with the same degree distribution as the real networks. We randomized each of the 141 tissue specific networks by the edge-shuffling method. We counted each two-node and three-node graphlets in all of the 141 randomized networks and computed the z-score.

3. RESULTS

In this section, we provide a statistically significant graphlet with differential recurrence between the normal and disease tissues of the prostate gland tissue (Figure 1). The z-score of the graphlet in the diseased prostate gland tissue network is 68.68; whereas, the z-score of the same graphlet is 7.63 in the normal network. Protein A and B are the members of the same protein complex and protein A catalyses a reaction before protein C. The state-change relations exhibit the feed-forward loop motif.

4. CONCLUSIONS

In this paper, we have proposed a bitwise encoding of multi-label directed edges for counting graphlets of size 2-3 with a hash-based approach. We have applied the proposed methodology on 141 tissue/disease specific networks and identified statistically significant graphlets. We have also shown that graphlets can be used for effective comparison of tissue specific networks. This study is a first attempt for counting multi-label directed graphlets in genome-scale integrated networks. In the future, computationally more efficient algorithms can be designed for counting graphlets of larger sizes. In addition, placement of graphlets in higher level systematic groups such as molecular complexes, signaling pathways, and metabolic networks will help molecular biologists interpret the results and construct novel biological hypotheses.

5. ACKNOWLEDGMENTS

This study is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) 1001 Program Grant #114E111 and by the Turkish Science Academy Young Scientists Award Program (BAGEP) 2014-2016.

6. REFERENCES

[1] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data. Nucleic Acids Research, 39(suppl 1):D685–D690, 2011.
[2] E. Demir, O. Babur, I. Rodchenkov, B. A. Aksoy, K. I. Fukuda, B. Gross, O. S. Sümer, G. D. Bader, and C. Sander. Using biological pathway data with paxtools. PLoS Comput Biol, 9(9):1–5, 09 2013.
[3] E. Demir, M. Cary, S. Paley, and et al. BioPAX - a community standard for pathway data sharing. Nature biotechnology, 28(9):935–942, 2010.
[4] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Research, 30(1):207–210, 2002.
[5] W. Hayes, K. Sun, and N. Pržulj. Graphlet-based measures are suitable for biological network comparison. Bioinformatics, 29(4):483–491, 2013.
[6] T. S. Keshava Prasad, R. Goel, K.andasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shaforeen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database - 2009 update. Nucleic Acids Research, 37(suppl 1):D767–D772, 2009.
[7] N. Pržulj. Biological network comparison using graphlet degree distribution. Bioinformatics, 23(2):e177–e183, 2007.
[8] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. Nucleic Acids Research, 34(suppl 1):D535–D539, 2006.