Graph-theory Based Simplification Techniques for Efficient Biological Network Analysis

Euiseong Ko*, Mingon Kang*, Hyung Jae Chang†, Donghyun Kim*‡
* Department of Computer Science, Kennesaw State University, Marietta, GA, USA.
{eko1, mkang9, donghyun.kim}@kennesaw.edu.
† Department of Computer Science, Troy University-Montgomery, Montgomery, AL, USA.
hjchang@troy.edu.
‡ Corresponding Author.

***Abstract*—The recent years have witnessed the remarkable expansion of publicly available biological data in the related research fields. Many researches in these fields often require massive data to be analyzed by utilizing high-throughput sequencing technologies. However, it is very challenging to interpret the data efficiently due to it high complexity. This paper introduces two new graph algorithms which aim to improve the efficiency of the existing methods for biological network data interpretation. In particular, the algorithms focus on the problem of how to simplify gene regulatory networks so that many existing algorithms can efficiently discover important connected components of a biological system in their own context as many times as they need. The performance of the proposed algorithms is compared with each other with gene expression data of glioblastoma brain tumor cancer.**

***Keywords*-Biological Network Analysis, Graph Algorithm, Greedy Algorithm, Optimization**

I. Introduction

Recently, a graph-based representation of biological systems plays a critical role in identifying molecular interactions in biomedical research. Biological network analysis is a powerful approach to providing the insight of complex biological systems and the function of cellular components. Biological networks typically consist of a large number of biological components such as genes, protein, and metabolic, and its topological analysis makes the interpretation of the complex biological networks be feasible.

Complex biological systems are often represented by biological networks. Protein-Protein Interaction networks (PPI) and Gene Regulatory Networks (GRNs) are ones of the most representative biological networks that most research has focused on. PPI show the physical interactions between proteins or metabolic and signaling pathways of a cell, where proteins and their interactions are represented by nodes and edges respectively [1], [2].

Although PPI are the most intensively analyzed through many studies, PPI are difficult to make analysis conclusions due to the heterogeneous data sources. GRNs are comprised of nodes of genes, and the genes are connected if the expression of one gene regulates expression of another one by either activation of inhibition [3]. A gene results in constructing a single protein. GRNs assume that the gene regulate other genes if the protein that the gene manufactures controls the rate at which other genes manufacture proteins. Moreover, genes are not independent, but acting collectively, where genes regulate each other.

The expression data for biological networks can be acquired from high-throughput technology such as microarray, sequencing techniques, mass spectrometry, and protein arrays. The high-throughput genetic technologies empowers to study how genes interact with each other. The potential discovery of important components in biological networks would help one to identify triggering biological mechanism and treatments for diseases.

Research in biological networks includes two processes: (1) network reconstruction, and (2) network analysis. Network reconstruction is a reverse-engineering problem that determines weights of edges between nodes from high-throughput data. We also call the problem as a *network inference* problem. Although there are multiple types of biological networks [4], such as boolean networks, probabilistic, Petri nets, we focus on a standard graph-based network model represented by an adjacency matrix $\mathbf{A}$ where p is a number of node, $\mathbf{A} \in \Re^{p \times p}$, and $A_{ii} = 0$.

The approaches of biological network inference are mainly three-fold: (1) correlation-based, (2) Bayesian-based, and (3) regression-based approaches. Correlation-based approaches identify the interactions between biological components (e.g., proteins in PPI and genes in GRNs) by using linear correlation (e.g., Pearson's correlation coefficient). The biological networks computes pairwise correlation ($-1 \leq R \leq 1$) or coefficient of determination ($0 \leq R^2 \leq 1$) between nodes and determines active/inactive interactions if the correlation coefficient is lower than a certain threshold. Therefore, the network is undirected graph and lacks the interpretation of casual effects between the biological components.

WGCNA (Weighted Gene Coexpression Network Analysis) is a representative method of the correlation-based approaches in GRNs [5]. Mutual information (MI) [6], maximum information correlation (MIC) [7], and conditional mutual information (CMI) [8], [9] have been also used to determine the edges of the biological networks. Bayesian-based approaches can infer casual relationship in a probabilistic manner [10], [11]. However, Bayesian-based approaches would be infeasible when the number of nodes is large. On the other hand, regression-based approaches infer the relationship of nodes by decomposing the whole network graph into p numbers of regression problems [12], [13]. Regression-based approaches often use LASSO solution for generating sparse graphs in GRNs, since it aims to infer the transcription regulatory interactions between genes.

Network analysis is an essential to analyze and interpret the biological network, once we obtain the biological network from the high-throughput data. Biological systems includes a large number of biological components, so the biological network may be extremely complicated to analyze.

There are three types of motifs: (1) feed-forward loops (FFL), single-input modules (SIM), and dense overlapping regulons (DOR) [14]. Maximum cliques have been used to discover subgraphs of the biological networks, which may be DOR motifs in biological networks [15], [16]. A clique is a subset of a graph, where its induced subgraph is complete. Finding the largest subset, which is the maximum clique, may provide biological interpretations of largest biological components on the biological system. However, some biological component groups of interests are not on the assumption.

In this paper, we aim to discover subgraphs that consist of important biological components connected each other without the strict criteria such as cliques in biological systems. We assume that the biological system is represented by an undirected graph where a node and an edge show a biological component and their interactions respectively. The discovery of subgraphs in a large scale of biological networks would provide interpretable analysis and visualization solutions for better understanding of the biological system. The main contribution of this paper includes (a) two greedy graph algorithms for this purpose, (b) their performance evaluation with glioblastoma brain tumor cancer data, and (c) the idea of protecting the privacy of the owners of the biological network.

The rest of this paper is organized as follows. Section II introduces the formal definition of our problem and our two new graph algorithms to simplify gene networks by capturing key components. Section III describes the result of our experimental results. Finally, we conclude this paper in Section IV.

II. Problem Definition and Proposed Algorithms

In this section, we first introduce the formal definition of our problem of interest as well as its complexity analysis. As the problem of interest is NP-hard, we propose two greedy heuristic algorithms in the following subsection.

A. Notations and Problem Definition

In this paper, $G = (V, E, w_E)$, where $w_E : E \rightarrow \mathbb{R}^*$ is an edge weighted connected graph, where $V = V(G)$ is the set of nodes in the graph and $E = E(G)$ is the set of edges in the graph. For any given node subset V', $G[V']$ is the subgraph of G induced by V'. We also notate $W(G[V'])$ is the sum of weight of edges in $G[V']$. Finally, $|V'|$ notates the number of nodes in the set.

Now, the formal definition of the problem of our interest is as follow.

Definition 1 (MT-GNSIP)**.** *Given an edge weighted graph $G = (V, E, w_E)$, where $w_E : E \rightarrow \mathbb{R}^*$ and a positive real number T, the Maximum Total-edge-weight Gene Network Subgraph Identification Problem (MT-GNSIP) is to identify a subset of nodes V' from V such that*

(a) Optimization Goal 1: $W(G[V'])$ is maximized,
(b) Optimization Goal 2: $|V'|$ is minimized,
(c) $W(G[V']) \leq T$, and
(d) $G[V']$ is connected.

The ultimate goal is to identify sub-graphs of strongly connected components in biological networks, where the lowest numbers of nodes are included. Since edge weights indicate the strength of relationship between nodes, the subgraph that maximizes the total edge weights may indicates the components that play critical roles in the biological system. Furthermore, the hyper-parameter T controls either the complexity of the sub-graph or the hierarchical components in the biological network.

One can easily see that Optimization Goal 1 and Optimization Goal 2 are opposition with each other, and therefore this problem is very difficult. That is, in order to optimize a solution toward Optimization Goal 2, the best possible solution would be an empty node set. However, such solution would be very bad with respect to Optimization Goal 1.

B. Two Greedy Graph Algorithms

In this section, we introduce our two different algorithms for MT-GNSIP, Greedy-Augmenter (Algorithm 1) and Greedy-Shrinker (Algorithm 2). The input of the both algorithms consists of $G = (V, E, w_E), T$ such that $V \leftarrow V(G)$, $E \leftarrow \{e | e \in E(G)$ and $w_E(e) \leq b\}$, w_E is the edge-weight function from $G = (V, E, w_E)$, and T is a threshold value. The output of the algorithms is a subset of nodes. As we mentioned earlier, each algorithm assumes that G is a connected graph and proceeds our discussion during the rest of this paper. In the final section, we discuss how to deal with the case in which G is not a connected graph.

Greedy-Augmenter. This algorithm initially picks a node from V and adds it to V'. Then, the algorithm iteratively identifies a node v_{max} outside V' but is neighboring to at least one node in V' such $W(G[V' \bigcup \{v_{max}\}])$ can be further maximized while it does not exceed the threshold level T. In each iteration, the greedy selection of such node reflects both optimization goals as the algorithm attempts to maximize the total edge weight sum of the induced graph of G by V' and at the same time, the size of V' (or equivalently the total number of iterations) can be minimized when $W(G[V' \bigcup \{v_{max}\}])$ cannot grow any further.

Greedy-Shrinker. This algorithm initially copies all nodes in V to V'. Then, it iteratively identifies a node v_{min} from V' such that the removal of the nodes does not affect the connectivity of $G[V']$ and $W(G[V'])$ can be minimally reduced. The loop finally ends once $W(G[V']) \leq T$ becomes true. By greedily selecting the minimum v_{min}, more number of nodes can be removed from V' (Optimization Goal 2).

Algorithm 1 Greedy-Augmenter $(G = (V, E, w_E), T)$

```
Let $r$ be a node in $V$ such that the total weight of the edges connected to $r$ is maximum. Set $V' \leftarrow \{r\}$.
loop
    for each node $v_i \in V \setminus V'$ do
        set $p_i \leftarrow 0$.
    end for
    for each node $v_i \in V \setminus V'$ do
        for each node $v_j \in V'$ do
            if $(v_i, v_j) \in E$ then
                $p_i \leftarrow p_i + w_E(v_i, v_j)$.
            end if
        end for
    end for
    find the maximum $p_{max}$ such that $W(G[V' \bigcup p_{max}]) \leq T$ and $G[V' \bigcup p_{max}]$ is connected.
    if no such $p_{max}$ exists then
        exit this loop.
    else
        set $V' \leftarrow V' \bigcup \{v_{max}\}$.
    end if
end loop
return $V'$.
```

Meanwhile, due to the greedy strategy, V' is gradually losing least important nodes in order to maintain $W(G[V'])$ high until $W(G[V'])$ is just below T (Optimization Goal 1). As a result, the design of this algorithm well reflects both of the optimization goals.

III. Experimental Results

We applied our methods to the Glioblastoma (GBM) brain tumor cancer data, which is available to download from TCGA (The Cancer Genome Atlas, https://cancergenome.nih.gov) database. The GBM data contains 528 patient samples, where there are 12,042 gene expression on each patient. Among the whole gene expression data, we considered only genes of the top 500 highest expression in average. The adjacency matrix (**A**) was generated by pairwise Pearson correlation coefficient, but the diagonal matrix of **A** was set to zeros (no self-loop). We finally made the network sparse by setting the edge weights less than 0.3 to zeros. Since the adjacency matrix graph has more than one connected component, we selected the one which with the highest edge total weight as an input of our algorithm.

Then, we performed the two proposed algorithms, greedy-augmenter and greedy-shrinker, to the GBM data. The hyper-parameter of T was set as 100.

The greedy-augmenter identified a sub-graph that consists 20 genes and 187 interactions between the genes, and the greedy-shrinker detected 19 genes and 136 edges. We assessed the experimental results with a protein-protein interaction database. The gene lists were introduced to the *String* database (http://string-db.org) [17], and compared the interactions of the genes with our findings.

Algorithm 2 Greedy-Shrinker $(G = (V, E, w_E), T)$

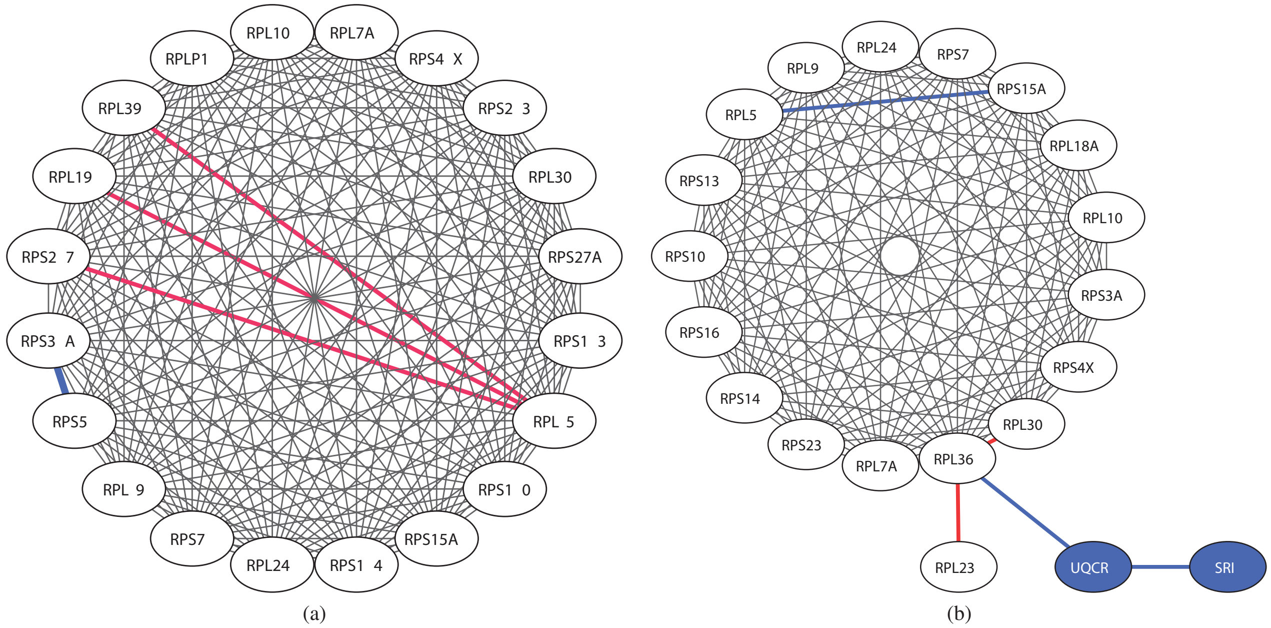
```
Set $V' \leftarrow V$.
loop
    for each node $v_i \in V'$ do
        set $p_i \leftarrow 0$.
    end for
    for each node $v_i \in V'$ do
        for each node $v_j \in V'$ such that $v_j \neq v_i$ do
            if $(v_i, v_j) \in E$ then
                $p_i \leftarrow p_i + w_E(v_i, v_j)$.
            end if
        end for
    end for
    find the minimum $p_{min}$ such that $W(G[V' \setminus p_{min}]) > T$ and $G[V' \setminus p_{min}]$ is still connected.
    if no such $p_{min}$ exists then
        find the minimum $p_{min}$ such that $W(G[V' \setminus p_{min}]) \leq T$ and $G[V' \setminus p_{min}]$ is still connected. Then, and remove $v_{min}$ from $V'$.
        exit this loop.
    else
        remove $v_{min}$ from $V'$.
    end if
end loop
return $V'$.
```

The sub-graphs that the forward/backward-search methods identified are depicted in Fig. 1. The blue nodes and edges indicate new genes and interactions that our proposed methods found but do not exist in the String database. In the other hand, the red shows those which are not identified by our methods but exist in the database. Surprisingly, most genes and their interactions that we discovered by the proposed methods are also observed in the String database as high scores ($>$ 0.9). The backward-search method proposed that two genes of UQCR and SRI may interact with the strongly connected other genes. Especially, SRI has been reported as a marker in GBM brain cancer [18], and it also has been claimed that UQCR is related to brain cells [19].

IV. Conclusion and Future Work

In this paper, we proposed new algorithms that identify subgraphs of biological networks, which may be important biological components. We discover the connected-subgraphs by maximizing the sum of edges while penalizing the number of node in the subgraph. We applied the proposed methods to the BGM brain tumor cancer data. The biological network analysis would provide interpretable solutions to understand complex biological systems and the interactions of the biological components in the system. Our experimental results show that the proposed algorithms are effective and promising.

In this paper, we assumed that the input graph is connected, which is not necessarily the case in practice. In order to deal with this situation, we may need to apply our algorithms for each connected component. In such case, it is important to



(a)

(b)

Fig. 1: Experimental result of (a) greedy-augmenter and (b) greedy-shrinker algorithm with the GBM data

distribute T among components. One idea to distribute T is based on the total edge weight of each connected component. As a future work, we plan to further investigate proper strategies for this purpose.

References

[1] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell, "Protein-protein interaction networks and biologywhat's the connection?" *Nature Biotechnology*, vol. 26, no. 1, pp. 69–72, 2008.

[2] H. G. Vikis and K. L. Guan, "Glutathione-s-transferase (GST)-fusion based assays for studying protein-protein interactions," in *Protein-Protein Interactions: Methods and Applications: Second Edition*, 2015, pp. 353–364.

[3] T. Schlitt and A. Brazma, "Current approaches to gene regulatory network modelling." *BMC.Bioinformatics*, vol. 8 Suppl 6, p. S9, 2007.

[4] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nat Rev Mol Cell Biol*, vol. 9, no. 10, pp. 770–780, 2008.

[5] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis." *BMC bioinformatics*, vol. 9, p. 559, 2008.

[6] G. Altay and F. Emmert-Streib, "Inferring the conservative causal core of gene regulatory networks." *BMC systems biology*, vol. 4, p. 132, 2010.

[7] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[8] K.-C. Liang and X. Wang, "Gene regulatory network reconstruction using conditional mutual information." *EURASIP journal on bioinformatics & systems biology*, vol. 2008, no. 1, p. 253894, 2008.

[9] X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information." *Bioinformatics (Oxford, England)*, vol. 28, no. 1, pp. 98–104, 2012.

[10] P. Li, C. Zhang, E. J. Perkins, P. Gong, and Y. Deng, "Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks," *BMC Bioinformatics*, vol. 8, p. 8 str., 2007.

[11] H. Njah and S. Jamoussi, "Weighted ensemble learning of Bayesian network for gene regulatory networks," *Neurocomputing*, vol. 150, no. PB, pp. 404–416, 2015.

[12] N. Omranian, J. M. O. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, "Gene regulatory network inference using fused LASSO on multiple data sets." *Scientific Reports*, vol. 6, p. 20533, 2016.

[13] D.-C. Kim, M. Kang, B. Zhang, X. Wu, C. Liu, and J. Gao, "Integration of DNA Methylation, Copy Number Variation, and Gene Expression for Gene Regulatory Network Inference and Application to Psychiatric Disorders," *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pp. 238–242, 2014.

[14] U. Alon, "Network motifs: theory and experimental approaches," *Nat Rev Genet*, vol. 8, no. 6, pp. 450–461, 2007.

[15] J. D. Eblen, "The Maximum Clique Problem: Algorithms, Applications, and Implementations (thesis)," *Bioinformatics Research and Applications*, pp. 1–100, 2011.

[16] M. P. Pradhan, K. Nagulapalli, and M. J. Palakal, "Cliques for the identification of gene signatures for colorectal cancer across population." *BMC systems biology*, vol. 6 Suppl 3, no. Suppl 3, p. S17, 2012.

[17] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. Von Mering, "STRING v10: Protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. D1, pp. D447–D452, 2015.

[18] T. Yokota, J. Kouno, K. Adachi, H. Takahashi, A. Teramoto, K. Matsumoto, Y. Sugisaki, M. Onda, and T. Tsunoda, "Identification of histological markers for malignant glioma by genome-wide expression analysis: Dynein, α-PIX and sorcin," *Acta Neuropathologica*, vol. 111, no. 1, pp. 29–38, 2006.

[19] Y. Hong, F. Piao, Y. Zhao, S. Li, Y. Wang, and P. Liu, "Subchronic exposure to arsenic decreased Sdha expression in the brain of mice," *NeuroToxicology*, vol. 30, no. 4, pp. 538–543, 2009.