Received November 23, 2017, accepted January 15, 2018, date of publication January 24, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2796565

An Efficient Network Motif Discovery Approach for Co-Regulatory Networks

JIAWEI LUO[1], LV DING[2], CHENG LIANG[3], AND NGUYEN HOANG TU[4]

[1]College of Computer Science and Electronic Engineering and the Collaboration and Innovation Center for Digital Chinese Medicine of 2011 Project of Colleges and Universities in Hunan Province, Hunan University, Changsha 410082, China
[2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[3]School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China
[4]Faculty of information and technology, Hanoi University of Industry, Hanoi 100803, Vietnam

Corresponding author: Jiawei Luo (luojiawei@hnu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61572180 and Grant 61602283.

ABSTRACT Co-regulatory networks, which consist of transcription factors (TFs), micro ribose nucleic acids (miRNAs), and target genes, have provided new insight into biological processes, revealing complicated and comprehensive regulatory relationships between biomolecules. To uncover the key co-regulatory mechanisms between these biomolecules, the identification of co-regulatory motifs has become beneficial. However, due to high-computational complexity, it is a hard task to identify co-regulatory network motifs with more than four interacting nodes in large-scale co-regulatory networks. To overcome this limitation, we propose an efficient algorithm, named large co-regulatory network motif (LCNM), to detect large co-regulatory network motifs. This algorithm is able to store a set of co-regulatory network motifs within a *G*-tries structure. Moreover, we propose two ways to generate candidate motifs. For three- or four-interacting-node motifs, LCNM is able to generate all different types of motif through an enumeration method. For larger network motifs, we adopt a sampling method to generate candidate co-regulatory motifs. The experimental results demonstrate that LCNM cannot only improve the computational performance in exhaustive identification of all of the three- or four-node motifs but can also identify co-regulatory network motifs with a maximum of eight nodes. In addition, we implement a parallel version of our LCNM algorithm to further accelerate the motif detection process.

INDEX TERMS Micro ribose nucleic acids (miRNAs), transcription factor, co-regulatory network motif, G-trie, parallel processing.

I. INTRODUCTION

In recent years, biological regulatory networks, including protein-protein interaction networks (PPIN) [1], [2], signal transduction networks (STNs) [3], [4], gene regulatory networks (GRN) [5]–[7], and metabolic networks (MN) [8], have become a hot area of research in computational biology. With the development of high-throughput technologies, the study of micro ribose nucleic acids (miRNAs), TFs, genes, and the regulatory relationships between these entities has produced a large amount of data [9]–[12]. Specifically, the co-regulatory network that combines miRNAs, TFs, and genes has become a popular research focus [13]–[16]. In contrast to a regulatory network that involves only one type of regulator, co-regulatory networks with multiple types of regulators are enriched with intricate biological regulatory relationships.

One approach to study biological regulatory networks is through network motif analysis. Network motifs are subgraphs that are statistically more significant within a given network than expected for a random network [17]. In general, if a subgraph g appears much more frequently in a given network G than in random graphs with similar degrees of distribution to G, the subgraph is considered a network motif. In colored networks involving multiple interacting node types, the topological structure of graph is ignored, but the node type and edge type are taken into consideration. The network motif containing node type and edge type information is called a colored network motif [18]. In this paper, we mainly focus on the discovery of co-regulatory motifs (colored network motifs in co-regulatory network) in large human co-regulatory networks of TFs and miRNAs to reveal their co-regulatory mechanisms. Co-regulatory network motifs discussed here are motif patterns that involve at least one TF, one miRNA and one target gene.

The general framework of the network motif identification algorithm consists of three main steps, that involve several graph theory methods, such as subgraph enumeration, graph isomorphism and random network shuffling

algorithm [16], [19], [20]. The identification of network motifs is a computationally intensive task. Even if the networks only contain a few thousand nodes, it may require several days to identify network motifs. Specifically, in co-regulatory networks, it would require significantly more time due to the increased information of node types and edge types.

Several studies related to co-regulatory network motif discovery have been published [16], [21]. These studies note that feedback loops (FBL), feed-forward loops (FFL), auto-regulation loops (ARL), bi-fans and single-input motifs (SIM) play important roles in molecular adjustment to ensure a stable physiological environment within humans. There are various methods designed to identify network motifs. For example, FANMOD [19] is one of the most widely used software programs to discover network motifs. This software is based on the ESU algorithm [22], which is very efficient and is able to avoid double enumeration of certain subgraphs. Moreover, FANMOD is capable of finding colored network motifs by adopting an edge-switching algorithm to generate randomized networks from the original network. However, Megraw *et al.* [21] traced the randomization process and discovered a large number of failed switches, which may result in insufficient network shuffling. Therefore, the author proposed the WaRSwap algorithm, which only maintains the network degree distribution. In contrast, FANMOD is required to keep the degree distribution of each node in the original network.

Network motif identification in a large-scale co-regulatory network is time consuming, as the process must enumerate a large number of co-regulatory subgraphs and determine graph isomorphism. Furthermore, the addition of the color attribute indicates that there are more classes of subgraphs and, therefore, the progress of co-regulatory motif identification will be more time-consuming. For this reason, previous exhaustive searching methods only focus their attention on co-regulatory motifs consisting of 3 or 4 nodes [16], [18]. This limitation prevents further investigation of the regulatory mechanism within the cell, especially for the intricate interplays between multiple types of regulators in gene regulation.

In this paper, we adopt a G-trie structure [23] to efficiently identify co-regulatory network motifs. A G-trie is a prefix tree data structure that is able to store a set of graphs; its efficiency benefits from reusing the information of subgraphs with common prefixes. We extend the G-trie structure to identify network motifs with sizes larger than 4 nodes in large co-regulatory networks. Specifically, we propose two sampling methods to generate candidate subgraph patterns: random walking and quick sampling. Moreover, we design a parallel version to further improve the computational efficiency of the large co-regulatory network motif (LCNM) algorithm. To determine the potential biological significance contained in co-regulatory motifs, we also analyze the cluster characteristic of the identified co-regulatory motifs.

The rest of the paper is organized as follows: section 2 presents a co-regulatory network motif identification algorithm based on G-trie structure, section 3 shows the experimental result, and section 4 presents conclusions and future directions.

II. METHOD

A. PRELIMINARIES

To identify co-regulatory network motifs in TF-miRNA co-regulatory networks, we first introduce the basic graph terminology. Here, $G(V, E)$ is defined as a co-regulatory network that involves miRNAs, TFs, and target genes together with the regulations among them. We define $V = \{V_m, V_t, V_g\}$ as the node set of the co-regulatory network, where V_m, V_t and V_g represent miRNAs, TFs and the target gene set, respectively. $T(u)$ represents the type of node u. Here, we simply use an integer to denote the miRNA, TF, and target gene vertex types. $E \subseteq (V \times V)$ is the set of directed edges. Each directed edge $e(u, v) \subseteq E$ represents a regulation between two biological molecules, such as $miRNA \rightarrow TF$, $miRNA \rightarrow gene$, $TF \rightarrow miRNA$, $TF \rightarrow TF$, or $TF \rightarrow gene$. $T(u, v)$ represents the edge types of $e(u, v)$. In addition, a co-regulatory subgraph G_k is a subgraph of size k. G_k is considered to be reduced from G, if $V(G_k) \in V(G)$, $E(G_k) \in E(G)$ and any pair of vertices u and a vertex set of the subgraph have all the edges that the same vertex set has in the complete graph G.

In the co-regulatory network, we adopt the definition of graph isomorphism from [24]. Two subgraphs G and H induced from a co-regulatory network are considered to display isomorphism if and only if there exists a one-to-one mapping $f : V(G) \rightarrow V(H)$. $\forall u, v \in V(G)$, if edge $\langle u, v \rangle \in E(G)$, then there is an edge $\langle f(u), f(v) \rangle \in E(H)$ and the node type of u, v is the same as that of node $f(u), f(v)$. In this paper, we follow the same definition of network motifs as proposed in [17] and use three standard statistical measures to evaluate the significance of the co-regulatory network motif m, i.e., observed frequency, Z_{score}, and P_{value}.

$$Z_{score}(m) = \frac{f_G(m) - f_R(\bar{m})}{\sigma(m)} \tag{1}$$

where

$$\sigma(m) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_G(m) - f_{R_i}(m))^2}$$

$$P_{value} = \frac{1 + \sum_{i=1}^{n} C(n)}{1 + n} \tag{2}$$

where

$$C(n) = \begin{cases} 1 & \text{if } f_R(m) \geq f_G(m) \\ 0 & \text{otherwise} \end{cases}$$

Here, n denotes the number of randomized graphs generated. As in the previous study, we generated 1000 randomized networks to evaluate the significance of the co-regulatory network motifs. $f_G(m)$ denotes the frequency of subgraph m in the original graph G. $f_R(\bar{m})$ and $\sigma(m)$ denote the average

and standard deviation of frequencies in these 1000 randomized networks, respectively. The cutoffs for the frequency, P_{value} and Z_{score} are set to 5, 0.01 and 2, respectively, as suggested in [17].

B. A NOVEL METHOD FOR LARGE CO-REGULATORY MOTIF DISCOVERY

Most existing motif discovery algorithms enumerate all subgraphs of a given size from a given network G and an ensemble of randomized graphs with the same degree distribution as G, which is feasible in small-scale networks. Nevertheless, these algorithms are not applicable to large-scale networks due to their high computational complexity, which limits discovery of larger motifs. As a matter of fact, in our previous study [16], we mainly searched for co-regulatory network motifs containing only three or four nodes. In addition, in a large co-regulatory network, the subgraph patterns grow exponentially as the subgraph scale increases. However there are only a minority of subgraph patterns that are motifs. Moreover, a frequent subgraph is not always a motif. In other words, previous motif identification algorithms usually waste much time on non-motif subgraph patterns.

Inspired by [25], we propose two methods to identify co-regulatory network motif in this paper: exhaustive counting and subgraph sampling.

- Exhaustive counting generates all subgraph patterns. In our previous work, the CoMoFinder algorithm was able to exhaustively generate all subgraph patterns with the given size.
- To obtain candidate subgraphs, with subgraph sampling, two sampling methods are adopted. The first approach, random walk, picks a node at random and takes a random walk until the subgraphs reach as many nodes as given. The second approach, quick sampling, is the ESU sampling algorithm [22]. However, we only generate 10 randomized networks and choose the graph patterns with larger Z-score, which have greater probability to be a motif.

C. G-TRIE STRUCTURE IN CO-REGULATORY NETWORKS

Now, the candidate subgraphs are prepared. To evaluate the significance of the set of candidate motifs, the next step is to construct the G-trie structure to store the subgraphs, as has been done in [23]. However, in this paper, the motif stored in G-trie is the co-regulatory network motif. Thus, the workflow to construct the G-trie should be modified to adapt to the new problem.

The G-trie data structure was first proposed by Ribeiro *et al.* [26] and is an efficient data structure to store a set of graphs, as shown in Figure 1. The G-trie is derived from a prefix tree structure. In this section, we use the terms node and vertex represent the nodes in the G-trie and graph vertices, respectively. A path from the root node to a leaf node represents a subgraph. The G-trie structure stores a set of graphs that share common subgraphs. As a result, the subgraph representation can be compressed, using less memory

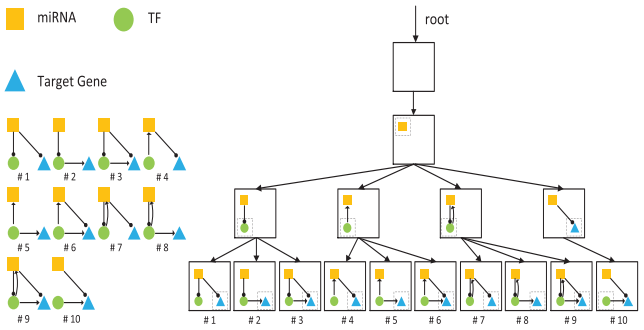

FIGURE 1. A G-trie structure storing 10 co-regulatory motifs.

to store maximum number of subgraphs. In addition, unlike previous algorithms that identify motifs one by one, the G-trie structure reuses the information of a common prefix subgraph. Therefore, the G-trie structure can substantially improve computational efficiency.

Motifs were inserted into the G-trie structure one by one. Because we want to construct a co-regulatory G-trie with fewer nodes, we give preference to the vertices with small vertex type label when calculating the canonical label. In other words, the algorithm ordered the vertex by miRNA, TF, and gene.

As mentioned above, the candidate co-regulatory subgraphs were generated in two ways: exhaustive enumeration and sampling. The G-trie structure was used to store these subgraph patterns. The G-trie construction process is detailed in Algorithm 1. In short, we selected graphs in the set S_G (line 2) and executed a series of procedures for each graph. The first procedure obtains a canonical form for the graph (line 3). The next procedure generates the symmetry-breaking condition role (lines 4). As mentioned in [27], we adopted conditions to generate the algorithm, which ensures that each subgraph is counted only once. In addition, we considered the node type when generating conditions. The AddNode function is an almost straightforward recursive procedure, which follows the path that corresponds to the graph being inserted and creates new G-trie nodes as needed (lines from 9 to 24).

D. USING G-TRIES IN CO-REGULATORY NETWORKS

Subgraphs with G-tries are queried by a recursive backtracking procedure. Algorithm 2 provides the details of the subgraph census process. Initially, we follow all G-trie root children and start with an empty partial match (line 1 and 2). We then find all candidate vertices to fill the position of that G-trie node (line 6). In this line, a set of candidate vertexes are generated, and the vertexes that meet the symmetry condition are designated with a node type. When located at a G-trie leaf, we can find a complete match to a subgraph and increment its frequency (line 9). If not, we continue as before, recursively following all possible G-trie paths from that point.

E. NETWORK RANDOMIZATION

As shown in our previous work [16], generation of randomized networks has a significant impact on motif

Algorithm 1 Creating a G-Trie From a Set of Co-Regulatory Subgraphs

Input: Candidate co-regulatory graph set S_G
Output: Co-regulatory G-trie T

```
1: T = empty co-regulatory G-trie
2: for G in S_G do
3:    Str = canomical form of G
4:    Cond = sysmmetry breaking condition of G
5:    AddNode(T.root, Str, Cond, 0, |V(G)|)
6: end for
7: Filter Condition of T
8: return T
9: Function AddNode(Node, Str, Cond, k, size)
10: Add relevant conditions of Cond to Node
11: Set the insert graph vertices type as the node type
12: if k = size then
13:    mark this Node as leaf node of G-trie
14: else
15:    for all children c of Node do
16:       if if c.connections = k-vertex of Str then
17:          AddNode(c, Str, Cond, depth + 1, size)
18:          return
19:       end if
20:    end for
21:    c = new G-trie node
22:    c.connections = k-vertex of Str
23:    AddNode(c, Str, Cond, depth + 1, size)
24: end if
```

Algorithm 2 Census of Subgraph of G-Trie T in Co-Regulatory Network G

Input: G-trie T, co-regulatory network G
Output: The frequency of each subgraph

```
1: for all children c of T.root do
2:    Census(c, ∅, G)
3: end for
4: return the frequency of each subgraph
5: Function Census(Node, V_used, G)
6: V_cand = candidates of V(G) that respect node type and
   condition of Node
7: for all v ∈ V_cand do
8:    if Node is the leaf node of G-trie then
9:       Node.frequency + +
10:   end if
11:   for all children c of Node do
12:      Census(c, V_used ⋃ v, G)
13:   end for
14: end for
```

identification. To improve the accuracy of our algorithm, we adopted the same edge exchange strategy proposed in our previous paper [16], which maintains the number of each type of regulation invariant. It has proven efficient to generate randomized networks from an original network. Experimental results show that the network randomization method adopted in [16] is able to avoid either 'under-shuffling' or 'over-shuffling' events during the randomization process.

F. PARALLEL IN COUNTING

Network motif identification is a computationally hard problem. The execution time of a sequential algorithm grows exponentially with increased motif size, especially in co-regulatory networks. Though computational complexity in subgraph enumeration improves with the use of G-tries, the entire process can be accelerated by implementing the algorithm in parallel. Therefore, in this paper, we present a parallel version of our algorithm based on the openMP library [28]. The procedure-code is described in Algorithm 3. We created threads in line 1 and output the number of each thread in line 2. Line 10 indicates parallel processing of the for-loop structure by dynamic scheduling. To remove the conflict of each thread and alter the variable frequency, we used a vector to replace the single frequency (line 13). Finally, the sum of the frequency vector is calculated in line 6.

Algorithm 3 Parallel Census of Subgraph of G-Trie T in Graph G

Input: G-trie T, Graph G, thread count *thread_count*
Output: The frequency of each subgraph

```
1: # pragma omp parallel num_threads(thread_count)
2: for all children c of T.root do
3:    thread = parallel thread number
4:    Census(c, ∅, G, thread)
5: end for
6: frequency = the sum of frequency vector
7: return the frequency of each subgraph
8: Function Census(Node, V_used, G, thread)
9: V_cand = candidates of V(G) that respect color and
   condition of Node
10: # pragma omp for schedule(dynamic)
11: for all v ∈ V_cand do
12:    if Node is the leaf node of G-trie then
13:       Node.frequency[thread] + +
14:    end if
15:    for all children c of Node do
16:       Census(c, V_used ⋃ v, G)
17:    end for
18: end for
```

III. EXPERIMENT

A. ENVIRONMENT

This study aims to increase the speed of identifying network motifs in co-regulatory networks and to identify larger-scale network motifs. To verify the efficiency of the algorithm proposed in this paper, we implemented our algorithm in C++. Experiments were performed on a general computer, which contains an Intel Xeon E3-1230 CPU with 4-cores and 8-threads and 8 GB memory.

B. DATA COLLECTION

To verify the efficiency and accuracy of our method, we adopted three co-regulatory networks from previous research. These co-regulatory networks are listed in Table 1 and include two small-scale, published co-regulatory networks: Glioblastoma multiform (GBM) and Alzheimer disease (AD). The largest co-regulatory network is derived from ENCODE project.

TABLE 1. The information of three co-regulatory networks adopted in this study.

Name	miRNAs	TFs	Genes	Regulations	Ref
GBM	99	142	167	4207	[29]
AD	388	412	2302	6040	[30]
ENCODE	736	119	15043	144473	[16]

TABLE 2. The CPU time of FANMOD, CoMoFinder and LCNM to identify 3 and 4 node co-regulatory motifs in the GBM network.

Size	FANMOD	CoMoFinder	LCNM
3	90.98	66.86	5.68
4	4494.81	4419.76	93.72

TABLE 3. The CPU time of FANMOD, CoMoFinder and LCNM to identify 3 and 4 node co-regulatory motifs in the AD network.

Size	FANMOD	CoMoFinder	LCNM
3	146.63	243.45	7.66
4	12007.93	4456.08	91.87

C. COMPARISON ALGORITHMS

In this paper, we adopted two algorithms for comparison: FANMOD [19], which is widely used in motif identification, and CoMoFinder [16], which is publicized in our previous paper. Additional algorithms exist, such as WaRSwap [21] which is designed to identify co-regulatory network motif, but this algorithm directly adopts the enumeration and classification process from FANMOD. For consistency, the duration of experiments is shown in seconds by default.

D. COMPARISON WITH PREVIOUS ALGORITHM

To evaluate the efficiency of our algorithm, we compared the running time of LCNN with CoMoFinder and FANMOD in three co-regulatory networks. Due to the limitation of computational time, previous research studied co-regulatory network motifs that only contain 3 or 4 nodes. In addition, identification of all 3- or 4-node motif types by our algorithm is highly efficient. Therefore, our comparison focuses on 3 and 4 node motifs. For a fair comparison, the parallel mode is not turned on for CoMoFinder and LCNM. Here, we compare the CPU time of the entire process of network motif identification, including time elapsed in the original graph and the 1000 randomized graphs. Compare Results are shown in Table 2, Table 3 and Table 4. LCNM shows the best performance in comparison with FANMOD and CoMoFinder.

TABLE 4. The CPU time of FANMOD, CoMoFinder and LCNM to identify 3 and 4 node co-regulatory motifs in the ENCODE network.

Size	FANMOD	CoMoFinder	LCNM
3	22008.29	13901.17	475.96
4	>1000h	>300h	26163.58

TABLE 5. The compression ratio and CPU time of LCNM without ordering of node types to identify 3-node graph patterns in 3 co-regulatory networks.

	GBM	AD	ENCODE
Compress Rate	40%	40%	40%
Time	11.73	14.18	588.71

TABLE 6. The compression ratio and CPU time of LCNM without ordering of node types to identify 4-node graph patterns in 3 co-regulatory networks.

	GBM	AD	ENCODE
Compress Rate	65.64%	65.14%	65.92%
Time	132.04	158.95	42868.40

TABLE 7. The compression ratio and CPU time of LCNM with ordering of node types to identify 3-node graph patterns in 3 co-regulatory networks.

	GBM	AD	ENCODE
Compress Rate	50%	50%	50%
Time	5.68	7.66	475.96

TABLE 8. The compression ratio and CPU time of LCNM with ordering of node types to identify 4-node graph patterns in 3 co-regulatory networks

	GBM	AD	ENCODE
Compress Rate	69.295%	68.92%	69.43%
Time	93.72	91.87	26163.58

E. PERFORMANCE IMPROVEMENT

Although G-trie is an efficient data structure to identify a set of graphs, its efficiency depends on the compression ratio [27] of the subgraphs stored within it. The compression ratio can be calculated with Eq. (3). By ordering the node types, we achieved better performance. The performance of G-trie without ordering of the node types is shown in Table 5 and Table 6. The performance of G-trie with ordering of the node types is shown in Table 7 and Table 8. The experimental results show that the compression ratio of LCNM is higher and the CPU time is shorter with node type ordering, which suggests that node type ordering improves the efficiency of LCNM.

$$\text{compression ratio} = 1 - \frac{\text{nodes in tree}}{\sum \text{nodes of stored graphs}} \quad (3)$$

F. PARALLEL PERFORMANCE

In our previous paper, we proposed a parallel version of the CoMoFinder algorithm. In this paper, we also propose a

parallel version of our algorithm and compare the efficiency of this algorithm. We identified 4-node network motifs in ENCODE co-regulated networks. The parallel version of the LCNM algorithm is designed to take full advantage of computing resources. Considering that a high-performance computer is difficult to obtain for most researchers, we only tested LCNM on a typical computer, containing an Intel Xeon E3-1230 4-cores and 8-threads and 8 GB memory. The wall clock time of CoMoFinder and LCNM is indicated in Table 9. Compared with the CoMoFinder algorithm, the LCNM algorithm showed efficient parallel performance.

TABLE 9. The wall clock time of CoMoFinder and LCNM to identify 4-node co-regulatory network motifs in ENCODE network with multiple threads.

Threads	1	2	3	4	6	8
CoMoFinder	1177.82	798.94	622.37	619.70	637.70	645.35
LCNM	29.25	15.71	11.24	9.29	8.80	8.20

TABLE 10. Co-regulatory network motifs identified by LCNM with random walk in the GBM network.

Size	5	6	7	8
Motifs	19	21	27	30
Time	535.23	4076.83	16029.86	149541.80

TABLE 11. Co-regulatory network motifs identified by LCNM with random walk in the AD network.

Size	5	6
Motifs	2	7
Time	1660.84	60936.32

G. DISCOVERY OF LARGER CO-REGULATORY NETWORK MOTIFS

We discovered larger co-regulatory network motifs by sampling connected subgraphs from the co-regulatory networks. To obtain the candidate motif type, we sampled 100 subgraphs from the original network. The number of motifs that we identified is indicated in Table 10 and Table 11. The large network motif examples are illustrated in Figure 3. Because of space constraints, we only show 3 types of motifs on each scale. Graphs from #1.1 to #4.3 in Figure 2 are generated from the GBM network. Graphs from #5.1 to #6.3 in Figure 3 are generated from the AD network. These motif patterns clearly show regulations between miRNAs, TFs and genes. For instance, in motif pattern #4.1, there are three miRNAs regulate three genes and one TF together, another miRNA regulate one of the three genes.

We also proposed a quick sampling method to generate candidate graph patterns, which we applied with the ESU sampling algorithm for 10 randomized graphs and chose the 100 graphs with the highest Z_{score}. The number of motifs that we identified is indicated in Table 12 and Table 13. Though

TABLE 12. Co-regulatory network motifs identified by LCNM with quick sampling in the GBM network.

Size	5	6	7
Motifs	30	42	35
Time	148.39	190.84	397.69

TABLE 13. Co-regulatory network motifs identified by LCNM with quick sampling in the AD network.

Size	5	6
Motifs	53	54
Time	81.07	44.48

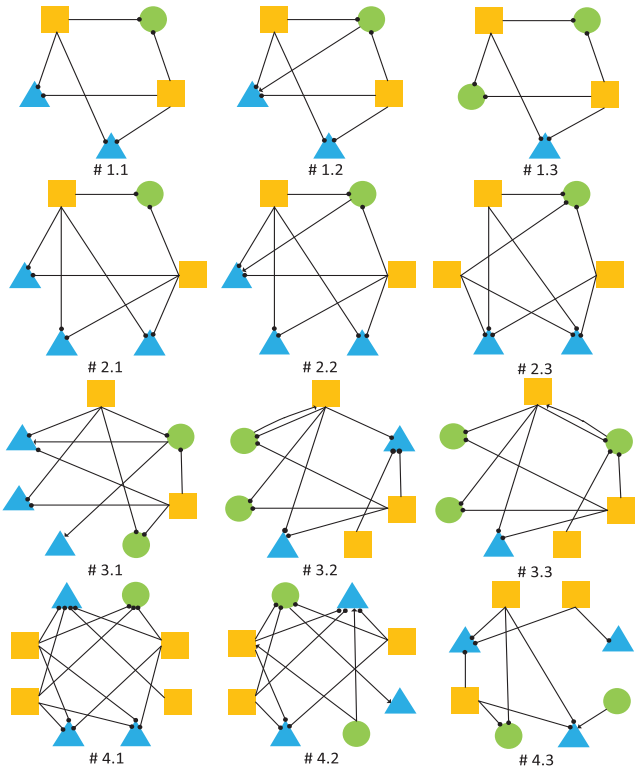

FIGURE 2. The large co-regulatory motif identified by LCNM in the GBM network.

the sampling time may be longer than random walk, it significantly increased the number of co-regulated motifs identified by LCNM. There is a phenomenon in which the motif is not always the highest frequency graph pattern. Therefore, choosing the graph patterns that have higher probability of being a motif will save much time in motif identification.

H. A CLUSTER OF CO-REGULATORY MOTIFS

To reveal the relationship between the co-regulatory network motifs, we analyzed the instances of the motifs that we identified. For example, the instances in network of motif #1.1 in Figure 3 gather into a cluster, which is shown in Figure 4. The cluster is formed by 15 motif instances that contain hsa-miR-124, hsa-miR-137, TCF4, TEAD1 and several

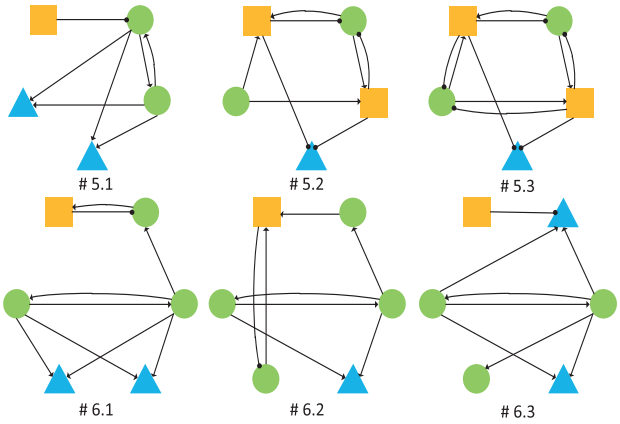

FIGURE 3. **The large co-regulatory motif identified by LCNM in the AD network.**

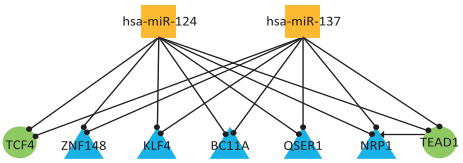

FIGURE 4. **A co-regulatory motif cluster which contains 15 motif instances of motif #1.1 in the GBM network.**

genes. The hsa-miR-137 has been observed to be enriched in the brain [31] and is related to schizophrenia [32]. Some researchers have also demonstrated that hsa-miR-137 acts as a tumor suppressor in several biological processes [32], [33], and hsa-miR-124 contains tumor suppressive function [34].

IV. CONCLUSION

In this paper, we proposed a novel algorithm to identify large network motifs based on the G-trie structure. Experimental results indicated that our method is efficient for the identification of large co-regulatory motifs. Moreover, larger network motifs provide new insights into co-regulatory networks. The advantage of the LCNM algorithm proposed in this paper consists of three components: (i) the identification of large network motifs in a short time period, (ii) lower computational complexity compared with previous methods, and (iii) an efficient parallel version of LCNM. However, generating candidate graph patterns by sampling is not applicable to identify all motif types in a co-regulatory network. Furthermore, the calculation time remains excessive in the context of a network with tens of thousands of nodes. Future studies will focus on parallel processing of this algorithm via supercomputer.

V. CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

REFERENCES

[1] B. Tian, Q. Duan, C. Zhao, B. Teng, and Z. He, "Reinforce: An ensemble approach for inferring PPI network from AP-MS data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[2] M. Li, Y. Lu, Z. Niu, and F.-X. Wu, "United complex centrality for identification of essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 370–380, Mar. 2017.

[3] M. Li, R. Zheng, Y. Li, F.-X. Wu, and J. Wang, "MGT-SM: A method for constructing cellular signal transduction networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[4] T. Wang, J. Xue, W. Tan, and B. Ye, "Learning and identifying the crucial proteins in signal transduction networks by a novel method," in *Proc. 9th Int. Conf. Comput. Sci. Edu.*, Aug. 2014, pp. 15–19.

[5] A. Zarnegar, P. Vamplew, A. Stranieri, and H. F. Jelinek, "A heuristic gene regulatory networks model for cardiac function and pathology," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 353–355.

[6] O. A. Arshad, P. S. Venkatasubramani, A. Datta, and J. Venkatraj, "Using Boolean logic modeling of gene regulatory networks to exploit the links between cancer and metabolism for therapeutic purposes," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 1, pp. 399–407, Jan. 2016.

[7] N. Avcu, N. Pekergin, F. Pekergin, and C. Guzelis, "Aggregation for computing multi-modal stationary distributions in 1-D gene regulatory networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[8] Z. Yao, B. Hu, X. Chen, Y. Xie, and L. Fang, "Modular reconfiguration of metabolic brain networks in health and cancer: A resting-state pet study," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2016, pp. 1040–1046.

[9] J. Luo, Q. Xiao, C. Liang, and P. Ding, "Predicting MicroRNA-disease associations using kronecker regularized least squares based on heterogeneous omics data," *IEEE Access*, vol. 5, pp. 2503–2513, 2017.

[10] P. Ding, J. Luo, C. Liang, J. Cai, Y. Liu, and X. Chen, "A novel group wise-based method for calculating human miRNA functional similarity," *IEEE Access*, vol. 5, pp. 2364–2372, 2017.

[11] G. Li, J. Luo, Q. Xiao, C. Liang, P. Ding, and B. Cao, "Predicting microrna-disease associations using network topological similarity based on deepwalk," *IEEE Access*, vol. 5, pp. 24032–24039, 2017.

[12] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, 2017.

[13] H. Ye *et al.*, "MicroRNA and transcription factor co-regulatory network analysis reveals miR-19 inhibits CYLD in T-cell acute lymphoblastic leukemia," *Nucl. Acids Res.*, vol. 40, no. 12, pp. 5201–5214, 2012.

[14] K. R. Chng *et al.*, "A transcriptional repressor co-regulatory network governing androgen response in prostate cancers," *EMBO J.*, vol. 31, no. 12, pp. 2810–2823, 2012.

[15] J. Luo, G. Xiang, and C. Pan, "Discovery of microRNAs and transcription factors co-regulatory modules by integrating multiple types of genomic data," *IEEE Trans. Nanobiosci.*, vol. 16, no. 1, pp. 51–59, Jan. 2017.

[16] C. Liang, Y. Li, J. Luo, and Z. Zhang, "A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and microrna co-regulatory networks in human," *Bioinformatics*, vol. 31, no. 14, pp. 2348–2355, 2015.

[17] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[18] P. Ribeiro and F. Silva, *Discovering Colored Network Motifs*. Berlin, Germany: Springer, 2014, pp. 107–118.

[19] S. Wernicke and F. Rasche, "FANMOD: A tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, 2006.

[20] S. Omidi, F. Schreiber, and A. Masoudi-Nejad, "MODA: An efficient algorithm for network motif discovery in biological networks," *Genes Genet. Syst.*, vol. 84, no. 5, pp. 385–395, 2009.

[21] M. Megraw, S. Mukherjee, and U. Ohler, "Sustained-input switches for transcription factors and microRNAs are central building blocks of eukaryotic gene circuits," *Genome Biol.*, vol. 14, no. 8, p. R85, 2013.

[22] S. Wernicke, "Efficient detection of network motifs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 3, no. 4, pp. 347–359, Oct./Dec. 2006.

[23] P. Ribeiro and F. Silva, "Querying subgraph sets with g-tries," in *Proc. 2nd ACM SIGMOD Workshop Databases Social Netw.*, 2012, pp. 25–30.

[24] B. D. McKay and A. Piperno, "Practical graph isomorphism, II," *J. Symbolic Comput.*, vol. 60, pp. 94–112, Jan. 2014.

[25] J. A. Grochow and M. Kellis, "Network motif discovery using subgraph enumeration and symmetry-breaking," in *Proc. Annu. Int. Conf. Res. Comput. Mol. Biol.*, 2007, pp. 92–106.

[26] P. Ribeiro, F. Silva, and L. Lopes, "Efficient parallel subgraph counting using g-tries," in *Proc. IEEE Int. Conf. Cluster Comput.*, Sep. 2010, pp. 217–226.

[27] P. Ribeiro, "Efficient and scalable algorithms for network motifs discovery," Ph.D. dissertation, Faculty Sci., Univ. Porto, Porto, Portugal, 2011.
[28] L. Dagum and R. Menon, "OpenMP: An industry standard API for shared-memory programming," *IEEE Comput. Sci. Eng.*, vol. 5, no. 1, pp. 46–55, Mar. 1998.
[29] J. Sun, X. Gong, B. Purow, and Z. Zhao, "Uncovering microrna and transcription factor mediated regulatory networks in glioblastoma," *PLoS Comput. Biol.*, vol. 8, no. 7, p. e1002488, 2012.
[30] W. Jiang *et al.*, "Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease," *Bioinformatics*, vol. 29, no. 20, pp. 2596–2602, 2013.
[31] M. L. Ou *et al.*, "Association between miR-137 polymorphism and risk of schizophrenia: A meta-analysis," *Genet. Mol. Res.*, vol. 15, no. 3, pp. 1–12, 2016.
[32] S. Wu *et al.*, "MicroRNA-137 inhibits *EFNB2* expression affected by a genetic variant and is expressed aberrantly in peripheral blood of schizophrenia patients," *EBioMedicine*, vol. 12, pp. 133–142, Oct. 2016.
[33] M. Neault, F. A. Mallette, and S. Richard, "miR-137 modulates a tumor suppressor network-inducing senescence in pancreatic cancer cells," *Cell Rep.*, vol. 14, no. 8, pp. 1966–1978, 2016.
[34] S. M. Wilting *et al.*, "Methylation-mediated silencing and tumour suppressive function of *hsa-miR-124* in cervical cancer," *Mol. Cancer*, vol. 9, no. 1, p. 167, 2010.

JIAWEI LUO received the Ph.D. degree in computer science from Hunan University in 2008. She is currently a Professor with the College of Computer Science and Electronic Engineering, Hunan University. She has published about 50 research papers in various international journals and proceedings of conferences. Her research interests include data mining, computational biology, and bioinformatics.

LV DING received the B.Sc. degree in mathematics from the China University of Mining and Technology, China, in 2015. He is currently pursuing the master's degree with the College of Computer Science and Electronic Engineering, Hunan University. His research interests include data mining and bioinformatics.

CHENG LIANG received the joint Ph.D. degree in computer science from Hunan University and the Donnelly Centre, University of Toronto, in 2014. She is currently an Assistant Professor with the School of Information Science and Engineering, Shandong Normal University, China. Her research interests include graph mining, semi-supervised learning, and big data analytics.

NGUYEN HOANG TU received the master's degree in computer science and the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include intrusion detection and prevention, vulnerability analysis, network security, bioinformatics, and data mining.