Bioinformatics, 31(14), 2015, 2348–2355
doi: 10.1093/bioinformatics/btv159
Advance Access Publication Date: 18 March 2015
Original Paper

Systems biology

A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and microRNA co-regulatory networks in human

Cheng Liang[1,†], Yue Li[2,†], Jiawei Luo[1,*] and Zhaolei Zhang[2,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, China and [2]Department of Computer Science, Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Igor Jurisica

Received on November 4, 2014; revised on January 3, 2015; accepted on March 13, 2015

Abstract

Motivation: Interplays between transcription factors (TFs) and microRNAs (miRNAs) in gene regulation are implicated in various physiological processes. It is thus important to identify biologically meaningful network motifs involving both types of regulators to understand the key co-regulatory mechanisms underlying the cellular identity and function. However, existing motif finders do not scale well for large networks and are not designed specifically for co-regulatory networks.

Results: In this study, we propose a novel algorithm CoMoFinder to accurately and efficiently identify composite network motifs in genome-scale co-regulatory networks. We define composite network motifs as network patterns involving at least one TF, one miRNA and one target gene that are statistically significant than expected. Using two published disease-related co-regulatory networks, we show that CoMoFinder outperforms existing methods in both accuracy and robustness. We then applied CoMoFinder to human TF-miRNA co-regulatory network derived from *The Encyclopedia of DNA Elements* project and identified 44 recurring composite network motifs of size 4. The functional analysis revealed that genes involved in the 44 motifs are enriched for significantly higher number of biological processes or pathways comparing with non-motifs. We further analyzed the identified composite bi-fan motif and showed that gene pairs involved in this motif structure tend to physically interact and are functionally more similar to each other than expected.

Availability and implementation: CoMoFinder is implemented in Java and available for download at http://www.cs.utoronto.ca/~yueli/como.html.

Contact: luojiawei@hnu.edu.cn or zhaolei.zhang@utoronto.ca

Supplementary information: supplementary data are available at *Bioinformatics* online.

1 Introduction

At the maturity of high-throughput technologies such as microarray and next-generation sequencing, a vast amount of genome-wide data have been generated to interrogate human regulatory networks at both transcriptional and post-transcriptional level (Barabasi and Oltvai, 2004; Gerstein *et al.*, 2012). The combination and orchestration between regulatory mechanisms at both levels are central to a precise gene expression program (Cheng *et al.*, 2011). In contrast to a regulatory network involving only one type of regulators, a co-regulatory network may involve transcription factors (TF), microRNAs (miRNA) and their (shared) target genes (Martinez and Walhout, 2009). The interplay between TF and miRNA underlies a coordinated circuitry responsible for a variety of cellular processes (Hobert, 2008); many of these processes are carried out by recurring structures called *network motifs*.

Network motifs are recurring patterns of connectivity, occurring significantly more frequently than expected (Alon, 2007; Milo *et al.*, 2002). Identification of network motifs in the context of co-regulatory networks can yield new mechanistic insights into biological system and offer crucial clues into gene regulation (Grochow and Kellis, 2007; Ideker *et al.*, 2011; Martinez and Walhout, 2009). For instance, it has been shown that miRNA-mediated feed-forward loops (FFLs) and feedback loops (FBLs) enhance the robustness of gene regulation in mammalian genomes (Ebert and Sharp, 2012; Herranz and Cohen, 2010; Tsang *et al.*, 2007). Furthermore, TF-mediated FFLs were theoretically proved to ensure the stability of gene expression against stochastic fluctuations (Riba *et al.*, 2014), whereas the TF ↔ miRNA FBLs are demonstrated to provide 'high flux capacity' to coordinate the high information flow that passes through the miRNA and TF within the FBLs (Martinez *et al.*, 2008).

Various methods have been developed to detect network motifs. However, only a handful of them could be applied on co-regulatory networks. As one of the widely used tools (Grochow and Kellis, 2007; Kashani *et al.*, 2009; Kashtan *et al.*, 2004; Khakabimamaghani *et al.*, 2013; Li *et al.*, 2012; Panni and Rombo, 2013), FANMOD is an efficient motif detection algorithm, which takes into account networks with different types of nodes and edges by exploiting the concept of graph isomorphism (Junttila and Kaski, 2007; McKay and Piperno, 2014; Wernicke, 2006). Although FANMOD has been widely applied to many kinds of biological networks (Gerstein *et al.*, 2012; Morgan and Soltesz, 2008; Roy *et al.*, 2010), its randomization scheme is limited to covering only a small range of subgraphs frequency reflected by the low-variance of the number of certain subgraphs in the random network ensemble (Beber *et al.*, 2012; Megraw *et al.*, 2013). Besides, FANMOD is not capable of utilizing multi-cores. WaRSwap is another network shuffling algorithm, which was proposed to discover network motifs in large TF-miRNA co-regulatory networks (Megraw *et al.*, 2013). Briefly, WaRSwap breaks the co-regulatory network into separate layers according to the node types and then shuffles the regulatory edges in each layer, while keeping the node degree distribution unchanged. Notably, WaRSwap is only a randomization technique that must be used accompanied with a motif discovery tool such as FANMOD, which substantially limits its application scope. Additionally, WaRSwap requires predefining values of several free parameters to adjust the sampling weights during the randomization process, which may affect the final outcomes. The limitations of these existing methods make it necessary to develop new method that can effectively detect reliable network motifs in large co-regulatory networks consisting of miRNAs, TFs and target genes.

In this study, we propose a novel network motif search algorithm called *CoMoFinder,* which is developed based on a parallel subgraph enumeration strategy to efficiently and accurately identify composite motifs in large TF-miRNA co-regulatory networks. Distinct from previous algorithms, CoMoFinder seeks for a maximum difference between a given co-regulatory network and its randomized counterpart to ensure sufficient shuffling during the randomization process, which provides a reliable background distribution of the random network ensemble for motif discovery. Compared with existing methods that were applied to the published co-regulatory networks, CoMoFinder achieved favorable performance in terms of accuracy, robustness and computational complexity. After having established its superior performance, we next applied CoMoFinder to systematically discover recurring 4-node motifs in human co-regulatory network derived from The Encyclopedia of DNA Elements (ENCODE) consortium. We discovered 44 composite motifs involving the previously well-studied 3-node FFLs and are corroborated by a number of biological evidences such as enrichment for Gene Ontology (GO) terms or canonical pathways. Together, we present a novel framework for network motif discovery, demonstrate its utility in analyzing large human co-regulatory networks and envision it being a useful tool for future analyses of various large biological networks.

2 Materials and methods

2.1 Notations

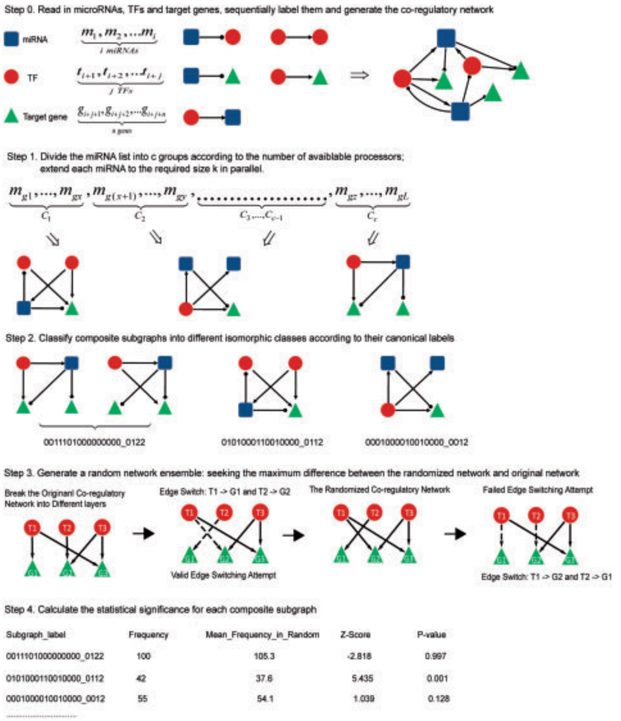
The main purpose of CoMoFinder is to detect all the composite network motifs in a given TF-miRNA co-regulatory network. Here, we define a network $G(V,E)$ that consists of miRNAs, TFs and their target genes as a TF-miRNA co-regulatory network. $V = \{V_m, V_t, V_g\}$ is a finite set of vertices, where V_m, V_t and V_g denote the vertices set of miRNAs, TFs and target genes, respectively. $E \subseteq (V \times V)$ is a finite set of directed edges, where $e(u,v) \in E$ indicates a regulation from a source vertex u to a target vertex v. Assuming no direct regulations between miRNAs, in total there are five types of regulations: $miRNA \to TF$, $miRNA \to gene$, $TF \to miRNA$, $TF \to TF$ and $TF \to gene$. A composite (or co-regulatory) subgraph $G_s = (V_s, E_s)$ of $G = (V,E)$ is then defined as a weakly connected subgraph that contains at least one vertex from each vertex set $\{V_m, V_t, V_g\}$. Notably, the smallest size k of the composite subgraphs in our study is 3, *i.e.* the number of distinct vertex types. For clarity, we simply use the integers {0, 1, 2} to denote the vertex types of miRNA, TF and target gene accordingly in our analysis. For example, a TF-mediated FFL which consists of three regulations (i.e. miRNA → TF/gene, TF → gene) could be denoted as 011001000_012, where the substring before '_' stands for the regulation relations among the three factors and the substring after '_' corresponds to the types of the three nodes (i.e. 012 encodes miRNA, TF and gene, respectively).

2.2 Method overview

The general framework of CoMoFinder can be divided into three sequential steps: (i) composite subgraph enumeration; (ii) grouping composite subgraphs into isomorphic classes and recording their number of occurrences and (iii) generating a random network ensemble and calculating the statistical significance for each composite subgraph class. Figure 1 presents an overall description of the algorithm.

2.3 Size-*k* composite subgraph enumeration and classification

Given a TF-miRNA co-regulatory network G, the algorithm CoMoFinder enumerates all of the size-*k* composite subgraphs. To use node indices to eliminate repetitive enumerations and extensions in the following analysis, our algorithm first groups the vertices of different types separately and sequentially. In particular, we first read all the miRNAs from the network, then all the TFs and last the target genes. Once the different types of nodes are arranged in order, CoMoFinder then extends every miRNA in the miRNA list to find all the composite subgraphs that contain at least one node from each node type. During the extension process, we employ a similar strategy


Step 0. Read in microRNAs, TFs and target genes, sequentially label them and generate the co-regulatory network
miRNA
$\underbrace{m_1, m_2, \ldots m_i}_{i \text{ miRNAs}}$
TF
$\underbrace{t_{i+1}, t_{i+2}, \ldots t_{i+j}}_{j \text{ TFs}}$
Target gene
$\underbrace{g_{i+j+1}, g_{i+j+2}, \ldots g_{i+j+n}}_{n \text{ genes}}$
Step 1. Divide the miRNA list into c groups according to the number of avaiblable processors;
extend each miRNA to the required size k in parallel.
$\underbrace{m_{g1}, \ldots, m_{gx}}_{C_1}, \underbrace{m_{g(x+1)}, \ldots, m_{gy}}_{C_2}, \underbrace{\ldots\ldots\ldots\ldots}_{C_3, \ldots, C_{c-1}}, \underbrace{m_{gz}, \ldots, m_{gL}}_{C_c}$
Step 2. Classify composite subgraphs into different isomorphic classes according to their canonical labels
0011101000000000_0122
0101000110010000_0112
0001000010010000_0012
Step 3. Generate a random network ensemble: seeking the maximum difference between the randomized network and original network
Break the Originanl Co-regulatory Network into Different layers
Edge Switch: T1 -> G1 and T2 -> G2
Valid Edge Switching Attempt
The Randomized Co-regulatory Network
Failed Edge Switching Attempt
Edge Switch: T1 -> G2 and T2 -> G1
T1
T2
T3
G1
G2
G3
Step 4. Calculate the statistical significance for each composite subgraph

Subgraph_label	Frequency	Mean_Frequency_in_Random	Z-Score	P-value
0011101000000000_0122	100	105.3	-2.818	0.997
0101000110010000_0112	42	37.6	5.435	0.001
0001000010010000_0012	55	54.1	1.039	0.128

Fig. 1. An overall workflow of CoMoFinder. Given a co-regulatory network involving miRNAs, TFs and target genes, CoMoFinder first divides the miRNA list into *c* independent groups, enabling multi-threaded computations. Based on the topological properties, each subgraph is then assigned with a refined canonical label associated with an isomorphic class. CoMoFinder then tries to generate random networks with maximum difference from the observed network. Finally, the significance of motif occurrences is tested relative to the background distribution constructed from the random network ensemble

as FANMOD (Wernicke, 2006) by recursively adding the valid neighbors to the current subgraph. We implicitly build a tree based on the neighborhood relationship, where the current miRNA is the root and the maximum depth of the tree is *k*. We then recursively descend the tree from the root nodes and choose possible child nodes at each level till the subgraph reaches the required size. There are three restrictions to choose a child node while descending the tree: (i) a child can be chosen if and only if it is in the exclusive neighborhood with respect to the current subgraph; (ii) all of the children in a particular tree must have numerical labels larger than that of the root node and (iii) only children of a specific node type can be chosen if there is a node type requirement. By recursively finding the valid neighbors, we extend the root node to the required size of composite subgraphs. In addition, we use a 'divide and conquer' approach by splitting the miRNA list into c groups according to the number of available processors and extending the c groups in parallel. The three aforementioned restrictions on extending the subgraphs ensure that the algorithm enumerates each composite subgraph exactly one time in either a linear or a parallel process. Moreover, by adding the node type restrictions during the enumeration process, the search space of our algorithm is significantly reduced compared with FANMOD. Algorithm 1 outlines the corresponding pseudocode.

After enumerating all of the size-*k* composite subgraphs in G, the algorithm needs to classify them into different isomorphic classes. Popular network motif discovery algorithms resort to graph isomorphism tools such as NAUTY (McKay and Piperno, 2014), SAUCY (Darga *et al.*, 2004) and Bliss (Junttila and Kaski, 2007). However, in the context of co-regulatory networks, we could use a more straightforward method to efficiently group composite subgraphs into isomorphic classes because (i) the size *k* of the composite network motifs in our study is moderate (i.e. 3 and 4), making exhaustive enumeration feasible and (ii) the number of distinct composite subgraph isomorphic classes is quite small compared with the number of enumerated composite subgraphs. To take advantage of this feature, CoMoFinder separates the isomorphism testing process from the enumeration process. It first enumerates all the subgraphs and directly groups them into classes without considering their isomorphism. After the enumeration process is finished, the algorithm then determines the isomorphism among subgraph classes and merges them accordingly. This feature is distinct from FANMOD and WaRSwap, which calculate the graph canonical labels for each enumerated subgraph and conduct billions of the graph isomorphism tests during the enumeration process. To obtain a refined canonical label for each subgraph class during the isomorphism testing process, we simply calculate all the combinations and select the lexicographically largest one for the adjacency matrix with the restriction that the nodes must be in the order of miRNAs at first, TFs at second and target genes at last. To distinguish different composite subgraphs with the same canonical labels, we also concatenate the node types for each node at the end of the subgraph labels (Fig. 1).

2.4 Network randomization

The aim of motif discovery is to compare the frequency of particular subgraphs in the input network with its frequency in randomized networks and see if they are significantly enriched/depleted in the given network (Megraw *et al.*, 2013). The generation of the randomized network ensemble is therefore a key step of the algorithm since it highly affects the statistical significance for each isomorphic class. Here, we propose a new network randomization strategy to maximize the difference between the observed and shuffled network while maintaining the basic network topology including the in/out degree for each node. Specifically, we first break the co-regulatory network into several subnetworks, which only contain certain types of nodes and edges. To keep the topology of the co-regulatory network unchanged as far as possible during the shuffling process, we also take bidirectional regulations between regulators as a distinct edge type. Therefore, a common co-regulatory network would be separated into seven network subtypes, corresponding to miRNA → TF/gene unidirectional regulations, miRNA/TF ↔ TF bidirectional regulations or TF → miRNA/TF/gene unidirectional regulations. Our algorithm then seeks for a maximum difference between the given network and each of the randomized subnetworks to avoid either 'under-shuffling' or 'over-shuffling' situation during the randomization process. By 'under-shuffling', we refer to the fact that during the edge swap procedure, only a small portion of the switchable edges are actually swapped; whereas for 'over-shuffling' situation, we meant that edges which are switched at first are switched back after a certain period of time. Both of the situations should be avoided as much as possible as we want the subnetworks to be sufficiently randomized. Suppose the total number of edges in a given network is E, the theoretical maximum number of different edges E_m between a given network and a fully shuffled network should be $0 \leq E_m \leq E$. For each subnetwork, we choose two edges (e.g. X1 → Y1, X2 → Y2) to swap only if the observed network does not contain any of the new edges after the swap (e.g. X1 → Y2, X2 → Y1) (Fig. 1); otherwise the switching is prohibited. We repeat this process until all edges are fully shuffled or the algorithm reaches a pre-defined number of iterations (default: 100). Thus, our approach ensures that the network is randomized sufficiently and provides a

Algorithm 1. Composite subgraph enumeration process

```
Input: miRNA list M, co-regulatory network G, composite subgraph size k;
Output: The number of size-k composite subgraphs in each isomorphic class;
Step 1: $M \rightarrow \{M_1, M_2, ..., M_C\}$; // C is the number of available processors
Step 2: for$(m_j$ in $M_i)$ $i = 1, 2, ..., C$ // Start parallelism
            $V_{Subgraph} \leftarrow \{m_j\}$, $V_{Extension} \leftarrow \{u \in N(\{m_j\}) : u > m_j\}$;
            Extend_subgraph$(V_{Subgraph}, V_{Extension}, m_j)$;
        end for
Step 3: for$(i$ in $1 : C)$
            $R_{total} \leftarrow R_i$;
            $R_{final} \leftarrow$ Isomorphism$(R_{total})$;
        end for
Function: Extend_Subgraph$(V_{Subgraph}, V_{Extension}, v)$
E1: While$(|V_{Extension}| > 0)$
E2:     $w \leftarrow V_{Extension}[0]$; $V_{Extension} \leftarrow V_{Extension} / w$;
E3:     $F_{Node\_Type} \leftarrow$ Check$(V_{Subgraph}, w)$ //Check node restrictions
E4:       if$(F_{Node\_Type} == -1)$ //"-1" means no node type restriction
            $V'_{Extension} \leftarrow V_{Extension} \cup \{u \in N_{excl}(w, V_{Subgraph}) : u > v\}$;
          else
            $V'_{Extension} \leftarrow V_{Extension} \cup$
            $\{u \in N_{excl}(w, V_{Subgraph}) : u > v : node\_type(u) == F_{Node\_Type}\}$;
          end if
E5:     $V'_{Subgraph} \leftarrow V_{Subgraph} \cup \{w\}$;
E6:     if$(|V'_{Subgraph}| == k - 1)$ $R_i \leftarrow R_i \cup$ Count_Subgraph$(V'_{Subgraph}, V'_{Extension})$;
        else        Extend_Subgraph$(V'_{Subgraph}, V'_{Extension}, v)$;
        end if
    end while
```

Table 1. Summary of the three co-regulatory networks used in this study

Datasets	miRNAs	TFs	Genes	Regulations
GBM	99	142	167	4207
AD	388	412	2302	6040
ENCODE	736	119	15 043	144 473

reliable background random network ensemble to evaluate the significance for all the composite subgraphs.

2.5 Motif evaluation

Three standard statistical measures are used in our approach to evaluate the significance of each subgraph found in the input network based on the observed frequency, Z score and P value (definitions are provided in Supplementary Information). In this study, we generated 1000 randomized networks for each observed network using the proposed edge-switching scheme and set the cutoffs for Z score, P value and frequency to 2, 0.01 and 5, respectively. Subgraphs that satisfy Z score > 2, P value < 0.01 and frequency $>$ 5 at the same time are considered as motifs.

2.6 Method comparisons

To compare the performance of our method, we applied two other leading methods in detecting network motifs: FANMOD and WaRSwap to the same testing data. FANMOD is one of the most popular motif discovery tools that can cope with colored networks. WaRSwap is based on FANMOD with modification on the randomization scheme to enable motif analysis on large multi-layer co-regulatory networks. The source code of FANMOD and WaRSwap was downloaded at http://theinf1.informatik.uni-jena.de/motifs/ and http://megraw.cgrb.oregonstate.edu/software/WaRSwap/, respectively. Default parameters were used for both tools. As the main focus of CoMoFinder is to find composite network motifs, we only chose network motifs consisting of miRNAs, TFs and genes from both methods to make fair comparisons.

2.7 Data collection

Because of the computational limitation of FANMOD and WaRSwap, we used two small-scaled published co-regulatory networks: glioblastoma multiforme (GBM) (Sun *et al.*, 2012) and Alzheimer disease (AD) (Jiang *et al.*, 2013) for method comparison. After established the model confidence, we then applied CoMoFinder to a much larger network derived from ENCODE project to elucidate novel mechanistic design in a more unbiased human co-regulatory network. The GBM data were downloaded from the Supplementary File, Supplementary Table S7 of Sun *et al.* and AD data were obtained from the Supplementary File S3 of Jiang *et al.* The processed ENCODE data were retrieved from the consortium website (enets2.Proximal_filtered.txt, enets10.TF-miRNA.txt and enets11.miRNA-gene.txt, http://encodenets.gersteinlab.org/). Details of the three datasets are listed in Table 1.

3 Results

3.1 Comparison of generated random network ensembles and identified networks motifs

We applied each method on GBM and AD co-regulatory networks and compared the overlap of network motifs (size 3 or 4) discovered

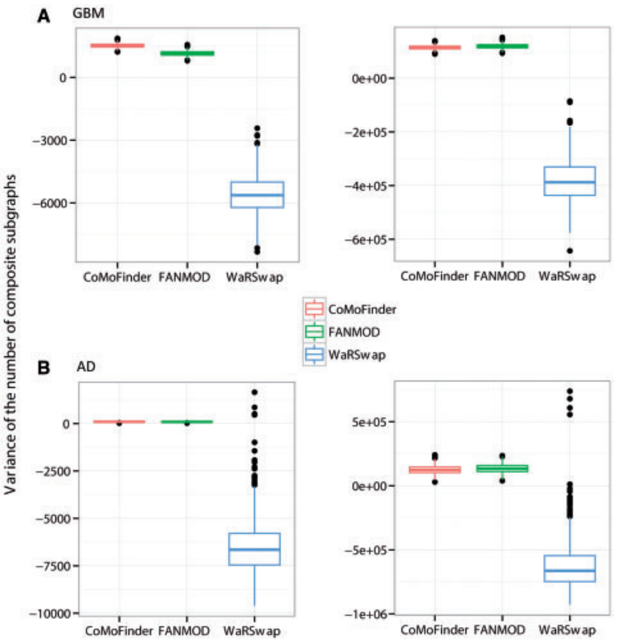


Fig. 2. Variance of the total number of composite subgraphs between the randomized networks and observed network. One thousand random networks were generated by different randomization schemes employed by CoMoFinder, FANMOD and WaRSwap, respectively. (**A**) GBM co-regulatory network. Left panel is the variance of the total number of composite subgraphs of size 3 and right panel is that of size 4. (**B**) Same as (A) but for co-regulatory network in AD

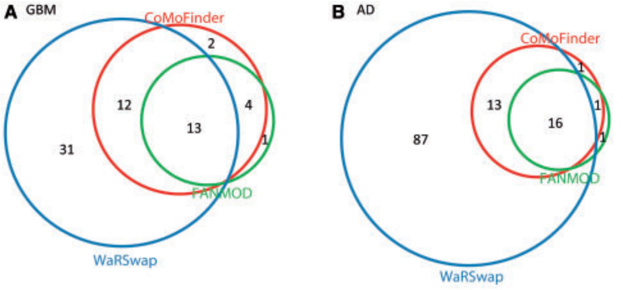


Fig. 3. Overlaps of network motifs among CoMoFinder, FANMOD and WaRSwap. The size of the circles is proportional to the number of the co-regulatory network motifs found by each method applied to (**A**) GBM and (**B**) the AD co-regulatory network

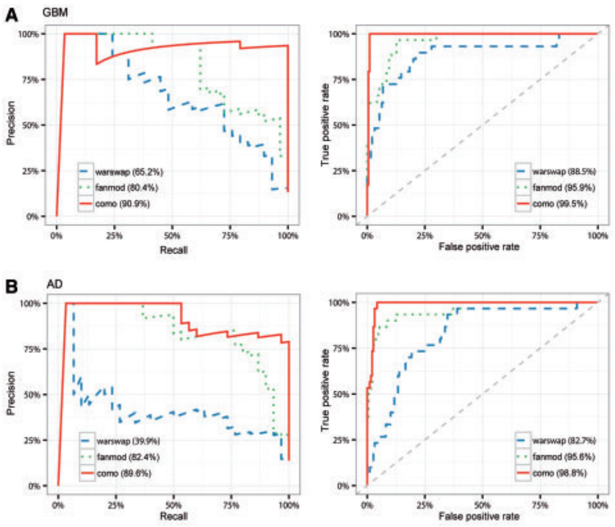


Fig. 4. Power analysis of motif discovery. The three methods were compared using (**A**) the GBM and (**B**) the AD co-regulatory network. PR (left) and ROC curves (right) were constructed using Z scores as metrics specific to each method. The percentages in the parentheses indicate the corresponding area under the PR or ROC curves

by them. For network motifs of size 3 and size 4, CoMoFinder discovered 10/10 and 219/218 different types of composite subgraphs for GBM/AD co-regulatory network, respectively. Among the 10/10 size-3 composite subgraphs, CoMoFinder only identified 1 motif for each dataset, which is in agreement with FANMOD and WaRSwap, whereas for the 219/218 size-4 composite subgraphs, CoMoFinder identified 31/31 network motifs, which differs substantially from those detected by FANMOD and WaRSwap. Because of the smaller number of 3-node motifs discovered, we hereafter only focused on the results derived from 4-node motif analysis below due to the much larger search space and the gain of statistical power. Because the frequency of the composite subgraphs highly affects the statistical significance of motif outputs, we first examined for each method the variance of the differences between the number of composite subgraphs N_t found in the 1000 shuffled networks and the observed network (Fig. 2, Supplementary Fig. S1). Here, we expect the random networks to have similar number of subgraphs to that of the observed network, such that the motifs identified in the observed network in the subsequent steps are not due to the change in search space size. Indeed, CoMoFinder and FANMOD both maintain a similar number of composite subgraphs in the randomized networks comparing with the observed networks; in contrast, the total number of composite subgraphs is substantially lower in the WaRSwap-randomized networks. These results suggest that WaRSwap may overestimate the significance of candidate subgraphs. Indeed, we found that WaRSwap predicted a much larger set of the motifs than CoMoFinder or FANMOD did (Fig. 3). Importantly, CoMoFinder-derived motif sets have significant overlap with the motif sets derived from the other two methods and identified more motifs than FANMOD. Thus, CoMoFinder appears to achieve a good balance between sensitivity and specificity comparing with FANMOD and WaRSwap.

3.2. Power analysis

To more rigorously evaluate the statistical power of WaRSwap, FANMOD and CoMoFinder, we next set out to construct an empirical gold standard motif set. As FANMOD identifies the least number of motifs, the results would bias toward FANMOD if we choose the intersection of the motifs from all three methods as gold standard. We thus select motifs that were identified by at least two of the three methods as the gold-standard motif set to avoid such bias. In total, we obtained 29 and 30 such motifs for GBM and AD, respectively. We then constructed for each method the precision/recall (PR) as well as the receiver operating characteristic (ROC) curves using the Z scores as metric. As shown in Figure 4, CoMoFinder outperformed FANMOD and WaRSwap on both test datasets in terms of area under the PR or ROC curves (AUROC or AUPRC). We also tested the performance of CoMoFinder by using motifs detected only by FANMOD or WaRSwap as an alternative gold-standard motif set and compared the AUPRC and AUROC of CoMoFinder with that of the other method (Supplementary Figs. S2 and S3). The results are consistent with Figure 4. Moreover, we constructed a synthetic network based on network motifs and compared the

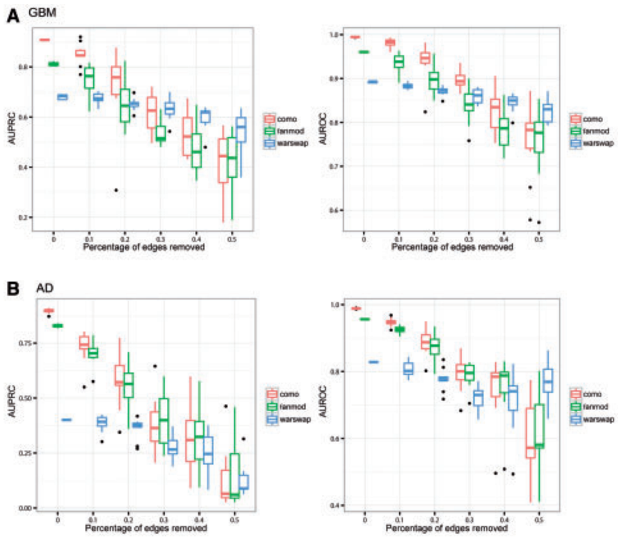


Fig. 5. Robustness test on networks with missing edges. Each method was repeatedly applied to the same networks with random edge removals. Decreases of AUPRC and AUROC corresponding to the random removal are used as metric to evaluate the robustness for each method performing on (**A**) the GBM and (**B**) the AD co-regulatory network

Table 2. The averaged computational time used for composite subgraph enumeration and classification process of size 3 and size 4 on three datasets of FANMOD and CoMoFinder based on five runs, respectively (time unit: seconds)

Methods	Datasets					
	GBM		AD		ENCODE	
	3	4	3	4	3	4
FANMOD	1.00	10.00	1.00	29.00	52.00	175 53
CoMoFinder	1.29	12.88	1.57	16.94	13.12	359.1

performance of three methods (Supplementary Information, Section 3). As expected, our method outperformed FANMOD and WaRSwap on the synthetic network as well. The superior performance of our algorithm is perhaps attributable to its ability to effectively control for systematic bias of the observed networks via a more sensible randomization scheme, which would otherwise produce misleading results.

3.3 Robustness analysis

To compare the robustness of three methods, we randomly removed 10–50% of the edges and re-evaluated the AUROC and AUPRC. CoMoFinder maintains its superior performance up to 30% edge removal (Fig. 5). Though the robustness of WaRSwap is generally consistent against edge removal, its performance remains at a lower level compared with CoMoFinder and FANMOD from 0 to 30% stage. FANMOD exhibits similar trends as CoMoFinder throughout the process, while its performance continually worse than CoMoFinder. Overall, CoMoFinder appears to be the most robust and accurate method among the three tested approaches.

3.4 Computational complexity

We next examined the time complexity of the three methods by evaluating their computational time used for subgraph enumeration and classification process. Because WaRSwap directly adopts these two processes from FANMOD, here we only compared CoMoFinder with FANMOD. Experiments were performed on a computer cluster which contains 32 Dual-Core AMD Opteron (tm) 8218 Processors and 64 GB of memory. For small-scaled co-regulatory networks such as GBM and AD, it took less than 30 s for both algorithms to finish enumeration and classification of composite subgraphs of size 3 or 4 (Table 2). For large co-regulatory network such as ENCODE, our program completed the subgraph census process of all the size-4 composite subgraphs in 360 s with 32 cores, while it took about 5 h for FANMOD to finish the same procedure (Table 2). Moreover, FANMOD stores all the searched subgraphs in memory and outputs them all at once when required, which would cause memory overhead with the exponentially increased number of subgraphs. In contrast, CoMoFinder is more memory efficient as it routinely outputs each of the searched subgraphs to intermediate files instead. Taken together, CoMoFinder is more efficient than FANMOD in searching for composite motifs in terms of both computational time and memory usage since it not only has much smaller search space but also can leverage multiprocessors.

3.5 Overlap between motifs discovered from GBM, AD and ENCODE

Encouraged by the aforementioned results, we next applied CoMoFinder to a much larger human TF-miRNA co-regulatory network derived from ENCODE project. After a full enumeration and classification of all the composite subgraphs of size 4 in the co-regulatory network, CoMoFinder found 256 size-4 subgraph types, of which 44 are considered as network motifs (Supplementary Table S1 and Fig. S4). Among these 44 recurring TF-miRNA motifs, 9 were also observed in the GBM and AD co-regulatory networks (Supplementary Fig. S5). We then tested the overlap significance between the motif sets detected in each of the three co-regulatory networks by hypergeometric test. The overall unique composite subgraph types in GBM ∪ ENCODE, AD ∪ ENCODE and GBM ∪ AD are 258, 258 and 242, whereas the number of common motifs between GBM/ENCODE, AD/ENCODE and GBM/AD is 9, 9 and 3, respectively. We observed a significant overlap between GBM (AD) and ENCODE datasets (hypergeometric test, P value = 0.02), suggesting that the common motifs we identified in ENCODE dataset are recurrent among diverse datasets.

3.6 Functional analysis of the identified motifs

To further ascertain the functional relevance of the 44 detected TF-miRNA co-regulatory motifs, we next examined whether they are more biologically meaningful than the remaining 212 non-motifs. To this end, we first merged the corresponding subgraph instances within each motif or non-motif based on the same miRNA and TF regulators to get a list of common gene sets and then counted the number of significant GO terms (biological process) or canonical pathways obtained from MSigDB (Subramanian *et al.*, 2005) for each gene set with False Discovery Rate (FDR) < 0.1 (hypergeometric P value adjusted by Benjamini–Hochberg method). Encouragingly, the 44 predicted motifs are enriched for significantly higher number of meaningful biological pathways or processes compared with non-motif instances (Fig. 6A).

To demonstrate the potential gain of mechanistic insights from our motif analysis, we further studied one of the 44 motifs, which is a composite bi-fan motif involving two genes that are co-regulated by one miRNA and one TF (Fig. 6B). In total, there are 258 166 distinct gene pairs with 5360 distinct genes involved in this motif. We then

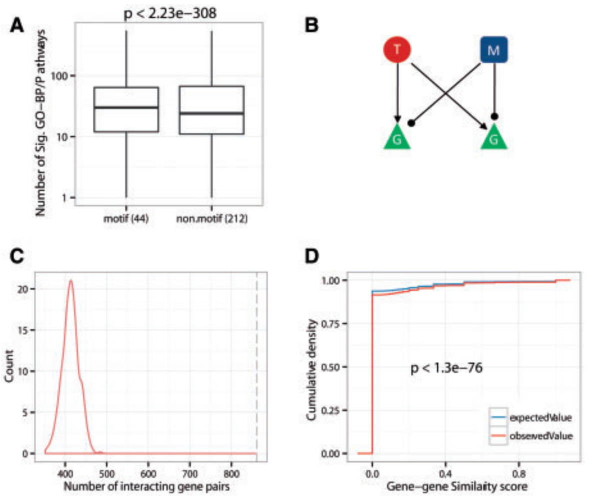


Fig. 6. (**A**) The number of significant GO (biological process) terms or canonical pathways from MSigDB enriched in motifs and non-motifs were compared. (**B**) The identified composite bi-fan motif. (**C**) Distribution of the randomly select gene pairs with interactions. The vertical dashed line is the observed value (861). (**D**) The cumulative distribution function of similarity score of randomly selected gene pairs (blue) and observed gene pairs (red). *P* value was calculated by Wilcoxon rank sum test

tested whether gene pairs involved in this motif structure tend to physically interact with each other. We first obtained 83 826 experimentally confirmed human protein-protein interactions (PPIs) from BioGrid database (Stark *et al.*, 2006). Among the 258 166 distinct gene pairs involved in this specific co-regulatory motif, 861 pairs share BioGrid-derived PPI. To ascertain the statistical significance of the observed overlap, we randomly selected the same number of gene pairs from the same gene set and counted how many gene pairs have physical interactions by chance. Indeed, the observed PPI count (861) is located at the far right tail of the background distribution, suggesting that the gene pairs involved in the motif have a strong tendency to interact with each other (Fig. 6C). We next tested whether gene pairs within this bi-fan motif tend to have more functional similarities measured by the following equation:

$$(A_{go} \cap B_{go})/\min\left(|A_{go}|, |B_{go}|\right) \qquad (1)$$

where A_{go} and B_{go} are the GO annotations shared by genes A and B. We first filtered out the genes that do not have mapped GO terms and their involved gene pairs. This resulted in 3153 genes and 100 373 gene pairs. We then repeated the same permutation test by randomly selecting the same number of gene pairs and calculating the functional similarities for them. Indeed, the result suggests that the gene pairs involved in the motif construct have higher level of functional similarities than the random selected pairs (Fig. 6D). Together, our motif analysis revealed a parsimonious and biologically relevant 'evolutionary principle' in regulatory network where genes that are co-regulated by TFs and miRNAs have strong tendency to physically interact or are more functionally related comparing with gene pairs at random.

4 Discussion

It has been appreciated that gene regulatory network in metazoans are complex and involve multiple layers and distinct type of regulators, e.g. regulations at the transcriptional or post-transcriptional level and regulations by TF or by miRNA (or RNA-binding protein). An important goal in systems biology is to accurately recapitulate co-regulatory networks to investigate how different types of regulators interact with each other. Accurate network motif discovery plays a pivotal role in studies of complex networks as they provide a systematic way to reveal the intrinsic co-regulatory patterns among multiple types of regulators. However, detecting network motifs in arbitrarily complex networks remains computationally intractable and challenging. In this study, we developed a novel de novo motif finding algorithm called CoMoFinder to discover co-regulatory network motifs each involving at least one TF, one miRNA and one target gene. Here, we define network motif as network patterns consisting of miRNAs, TFs and target genes that occur significantly in higher frequencies than expected. The novelty of our algorithm is 3-fold: (i) we first proposed an efficient enumeration process specifically for finding co-regulatory subgraphs of moderate size; (ii) we then introduced a refined canonical labeling method together with a new graph isomorphism testing strategy to effectively classify these composite subgraphs into different isomorphic groups and (iii) we also proposed a new randomization strategy by breaking the co-regulatory network into different layers and seeking for the maximum difference between the original network and the randomized network in each layer. Notably, FANMOD (Wernicke, 2006) and WaRSwap (Megraw *et al.*, 2013) use canonical labeling generated by a practical graph isomorphism algorithm NAUTY, which takes the diagonal of the adjacency matrix to denote the node types. Different from these previous approaches, our refined canonical labeling method can easily differentiate composite subgraphs with or without self-regulations/interactions in co-regulatory networks. Furthermore, our randomization algorithm is applicable to co-regulatory networks containing multiple types of edges including undirected edges such as PPI and directed repression/activation edges. In contrast, WaRSwap can only be applied on directed co-regulatory networks.

To demonstrate the utility of CoMoFinder, we compared it with FANMOD and WaRSwap on published co-regulatory networks of GBM (Sun *et al.*, 2012) and AD (Jiang *et al.*, 2013). Because the range of co-regulatory subgraphs and their isomorphic classes are the same, the performance of each method solely depends on their constructed background network distributions. By examining the distributions of composite subgraphs generated by each randomization technique, we found that CoMoFinder and FANMOD could maintain a similar number of composite subgraphs as in the observed network, whereas WaRSwap-derived random networks lack a substantial number of composite subgraphs. The results thus indicate that WaRSwap tends to generate more statistically significant network patterns as network motifs improperly comparing with CoMoFinder and FANMOD. Based on the overlaps of motifs found by each method, CoMoFinder achieved a good balance comparing with FANMOD and WaRSwap, which appear to be too stringent and too lenient, respectively. Moreover, we conducted rigorous power analysis using an empirical gold standard motif set. Encouragingly, CoMoFinder achieved consistently the greatest area under the curves. We also tested the robustness of each method upon random removals of edges in the given networks and showed that CoMoFinder is relatively robust up to 30% of edge removal. Together, the results corroborate the reliability and stability of CoMoFinder. We then applied CoMoFinder to a human co-regulatory network derived from ENCODE project, which is much larger than the GBM and AD networks. We detected 44 co-regulatory network motifs of size 4 and found a significant overlap between motifs from ENCODE and GBM/AD dataset. The functional analysis revealed that the 44 motifs are significantly enriched for higher number of meaningful biological pathways or processes compared with non-motif instances. Among the 44 motifs, 35 of them involve two TFs, one miRNA and one target gene; eight of them involve one TF, two miRNAs and one target gene; only one of them involve two genes with one TF and one miRNA. We further investigated the latter motif and found that the gene pairs involved in this motif tend to physically interact with each other and have more functional similarities. In terms of computational time efficiency, CoMoFinder leverages multicores of modern CPU, which can provide substantial speed-up in running time over single-threaded algorithms such as FANMOD and WaRSwap. In particular, CoMoFinder could gain an almost linear speed-up as increase of the CPU cores.

In summary, CoMoFinder is a reliable and efficient tool that can detect co-regulatory network motifs in large-scale datasets. Nevertheless, as the main focus of our method is to find network motifs consisting of miRNAs, TFs and genes, one of the limitations of CoMoFinder is that it is only applicable to co-regulatory networks involving these elements up to now. We also want to stress that the miRNAs examined in our work represent both miRNA precursors (mir) and the mature forms (miR). However, multiple distinct precursors may give rise to identical mature sequences that target the same set of genes, while regulated by very distinct set of TFs. Since precursors typically reside within gene introns and are often subject to co-regulation with their host gene, the TF → miRNA regulatory relations usually involve host genes and precursor (i.e. TF → gene → mir → miR). This also implies the existence of miR → gene → mir regulation, which would imply an indirect miRNA → miRNA regulation. Thus, it would be ideal to consider miR and mir as two separate node types. Although in this work we did not take the indirect regulations into consideration, CoMoFinder is able to deal with this type of regulations, which can be explicitly specified by users in the input co-regulatory network. The general framework of CoMoFinder will be further extended to other types of composite networks with arbitrary numbers of node types in the future.

Funding

This work was supported by Natural Sciences and Engineering Research Council (NSERC) Canada Graduate Scholarship [to Y.L.], the National Natural Science Foundation of China [61240046 to J.L.], Hunan Provincial Natural Science Foundation of China [13JJ2017 to J.L.], Ontario Research Fund – Global Leader (Round 2) [to Z.Z.] and an NSERC grant [to Z.Z.].

Conflict of Interest: none declared.

References

Alon,U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, **8**, 450–461.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Beber,M.E. *et al.* (2012) Artefacts in statistical analyses of network motifs: general framework and application to metabolic networks. *J. R. Soc. Interface*, **9**, 3426–3435.

Cheng,C. *et al.* (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol*, **7**, e1002190.

Darga,P.T. *et al.* (2004) Exploiting Structure in Symmetry Detection for CNF, *DES AUT CON*, 530–534.

Ebert,M.S. and Sharp,P.A. (2012) Roles for microRNAs in conferring robustness to biological processes. *Cell*, **149**, 515–524.

Gerstein,M.B. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.

Grochow,J.A. and Kellis,M. (2007) Network motif discovery using subgraph enumeration and symmetry-breaking. *RECOMB*, 92–106.

Herranz,H. and Cohen,S.M. (2010) MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. *Genes Dev.*, **24**, 1339–1344.

Hobert,O. (2008) Gene regulation by transcription factors and microRNAs. *Science*, **319**, 1785–1786.

Ideker,T. *et al.* (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, **144**, 860–863.

Jiang,W. *et al.* (2013) Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease, *Bioinformatics*, **29**, 2596–2602.

Junttila,T. and Kaski,P. (2007) Engineering an efficient canonical labeling tool for large and sparse graphs. *ALENEX*, 135–149.

Kashani,Z.R. *et al.* (2009) Kavosh: a new algorithm for finding network motifs. *BMC Bioinformatics*, **10**, 318.

Kashtan,N. *et al.* (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, **20**, 1746–1758.

Khakabimamaghani,S. *et al.* (2013) QuateXelero: an accelerated exact network motif detection algorithm. *PloS One*, **8**, e68073.

Li,X. *et al.* (2012) NetMODE: network motif detection without nauty. *PloS One*, **7**, e50093.

Martinez,N.J. and Walhout,A.J.M. (2009) The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays*, **31**, 435–445.

Martinez,N.J. *et al.* (2008) Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Res.*, **18**, 2005–2015.

McKay,B.D. and Piperno,A. (2014) Practical graph isomorphism. *II. J. Symb. Comput.*, **60**, 94–112.

Megraw,M. *et al.* (2013) Sustained-input switches for transcription factors and microRNAs are central building blocks of eukaryotic gene circuits. *Genome Biol.*, **14**, R85.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Morgan,R.J. and Soltesz,I. (2008) Nonrandom connectivity of the epileptic dentate gyrus predicts a major role for neuronal hubs in seizures. *Proc. Natl Acad. Sci. USA*, **105**, 6179–6184.

Panni,S. and Rombo,S.E. (2013) Searching for repetitions in biological networks: methods, resources and tools. *Brief. Bioinform*, **16**, 118–136.

Riba,A. *et al.* (2014) A combination of transcriptional and microRNA regulation improves the stability of the relative concentrations of target genes. *PLoS Comput. Biol.*, **10**, e1003490.

Roy,S. *et al.* (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797.

Stark, C. *et al.* (2006) BioGRID: a general repository for interaction datasets, *Nucleic Acids Res.*, **34**, D535–D539.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Sun,J. *et al.* (2012) Uncovering MicroRNA and Transcription Factor Mediated Regulatory Networks in Glioblastoma, *PLoS Comput. Biol.*, **8**, e1002488.

Tsang,J. *et al.* (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol. Cell*, **26**, 753–767.

Wernicke,S. (2006) Efficient detection of network motifs. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 347–359.