

Sepideh Sadegh¹ / Maryam Nazarieh^{1,2} / Christian Spaniol³ / Volkhard Helms¹

Randomization Strategies Affect Motif Significance Analysis in TF-miRNA-Gene Regulatory Networks

¹ Center for Bioinformatics, Saarland University, Saarbruecken, Germany, E-mail: sepideh.saadegh@gmail.com² Graduate School of Computer Science, Saarland University, Saarbruecken, Germany³ Department of Psychiatry and Psychotherapy, Saarland University Hospital, Homburg, Germany**Abstract:**

Gene-regulatory networks are an abstract way of capturing the regulatory connectivity between transcription factors, microRNAs, and target genes in biological cells. Here, we address the problem of identifying enriched co-regulatory three-node motifs that are found significantly more often in real network than in randomized networks. First, we compare two randomization strategies, that either only conserve the degree distribution of the nodes' in- and out-links, or that also conserve the degree distributions of different regulatory edge types. Then, we address the issue how convergence of randomization can be measured. We show that after at most $10 \times |E|$ edge swappings, converged motif counts are obtained and the memory of initial edge identities is lost.

Keywords: Feed-forward loop, Network randomization, Target gene, Transcription factor, Transcriptional regulation

DOI: 10.1515/jib-2017-0017

Received: March 21, 2017; **Revised:** April 27, 2017; **Accepted:** May 2, 2017


1 Introduction

Gene-regulatory networks (GRNs) are typically formulated as directed mathematical graphs whereby nodes stand for target genes, transcription factors (TFs), and microRNAs and edges stand for activating or repressing regulatory interactions. By edges we refer to directed edges here. TFs either activate or repress the transcription of target genes. MicroRNAs typically induce the degradation of messenger RNAs of their target genes. Hence, modern GRNs address the regulation of messenger RNA levels at transcriptional and post-transcriptional levels [1], [2]. Our group recently introduced a webserver termed TFmiR [1] that enables users to construct and analyze disease-specific TF and miRNA co-regulatory networks. Please see the methods section for more details on TFmiR.

Shen-Orr and Alon were the first to identify regulatory motifs in a GRN of *E.coli* that only consisted of TFs and target genes [3]. They discovered that feed-forward loops (FFLs) involving two TFs whereby TF1 regulates TF2 and both TFs jointly regulate a target gene are statistically significantly enriched in real GRNs with respect to randomized GRNs. Besides, they also discovered that single-input modules and densely overlapping regions are enriched too, but we will focus on FFL-type motifs here. Recently, several authors have expanded the concept of FFL-motifs to GRNs with TFs, miRNAs, and target genes [1], [2], [4]. In this context, proper randomization of GRNs becomes even more important for determining which FFL motifs are enriched in the real GRN. In our original TFmiR paper [1], we did not distinguish between the three possible types of regulatory links, TF → target gene, TF → miRNA, and miRNA → target gene, during randomization. However, Ohler and co-workers recently pointed out that an edge-type preserving randomization strategy may be beneficial whereby switching of edge end-points only takes place between two edges that both belong to either one of the three groups of regulatory links [4].

Another important technical question is how to quantify proper randomization. In our original TFmiR paper, we randomized $2 \times |E|$ times, where $|E|$ is the number of links in the GRN. It was argued that $100 \times |E|$ switches of edge end points ensure proper randomization [5]. Based on two GRNs with different link densities, we present here a thorough analysis what motifs are statistically enriched in these GRNs under the edge-type conserving and non-conserving randomization strategies and how proper randomization can be quantified. For comparison, we also used the established motif-discovery tool FANMOD [6].

Sepideh Sadegh is the corresponding author.

 ©2017 Sepideh Sadegh et al., published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

2 Related Works

There exist many motif finding tools including the well-known tools mfinder [7] and FANMOD [6]. mfinder detects network motifs either by full enumeration of subgraphs, or by sampling of subgraphs for estimation of subgraph concentrations. The latter method is faster but has a bias for sampling certain subgraphs more than others [8]. mfinder provides several methods to generate random networks including the switching method, the stub method, and the “go with the winners” algorithm [5]. FANMOD uses an algorithm called RAND-ESU [8] that enables quick and accurate estimation of the total number of size-k subgraphs in a given network. A new randomization algorithm named WaRSwap [4] provides a practical network motif discovery method for large multi-layer networks such as co-regulatory networks. However, this technique must be used together with a motif discovery tool such as FANMOD, which limits its applicability. WaRSwap generates randomized networks by preserving the in-degree distribution of target nodes with respect to each source-target type rather than the exact in-degrees. This randomization method seems to be more compatible with multi-layer networks than the universal method where only the in- and out-degree of nodes are conserved.

3 Materials and Methods

3.1 Types of 3-Node Motifs in miRNA-TF Synergistic Regulatory Networks

miRNA and TF co-regulatory networks contain four types of regulations, $TF \rightarrow Gene$, $TF \rightarrow miRNA$, $miRNA \rightarrow Gene$ and $miRNA \rightarrow TF$, that can be combined in ten different ways as 3-node motifs, see Figure 1. Eight of these are synergistic motifs consisting of two different types of regulators (miRNA and TF), and their directly/indirectly synergistically regulated target gene (first two rows of Figure 1). The last two motifs, where the target gene is not cooperatively regulated, are not studied here.

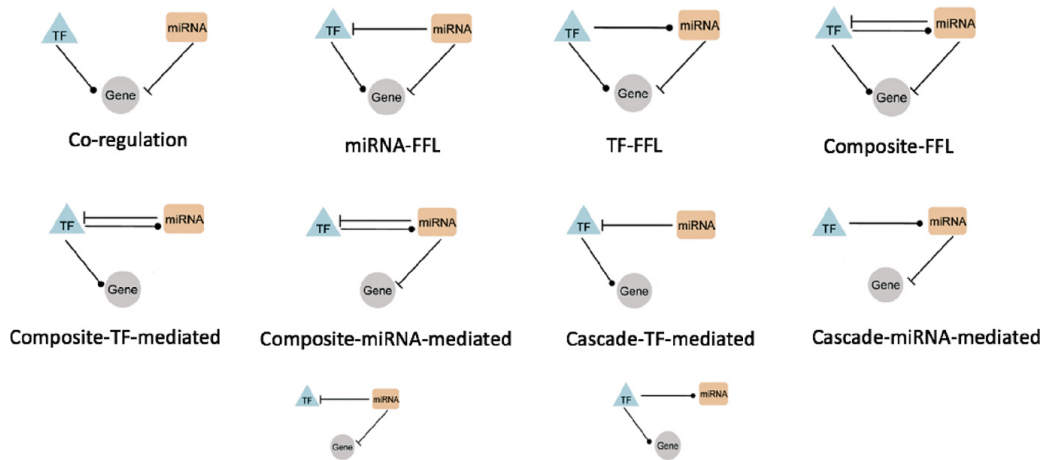


Figure 1: 3-node motifs in miRNA & TF synergistic regulatory networks. In FFLs the gene is regulated via two paths: (1) a direct regulation by a main regulator (TF/miRNA) and (2) an indirect regulation through an intermediate regulator (miRNA/TF) which is itself regulated by the main regulator. Composite-TF/miRNA-mediated: mutual regulation of TF and miRNA besides regulation of the target gene by only one of them. Cascade-TF/miRNA-mediated: are non-loop forms, including an indirect effect of the main regulator (TF/miRNA) on the target gene only via another type of regulator (miRNA/TF). The 3rd row shows two non-cooperative motifs where the target gene is not cooperatively regulated.

3.2 Data Sets

We used miRNA and TF co-regulatory networks for two different complex diseases as input to our motif finding tool. The first network is associated with breast cancer (BC) [1] and the second network with glioblastoma multiforme brain tumor (GBM) [9]. Table 1 lists topological properties of the two networks. The GBM network is about four times denser than the BC networks.

Table 1: Density of Networks.

	$ E $	$ V $	density
BC-complete	378	258	0.0057
BC-disease	297	206	0.0070
GBM	4248	408	0.0256

In a study on breast cancer using gene and miRNA expression data from The Cancer Genome Atlas (TCGA) portal, Hamed et al. [10] identified 1262 genes and 121 miRNAs that are deregulated in cancer tissue with respect to matched normal tissue. With the TFmiR web server [1] we identified regulatory interactions for the provided lists of up- and down-regulated genes and miRNAs using data from established and curated regulatory databases of both predicted and experimentally validated interactions. The resulting network is termed BC-complete in Table 1. Then we used TFmiR to intersect this global network with genes associated with “breast neoplasms” based on the human miRNA disease database (HMDD) [11] and DisGeNET, a database for gene-disease association [12]. This gave the breast cancer-specific subnetwork that we termed BC-disease.

A co-regulatory network for GBM with 415 genes and 124 mature GBM-related miRNAs was retrieved from Sun et al. [9], who used a similar approach for constructing GRNs to the approach used in TFmiR. 428 human TFs were retrieved from the TRANSFAC database [13]. The regulatory interactions between a TF, a miRNA, and a target gene were predicted using computational approaches.

The main difference between the two networks considered here is in the last step. In the GBM network, the authors included only miRNA-TF co-occurring pairs that are significant based on the hypergeometric test. In contrast, TFmiR does not check for significance here. Another difference is that in building the GBM-specific co-regulatory network only predicted interactions were utilized, while in the BC-complete/disease co-regulatory networks both predicted and experimentally validated interactions were taken into account.

3.3 Motif Discovery Process

The sequential steps for the motif discovery are as follows: A subgraph census is conducted for the types of desired motifs on the original network. An ensemble of N similar random networks is generated and subgraph enumeration is applied to each of these networks. Finally, after calculating the frequency of each type of subgraph in all networks (original and randomized), its significance metrics are calculated, with the over-represented subgraphs being reported as motifs. For the purpose of comparing two randomization strategies, we implemented the entire process of motif finding as an in-house Cytoscape App [14], which is an OSGi Bundle style App. This functionality will be made publicly available in the next release of TFmiR.

3.3.1 Enumeration of Desired Subgraphs

Typical algorithms for enumeration of subgraphs work on a connectivity matrix C , whose elements (C_{ij}) are equal to 1 if regulator i regulates target j and 0 otherwise. Then, they scan all n by n submatrices of C , that represent topologies of each desired type of size n motif. We modified this typical subgraph enumeration algorithm by using the data models in Cytoscape (namely CyNetwork and CyTable).

3.3.2 Generating Random Networks

Randomization of networks must be conducted such that sampling is done as uniformly as possible from the collection of all obtainable random networks. Megraw et al. [4] suggested that low-variance distributions of motif counts in randomized networks (<1) are a sign of inadequate randomization, and that they can happen due to edge switching in large multi-layer networks. To evaluate the adequacy of sampling and uniformity of random networks generation, variances of subgraph count of all types of possible motifs in the randomized networks should be considered (see Section 4.3.1).

The key aspect in assessing the statistical over-representation of motifs is to generate the random networks in a way so that their characteristics are as similar as possible to the original network. The method using swapping of end-points ensures that each node in the randomized networks has the same number of incoming and outgoing edges (in- and out-degree) as the corresponding node in the real network. The universal method used for this purpose is the so-called “switching method”, employed for the first time in the field of motif detection by Shen-Orr et al. [3]. By construction, it strictly conserves the degree distribution of the graph and even of each node. The algorithm generates a Markov chain of states by randomly selecting a pair of edges ($A \rightarrow B, C \rightarrow D$)

and swapping their endpoints to create the new edges ($A \rightarrow D$, $C \rightarrow B$). Creation of self-edges and multiple edges are not allowed and considered as failed attempts of switching. This process is repeated $Q \times |E|$ times, where $|E|$ is the number of edges in the graph and Q is chosen large enough so that the Markov chain shows good mixing. Milo et al. [5] found that for many networks, values of around $Q = 100$ appear to be more than adequate. In our approach failed attempts are not counted, i.e. we repeat swapping as many times as needed to reach $Q \times |E|$ successful attempts. This algorithm returns a shuffled version of the original network as a randomized network. We suggest some measures to assess proper mixing in Section 3.4.

Different Strategies

We modified the switching method to consider additional features of miRNA-TF synergistic regulatory networks (different node and edge types). To deal with networks with multiple types of connections, we use the terminology introduced by Yeager-Lotem et al. [15] where the extended degree of a node stands for the number of edges per type that point to/from a node. Two nodes have the same extended degree if they have the same number of incoming and outgoing edges for each edge type.

Based on this definition, one can develop a new switching strategy, which allows only swapping endpoints of edges with the same regulatory relationship among miRNAs, TFs, and target genes. Hence, we distinguish a “conserving method” that conserves the extended degree of nodes, i.e. edges are switched only between edges of the same type, and a “non-conserving method” that does not conserve the extended degree of nodes, i.e. switching is done without considering the edge type, consequently the frequency of each edge type is not conserved. The latter method is equivalent to the original switching method. Note that the non-conserving method can also create new edge types, which did not exist in the original network, unless this is prevented (such as $TF \rightarrow TF$, or $miRNA \rightarrow miRNA$ edges).

An efficient algorithm for the conserving method can be implemented by grouping network edges of different edge types into different lists and then randomly selecting the second edge from the edge list of the first selected edge type. This helps to improve the efficiency of the randomization algorithm in terms of run-time.

3.3.3 Comparison of Real and Randomized Networks by Significance Metrics

The goal of network motif discovery is to determine which subgraph types occur in the original network at significantly higher frequencies than in random networks. For this, the occurrence of a particular subgraph in the network of interest is compared to the distribution of counts for the same subgraph over a set of randomized networks using p -value and z -score. The p -value represents the probability of a motif to appear an equal or greater number of times in a random network than in the original network [16]. This probability should be smaller than a determined probability threshold to reject the null hypothesis. This can be empirically determined using a large number of randomized networks:

$$p\text{-value} = \frac{N_{rh}}{N_t}$$

where N_{rh} is the number of random networks in which a certain motif type is acquired more than or equal to its number in the real network and N_t is the total number of randomized networks. It has been suggested [1] that $N_t = 100$ is sufficiently large. Alternatively, let f_{real} be the frequency in the real network and f_{rand} be the frequency in a random network. We can then define the z -score as follows (with σ being the standard deviation):

$$z\text{-score} = \frac{(f_{\text{real}} - \bar{f}_{\text{rand}})}{\sigma(f_{\text{rand}})}, \quad \sigma(f_{\text{rand}}) = \sqrt{\frac{\sum_i (f_{\text{rand}_i} - \bar{f}_{\text{rand}})^2}{N_t}}$$

Subgraphs with z -score ≥ 2 and p -value < 0.05 are considered significant motifs as was previously done [3], [4], [6].

3.4 Measures for Proper Mixing of Randomized Networks

One general drawback of randomizing networks by the switching method is that there is no measure of how long one needs to iterate over the “select two edges and swap their endpoints” routine to attain well randomized networks. Here we propose two measures to characterize whether randomized networks are properly mixed.

First, we measure the similarity of networks before and after randomization. Ideally, edges should be switched until there are no common edges between the original network and each randomized network. In

other words one should search for the maximum difference between the given network and each randomized network to avoid the situation termed “under-shuffling” [17]. Under-shuffling means that only a small fraction of the switchable edges were swapped. We defined a *similarity metric* to measure how similar is the ensemble of randomized networks to the original network in terms of common edges:

$$\text{Similarity} = \frac{\langle \text{Sim} \rangle}{|E|}$$

Here, *Sim* is the number of common edges between the original and a particular randomized network, $\langle \text{Sim} \rangle$ is its average in all randomized networks, and $|E|$ is the total number of edges in the original network. Lower *Similarity* values indicate better randomization. The lowest possible value of zero happens in case of no common edges. This definition considers the size of the network as well as the number of randomized networks. This enables comparison of the similarity metrics of randomization approaches applied to different given networks. A value close to zero indicates that under-shuffling is avoided.

Another measure is the convergence of subgraph counts during randomization. For $0.01 \times |E|$ to $100 \times |E|$ randomization iterations, we recorded how often the investigated subgraph types occurred in the random networks and checked whether this number converged to a specific value or whether it did not follow any pattern and changed erratically.

4 Results

4.1 Synergistic 3-node Motifs

Table 2 shows which co-regulatory 3-node motifs were significantly enriched in the real GRN vs. randomized GRNs when either the edge-type conserving randomization strategy was applied or the non-conserving one. $100 \times |E|$ iterations were used for this part of our study. In the BC-complete network, no significant motif is found by the conserving method. In contrast, the composite-miRNA-mediated and cascade-miRNA-mediated motifs are reported as significant by the non-conserving method. In the BC-disease network, only the co-regulation type is identified as significant by the conserving method, whereas the non-conserving method gives the same significant motifs as for the BC-complete network. In the GBM network, TF-FFL and miRNA-FFL are reported as significant by the conserving method; whereas by the non-conserving method all types of subgraphs except co-regulation are identified as significant. In all three networks, subgraphs of types composite-miRNA-mediated and cascade-miRNA-mediated are identified as statistically significant by the non-conserving method. All subgraphs meeting the *p*-value criterion (Table 2) also met the z-score criterion. Note that *p*-value is a probability and can be slightly different in each run of the algorithm due to the generation of different randomized networks.

Table 2: *p*-values for different 3-node subgraphs in the considered networks when either the non-conserving or the conserving randomization strategy is used.

Subgraph type	BC-complete		BC-disease		GBM	
	Non-cons.	Cons.	Non-cons.	Cons.	Non-cons.	Cons.
Co-regulation	0.77	0.33	0.96	0.04	1.00	1.00
TF-FFL	1.00	0.77	0.98	0.61	0.00	0.00
miRNA-FFL	0.77	0.86	0.69	1.00	0.00	0.00
Composite-FFL	0.29	0.12	0.45	0.50	0.00	0.68
Composite-TF-Med.	0.62	0.26	0.66	0.42	0.00	0.55
Composite-miRNA-Med.	0.00	0.50	0.01	0.53	0.00	0.62
Cascade-TF-Med.	0.47	0.69	0.16	0.26	0.00	0.84
Cascade-miRNA-Med.	0.00	0.54	0.01	0.49	0.00	1.00

Significant motifs are marked in bold.

The non-conserving method leads to detecting more subgraph types as significant compared with the conserving method. In randomization by the conserving method, swapping happens between all edges of the same type, hence the chance of having the same types of subgraphs in randomized networks compared to the original network is not decreased that much. This could result in higher p -values and consequently fewer subgraphs will show significant differences.

In the GBM network, we found 3-node FFLs to be significant while in BC networks they are not. One reason for this could be the higher density of the GBM network than the BC networks.

4.2 Motif Finding with FANMOD

FANMOD also employs the “switching method” for randomization of network. The randomization step can optionally keep the extended degree of a node constant by exchanging edges only with edges of the same type. This is equal to randomization by the conserving method in our approach. The same number of swappings per edge ($Q = 100$) and the same criteria for p -value and z -score were chosen for both tools.

FANMOD gave similar motif finding results for all three miRNA-TF co-regulatory networks (Figure 2) to those of our tool. The few dissimilarities can be due to slight differences of the randomization algorithms. In the routine of randomly selecting edges for swapping, we only count successful attempts until a pre-defined number of iterations is reached whereas FANMOD tries a limited pre-defined number of times to find an appropriate candidate for swapping irrespective of whether this is successful or not. By inspecting the output file of FANMOD we found that $\sim 80\%$ of the attempts were successful when randomization was done with the conserving method and $\sim 50\%$ with the non-conserving method.

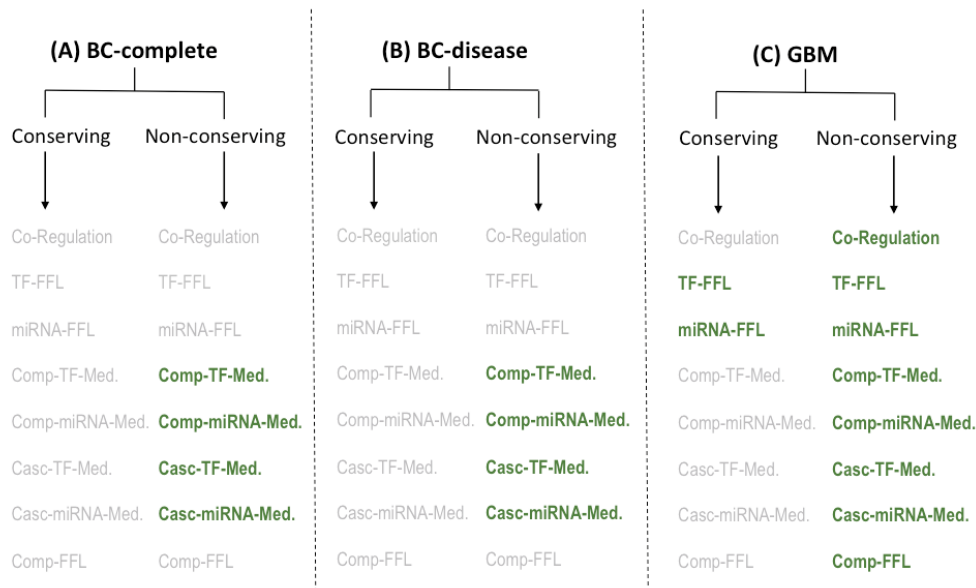


Figure 2: Significant 3-node motifs (highlighted in green) detected by the FANMOD tool with two different randomization strategies.

(A) BC-complete. (B) BC-disease. (C) GBM.

4.3 Validation of Randomization

4.3.1 Uniform Sampling of Randomized Networks

Megraw et al. [4] observed many failed switches during the execution of FANMOD randomization, which was also the case here. As mentioned, our approach counts only the number of successful attempts until a pre-defined number of iterations is reached. By close inspection of the resulting background histogram of significant motifs, we observed in the miRNA-TF synergistic regulatory networks of BC and GBM a high variance of count distributions of subgraphs in randomized networks for all significant 3-node motifs (Table 3 and Table 4). This indicates an adequate randomization of networks in our approach. For the GBM network, much higher

variances were obtained than for the BC networks. It is suggestive to attribute this to the higher density of the network.

Table 3: Variance of count distributions of subgraphs in randomized networks for the BC-complete/-disease networks.

Subgraph type	Non-conserving method		Conserving method	
	BC-disease	Variance BC-complete	BC-disease	Variance BC-complete
Co-regulation	14.9	32.0	6.7	10.3
TF-FFL	5.5	7.5	2.4	2.6
miRNA-FFL	4.2	4.9	7.0	8.4
Composite-FFL	1.7	1.5	2.0	1.7
Composite-TF-Med.	289.8	339.1	177.6	130.0
Composite-miRNA-Med.	85.1	74.0	391.9	458.5
Cascade-TF-Med.	592.3	880.8	174.8	134.4
Cascade-miRNA-Med.	394.6	433.4	435.9	522.6

Significant motifs are marked in bold.

Table 4: Variance of count distributions of subgraphs in randomized networks for GBM network.

Subgraph type	Non-conserving method	Conserving method
	Variance	Variance
Co-regulation	18,335.7	4729.6
TF-FFL	4340.9	3298.5
miRNA-FFL	1526.1	2344.3
Composite-FFL	277.0	890.4
Composite-TF-Med.	10,282.2	38,169.3
Composite-miRNA-Med.	1690.0	6153.5
Cascade-TF-Med.	53,127.1	37,486.1
Cascade-miRNA-Med.	35,798.9	7742.2

Significant motifs are marked in bold.

4.3.2 Measures for Proper Mixing of Randomized Networks

Similarity Metric

Two sets of 100 randomized networks were generated from the BC-complete network using in one case the edge-type conserving strategy and in the other case the non-conserving randomization strategy. Between $0.01 \times |E|$ and $100 \times |E|$ iterations of edge swapping ($Q \times |E|$) were carried out. Figure 3 shows the similarity between original and randomized GRNs for varying Q . The number of iterations required to reach values close to zero depends on the randomization strategy. For the non-conserving method, the similarity metric reaches zero at fewer iterations ($Q=7$ for BC-complete and $Q=8$ for GBM) than the conserving method ($Q=14$ for BC-complete and $Q=15$ for GBM). Results for the GBM network are very similar to those for the BC-complete network, only slightly more iterations are needed to reach zero. Both methods of randomization for both networks reach similarities below 0.01 after $Q=3$ iterations. This means that after $3 \times |E|$ iterations, less than 1% of the edges in the ensemble of randomized networks are in common with the original network. This low percentage of similarity seems to be a good threshold for choosing a proper Q for our randomization method.

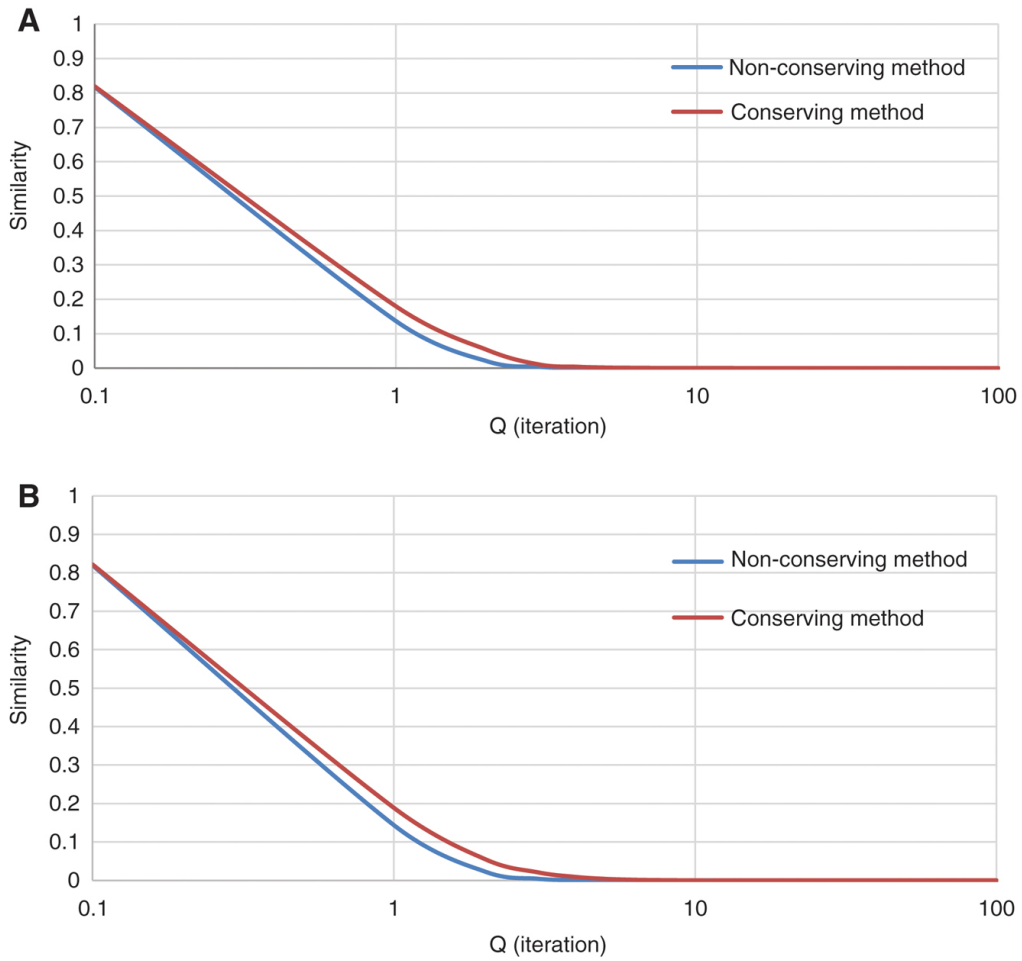


Figure 3: Similarity metric vs. number of iterations for (A) the BC-complete and (B) the GBM networks.

Convergence of Subgraph Counts

Figure 4 shows how often subgraph types occurred in the set of randomized BC-complete networks after randomization when Q was varied between 0.01 and 100. With the non-conserving method (Figure 4), the total subgraph count converged to a fixed value after Q = 10 iterations and did not change erratically thereafter. With the conserving method (Figure 4), the total number of subgraphs found in the randomized networks was quite stable over the whole range of $0.1 < Q < 100$.

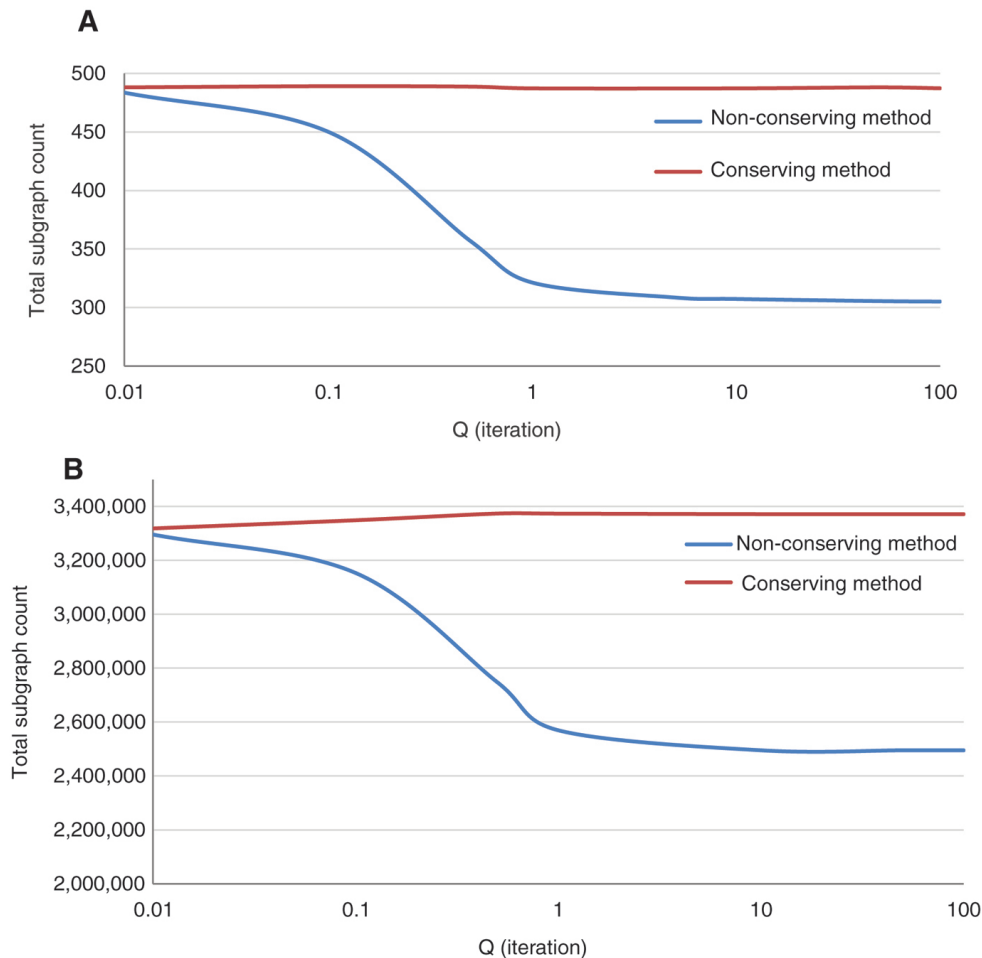


Figure 4: Total number of subgraphs vs. number of iterations for (A) the BC-complete and (B) the GBM networks.

Our empirical findings for both BC and GBM networks suggest that $Q = 1$ is adequate to obtain properly mixed randomized networks by the conserving method; whereas for the non-conserving method $Q = 10$ appears suitable for ensuring good mixing of the randomized networks. Evaluation of network similarity for the BC-complete network suggests $Q \sim 3$ as a good balance for both conserving and non-conserving methods of randomization.

4.4 Network Centrality of Gene and miRNA Sets

Here, we analyzed the overlap between the genes and miRNAs participating in the enriched 3-node motifs (here termed motif nodes) and the most central genes and miRNAs with respect to degree, betweenness, and closeness centralities (here termed central nodes). Using either edge-type conserving or non-conserving randomization gave 26 and 130 genes and miRNAs in enriched motifs, respectively. These sets were compared to sets with the same number of most central genes and miRNAs. The centralities were measured using the *igraph* package [18], considering only out-degree of nodes in the directed network. The motif nodes identified by the conserved method had the highest overlap with the central nodes defined according to closeness and betweenness centrality, respectively (Figure 5A). In contrast, the motif nodes defined with the non-conserved method showed a similar overlap with the central nodes identified by all three centralities (Figure 5B). This latter observation can be explained by noting that only 57 genes and miRNAs have out-degree greater than or equal to 1 in the BC disease networks. The overlap of around 45% with the central nodes means that essentially all these 57 motif nodes are hub nodes in this network. A larger fraction of hub nodes (up to around 60%) exists in the smaller set of 26 motif nodes defined by the conserving method.

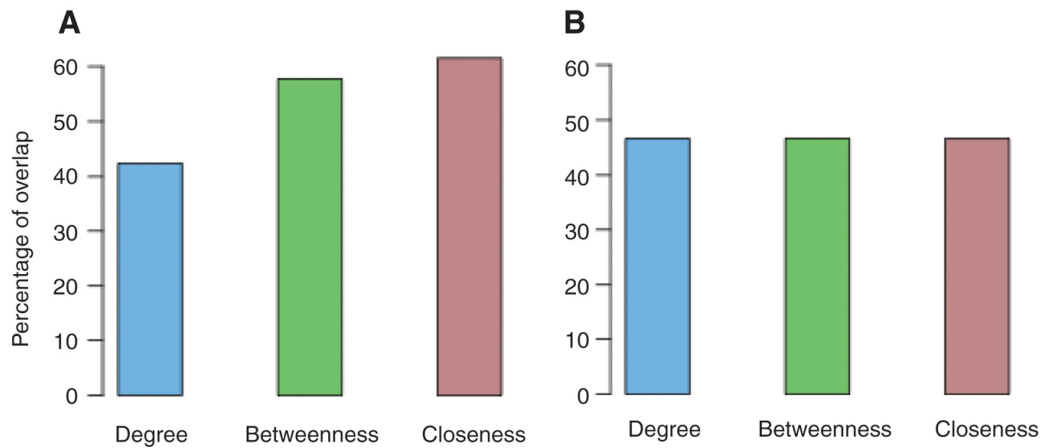


Figure 5: Overlap of most central nodes (according to three different centralities) with the set of genes and miRNAs in the statistically enriched motifs. (A) Conserved. (B) Non-conserved.

Next, we analyzed the overlap of the 26 motif nodes identified by the conserving method with a smallest connected set of key regulatory genes and miRNAs that dominates the network. For this, we solved the ILP formulation of the respective minimum connected dominating set (MCDS) in the largest strongly connected component of this network [19] and obtained an MCDS of seven genes and miRNAs. Among these, *TGFB1*, *TP53*, *ESR1*, and *hsa-mir-22* belong to the motif nodes.

4.5 Biological Relevance of the Detected Motifs

The biological relevance of the genes among the motif nodes obtained by the conserved and non-conserved randomization methods was evaluated based on the functional categories in GO Direct using the enrichment analysis via DAVID (version 6.8) [20]. *p*-values below the threshold 0.05 obtained by the hypergeometric test were adjusted for multiple testing using the Benjamini & Hochberg (BH) procedure [21]. Both methods returned almost the same number of significant GO terms, mostly involving transcription and apoptotic processes, although the non-conserving method considered 104 genes versus 14 genes considered by the conserving method.

5 Discussion

If the network of interest contains more than one node or edge type, different randomization strategies can be applied for motif discovery. In this study, different strategies led to quite different enriched 3-node motif types.

The reason why FFLs were statistically significantly enriched only in the GBM network could originate from the difference in constructing the GBM and BC networks, where only significant TF-miRNA co-occurring pairs were considered in the regulatory network of GBM. This means that the TF → gene ← miRNA triad is enriched *a priori* in this network. Our study suggests that the way of network construction and also the density of the network may affect the results of motif finding. For the considered BC-networks, only subgraphs of types other than FFLs were found to be significantly enriched. Our motif finding tool identified composite-miRNA-mediated and cascade-miRNA-mediated as statistically significant motifs (by the non-conserving method). Although the results are similar in BC-networks, the conserving method identified the co-regulation motif type to be significant in the filtered BC-disease network that was not found significant in the BC-complete network. We thus speculate that motif searches in filtered (i.e. more specific) networks may identify biologically more meaningful motifs.

We suggest variance of motif counts and similarity of original and randomized networks as suitable auxiliary measures to judge whether randomization generates properly mixed networks. Our study suggests that the density of networks does not affect the minimum required *Q* to obtain properly mixed randomized networks.

In conclusion, the non-conserving method leads to detecting more subgraph types as being statistically significant compared with the conserving method. For the 2.5 networks studied here, we noticed that (a) the conserving randomization method identified significant motifs containing a larger fraction of the most central nodes (Figure 5) than the non-conserving method, and (b) both methods gave the same number of significant Gene Ontology terms, although the conserving method considered much fewer genes for this than the non-

conserving method. Certainly, the same analysis should be extended to a representative number of comparable GRNs. So far, it seems that the conserving method gives biologically more meaningful results.

Acknowledgement

M.N. was supported by Deutsche Forschungsgemeinschaft via SFB 1027.

Conflict of Interest Statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

- [1] Hamed M, Spaniol C, Nazarieh M, Helms V. TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. *Nucleic Acids Res.* 2015;43:W283–288.
- [2] Zhang H, Kuang S, Xiong X, Gao T, Liu C, Guo A. Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief Bioinform.* 2015;16:45–58.
- [3] Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet.* 2002;31:64–68.
- [4] Megraw M, Mukherjee S, Ohler U. Sustained-input switches for transcription factors and microRNAs are central building blocks of eukaryotic gene circuits. *Genome Biol.* 2013;14:R85.
- [5] Milo R, Kashtan N, Itzkovitz S, Newman ME, Alon U. On the uniform generation of random graphs with prescribed degree sequences., 2004 Available from: <https://arxiv.org/abs/cond-mat/0312028v2>.
- [6] Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics.* 2006;22:1152–1153.
- [7] Kashtan N, Itzkovitz S, Milo R, Alon U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics.* 2004;20:1746–1758.
- [8] Wernicke S. A Faster algorithm for detecting network motifs. In: Casadio R, Myers G, editor(s). *Algorithms in bioinformatics*. Springer, 2005:165–177.
- [9] Sun J, Gong X, Purow B, Zhao Z. Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput Biol.* 2012;8:e10024888.
- [10] Hamed M, Spaniol C, Zapp A, Helms V. Integrative network-based approach identifies key genetic elements in breast invasive carcinoma. *BMC Genomics.* 2015;16:S2.
- [11] Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W. An analysis of human microRNA and disease associations. *PLoS ONE.* 2008;3(10):e34203.
- [12] Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics.* 2010;26:2924–2926.
- [13] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006;34:D108–110.
- [14] Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13:2498–2504.
- [15] Yeager-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, et al. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proc Natl Acad Sci USA.* 2004;101:5934–5939.
- [16] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science.* 2002;298:824–827.
- [17] Liang C, Li Y, Luo J, Zhang Z. A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and miRNA co-regulatory networks in human. *Bioinformatics.* 2015;31:2348–2355.
- [18] Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal.* 2006;Complex Systems:1695.
- [19] Nazarieh M, Wiese A, Will T, Hamed M, Helms V. Identification of key player genes in gene regulatory networks. *BMC Sys Biol.* 2016;10:88.
- [20] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2008;4:44–57.
- [21] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological).* 1995;57:289–300.