Computers in Biology and Medicine 107 (2019) 1–9

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Detecting list-colored graph motifs in biological networks using branch-and-bound strategy

Yiran Huang, Cheng Zhong*

*School of Computer and Electronics and Information, Guangxi Key Laboratory of Multimedia Communications Network Technology, Guangxi University, Nanning, 530004, China*

ARTICLE INFO

*Keywords:*
Functional motif
List-colored graph
Biological networks
Branch-and-bound strategy

ABSTRACT

In this work, we study the list-colored graph motif problem, which was introduced to detect functional motifs in biological networks. Given a multi-set *M* of colors as the query motif and a list-colored graph *G* where each vertex in *G* is associated with a set of colors, the aim of this problem is to find a sub-graph of *G* whose vertex set is colored exactly as motif *M*. To solve this problem, we present a heuristic method to efficiently and accurately detect list-colored graph motifs in biological networks using branch-and-bound strategy. We transform the detection of list-colored graph motif to the search of connected induced sub-graphs in list-colored graph, where the vertices in the sub-graph are assigned to distinctive colors of query motif. This transformation enables our method to accurately discover the occurrences of query motif without enumerating and verifying all sub-graphs. Furthermore, a new initial vertex selection strategy based on the colors of vertices is proposed to accurately determine the search scope of motifs. Experiments conducted on metabolic networks and protein-interaction networks demonstrate that our method can achieve better performance in accuracy and efficiency in comparison to other existing methods.

## 1. Introduction

An important and challenging task in systems biology is to understand the structures of all molecules and their interactions in a system level [1]. There has been a variety of biological interactions expressed in metabolic networks and protein-protein interaction (PPI) networks. In order to mine interesting features from these networks, some measurements have been proposed in recent years [2]. Milo et al. [3] proposed that some specific sub-graphs appeared with high frequency in the studied networks, and named such sub-graphs with overrepresented topological patterns as network motifs. Recently, the detection of network motif has become a focus of bioinformatics and systems biology. For example, it has been reported that network motifs have functional significance in metabolic networks [4,5] or protein–protein interaction networks [6]. Network motifs also have significance in some other fields, such as electronic circuit analysis [7] and software architecture design [8].

However, the original Definition of network motif is purely topological and disregards the nature of the components in a motif [9]. Such a topological definition of motif appears to be not adapted to metabolic networks since similar topologies can give rise to very different functions [9]. Subsequently, Lacroix et al. [9] introduced a novel definition for functional or topology-free network motifs and proposed the problem of finding functional motifs in biological networks. In the new definition and problem, a functional motif is a multi-set of desired functionalities and the biological networks are represented as graphs where each vertex denotes a biological molecule and each edge represents an interaction between two molecules [4,9]. Each vertex in such a graph is associated to a set of functionalities, each of the functionalities is represented as a color, and each vertex is marked by a set of distinctive colors. Such attributed graph is called list-colored graph which means that every vertex is assigned with a set of colors [4,10]. The problem of finding functional motifs in biological networks, which is also called list-colored graph motif problem, can be modeled as finding a sub-graph in list-colored graphs [11]. More specifically, for the list-colored graph motif problem, a biological network is denoted as a list-colored graph, and a functional motif is a multi-set of desired functionalities and is represented as a multi-set *M* of colors. And its aim is to find a connected induced sub-graph of list-colored graph such that the vertices in the sub-graph have exactly the colors specified by *M*.

Historically, the search of “purely” topological motifs could be modeled as a sub-graph isomorphism problem [9], which aims to detect the motif with specified topology in the network [12]. However, as Lacroix et al. [9] stated that the problem we study is different from sub-graph isomorphism because the topology in our discussion problem is not specified for the motif.

The list-colored graph motif problem was first introduced to find functional motifs in metabolic networks [9,11]. In this problem, the given

* Corresponding author.
*E-mail address:* chzhong@gxu.edu.cn (C. Zhong).

https://doi.org/10.1016/j.compbiomed.2019.01.025
Received 9 December 2018; Received in revised form 27 January 2019; Accepted 27 January 2019

metabolic network is a reaction graph where each vertex denotes a reaction and two reactions are connected if the products of one reaction are the inputs of the other reaction or vice versa. And the functionalities of reactions could be represented by reaction types and each reaction type is represented as a color [9,10]. A given functional motif is a multi-set of reaction types and is represented as a multi-set of colors. Since a reaction could be classified as an instance of more than one reaction type, the vertices in the reaction graph may have more than one color and the reaction graph is a list-colored graph [9,10]. The goal of detecting functional motifs in metabolic networks is to find a connected induced sub-graph in the reaction graph such that the vertices (reactions) in the sub-graph have exactly the reaction types (colors) specified by functional motif [9,10].

During the past few years, considerable efforts have been devoted to find list-colored graph motifs in biological networks [4,9–17]. Besides introducing the concept of list-colored graph motif for metabolic networks, Lacroix et al. [9] proposed a method called MUTOS to find functional motifs in metabolic networks by breadth first search. As the size of list-colored graph grows, the implementation of breadth first search requires tremendous computing resources and takes an unusually long time. Fellows et al. [13] proposed a FPT algorithm to find connected motifs in sparse vertex-colored graphs where each vertex is marked by a single color, and proved that in this special case, the problem is NP-complete. Rizzi et al. [12] discussed the variant of list-colored graph motif where some colors of the motif can be removed from the solution, and proposed inapproximability results for the optimization variants of list-colored graph motif. Moreover, Rizzi et al. [12] also studied a variant of list-colored graph motif problem where the connectivity constraint is substituted by the graph modularity, and proved that the problem remains NP-complete.

Some researchers have tried to solve the problem of finding functional motifs in list-colored graph by multi-linear detection. Guillemot and Sikora [14] converted the list-colored graph motif problem to the multi-linear detection problem. Koutis et al. [15] presented an extended version of multi-linear detection problem called constrained multi-linear detection problem, and proposed an efficient algorithm to solve this problem.

In addition, Bruckner et al. [16] investigated a particular case of list-colored graph motif problem where the motif can contain uncolored vertices, and proposed an algorithm called TORQUE to solve this problem by integer linear programming and color coding [18]. Based on linear pseudo-Boolean optimization, Blin et al. [17] proposed an algorithm called GraMoFoNe to solve the list-colored graph motif problem and some of its extensions. By using color-coding and dynamic programming, Betzler et al. [10,19] presented the algorithms to find the list-colored graph motifs, but their algorithms do not enumerate all occurrences of the list-colored graph motifs and fail to return the large-sized motifs in time limit.

In order to improve the efficiency of finding functional motifs, CeFunMO [4] used a greedy strategy to find functional motifs based on a measure of vertex centrality, namely color-centrality, and returned the connected sub-graphs with high color-centralities as functional motifs. Recently, Rudi et al. [11] designed an algorithm called RANGI to detect functional motifs by sub-graph enumeration. RANGI first enumerates all candidate motifs in the input graph and uses some heuristic techniques to prune the search space during the enumeration. And then RANGI checks the validity of candidate motifs by bipartite maximum matching algorithm. Compared with the method in Ref. [10], RANGI is much faster and is able to return all motifs of the maximum size in the list-colored graph, but it could be computationally hard to search all candidate motifs as the size of input graph grows.

In this paper, we propose a Motif Search for Biological Networks (MSBN) method to solve the list-colored graph motif problem in biological networks. We transform the detection of list-colored graph motif to the search of the connected induced sub-graph in list-colored graph, where the vertices in the sub-graph are assigned to distinctive colors of query motif. This transformation enables us to accurately discover the occurrences of query motif in the graph without enumerating and verifying all sub-graphs. Due to the fact that the efficiency and accuracy of functional motif detection are directly affected by the initial vertex selection in sub-graph enumeration, we propose a new strategy for choosing initial vertices based on the colors of vertices, which accurately determines the initial vertices in a certain scope. We evaluated the utility of our method on metabolic networks and protein-interaction networks. The experimental results show that compared with other existing methods, our method achieved favorable performance in terms of running time and accuracy.

We organize the rest of this paper as follows. Section 2 describes the list-colored graph motif problem in biological networks, and presents our heuristic method MSBN. Section 3 gives the experimental results. Section 4 concludes the paper.

## 2. Method

### 2.1. Preliminaries

In this work, a biological network is represented as a simple graph $G$ with vertex set $V(G)$ and edge set $E(G)$. Let the number of vertices of $G$ be $|G|$. For a metabolic network, the graph $G$ is a directed graph and each vertex in $G$ denotes a reaction, and two reactions (vertices) $A$ and $B$ are connected with a direct edge from $A$ to $B$ if the products of reaction $A$ are the inputs of reaction $B$ or vice versa. For a protein-interaction network, the graph $G$ is an undirected graph, and each vertex denotes a protein and each edge represents the interaction between two proteins. The graph $H$ is a sub-graph of $G$, if and only if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. A color set of vertex $v$ is represented by col($v$).

A list-colored graph $G_l$ for biological network is a directed or undirected graph where each vertex $v$ is associated with the color set col($v$) [9]. A functional motif is a multi-set of colors. An occurrence of motif in the list-colored graph is a connected vertex set labeled by the colors of motif [9]. Let multi-set $M = \{c_1, c_2, \ldots, c_i, \ldots, c_{|M|}\}$ be a functional motif and let $G_{sub}$ be a sub-graph of $G_l$ with the size of $|M|$ [9], where $c_i$ denotes a color of motif $M$ and $|M|$ is the number of colors of motif $M$, $1 \le i \le |M|$. Let $V_{neighbor}$ be the neighbor set of the vertices in $G_l$.

Given vertex sets $R_1, \ldots, R_i, \ldots, R_{|M|}$, $R_i$ is the vertex set that includes all vertices with color $c_i$ in $G_l$, and $c_i$ is marked as the color of $R_i$. We can define $R_i = \{v | v \in V(G_l) \wedge c_i \in \text{col}(v)\}, 1 \le i \le |M|$. And $R_{\min}$ is the vertex set whose vertex number is minimum in $R_1, \ldots, R_i, \ldots, R_{|M|}$.

Let $H(G_{sub}, M)$ denote a bipartite graph where the vertex set of $H(G_{sub}, M)$ is $G_{sub} \cup M$ and there is an edge between a vertex $v$ of $G_{sub}$ and a vertex $v'$ of $M$ if and only if $v$ has $v'$ as one of its colors [9]. The exact occurrence of motif in list-colored graph is defined as:

**Definition.** An exact occurrence of a motif $M$ in list-colored graph $G_l$ is a connected induced sub-graph $G_{sub}$ of size $|M|$ in $G_l$ such that $H(G_{sub}, M)$ has a perfect matching [9].

Next we describe the list-colored graph motif problem.

Problem statement: Given a list-colored graph $G_l$ and a multi-set of colors as the motif $M$, the list-colored graph motif problem is to find all exact occurrences of motif $M$ in $G_l$.

Generally, this problem can be solved by enumerating sub-graphs in two steps. First, we can enumerate all possible connected induced sub-graphs of $G_l$. Then we can verify whether these sub-graphs are the exact occurrences of motif $M$ in $G_l$. When verifying the enumerated sub-graphs, the connected induced sub-graph $G_{sub}$ is an exact occurrence of motif $M$ in $G_l$ if there is a one-to-one mapping $f$ from $G_{sub}$ to $M$ such that $\forall v \in G_{sub}$: $f(v) \in \text{col}(v)$ [10]. However, the enumeration of all possible sub-graphs could be computationally hard as the size of input graph grows, which makes it difficult to accurately find complex functional motifs in large biological networks. In next section, we will describe our solution for this problem.

### 2.2. Our algorithm

Our presented heuristic algorithm includes two main steps: (1) Choosing the initial vertices for enumerating sub-graphs of input graph. (2) Discovering functional motifs in the procedure of sub-graph enumeration.

Algorithm 1 describes our proposed Motif Search for Biological Networks (MSBN) algorithm.

**Algorithm 1.** MSBN

**Input:** A list-colored graph $G_l$=($V_l$,$E_l$) as the input graph, a multi-set $M$={$c_1$,$c_2$,...,$c_i$,...,$c_{|M|}$} of colors as the query motif, where $c_i$ is a color of $M$ and $|M|$ is the color number of $M$, $1 \le i \le |M|$, the vertex set $R_{min}$;

**Output:** A set of colored motifs $G_{motif}$={$G_1$,$G_2$,...,$G_j$,...,$G_{|Rmin|}$}, where $G_j$ is a motif of size $|G_j|$ in $G_l$ and $|R_{min}|$ is the number of vertices in $R_{min}$, $2 \le |G_j| \le |M|$, $1 \le j \le |R_{min}|$;

1. **for** each vertex $v$ in $G_l$ **do**
2. **if** color $c_i$ is in col($v$) where $c_i \in$ {$c_1$, $c_2$,...,$c_{|M|}$} and $1 \le i \le |M|$ **then**
3. Add $v$ to $R_i$ where $R_i \in$ {$R_1$, $R_2$,…,$R_{|M|}$};
4. **if** the number of vertices in $R_i$ is minimum in {$R_1$,…,$R_i$,…,$R_{|M|}$ } **then**
5. Replace $R_{min}$ with $R_i$;
   end if
   end if
   end for
6. The initial vertex set $V_{initial} \leftarrow R_{min}$;
7. **for** each vertex $v_j$ in $R_{min}$ where $1 \le j \le |R_{min}|$ **do**
8. $V_{neighbor} \leftarrow \Phi$;
9. $G_{sub} \leftarrow \Phi$;
10. Add $v_j$ to $G_{sub}$;
11. Mark $v_j$ as visited;
12. Put the unvisited neighbors of $v_j$ in $V_{neighbor}$;
13. **for** each unvisited vertex $v_k$ in $V_{neighbor}$ where $1 \le k \le |V_{neighbor}|$ **do**
14. Mark $v_k$ as visited;
15. Use $G_{sub}$ and vertex $v_k$ to make an extended sub-graph $G_e$ of $G_{sub}$;
16. Use bipartite maximum matching algorithm to verify whether each vertex of $G_e$ can be assigned to a distinctive color of motif $M$;
17. **if** each vertex of $G_e$ can be assigned to a distinctive color of motif $M$ **then**
18. Add $v_k$ to $G_{sub}$;
19. Put the unvisited neighbors of $v_k$ in $V_{neighbor}$;
    end if
    end for
20. **if** $2 \le |G_{sub}| \le |M|$ **then**
21. Add $G_{sub}$ to $G_{motif}$;
    end if
    end for
22. **Return** $G_{motif}$.

In the subsections 2.2.1 and 2.2.2, we discuss our heuristic algorithm in detail.

### 2.2.1. Choosing initial vertices for sub-graph enumeration

In the list-colored graph motif problem, choosing initial vertex directly affects the accuracy and efficiency of the sub-graph enumeration algorithm. In the subsection 2.2.1, we elaborate on how to select the initial vertex to improve the efficiency and accuracy of sub-graph enumeration.

Recall that $R_i = \{v | v \in V(G_l) \wedge c_i \in \mathrm{col}(v)\}$ is the vertex set that includes all vertices with color $c_i$ in list-colored graph $G_l$, and color $c_i$ is marked as the color of $R_i$. Combining the sub-graph enumeration and this classification, we observe that there exists a one-to-one correspondence between $R_i$ and the color of motif $M$, and derive the following fact.

Fact 1: If there is an exact occurrence of motif $M$ in list-colored graph $G_l$, in this occurrence, the vertex that has the color of $R_i$ comes from $R_i$, $1 \le i \le |M|$.

Fact 1 implies that, in the exact occurrences of motif $M$, all vertices that have the color of $R_i$can only be found in $R_i$, and we can infer the exact occurrence of motif $M$ from the vertices of $R_1$, …,$R_i$, …,$R_{|M|}$.

Based on Fact 1, we derive the following proposition.

**Proposition 1.** *If motif M can be found in the enumeration of the sub-graph of size $|M|$, all exact occurrences of motif M are contained in the sub-graphs of size $|M|$ and these sub-graphs can be enumerated from the vertices of $R_i$,$1 \le i \le |M|$.*

Proof: Because the exact occurrence of motif $M$ is a connected induced sub-graph $G_{sub}$ of size $|M|$ in $G_l$, all exact occurrences of motif $M$ are contained in the sub-graphs of size $|M|$ in $G_l$. Furthermore, because the vertex that has the color of $R_i$ in the occurrence of motif $M$ comes from $R_i$ in $G_l$ (Fact 1), we can enumerate the sub-graphs which contains the exact occurrences of motif $M$ from the vertices of $R_i$.□

Note that if we can determine a vertex of the exact occurrence of a motif, we can directly find this occurrence by enumerating sub-graphs from this vertex. Thus Proposition 1 implies that we can detect all exact occurrences of motif $M$ by enumerating the sub-graphs of size $|M|$ from the vertices of $R_i$.

Inspired by Proposition 1, our algorithm MSBN first groups the vertices of graph $G_l$ into $|M|$ vertex sets $R_1$, …,$R_i$, …,$R_{|M|}$ by the colors of motif $M$ (lines 1–3). Meanwhile, MSBN determines $R_{\min}$ (lines 4–5). Because $R_{\min}$ is the vertex set whose vertex number is minimum in $R_1$, …,$R_i$, …,$R_{|M|}$, we can use $R_{\min}$ as the initial vertex set $V_{initial}$ (line 6) and start the enumeration of the sub-graph $G_{sub}$ from the vertices in $R_{\min}$ to obtain the exact occurrences of motif $M$. Thus, by using this initial vertex selection strategy, we can accurately determine the initial vertices in a certain scope and improve the efficiency of motif detection.

### 2.2.2. Discovering functional motifs during sub-graph enumeration

After the initial vertex set has been determined, in this subsection, we will elaborate on our solution for the discovery of functional motifs in the procedure of sub-graph enumeration. Based on the Definition of the exact occurrence of motif $M$, there exists the following proposition.

**Proposition 2.** *In the list-colored graph motif problem, the exact occurrence of motif M in list-colored graph $G_l$ is a connected induced sub-graph $G_l'$ of size $|M|$ and each vertex of $G_l'$ can be assigned to a distinctive color of motif M.*

Proof: According to the Definition of the exact occurrence of motif, the exact occurrence of motif $M$ in $G_l$ is a connected induced sub-graph $G_l'$ of size $|M|$ such that bipartite graph $H(G_l', M)$ has a perfect matching. Because bipartite graph $H(G_l', M)$ has a perfect matching, each vertex of $G_l'$ can be assigned to a distinctive color of motif $M$.□

Proposition 2 implies that, in order to find the exact occurrence of motif $M$ in graph $G_l$, we can select the vertices based on the colors of

motif $M$ and the colors of the vertices in $G_l$ to construct candidate motifs when enumerating sub-graphs.

Next, we discuss how to transform the detection of list-colored graph motif to the search of the connected induced sub-graph of $G_l$ based on Proposition 2.

At the beginning of enumerating sub-graphs, MSBN selects an unvisited vertex $u$ from $R_{\min}$, and produces an empty sub-graph $G_{sub}$ and an empty vertex set $V_{neighbor}$ for this vertex (lines 7-9). Then MSBN adds vertex $u$ as the first vertex in $G_{sub}$ (line 10), and marks $u$ as visited (line 11) and puts the unvisited neighbors of $u$ in $V_{neighbor}$ (line 12).

In a recursive manner, MSBN first selects an unvisited vertex $v_k$ from $V_{neighbor}$ and marks this vertex as visited (lines 13–14). In lines 16–17, MSBN uses $G_{sub}$ and $v_k$ to make an extended sub-graph $G_e$ of $G_{sub}$, and uses Hopcroft-Karp bipartite maximum matching algorithm [20] to check if each vertex of $G_e$ can be assigned to a distinctive color of motif $M$. Fig. 1 shows an example of bipartite graph $G_B$ for this checking.

In Fig. 1, the bipartite graph $G_B$ consists of graph $G_e$ and the colors of vertices of $G_e$. Every vertex $v$ in $G_e$ is connected to its colors col($v$). If there is a maximum matching in bipartite graph $G_B$, each vertex of $G_e$ can be assigned to a distinctive color of motif $M$. That is, $G_e$ is a valid list-colored graph motif and the detected candidate motif is validated here. Then MSBN adds vertex $v_k$ to $G_{sub}$ (line 18) and puts the unvisited neighbors of this vertex in $V_{neighbor}$ (line 19). In the enumeration of sub-graph $G_{sub}$, this procedure (lines 13–19) will be repeated until no vertex can be added to $G_{sub}$. Thus, inspired by Proposition 2, MSBN selects vertices based on the colors of motif and the colors of the vertices in $G_l$ to construct and validate candidate motifs when enumerating $G_{sub}$.

At the end of the enumeration, $G_{sub}$ is a list-colored graph motif. If the size of $G_{sub}$ is $|M|$, according to Proposition 2, $G_{sub}$ is the exact occurrence of motif $M$ in graph $G_l$. After finishing this sub-graph enumeration, MSBN will continue to enumerate other similar sub-graphs from the vertices of $R_{\min}$ in the same way until all vertices of $R_{\min}$ are visited. Then, the functional motifs are discovered once the sub-graph enumeration is finished; thereby we integrate the process of candidate motif validation into the sub-graph enumeration. This enables us to accurately discover the occurrences of query motif without enumerating and verifying all sub-graphs.

Finally, MSBN keeps all detected list-colored graph motifs (lines 20–21) and reports these motifs as the search results (line 22).

The overview of our method is summarized in Fig. 2.

### 2.3. Time complexity

In the following, we discuss the time complexity of Algorithm 1. (1) As can be seen from Algorithm 1, the "for" loop in line 1 will repeat at most $|G_l|$ times to determine the initial vertex set $V_{initial}$ (lines 1–6). (2) It requires at most $|R_{\min}|$ times to repeat the "for" loop in line 7to enumerate $|R_{\min}|$ sub-graphs. For each sub-graph enumeration, the inner "for" loop in line 13 will be repeated at most $|G_l|$ times because it requires visiting at most $|G_l|$ vertices to find valid vertex to enumerate candidate sub-graph $G_{sub}$. Meanwhile, in each inner "for" loop in line 13, the dominate part is finding bipartite maximum matching using Hopcroft-Karp algorithm with $O(|M|^{5/2})$ time [20]. Hence, in lines 7–21, the time complexity for finding motif $M$ through enumerating sub-graphs from $R_{\min}$ is $O(|R_{\min}| \times |G_l| \times |M|^{5/2})$. Consequently, the total time complexity for MSBN is $O(|G_l| + |R_{\min}| \times |G_l| \times |M|^{5/2}) = O(|R_{\min}| \times |G_l| \times |M|^{5/2})$.

## 3. Results

In this section, we will evaluate the experimental performance of MSBN on searching list-colored graph motifs in biological networks. MSBN is implemented in C++. MOTUS [9] is a representative motif detection algorithm which is developed to find list-colored graph motifs in metabolic networks using breadth first search. And RANGI [11] is another recently proposed algorithm that is capable of finding list-colored graph motifs in metabolic networks using heuristic search. Since RANGI [11] and MOTUS [9] are the two available frameworks that are able to detect list-colored graph motifs in metabolic networks, we choose MOTUS and RANGI as the baselines to verify the performance of MSBN on finding list-colored graph motifs in metabolic networks.

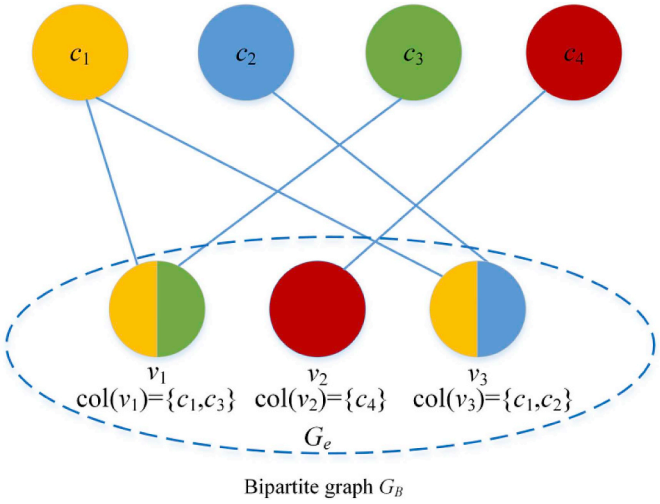


**Fig. 1.** An example of bipartite graph $G_B$ for checking the matching between vertices and colors. The first part of bipartite graph $G_B$ is graph $G_e$ ($G_e$ is encircled with blue dashed line), and the other part is the colors of the vertices of $G_e$. Every vertex $v$ in $G_e$ is connected to its colors col($v$).

In subsection 3.1, we will compare the performance of MSBN, RANGI and MOTUS on searching list-colored graph motifs in metabolic networks.

Although the list-colored graph motif problem has been first introduced in finding functional motifs in metabolic networks [9,11], some efforts such as RANGI [11], GraMoFone [17], and CeFunMO [4], have tried to find list-colored graph motifs in protein-interaction networks. GraMoFone is an available method which is designed to find functional motifs in protein-interaction networks using Linear Pseudo-Boolean optimization. Besides finding motifs in metabolic networks, RANGI is also able to find functional motifs in protein-interaction networks as well. Lastly, CeFunMO is the most recently proposed method for detecting functional motifs in protein-interaction networks using greedy search strategy. Based on the availability of the softwares, we choose RANGI, GraMoFone and CeFunMO as the baselines to evaluate the performance of our method on finding list-colored graph motifs in protein-interaction networks.

In subsection 3.2, we will compare the experimental performance of MSBN, RANGI, GraMoFone and CeFunMO on searching list-colored graph motifs in protein-interaction networks.

### 3.1. Functional motif detection in metabolic networks

From the metabolic network data of the KEGG database retrieved and reformatted by Ay et al. [21], we use the metabolic networks of four species *Homo sapiens* (*hsa*), *Musmusculus* (*mmu*), *Escherichia coli* (*eco*) and *Agrobacterium tumefaciens* (*atc*) as the test data to evaluate the performance of motif detection methods on finding functional motifs in metabolic networks. *hsa* and *mmu* are two representative species from the eukaryota domain, while *eco* and *atc* are two representative species from the bacteria domain [21]. Besides the real metabolic network data, we also include the synthetic metabolic networks of *hsa*, *mmu*, *eco* and *act* as the test data as well. Abaka et al. [22] produced these synthetic metabolic networks by the metabolism categories and metabolic network data provided in the KEGG database.

In this section, *hsa-mmu* means that we use the sub-graphs of the

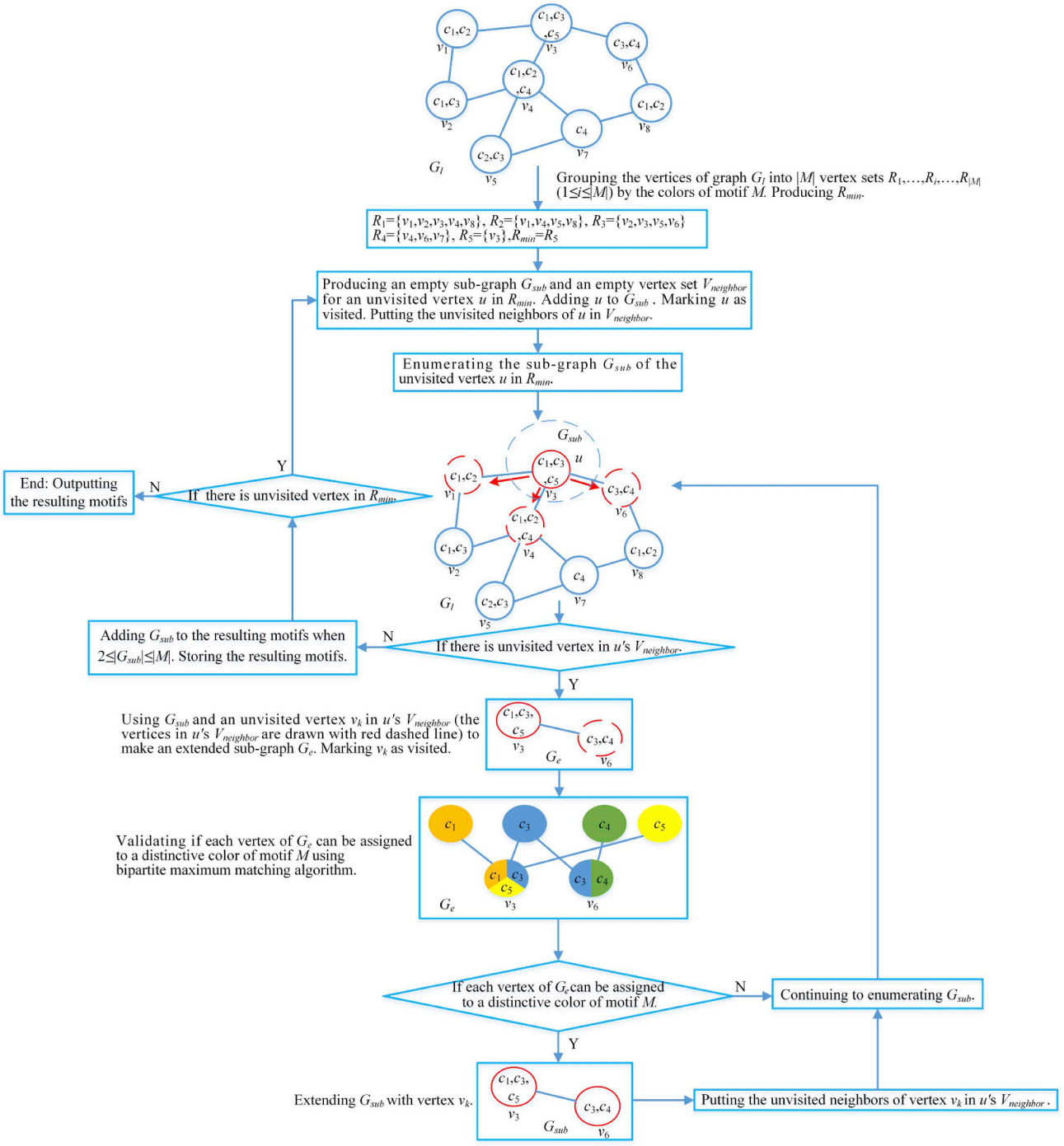


**Fig. 2.** The overview of MSBN. A sample list-colored graph $G_l$ is given and $c_1$, $c_2$, $c_3$, $c_4$, $c_5$ are the colors of motif $M$. The sub-graph $G_{sub}$ in $G_l$ is encircled with blue dashed line. In $G_l$, the vertex $u$ in $G_{sub}$ is drawn with red line, and the neighbors of $u$ are drawn with red dashed line.

metabolic networks of *mmu* as the query motifs and use the corresponding metabolic networks of *hsa* as the targets, and *eco-atc*, *hsa-atc* and *eco-mmu* have similar meanings. For each query, we use motif detection methods to search the exact occurrences of a query motif in the target. In the query motif, each different reaction is assigned to a distinctive color [9]. Given a reaction $A$ of the query and a reaction $B$ of the target, if the input or output compounds of $A$ are identical to the input or output compounds of $B$, we assign the color of $A$ to $B$ [9]. In contrast, if there is no reaction that has the same input or output compounds with reaction $B$ in the query, we do not assign any color to reaction (vertex) $B$. In this way, each metabolic network is regarded as a list-colored graph, and the colors assigned to the reactions of query motif consist of the color set of query.

The experimental comparisons are conducted based on the following five criteria.

a. Sensitivity Sn = $tp/(tp + fn)$ where true positives ($tp$) are the reactions found in both query motif and resulting motif, false negatives ($fn$) are the reactions in query motif but not in resulting motif.
b. Positive prediction value $PPV = tp/(tp + fp)$ where false positives ($fp$) are the reactions not in query motif but in resulting motif.
c. Accuracy $Ac = (Sn + PPV)/2$. Higher accuracy of resulting motifs indicates that the resulting motifs are closer to query motifs, which demonstrates the ability of recovering query motifs for motif

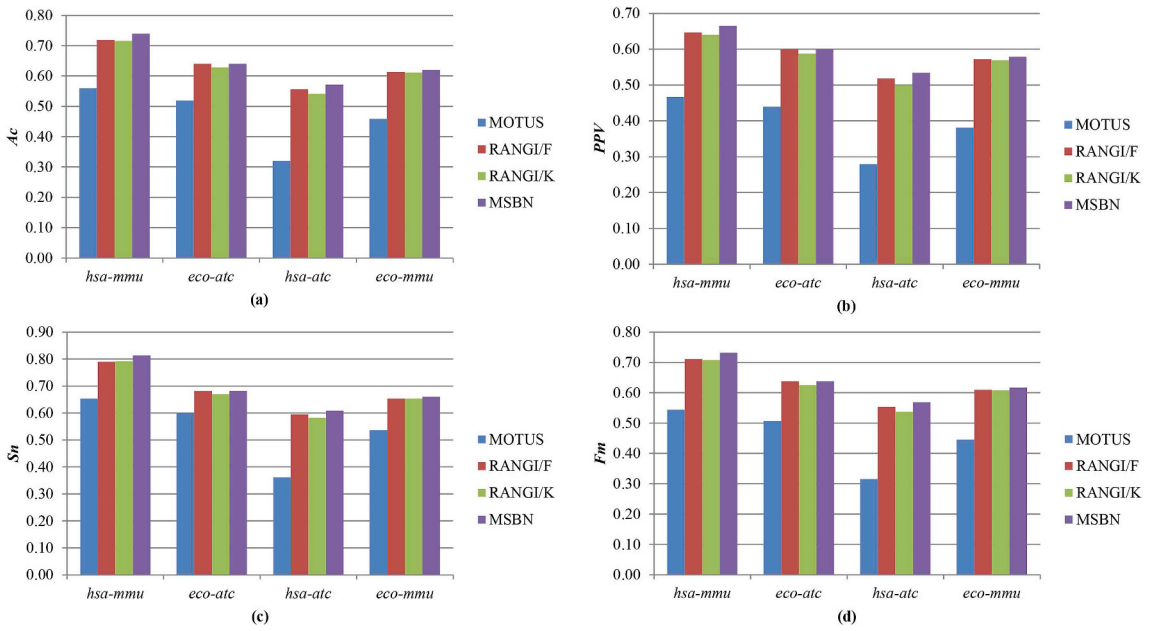


**Fig. 3.** *Ac*, *PPV*, Sn and *Fm* of the resulting motifs detected in metabolic networks.

**Table 1**
*RCR* of the resulting motifs detected in metabolic networks.

| Methods | *RCR* | | | |
|---|---|---|---|---|
| | *hsa-mmu* | *eco-atc* | *hsa-atc* | *eco-mmu* |
| MOTUS | 0.830 | 0.758 | 0.648 | 0.591 |
| RANGI/F | 0.859 | 0.828 | 0.735 | 0.772 |
| RANGI/K | 0.857 | 0.818 | 0.737 | 0.768 |
| MSBN | **0.875** | **0.836** | **0.751** | **0.774** |

The best performer is marked in bold.

detection method.

d. *F*-measure $Fm = (2 \times PR \times RC)/(PR + RC)$ where *PR* is the precision and $PR = PPV$, *RC* is the Recall and $RC = Sn$, and Recall is the proportion of positive cases [20].
e. Reaction correctness ratio (*RCR*) is the percentage of the correct vertices of resulting motif in query motif. The average value of *RCR* is calculated by the following equation:

$$RCR = \frac{1}{N}\sum_{i=1}^{N}\frac{V_i}{V_M}$$

where we use measurement FGC (functional group conversion category), which was previously used to evaluate the biochemical relevance of reactions in metabolic networks in Ref. [22], to measure the biochemical relevance of the vertices (reactions) in resulting motifs. Based on functional group conversion categories for reactions, the KEGG database provides FGC hierarchy [23], which divides the reactions into 8 categories and partitions every category into 5 levels [22]. The same functional group undergoes the same or similar chemical reaction(s) [24,25]. That is if the reactions are included in the same level of the FGC hierarchy under the same category, these reactions are considered to be biochemically relevant [22]. The higher level of the FGC hierarchy is, the better biochemical relevance of the reactions is. Recall that a color of the vertex (reaction) in resulting motif corresponds to a reaction of query motif. In the resulting motif, if a vertex (reaction) and one of its colors (reactions) are included in the 5th level of the FGC hierarchy under the same category, these two reactions are considered to be biochemically relevant and this vertex is called correct vertex (reaction). In the equation for calculating the average value of *RCR*, $V_i$ is the number of the correct vertices in the *i*th resulting motif, *N* is the number of resulting motifs and $V_M$ is the number of the colors in query motif. Higher values of *RCR* of resulting motifs indicate that the biochemical relevance between the reaction of resulting motifs and the reaction of query motifs is better.

In the experiment of detecting motifs in metabolic networks, we ran three algorithms MSBN, RANGI and MOTUS on the computer with an Intel Xeon E5620 CPU and 40 GB RAM. The running operating system is Linux. MOTUS used the search mode to get all occurrences of the motif. RANGI has two different modes “RANGI/F” and “RANGI/K”. “RANGI/F” and “RANGI/K” indicate the RANGI algorithm with FANMOD [26] and Kavosh [27] enumeration algorithms, respectively.

In the following, we will evaluate the performance of MSBN, RANGI and MOTUS on finding functional motifs in metabolic networks.

Fig. 3 summarizes the accuracy (*Ac*), positive prediction value (*PPV*), sensitivity (*Sn*) and *F*-measure (*Fm*) of the resulting motifs for all compared methods.

As shown in Fig. 3, both MSBN and RANGI perform better than MOTUS, while the results of MSBN obtain relatively good performances in accuracy (*Ac*), positive prediction value (*PPV*), sensitivity (*Sn*) and *F*-measure (*Fm*). These results indicate that the resulting motifs of algorithms MSBN and RANGI are more similar to the query motifs than those of algorithm MOTUS, and the ability of recovering query motifs for MSBN is better in comparison with MOTUS and RANGI.

Table 1 presents the values of *RCR* of the resulting motifs detected in metabolic networks.

From Table 1, we can observe that, the values of *RCR* of algorithms MSBN and RANGI are higher than those of algorithm MOTUS. Moreover, MSBN performs the best with the highest values of *RCR* for the resulting motifs detected in the metabolic networks of *hsa* and *eco*. This illustrates that MSBN tends to generate functional motifs with more biochemically relevant reactions comparing with MOTUS and RANGI.

Consequently, these results indicate that our method is an effective method for detection of functional motifs in metabolic networks.

Furthermore, for each test, following literature [4,11], all programs were given 10 min to find functional motifs in metabolic networks. For each method, the instance taking longer than 10 min or running out of memory was assumed to be unsolved. Table 2 lists the percentage of the solved instances.

As shown in Table 2, both MSBN and RANGI solved more instances

**Table 2**
Comparison of the percentage of the solved instances.

| Methods | The ratio of solved instances | | | |
|---|---|---|---|---|
| | *hsa-mmu* | *eco-atc* | *hsa-atc* | *eco-mmu* |
| MOTUS | 0.35(38/110) | 0.125(13/104) | 0.24(23/94) | 0.25(23/92) |
| RANGI/F | **0.44(48/110)** | **0.40(42/104)** | **0.32(30/94)** | **0.37(34/92)** |
| RANGI/K | 0.43(47/110) | 0.38(40/104) | 0.29(27/94) | 0.36(33/92) |
| MSBN | **0.44(48/110)** | **0.40(42/104)** | **0.32(30/94)** | **0.37(34/92)** |

The best performer is marked in bold.

**Table 3**
Average running time for the solved instances.

| Methods | Average running time (seconds) | | | |
|---|---|---|---|---|
| | *hsa-mmu* | *eco-atc* | *hsa-atc* | *eco-mmu* |
| MOTUS | 21.513 | 0.273 | 2.385 | 0.240 |
| RANGI/F | 0.680 | 0.953 | 0.799 | 0.680 |
| RANGI/K | 0.273 | 10.611 | 2.865 | 0.863 |
| MSBN | **0.156** | **0.176** | **0.069** | **0.185** |

The best performer is marked in bold.

than MOTUS. These results show that within given time, MSBN and RANGI are more capable of finding functional motifs in metabolic networks than MOTUS.

Table 3 shows the average running time for the solved instances in the motif detection of metabolic networks.

In Table 3, we can see that the average running times of MSBN for the solved instances are less than that of MOTUS and RANGI in the motif detection of metabolic networks. This demonstrates that our heuristic motif search strategy, which determines the initial vertices in a certain scope and integrates the process of motif validation into subgraph enumeration, helps to reduce the search space and improve the efficiency of the algorithm.

In summary, these results demonstrate that MSBN is more time-efficient and is capable of achieving higher accuracy than the compared approaches in finding functional motifs in metabolic networks.

### 3.2. Functional motif detection in protein-interaction networks

In this section, we verify the performance of MSBN on finding functional motifs in protein-interaction networks. Following [11,17], we have also used the protein-interaction network data which was assembled by Bruckner et al. [17] as the test data. These data included 3 protein-interaction networks of fly (6650 proteins, 39936 interactions), yeast (5430 proteins, 39936 interactions), and human (7915 proteins, 21275 interactions), as well as 6 different sets of protein complexes of mouse (217 complexes), fly (144 complexes), human (969 complexes), rat (200 complexes), bovine (19 complexes) and yeast (297 complexes).

In the experiment of discovering functional motifs in protein-interaction networks, each protein complex is considered as a query and each protein in a protein complex is assigned with a color [4]. And the colors of query proteins are assigned to the proteins in the protein-interaction network whose BLAST *E*-values are lower than $10^{-7}$ [11]. In this way, each protein-interaction network is regarded as a list-colored graph [4].

#### 3.2.1. Comparison with existing methods

In the experiment of detecting motifs in protein-interaction networks, algorithms MSBN, RANGI [11], GraMoFone [17], and CeFunMO [4] were run on the computer with an Intel core i7 CPU and 16 GB RAM. The running operating system is Windows 7. Based on whether the query size or the motif size is used as input, GraMoFone has two versions which correspond to GRAM/M or GRAM/Q respectively. As

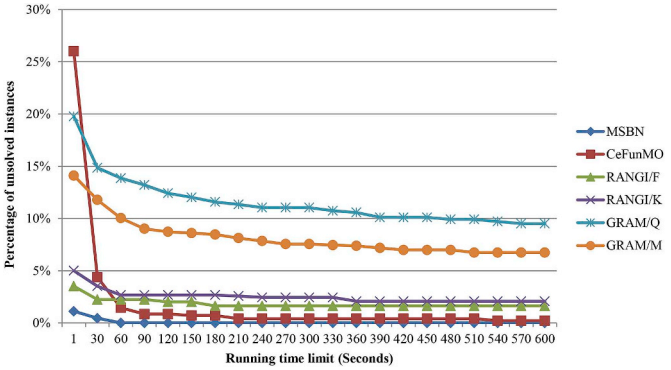


**Fig. 4.** Percentage of unsolved instances in different running time limits.

**Table 4**
Average running time (seconds) for different query size (|Q|).

| Methods | \|Q\| < 16 | 16 < =\|Q\| < 32 | 32 < =\|Q\| < 64 | \|Q\| > = 64 |
|---|---|---|---|---|
| MSBN | **0.00285** | **0.011** | **0.527** | **5.841** |
| CeFunMO | 6.234 | 1.887 | 9.754 | 98.852 |
| RANGI/F | 0.047 | 0.197 | 42.325 | 223.874 |
| RANGI/K | 1.898 | 0.971 | 58.724 | 275.648 |
| GRAM/Q | 38.753 | 121.519 | 154.318 | 340.286 |
| GRAM/M | 37.897 | 200.177 | 236.989 | 517.758 |

The best performer is marked in bold.

mentioned in section 3.1, RANGI has two different modes RANGI/F and RANGI/K.

In the experiment, following [4,11], all methods were given specified time to find functional motifs in protein-interaction networks. When a method took longer time than the specified time or ran out of memory in an instance, this instance was assumed to be unsolved for this method. Following [4], the results are restricted to the instances whose running times are non-zero for at least one of these comparative methods. These instances included 447 list-colored graphs. Fig. 4 presents the percentage of unsolved instances for each method in different running time limits.

As shown in Fig. 4, the percentages of unsolved instances of MSBN are of the lowest for different running time limits. And the running times of MSBN for all instances are less than 10 min whereas none of the other methods solve all the instances in 10 min (one instance of CeFunMO consumed more than 10 min).

These results show that MSBN is capable of finding functional motifs in protein-interaction networks within given time.

Furthermore, Table 4 shows the average running time of different algorithms for different size of query (|Q|).

As can be seen from Table 4, MSBN consumed less time than the comparative methods for different size of query. Table 5 summarizes the unsolved instances of RANGI (the time limit is 10 min) and the maximum size of detected motifs for each instance.

As shown in Table 5, over all 7 instances, MSBN solved all instances in less than 80 s, and MSBN took much less time than CeFunMO for these 7 instances.

#### 3.2.2. Functional enrichments of the motifs in protein-interaction networks

In order to measure the functional enrichment of detected motifs, following [11], we also used the GO:TermFinder tool [28] to find statistically significant gene ontology (GO) terms shared by the proteins in detected motifs. A GO term is assumed to be significant, if its corrected *p*-value using the Bonferoni method is lower than a threshold of 0.05 [11]. For the specified threshold, a discovered motif with at least one significant term is considered as a functionally enriched motif [4].

Table 6 shows the percentages of functionally enriched motifs found

**Table 5**
The protein-interaction networks that RANGI did not detect any motif in 10 min.

| List-colored graph name | Number of vertices | Number of edges | CeFunMO running time (sec) | MSBN running time (sec) | CeFunMO maximum size of detected motifs | MSBN maximum size of detected motifs |
|---|---|---|---|---|---|---|
| y_h.616.104 | 145 | 373 | 6.77 | 0.524 | 42 | 42 |
| y_m.152.50 | 120 | 475 | 12.541 | 12.172 | 45 | 36 |
| y_h.444.79 | 119 | 281 | 7.98 | 1.37 | 33 | 32 |
| h_y.11.367 | 661 | 882 | 684.385 | 64.033 | 81 | 82 |
| h_y.149.132 | 574 | 873 | 192.841 | 7.96 | 52 | 53 |
| f_y.11.367 | 446 | 163 | 132.358 | 76.748 | 65 | 66 |
| y_h.470.142 | 240 | 575 | 72.143 | 15.321 | 53 | 50 |

**Table 6**
Percentage of functionally enriched motifs found by MSBN.

| Protein-interaction network | Percentage of enriched motifs |
|---|---|
| Fly | 336/584(57.53%) |
| Human | 2225/4010(81.10%) |
| Yeast | 6310/6451(97.81%) |

by MSBN in three protein-interaction networks and the threshold of *p*-value is set to 0.05.

Furthermore, Fig. 5 shows the percentages of functionally enriched motifs detected by MSBN in these three protein-interaction networks for different threshold values.

The results of Table 6 and Fig. 5 demonstrate that MSBN is capable of detecting functionally enriched motifs in each protein-interaction network.

## 4. Conclusion

Accurate list-colored graph motif detection plays an important role in studies of biological networks as it offers a systematic way for uncovering the intrinsic biological interactions among the molecules of biological networks. However, discovering list-colored graph motifs in biological networks remains computationally challenging and intractable. Here, we present a heuristic motif search method MSBN to find list-colored graph motif in biological networks. MSBN transforms the detection of list-colored graph motif to the search of connected induced sub-graph in list-colored graph, where the vertices in the sub-graphs are assigned to distinctive colors of motif. Distinguishing from the well-known method RANGI, which enumerates all possible sub-graphs to detect list-colored graph motifs, this transformation enables us to accurately uncover the occurrences of query motif in list-colored graph without enumerating and verifying all sub-graphs.

On the other hand, since the efficiency and accuracy of functional motif detection are directly affected by the initial vertex selection in sub-graph enumeration, we propose a new strategy for choosing initial vertices based on the colors of vertices in the MSBN method, which accurately determines the search scope of motifs. To verify the utility of MSBN, we conducted the experiments of detecting functional motifs on metabolic networks and protein-interaction networks. The experimental results indicate that MSBN is superior to other existing methods in terms of processing time and accuracy. This enables us to accurately detect complex functional motifs in large biological networks easier, which can facilitate future study on network motif and its related areas in systems biology. MSBN thus can be an effective alternative to existing methods for detecting list-colored graph motifs in biological networks.

Generally, heuristic search tries to produce locally optimal solutions by available information and it is possible to lose globally optimal solution. MSBN is a heuristic method and not an exact method. To balance the efficiency and accuracy of our algorithm, we adopted heuristic search strategy for discovering functional motifs, which may lead to missing some possible functional motifs. The list-colored graph motif problem is a topology-free problem. We plan to incorporate the topology information of biological networks into the list-colored graph motif detection, which may help our method to accurately find more possible functional motifs.

Furthermore, traditional genome-based phylogenetic methods typically reconstruct phylogenetic relationships of organisms based on genomic and genetic data. Genomic information is difficult to clarify and does not necessarily signify function whereas functional motifs are classified on biological function and not rigorously on genomic information. It would be of our interest to combine genomic information

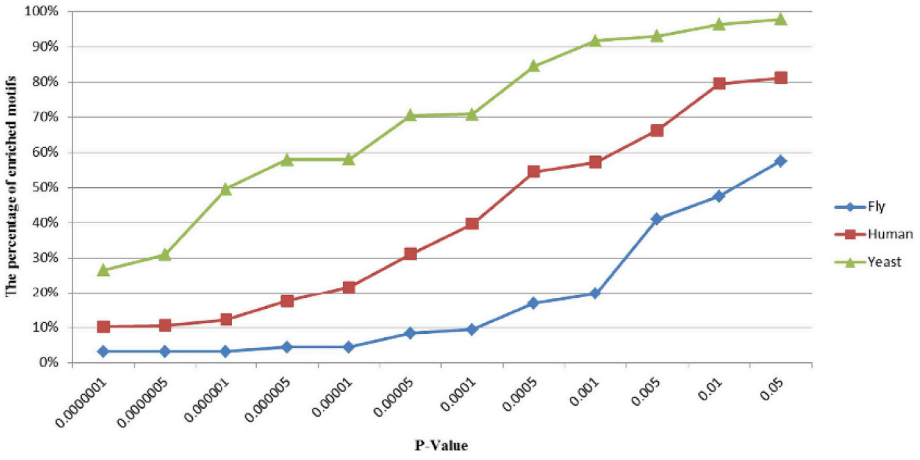


**Fig. 5.** The percentage of functionally enriched motifs found by MSBN (*p*-value less than specific cut-off).

with functional motifs to analyze evolutionary relationships in the future. This may provide a useful complement to traditional phylogenetic reconstruction and offer a more comprehensive understanding of evolution.

## Conflict of interest

None Declared.

## Availability of the software

Algorithm MSBN is implemented in C++. The DOI information to access the data and program in FigShare is 10.6084/m9.figshare.7429910.

## Acknowledgment

We thank the editor and anonymous reviewers for their constructive comments, which greatly help us improve our manuscript.

This work is supported by the National Natural Science Foundation of China under Grant No. 61862006, the National Natural Science Foundation of China under Grant No. 61462005, and the Basic Capability Improvement Project of Young University Teachers of Guangxi under Grant No.2018KY0026.

## References

[1] W. Kim, M. Li, J. Wang, Y. Pan, Biological network motif detection and evaluation, BMC Syst. Biol. 5 (3) (2011) 1–13.
[2] P. Ribeiro, F. Silva, L. Lopes, Parallel discovery of network motifs, J. Parallel Distr. Comput. 72 (2) (2012) 144–154 2//.
[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[4] M. Kouhsar, Z. Razaghi-Moghadam, Z. Mousavian, A. Masoudi-Nejad, CeFunMO: a centrality based method for discovering functional motifs with application in biological networks, Comput. Biol. Med. 76 (2016) 154–159 9/1/.
[5] Y. Huang, C. Zhong, H. Lin, J. Wang, Y. Peng, Reconstructing phylogeny by aligning multiple metabolic pathways using functional module mapping, Molecules 23 (2) (2018) 486.
[6] I. Albert, R. Albert, Conserved network motifs allow protein–protein interaction prediction, Bioinformatics 20 (18) (2004) 3346–3352 December 12, 2004.
[7] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, U. Alon, Coarse-graining and self-dissimilarity of complex networks, Phys. Rev. 71 (1) (2005) 01612701/21/.
[8] S. Valverde, R.V. Solé, Network motifs in computational graphs: a case study in software architecture, Phys. Rev. 72 (2) (2005) 02610708/08/.
[9] V. Lacroix, C.G. Fernandes, M.-F. Sagot, Motif search in graphs: application to metabolic networks, IEEE ACM Trans. Comput. Biol. Bioinf. 3 (4) (2006) 360–368.
[10] N. Betzler, R. v. Bevern, M.R. Fellows, C. Komusiewicz, R. Niedermeier, Parameterized algorithmics for finding connected motifs in biological networks, IEEE ACM Trans. Comput. Biol. Bioinf. 8 (5) (2011) 1296–1308.
[11] A.G. Rudi, S. Shahrivari, S. Jalili, Z.R.M. Kashani, RANGI: a fast list-colored graph motif finding algorithm, IEEE ACM Trans. Comput. Biol. Bioinf. 10 (2) (2013) 504–513.
[12] R. Rizzi, F. Sikora, Some results on more flexible versions of graph motif, Theor. Comput. Syst. 56 (4) (May 01, 2015) 612–629.
[13] M.R. Fellows, G. Fertin, D. Hermelin, S. Vialette, Upper and lower bounds for finding connected motifs in vertex-colored graphs, J. Comput. Syst. Sci. 77 (4) (2011) 799–811 2011/07/01.
[14] S. Guillemot, F. Sikora, Finding and counting vertex-colored subtrees, Algorithmica 65 (4) (2013) 828–844.
[15] I. Koutis, Constrained multilinear detection for faster functional motif discovery, Inf. Process. Lett. 112 (22) (2012) 889–892 11/30/.
[16] S. Bruckner, F. Hüffner, R.M. Karp, R. Shamir, R. Sharan, Topology-free querying of protein interaction networks, J. Comput. Biol. 17 (3) (2010) 237–252 2010/03/01.
[17] G. Blin, F. Sikora, S. Vialette, GraMoFoNe: a Cytoscape Plugin for Querying Motifs without Topology in Protein-Protein Interactions Networks, (2010), pp. 38–43.
[18] N. Alon, R. Yuster, U. Zwick, Color-coding, J. ACM 42 (4) (1995) 844–856.
[19] N. Betzler, M.R. Fellows, C. Komusiewicz, R. Niedermeier, Parameterized algorithms and hardness results for some graph motif problems, Combin. Pattern Matching (2008) 31–43.
[20] G. Hripcsak, A.S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, J. Am. Med. Inf. Assoc. 12 (3) (2005) 296–298.
[21] F. Ay, M. Kellis, T. Kahveci, SubMAP: aligning metabolic pathways with subnetwork mappings, J. Comput. Biol. 18 (3) (2011) 219–235.
[22] G. Abaka, T. Bıyıkoğlu, C. Erten, CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways, Bioinformatics 29 (13) (2013) i145–i153.
[23] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res. (2011) gkr988.
[24] J. March, Advanced Organic Chemistry: Reactions, Mechanisms, Structure, McGraw-Hill, 1968.
[25] Y. Huang, C. Zhong, H.X. Lin, J. Wang, A method for finding metabolic pathways using atomic group tracking, PLoS One 12 (1) (2017) e0168725.
[26] S. Wernicke, F. Rasche, FANMOD: a tool for fast network motif detection, Bioinformatics 22 (2006).
[27] Z.R.M. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E.S. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, A. Masoudi-Nejad, Kavosh: a new algorithm for finding network motifs, BMC Bioinf. 10 (1) (2009) 1–12.
[28] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, G. Sherlock, GO:TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, Bioinformatics 20 (18) (2004) 3710–3715.