Computers in Biology and Medicine 76 (2016) 154–159

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

CeFunMO: A centrality based method for discovering functional motifs with application in biological networks

Morteza Kouhsar [a,*], Zahra Razaghi-Moghadam [b,*], Zaynab Mousavian [a], Ali Masoudi-Nejad [a]

[a] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[b] Faculty of New Sciences and Technology (FNST), University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 24 April 2016
Received in revised form
12 July 2016
Accepted 17 July 2016

Keywords:
Biological network
Centrality
Functional motif
List-colored graph
Protein complex

ABSTRACT

Detecting functional motifs in biological networks is one of the challenging problems in systems biology. Given a multiset of colors as query and a list-colored graph (an undirected graph with a set of colors assigned to each of its vertices), the problem is reduced to finding connected subgraphs, which best cover the multiset of query. To solve this NP-complete problem, we propose a new color-based centrality measure for list-colored graphs. Based on this newly-defined measure of centrality, a novel polynomial time algorithm is developed to discover functional motifs in list-colored graphs, using a greedy strategy. This algorithm, called CeFunMO, has superior running time and acceptable accuracy in comparison with other well-known algorithms, such as RANGI and GraMoFoNe.

1. Introduction

In recent years, substantial interest has been shown in analysis of biological networks by means of extracting notable patterns, such as network motifs, within them [1,2]. During the earlier years, the concept of motif was mainly used to address recurring patterns of interconnections in networks, namely, topological motifs [3–5]. Later on, the notion of functional motif was introduced by Lacroix et al. [6] based on the functionality of various patterns within networks.

Generally, biological networks are modeled as graphs, in which, each vertex represents a biological element and each edge stands for an interaction between two such elements [7,8]. Moreover, in order to account for the functionalities of each biological element in these networks, the corresponding vertex is marked by a set of distinctive colors. Given a specific list of desired functionalities, which is represented by a multiset of colors, a functional motif within a network is defined by a connected subgraph which best covers this multiset of designated colors [9].

In the past decade, many computational methods have been developed to detect functional motifs in biological networks [9–13]. In the work of Lacroix et al. [6] on metabolic networks, the problem of finding functional motifs in list-colored graphs was proved to be NP-complete. Dondi et al. [11] investigated a problem, in which, the target network is a tree with maximum degree 3 and each color of the motif occurs at most twice in the network, and proved APX-hardness of this problem. Later, Fellows et al. [14] have developed the FPT algorithm to search within simplified list-colored graphs, in which, only a single color is assigned to each vertex. They have also proved that in this particular case, the problem is NP-complete. Guillemot and Sikora [15] have reduced the problem of finding functional motifs in list-colored graphs to a multi-linear detection problem. In [16], Koutis used an extended version of multi-linear detection problem, called constrained multi-linear detection to reformulate this problem, and then proposed a fast algorithm to solve it. Bruckner et al. [12] studied a special case, in which, motifs with possibly uncolored vertices were considered. Their proposed algorithm, called TORQUE, is based on integer linear programming and color-coding. Blin et al. [13] reduced the same problem to a linear pseudo-boolean optimization problem and subsequently introduced the GraMoFoNe algorithm, which experimentally outperformed TORQUE. Recently, Rudi et al. [9] proposed a new branch-and-bound algorithm called RANGI to solve the problem. This algorithm exploited two subgraph enumeration algorithms [17,18] to discover functional motifs. In RANGI, some heuristic techniques were used to reduce the size of search space.

To search for functional motifs in biological networks, a new algorithm called CeFunMO (**Ce**ntrality-based **Fu**nctional **MO**tif

* Corresponding authors.
E-mail addresses: m.kouhsar@ut.ac.ir (M. Kouhsar), razzaghi@ut.ac.ir (Z. Razaghi-Moghadam), zmousavian@ut.ac.ir (Z. Mousavian), amasoudin@ibb.ut.ac.ir (A. Masoudi-Nejad).

http://dx.doi.org/10.1016/j.compbiomed.2016.07.009
0010-4825/

detection algorithm) is presented in this paper. CeFunMO adopts a greedy strategy to solve the problem based on a newly-defined measure of vertex centrality (color-centrality), specifically designed for the purpose of finding functional motifs. The value assigned to each vertex by color-centrality measure, depends on two factors, namely, its connectivity with other vertices in the network and its appointed set of colors. Based on this new definition of centrality, the vertices with higher centralities are considered as more appropriate candidates in constructing functional motifs. In accordance with this key idea, CeFunMO detects the connected subgraphs with higher color-centralities as functional motifs.

In order to investigate the applicability and efficiency of CeFunMO, this algorithm was applied on three well-known protein-protein interaction networks, assembled by Bruckner et al. in [12]. In these experiments, protein complexes are fed as queries to CeFunMO, with the goal of finding maximal matches in aforementioned networks. These experiments indicate that CeFunMO leads to acceptable results in terms of accuracy, compared to established approaches such as RANGI [9] and GraMoFoNe [13], while proving to be more effective given desired queries of larger scale.

2. Terminology

In this article, biological networks are represented as simple undirected graphs, each denoted as a pair $G=(V, E)$, where V and E stand for the set of vertices (biological elements) and the set of edges (interactions among those elements), respectively. A graph $H=(V_H, E_H)$ is called a subgraph of G, if and only if $V_H \subseteq V$ and $E_H \subseteq E$. A subgraph $G[M]$ is said to be an induced subgraph of G on a vertex set M, if it contains all edges in G, whose endpoints are both in M [8]. The neighbor set of a vertex v in graph G, denoted by $N_G(v)$, is defined as follows:

$$N_G(v) = \{u \in V \mid (u, v) \in E\}. \quad (1)$$

For a subset of vertices V', its neighbor set is denoted by $NG(V')$ and defined by:

$$N_G(V') = \bigcup_{v \in V'} N_G(v). \quad (2)$$

A list-colored graph, $G_l=(V_l, E_l)$, is an undirected graph in which a set of colors is assigned to each vertex. $col(v)$ denotes the set of colors assigned to vertex v, referred to as color set of v. For a subset of vertices $V' \subseteq V_l$, the color set of V' is defined by:

$$col(V') = \bigcup_{v \in V'} col(v). \quad (3)$$

Given a vertex v in a list-colored graph G_l and a color $c \in col(v)$, let $V_{v,c}$ be a subset of V_l which contains the vertex v and all other vertices which have at least one color other than c in their color sets ($V_{v,c}=\{v\} \cup \{u \in V_l \mid col(u)-\{c\} \neq \emptyset\}$). $V_{v,c}$ is defined with the aim of providing us with a set of vertices, which may be used in composing a motif containing the vertex v colored with c. In this paper, we propose a color-based variation of the closeness centrality [19] for each pair of vertex v and color c, to rank vertices and their assigned colors for the purpose of constructing motifs. This centrality measure, which we call color-centrality, is defined by:

$$CC(v, c) = \left(\sum_{c' \in \{col(V_{v,c})-\{c\}\}} \frac{1}{\sum_{\{u \in \{V_{v,c}-\{v\}\} \mid c' \in col(u)\}} dist(u, v)} \right) * \left(|V_{v,c}|-1\right) \quad (4)$$

where $dist(v, u)$ is the length of shortest path between v and u in $G[V_{v,c}]$. The calculation of this centrality measure is illustrated on a small graph, as an example, in Fig. 1.

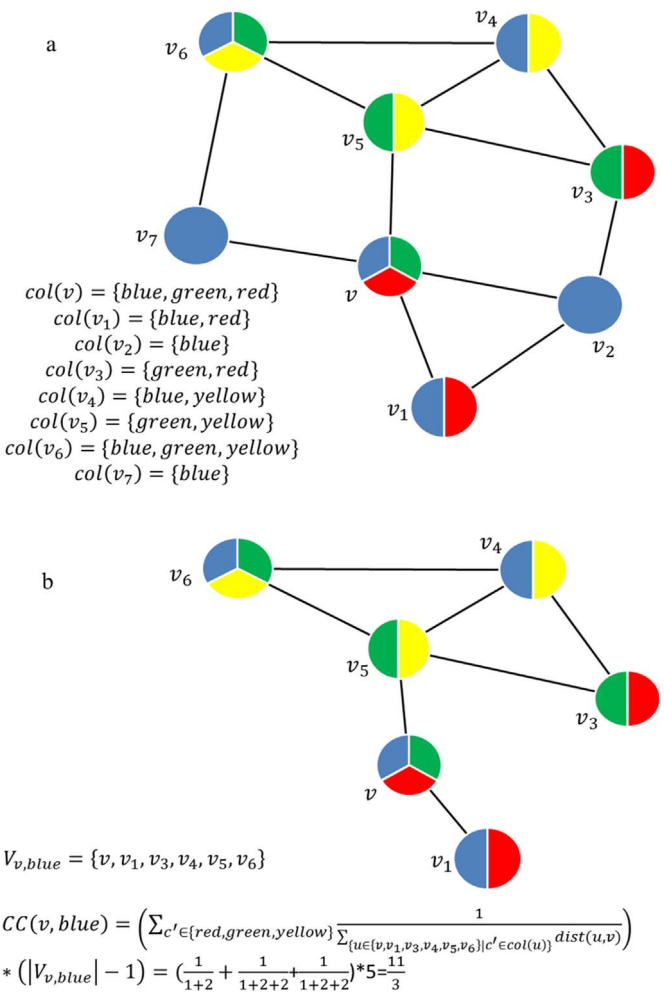

Fig. 1. An example of calculating color-centrality in a sample list-colored graph. (a) A list-colored graph with eight vertices and four colors, (b) the induced subgraph of $V_{v,blue}$, the color-centrality of the vertex v and color *blue* ($CC(v, blue)$) is $\frac{11}{3}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Method

As mentioned earlier, in the problem of functional motif detection, a list-colored graph $G_l=(V_l, E_l)$ (a biological network with some functionalities assigned to each vertex as colors) is given and a set of colors Q is specified as query. The goal is to detect all connected subgraphs $G'=(V', E')$ in G_l (functional motifs), where the color set of G', i.e. $col(V')$, covers the maximum number of colors in Q. The proposed algorithm, CeFunMO, uses a greedy approach to identify such functional motifs vertex by vertex. To this end, the notion of color-centrality measure (defined in section Terminology) is utilized and pairs of vertices and colors in G_l are ranked based on this new measure. Subsequently, vertices with higher color-centrality measures are selected to construct functional motifs. The pseudocode of CeFunMO, as shown in Algorithm 1, consists of two main steps:

1. Ranking pairs of vertices and colors with color-centrality measure.
2. Detecting functional motifs in a greedy approach.