Methods 102 (2016) 57–63

Contents lists available at ScienceDirect

Methods

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

Minimum dominating set-based methods for analyzing biological networks

Jose C. Nacher [a,*], Tatsuya Akutsu [b,*]

[a] Department of Information Science, Faculty of Science, Toho University, Miyama 2-2-1, Funabashi, Chiba 274-8510, Japan
[b] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form 16 November 2015
Accepted 16 December 2015
Available online 7 January 2016

Keywords:
Minimum dominating set
Complex networks
Protein-protein interaction networks
Network controllability

ABSTRACT

The fast increase of 'multi-omics' data does not only pose a computational challenge for its analysis but also requires novel algorithmic methodologies to identify complex biological patterns and decipher the ultimate roots of human disorders. To that end, the massive integration of omics data with disease phenotypes is offering a new window into the cell functionality. The minimum dominating set (MDS) approach has rapidly emerged as a promising algorithmic method to analyze complex biological networks integrated with human disorders, which can be composed of a variety of omics data, from proteomics and transcriptomics to metabolomics. Here we review the main theoretical foundations of the methodology and the key algorithms, and examine the recent applications in which biological systems are analyzed by using the MDS approach.

Contents

1. Introduction

The rapid technological developments of 'multi-omics' based methodologies are providing an increasing amount of data on the fundamental constituents of the cells, such as genes (genomics), RNAs (transcriptomics), proteins (proteomics) and metabolites (metabolomics). Computational and system biologists are, therefore, having a unique opportunity to design novel algorithmic and mathematical-based methods to analyse, identify and extract biological knowledge from the data [1]. Many computational techniques have then been proposed to analyse biological phenomena based on the collected experimental data. Among them, network approaches are becoming relevant because they use knowledge not only from the individual molecules, but also from the complex web of interactions between the components of the cell [2]. This detail is relevant because biological functions and diseases cannot

* Corresponding authors.
E-mail addresses: nacher@is.sci.toho-u.ac.jp (J.C. Nacher), takutsu@kuicr.kyoto-u.ac.jp (T. Akutsu).

http://dx.doi.org/10.1016/j.ymeth.2015.12.017
1046-2023/

be associated with a single molecule. The functional patterns emerge through complex associations between life molecules. Similarly, disease phenotypes are the result of pathobiological processes that occur in a biological pathway, or in a larger scale, in the human interactome, which represents the entire map of all molecular interactions in a cell [3,4]. Network pharmacology is also becoming more relevant since drugs efficacy depends on their interactions with molecules located in a large network [5,6]. As a result, when a right target is selected, the drug may enhance its effects via network propagation but, in the opposite case, it also may lead to unwanted side effects. Strategies to target hubs, to use disease modules to identify new targets as well as to disrupt strategic locations in disease pathways are among the present and future directions using cellular network concepts [3,7,8].

Network biology, therefore, rapidly emerged as a branch of computational biology that focuses on analyzing biological data using a network representation. Several types of interactions (protein-protein, chemical reactions, transcription-regulation) can define specific network levels which can be investigated using algorithmic tools and metrics. Protein networks are defined by proteins that are physically binding to each other. Metabolic networks consist of metabolites and chemical reactions where the latters are catalyzed by enzymes, and regulatory networks are represented by transcriptional factors that regulate genes. Evolutionary models, extraction of biological motifs, disease-gene identification and disease module prediction are some examples of the application of the network techniques across different network levels [3].

In this review, we focus on a novel algorithmic approach that is showing promising results to analyze complex biological networks. The Dominating Set (DS) and minimum dominating set (MDS) concepts emerged from classical graph theory several decades ago. The associated algorithms and their variants have been applied to assist a rich variety of problems from computer and wireless communication networks to social systems [9]. However, the application of MDS to specific complex network patterns such as scale-free networks had not been explored except one for web graphs [10]. This scale-free network structure is important because it seems ubiquitous in many biological systems. From protein-protein interaction networks to metabolic networks, the degree distribution of the network follows a power-law (i.e., the probability that a node with k links follows $P(k) \propto k^{-\gamma}$). Networks with this degree distribution are called scale-free networks. An application of the MDS methodology to network controllability analysis unveiled the conditions necessary to achieve full controllability and showed that a small fraction of nodes can control the entire network [11]. This finding later inspired Wuchty to investigate in more detail controllability in protein interaction networks [12]. Unexpectedly, his findings showed that not only the optimized subset of proteins within the MDS can reach/control any protein of the non-MDS, but also that the proteins in the MDS are enriched with unique biological functions and features, such as cancer-related, and virus targeted-genes. This review assembles the state of the art on the MDS algorithmic approaches that are having an impact on analyzing a wealth of 'omics' systems, from drug-target and protein networks to non-coding RNA interactions.

2. Methods

2.1. Minimum dominating set

The minimum dominating set is a well-known concept in graph theory and computer science [9]. In this paper, we assume that networks are represented as graphs. A graph $G(V,E)$ consists of a set of nodes V and a set of edges E, where a node represents some object and an edge represents a relation, or existence of a relation, between two objects. For example, in a protein-protein interaction (PPI) network, a node corresponds to a protein and an edge corresponds to an interaction between two proteins. Each edge has a direction in directed networks, whereas it does not (i.e., each edge is bi-directional) in undirected networks. Therefore, each edge is represented by a pair of nodes. For example, (u,v) represents an edge between nodes u and v. It is to be noted that (u,v) means an edge directed from u to v in directed networks whereas we do not distinguish (u,v) from (v,u) in undirected networks. Hereafter, networks (resp., graphs) mean undirected networks (resp., graphs) unless otherwise stated.

For a graph $G(V,E)$, a subset of nodes $S \subseteq V$ is called a *dominating set* (DS) if every node in V is either an element of S or is adjacent to an element of S. That is, for any node $v \in V$, $v \in S$ holds or there exists a node $u \in S$ such that $(u,v) \in E$. We say that v is *dominated* by u if $(u,v) \in E$. Then, S is a dominating set if each node in V is either in S or is dominated by some node in S. A dominating set with the minimum number of elements is called a *minimum dominating set* (MDS). As discussed later, MDSs are not necessarily uniquely determined (i.e., there may exist multiple MDSs for a given graph $G(V,E)$).

Fig. 1 illustrates an MDS and a DS, where dark gray circles represent nodes in an MDS or a DS. In Fig. 1, node a is dominated by node b, and node c is dominated by nodes b and e, In this case, the MDS is not uniquely determined: $S = \{a,d,e\}$ is also an MDS.

2.2. Computation of MDS

Although MDS is a very important concept in graph theory, it is known that computation of an MDS is NP-hard, which means that it is not plausible that there exists a theoretically efficient (i.e., polynomial-time) algorithm to exactly compute an MDS. However, NP-hardness does not necessarily mean that there does not exist a practically fast algorithm. Actually, many algorithms have been proposed for exactly computing an MDS. From a theoretical viewpoint, extensive studies have been done on development of $O(\alpha^n)$ time exact algorithms for smaller constants α [13–15], where n is the number of nodes in graph G. To our knowledge, the current best α is 1.4689 [15]. However, no implementation results are included in most of these studies and it seems that these algorithms are not practically useful. From a practical viewpoint, many heuristic methods have also been proposed using such techniques as simulated annealing, genetic algorithms, and ant colony optimization (see [16] and its references). However, these methods are not guaranteed to output exact solutions, or there is no theoretical guarantee on the quality of solutions. Among exact computational methods, the most widely used one would be that based on *integer linear programming* (ILP).

In the ILP-based method, an instance of the MDS problem is transformed into an integer linear program in a simple manner. Although ILP is also an NP-hard problem, there exist practical solvers such as CPLEX [17] and Gurobi [18] that can solve large-scale ILP instances and thus we can utilize them. In this method, we assign a 0–1 variable y_v to each node v in V, where $y_v = 1$ (resp., $y_v = 0$) indicates that v is in an MDS (resp., not in an MDS). From a given graph $G(V,E)$, we construct an integer linear program as follows.

$$\begin{aligned} &\text{minimize} \quad \sum_{v \in V} y_v, \\ &\text{subject to} \quad y_v + \sum_{(u,v) \in E} y_u \geqslant 1 \quad \text{for all } v \in V, \\ &y_v \in \{0,1\} \quad \text{for all } v \in V. \end{aligned} \tag{1}$$

Then, an MDS is given by the set $S = \{v | y_v = 1\}$. The objective function means that the number of nodes with value 1 (i.e., the number