Neurocomputing 206 (2016) 3–12

Contents lists available at ScienceDirect

Neurocomputing

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

Systematic tracking of coordinated differential network motifs identifies novel disease-related genes by integrating multiple data

Kai Shi [a,b], Lin Gao [a,*], Bingbo Wang [a]

[a] School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, 710126, China
[b] College of Science, Guilin University of Technology, Guilin, Guangxi, 541004, China

ARTICLE INFO

Article history:
Received 9 August 2015
Received in revised form
21 November 2015
Accepted 9 December 2015
Available online 2 June 2016

Keywords:
Disease genes
Classification
Network motifs
Subgraph
Data integration

ABSTRACT

Recently, one of the most hotspots in system biology is exploring the disease pathogenesis by integrating different omics data. A lot of methods are developed to identify disease genes for an indepth understanding of a given disease or a biological process. However, most of them do not sufficiently consider the relationship between epigenetic and expressional changes in deregulated genes. Here, we propose a network based approach to identify disease related genes by properly combining the network topological characteristic and the biological characteristic. Our approach identifies network motifs with coordinated changed pattern, differential-methylation and differential-expression, in the context of a human signaling network by integrating DNA methylation and gene expression data. For validation, we do experiments by using colorectal cancer data sets, the results show that the classification performance of our approach outperforms the existing method. The screened network motifs and predicted genes are almost epigenetically deregulated, which are highly associated with colorectal cancer development. Furthermore, functional enrichment analysis reveals that the functions they enriched in are hallmarks of cancer. We not only provide a method for identification of disease related genes but also add a new perspective to integrate heterogeneous data and mine subgraph with significant biological characteristics pattern.

1. Introduction

One of the major challenges in disease research is systems-level deciphering the underlying mechanisms of disease development and progression. Complex diseases are usually recognized as the result of mutation in multiple genes and their interplay rather than the individual gene [1,2]. How to identify them is a hard task for contemporary biology and medicine. The identification of disease-associated genes is the primary step towards the understanding of disease pathogenesis. Over the past decades, advances in high-throughput techniques have resulted in a wealth of data. Large-scale cancer genomic projects, such as the Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and the Catalogue Of Somatic Mutations In Cancer (COSMIC), accelerate the accumulation of data. So we can comprehensively explore the genetic and epigenetic basis of cancer. This has led to the need for developing new approaches or tools to integrate various data [3].

To address the problem of identifying disease genes, tremendous efforts have been made, such as gene set-based methods, network-based methods, and integration-based methods [4,5]. Some of them look for genes that are differentially expressed, are with significant network characteristic or are associated with disease in literature. Chuang et al. [6] presented a protein-network-based approach for identifying markers of metastasis, which significantly improved the marker sets across different data sets. Zhang et al. [7] proposed a network motif-based approach to identify disease genes, which integrated biological network topology and gene expression data not as individual genes but as network motifs. Chen et al. [8] also proposed a method based on network motifs to select classification features to distinguish breast cancer samples from normal samples, which had higher classification performance than mutual information method or the individual gene sets method. In another study, Wen et al. [9] developed a network-based approach to identify putative causal module biomarkers by integrating heterogeneous information of CRC. Functional enrichment analysis validated the identified modules were strongly related to hallmarks of cancer. Kim et al. [10] integrated different biological resources such as the epigenomic, transcriptomic and protein interactome data to identify glioblastoma prognostic biomarkers using gene expression and

* Corresponding author.
E-mail address: lgao@mail.xidian.edu.cn (L. Gao).

http://dx.doi.org/10.1016/j.neucom.2015.12.120
0925-2312/© 2016 Elsevier B.V. All rights reserved.

DNA methylation-based networks. Jiao et al. [11] proposed a novel functional supervised method to identify differential gene expression modules by integrative analysis of DNA methylation and matched or unmatched gene expression data.

Recently, biological networks analysis has been paid more attention in system biology. Especially, network local measure, network motif, has been widely researched [12,13]. The network motif is usually considered as a structural unit which appears more frequently in the real network than in the random network. Motifs widely appear in the real world network, such as food webs, electronic circuits, gene regulatory network, and so on. In system biology, these motifs perform specific functions enabling regulated cellular responses. It is important to screen the significant motifs in the context of disease condition.

In this work, we propose a network based approach to identify disease related genes by properly combining the network topological characteristic and the biological characteristic. It integrates the DNA methylation data, gene expression data and a human signaling network. Researches show that the biological networks represent valuable tools for understanding the complex biological system, since analyzing the properties of the network may generate new biological hypotheses, such as finding coherent gene modules or disease genes [14], screening cancer-related marketing centrality motifs [15]. At the same time, as mentioned in [11], epigenetic mechanisms play an important role in cellular differentiation and disease progression [16,17]. For example, DNA methylation aberration causing disease progression has been mentioned in many researches [18,19], and it directly corresponds to changes in gene expression [19]. Genetic and epigenetic mechanisms together cause the deregulation of gene expression. Motivated by these, we hypothesize that network motifs, in which the genes are topologically significant and coordinated high differential methylation, high differential expression, may have a strong association with disease. Based on this underlying hypothesis, we conduct an approach to screen network motifs with significant network centrality and DNA methylation, gene expression coordinated changed characteristic as biomarkers of disease. In our approach, topological centrality score (*TCscore*) is defined as the mean network centrality value of nodes in a motif; coordinated changed score (*CCscore*) is defined as the mean of differential changes strength of a motif, which presents coordinated changed pattern. The ultimate score (*UTscore*) of network motif is the combination of *TCscore* and *CCscore*. Unlike [8,11,15], our approach not only integrates genetic and epigenetic data, but also considers the topological characteristic of network. It is also important to note that our method differs from Wen et al. [9] as its mathematical model of module identifying is complex and the optimal parameter is also hard to confirm, whereas our method is simple and the parameter choice is more flexible. It is hard to compare to algorithms mentioned above, which are based on different situation. So we just compare the classification performance with Wen et al. [9]. Our method is a more generalized framework to integrate omics data for uncovering new insights into tumor biology.

2. Materials and methods

2.1. Overview of our approach

As a brief introduction to our approach, three-node subgraphs, network motifs, as the basic skeleton are mined in a human signaling network by a motif mining tool. After computing the *TCscore* and *CCscore* for each motifs, the ultimate score-*UTscore* is defined as the combination of the two scores. Finally, we select the top ranked motifs with the higher *UTscore* as the potential markers of cancer. The final selected motifs and corresponding genes are used separately as classification features to distinguish cancer samples from normal samples. The following are the main steps. For additional details please refer to the workflow of our approach shown in Fig. 1.

Step 1. An integrated common signaling network is obtained by extracting common genes from the DNA methylation matrix, the gene expression matrix and the dealt human signaling network adjacent matrix.
Step 2. Mining three-node motifs in the common signaling network.
Step 3. Computing *TCscore* and *CCscore* of the significant motifs and defining *UTscore* of a motif as a linear superposition of *TCscore* and *CCscore*.
Step 4. Coherent significantly different motifs are screened by ranking the *UTscore*.
Step 5. Finally, evaluating classifying performance of coherent significant motifs and corresponding candidate genes.

2.2. Resources and datasets

To validate the effectiveness of our approach, we focus on colorectal cancer (CRC) data sets. The Gene Expression Omnibus (GEO) database is one of the largest public functional genomics data repositories, which provide tools to query and download experiments and curated gene expression profiles [20]. We download gene expression data and DNA methylation data of CRC GSE25070 and GSE25062 [21], respectively from GEO. A total of 26 pairs gene expression profiling of colorectal tumors and adjacent no-tumor tissue samples from GSE25070, and 29 pairs DNA methylation profiling of colorectal tumors and adjacent no-tumor tissue samples from GSE25062 are available. The probe-level data is preprocessed, log 2 transformed and normalized the same as described in the original paper [21]. At the same time, we average the intensity of probes which corresponded to the same gene. And gene expression data of CRC GSE24514, and GSE8671 from GEO are as independent validation datasets.

In addition, we download the human signaling network (version 6) from a previous research [22], which includes BioCarta, CST Signaling pathways, Pathway Interaction database (PID), iHOP and many review papers on cell signaling. And it contains more than 6000 proteins, 63,000 relations, which refer to three types of interactions—activation, inhibition and physical interaction.

2.3. Constructing the integrated common signaling network

The integration of the DNA methylation data, the gene expression data and the human signaling network is the first step of our approach. These data representing three different profiles of the interested cancer have different gene sets. So it is necessary to get the uniform gene sets. We concern the overlapping gene set by intersecting genes of the DNA methylation data and the gene expression data, and then map those genes to the human signaling network for extracting the maximally connected component as the integrated common signaling network.

2.4. Mining network motifs

Network motifs are often known as small, repeated biological units within molecular networks, such as signaling networks, transcriptional regulatory networks and metabolic networks. At the same time, we treat them as “building blocks” of networks. Exploring the network motifs, we can learn more characteristics of function [12,13]. There are several algorithms and tools that have been developed for efficiently detecting network motifs [23].