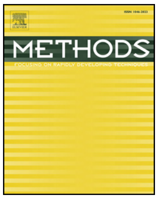
Contents lists available at ScienceDirect

Methods

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



In silico experiment system for testing hypothesis on gene functions using three condition specific biological networks

Chai-Jin Lee[a], Dongwon Kang[b], Sangseon Lee[b], Sunwon Lee[c], Jaewoo Kang[c], Sun Kim[a,b,d,*]

[a] *Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea*
[b] *Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea*
[c] *Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea*
[d] *Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea*

ARTICLE INFO

Keywords:
Gene regulatory networks
Network analysis
Computational biology

ABSTRACT

Determining functions of a gene requires time consuming, expensive biological experiments. Scientists can speed up this experimental process if the literature information and biological networks can be adequately provided. In this paper, we present a web-based information system that can perform *in silico* experiments of computationally testing hypothesis on the function of a gene. A hypothesis that is specified in English by the user is converted to genes using a literature and knowledge mining system called BEST. Condition-specific TF, miRNA and PPI (protein-protein interaction) networks are automatically generated by projecting gene and miRNA expression data to template networks. Then, an *in silico* experiment is to test how well the target genes are connected from the knockout gene through the condition-specific networks. The test result visualizes path from the knockout gene to the target genes in the three networks. Statistical and information-theoretic scores are provided on the resulting web page to help scientists either accept or reject the hypothesis being tested.

Our web-based system was extensively tested using three data sets, such as E2f1, Lrrk2, and Dicer1 knockout data sets. We were able to re-produce gene functions reported in the original research papers. In addition, we comprehensively tested with all disease names in MalaCards as hypothesis to show the effectiveness of our system. Our *in silico* experiment system can be very useful in suggesting biological mechanisms which can be further tested *in vivo* or *in vitro*.

Availability: http://biohealth.snu.ac.kr/software/insilico/.

1. Introduction

Important regulators such as TF (transcription factor) genes have system-wide effects on many genes, often resulting in significant changes in phenotypes [1]. To understand the role of TF, it is a common practice to use model organisms, e.g., mouse with the TF knocked out. Subsequently, sequencing technologies are used to measure changes in gene expression levels at the whole cell level. A common practice for the analysis of transcriptome data is to perform the DEG (Differentially Expressed Gene) analysis to measure the system-wide effects of a TF. However, the DEG analysis has several limitations. First, there are too many DEGs, up to several thousands, depending on the criteria for beings DEGs. More importantly, the DEG analysis do not explain how a TF affects DEGs since connections from the TF to DEGs are unknown. In addition, users do not have any control on the DEG analysis process except changing the cut-off values, even when the user has a good hypothesis on which biological mechanisms or pathways are likely to be affected by knocking out the TF.

Recently, there have been significant advances in bioinformatics technologies. Among them, a number of biological networks have been constructed using experimental data and/or computational methods. Thus, it is possible to use networks to investigate the system-wide effects of a TF by following edges of networks from the TF to all other genes. In addition, literature mining technologies have been advanced significantly and they were used in the recent research projects [2–4]. These literature mining technologies are now powerful enough to identify relationship between specific hypothesis of the user, e.g., disease names or certain biological pathways, and genes that are reported to be relevant to the hypothesis in the literature. By leveraging these recent advances, we developed a novel information system that can be used to perform *in silico* experiments for testing on functions of a TF.

[*] Corresponding author at: Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea.
E-mail address: sunkim.bioinfo@snu.ac.kr (S. Kim).

https://doi.org/10.1016/j.ymeth.2018.05.003
Received 7 February 2018; Received in revised form 30 April 2018; Accepted 3 May 2018

2. Methods

An *in silico* experiment is performed as follows. Given a user provided transcriptome and miRNA data from a knockout mouse experiment, the user can specify his/her hypothesis in English. The current system may not handle free-style sentences, thus a set of nouns are to be specified as input. Then, the hypothesis is translated to a set of genes using our literature knowledge mining system, BEST [5]. We call these genes *target genes*. Then, connections from the knockout gene to the target genes are constructed and evaluated by condition-specific networks that are instantiated by gene expression data. Three condition-specific networks are TF, miRNA and PPI (protein-protein interaction) networks. The connectivity between the regulator gene such as the knocked out TF and the target genes is determined by computing shortest paths. Intuitively, more target genes are reachable from the TF, an *in silico* experiment accepts or supports the user hypothesis while fewer connections would reject the hypothesis. Of course, our literature based experiment is not meant to use to determine the function of a TF since biological experiments should be performed to confirm the functions. However, our system allows the user to exploit his/her expert knowledge to explore potential functions of a TF, which, we expect, will reduce the burden of scientists significantly so that much smaller number of *in vivo* or *in vitro* experiments can confirm the function of a TF.

2.1. User input

- Transcriptome (mRNA expression) data
- microRNA expression data
- Regulator gene name (ex. knockout gene name)
- Hypothesis (ex. disease, pathway or gene) specified by a set of nouns

(Our tool supports raw data of microarray, pre-processed data of microarray and pre-processed data of RNA-seq. We do not support raw data of RNA-seq.)

2.2. Output result

- DEG analysis result
- Candidate target genes related with regulator gene and hypothesis
- Network of regulator gene and target gene within TF, microRNA and PPI network
- Statistical and informational test results

3. Workflow

The workflow of our system is shown in Fig. 1. Each step and the workflow is explained in detail in this section. To help understand readers the workflow, we will define genes in three categories. A *regulator gene* is a TF gene that is knocked out in the biological experiment. *Mediator* or *network genes* are genes in the condition specific TF network, miRNA network and PPI network. *Target genes* are genes that are relevant to the hypothesis or pathway that the user specified. These genes are called as targets since our *in silico* experiment is to test how well a regulator gene is connected to the target genes via network genes.

3.1. Step 1. Select target genes from hypothesis

3.1.1. 1–1: DEG analysis of miRNA and mRNA

The input mRNA and miRNA expression data are analyzed using `limma` [6] for microarray and `DEseq2` [7] for RNA-sequencing data. DEGs are selected based on the log2 fold change value and the p-value are calculated for the expression level of each gene.

3.1.2. 1–2: Search target genes related with the hypothesis from the literature

The user needs to provide, as input, the regulator gene name and hypothesis that is specified in English. The validity of an input gene name can be checked by clicking the check button on the web page. Instead of specifying a hypothesis, the user can select a KEGG [8] pathway from the list of KEGG pathways that are provided at the web page. With a user specified hypothesis, the BEST system converts the hypothesis to a set of genes. When the user selects a pathway name, genes in the pathway is selected as candidate target genes by searching the KEGG pathway database.

3.1.3. 1–3: Select top DEGs as target genes

Using the DEG analysis (Step 1–1) results, top 10 target DEGs in terms of gene expression level changes are selected from the candidate target genes determined by BEST tool or KEGG pathway DB (Step 1–2) with p-value. The top 10 target DEGs consist of 5 up-regulated DEGs and 5 down-regulated DEGs.

3.2. Step 2. Condition-specific TF, miRNA and PPI network generation by the DEG set

3.2.1. 2–1. Extract all edges of selected 10 target DEGs in the three network DB

Template TF network, miRNA network, and PPI network are instantiated by the gene expression information from the microarray or RNA-seq experiments. However, these networks are too big to be displayed on a web page. Thus, genes that are directly or indirectly connected to the 10 target genes are chosen since genes that do not have connection to the target genes are not relevant. A miRNA network database was obtained from TargetScan [9], and STRING [10] was used as a PPI network. A mouse TF network database was created using NARROMI [11].

3.2.2. 2–2. Select the edges of DEGs

Since we are interested in how the regulator gene affected the target genes, gene expression levels should be different in the expression value in the control vs. treated experiment. Thus, only edges incident to DEGs are selected and the others are removed from the network.

3.3. Step 3. Performing in silico experiment

3.3.1. 3–1: Compute shortest path between the regulator gene and target genes in the networks

Our system computes shortest path in the networks between the user provided regulator gene and 10 target genes from the hypothesis that is being tested.

3.3.2. 3–2: Visualize networks in graph

The networks computed in the previous steps are visualized using `Cytoscape` [12]. The regulator gene is located at the top position, and the target genes are located at the bottom, and network genes are grouped into TF, miRNA, and PPI networks. Edges in the networks show how each genes are connected according to changes in expression level. The expression level of each gene was visualized in color according to the amount of change. Up-regulated genes are in red and down-regulated genes are in blue. In addition, a coding gene is denoted by a circle and a non-coding gene is denoted by a diamond shape. The blue colored edge represents the miRNA network, the purple colored edge represents the PPI network, and the yellow colored edge represents the TF network. Clicking a specific gene shows a list of connected genes.

3.4. Step 4. Evaluation of the user hypothesis

3.4.1. 4–1: A statistical evaluation by random permutation

To evaluate network connections between the regulator gene and