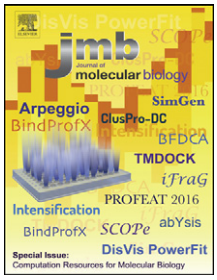

BFDCA: A Comprehensive Tool of Using Bayes Factor for Differential Co-Expression Analysis

Duolin Wang[1,2], **Juexin Wang**[2], **Yuexu Jiang**[1,2], **Yanchun Liang**[1,2] and **Dong Xu**[1,2]

*1 - **College of Computer Science and Technology,** Jilin University, Changchun, China 130012*
*2 - **Department of Computer Science and Christopher S. Bond Life Sciences Center,** University of Missouri, Columbia, MO 65211, USA*

***Correspondence to Dong Xu:** Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. xudong@missouri.edu*
http://dx.doi.org/10.1016/j.jmb.2016.10.030
Edited by Michael Sternberg

Abstract

Comparing the gene-expression profiles between biological conditions is useful for understanding gene regulation underlying complex phenotypes. Along this line, analysis of differential co-expression (DC) has gained attention in the recent years, where genes under one condition have different co-expression patterns compared with another. We developed an R package Bayes Factor approach for Differential Co-expression Analysis (BFDCA) for DC analysis. BFDCA is unique in integrating various aspects of DC patterns (including Shift, Cross, and Re-wiring) into one uniform Bayes factor. We tested BFDCA using simulation data and experimental data. Simulation results indicate that BFDCA outperforms existing methods in accuracy and robustness of detecting DC pairs and DC modules. Results of using experimental data suggest that BFDCA can cluster disease-related genes into functional DC subunits and estimate the regulatory impact of disease-related genes well. BFDCA also achieves high accuracy in predicting case-control phenotypes by using significant DC gene pairs as markers. BFDCA is publicly available at http://dx.doi.org/10.17632/jdz4vtvnm3.1.

Introduction

To infer a set of genes associated with a particular condition or phenotype, comparing gene-expression profiles between different biological conditions or phenotypes is a common approach [1,2]. The genes that are found with significant differential expression (DE) are often useful for understanding the gene regulation underlying the complex phenotypes and may serve as biomarker candidates. For the last two decades, the analyses of gene-expression data have been dominated by univariate-based methods, which analyze expression data under the assumption that genes are independent. However, such methods may overlook the interactions and coordination among genes in performing their functions [3]. In particular, the small changes in multiple differentially regulated genes may not be captured by single gene test (e.g., *t*-test or ANOVA). This is especially true for transcription factors, which are often stably expressed at baseline levels [4] but have major causal regulatory effects in biological processes. For example, myostatin (GDF8) was not a DE gene between Piedmontese and Wagyu cattle, but it is a differential hub gene in a DE network [5]. Another example is the "switching mechanism", where a gene can be an activator or a suppressor depending on another gene's activities under different conditions. A well-known case is the Max gene, a transcription factor, which activates transcription when it heterodimerizes with Myc and represses transcription when it heterodimerizes with Mad [6].

Due to the facts above, a promising approach of differential co-expression (DC) analysis emerged in the recent years. It seeks gene modules that are co-expressed under one condition but have different co-expression patterns in another condition. In addition to investigating gene-expression changes from single genes independently, DC takes gene–gene interactions into consideration and examines the changes in gene correlations between different conditions. Thus, DC can reveal the regulatory impact of a

gene on other genes via identifying condition-specific patterns in the context of regulatory relationships. There are various scenarios of DC presented between two genes, and three extreme cases are summarized [3,7] as follows: (1) Shift, the correlation of genes does not change between two conditions, but expression values under one condition are clearly different from those under the other (Fig. 1a); (2) Cross, genes are positively correlated under one condition but negatively correlated under the other (Fig. 1b); (3) Re-wiring, genes are positively (or negatively) correlated under one condition while not correlated (or less correlated) under the other (Fig. 1c). More complex changes in co-expression patterns often occur, which, from a mathematical perspective, can be represented as different distributions of gene-expression profiles between two potentially co-expressed genes under different conditions.

Several computational approaches have been developed for DC analysis. They can be divided into two categories: targeted and untargeted approaches. In targeted approaches, for example, Gene Set Co-expression Analysis [8] and Gene Sets Net Correlations Analysis [9], predefined gene sets are required to test the correlation changes between conditions. In untargeted approaches, no predefined gene sets are required and gene modules are detected on the basis of their DC profiles. Unlike targeted approaches, untargeted approaches are capable of detecting novel modules. A group of untargeted approaches aimed at detecting DC pairs, such as EBcoexpress, an empirical Bayesian approach for identifying DC gene pairs [10], and ROS-DET, a robust detector of switching mechanisms [11]. Another group of untargeted approaches aimed at detecting DC modules, such as DiffCoEx [12], which uses the Weighted Gene Coexpression Network Analysis

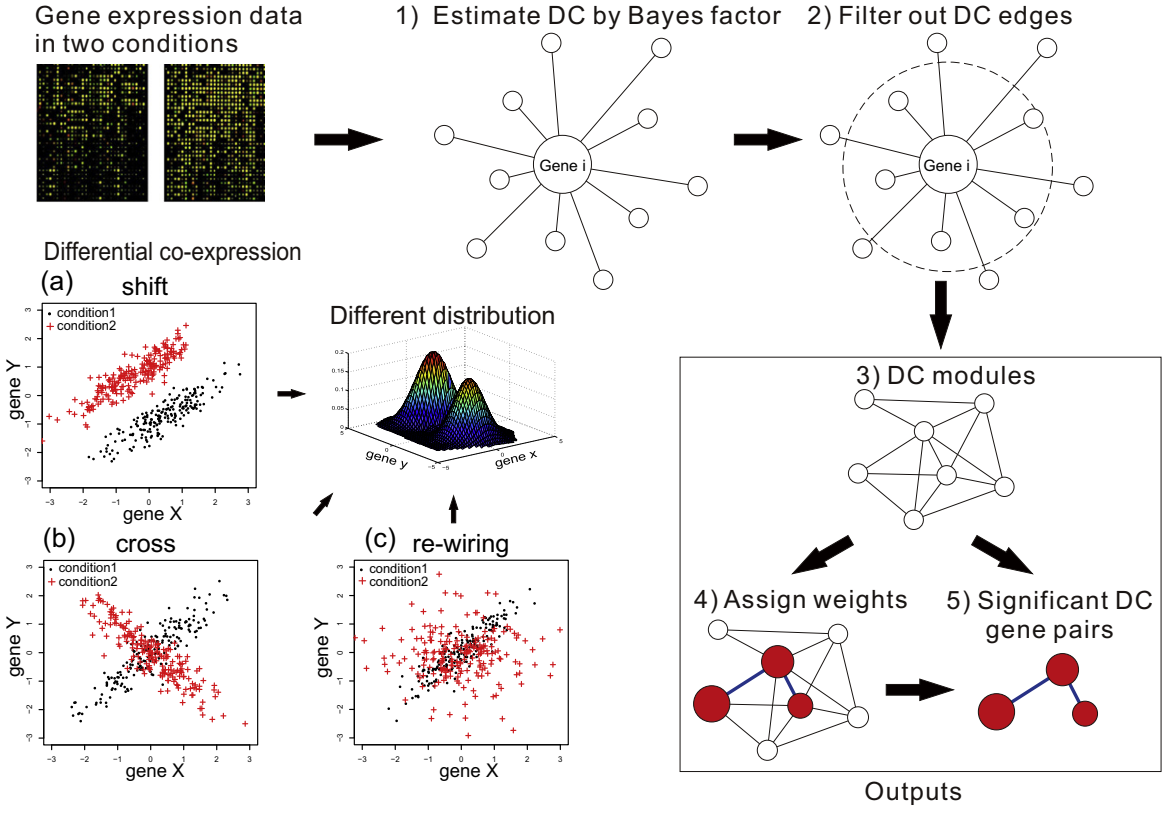

Fig. 1. Overview of the algorithm used by BFDCA. The input (left) is a set of gene-expression profiles from two classes of samples. The algorithm performs the following: (1) estimating the DC of gene pairs by calculating Bayes factors, (2) filtering DC edges whose Bayes factor values are too small; (3) inferring DC modules using the remaining edges, (4) assigning each DC with a weight to reflect its regulatory importance, (5) selecting significant DC gene pairs according to their end-node importance and edge values. (a–c) represent shift, cross, and re-wiring scenarios of differentially co-expressed gene pair. *x*-axis and *y*-axis represent the gene-expression levels of gene X and gene Y, respectively. Dots in different colors and shapes represent samples under two different conditions. Taken together, these scenarios can be interpreted as different distributions of gene-expression profiles between two potentially co-expressed genes under different conditions.