Biomedical Data Mining

Extended Abstract

Descriptive and predictive analysis of gene co-expression networks

Harun Pirim

Abstract

Ample availability of gene co-expression data challenges researchers to find and apply unique approaches for extracting biological information to infer about gene functions or predict gene-disease relations. By means of reliable co-expression network construction techniques compiled networks exist and they require further predictive analysis to focus on the genes or groups of the genes exhibiting certain patterns reflected on the co-expression networks. We propose an integrated network analysis where social network descriptive analysis techniques are borrowed to summarize some structural features of the network. Then, the features are employed in optimization models to find groups of genes exhibiting certain patterns.

A key objective in biological research is to systematically identify all molecules within a living cell and how they interact. However, the functions of many genes are still not understood, a situation that has only become more complex with the recent identification of many novel non-coding genes. With the development of high-throughput technologies including microarrays and RNA sequencing (RNA-seq), and their respective data-analysis methods, the functional status of a gene can now be identified from a systematic perspective. One method to infer gene function and gene–disease associations from genome-wide gene expression is co-expression network analysis, an approach that constructs networks of genes with a tendency to co-activate across a group of samples and subsequently interrogates and analyses this network.

Gene co-expression networks can be used for various purposes, including candidate disease gene prioritization, functional gene annotation and the identification of regulatory genes. However, co-expression networks are effectively only able to identify correlations; they indicate which genes are active simultaneously, which often indicates they are active in the same biological processes, but do not normally confer information about causality or distinguish between regulatory and regulated genes. An increasingly used method that goes beyond traditional co-expression networks is differential co-expression analysis. This approach identifies genes with varying co-expression partners under different conditions, such as disease states, tissue types and developmental stages, because these genes are more likely to be regulators that underlie phenotypic differences. The regulatory roles of such genes can be further investigated by integrating data types such as protein–protein interactions, methylome data, interactions between transcription factors (TFs) and their targets, and with sequence motif analysis of co-expressed genes. This aids in the identification of regulatory elements such as TFs, expression quantitative trait loci (eQTLs) and methylation patterns that affect the expression and composition of co-expression modules.

Gene expression and regulation can be highly tissue-specific, and most disease-related genes have tissue-specific expression abnormalities. The increased availability of expression data for multiple tissues has allowed for differential co-expression analysis, which can identify both tissue-specific signatures and shared co-expression signatures. These tissue-specific signatures can be disrupted in tissue-specific diseases and would not be detected in analyses aggregating multiple tissues. Even when no sample classification is available, subpopulation-specific modules can be resolved, an approach that has been particularly successful in classifying different cancer subtypes to provide prognostic markers. Differential co-expression analysis is also useful for analysing data sets in which the subpopulations are unknown, e.g. large-scale single-cell RNA-seq data. While differential co-expression methods are sensitive to noise, they are becoming more effective with the increase in RNA-seq data quantity and quality.

Harun Pirim
King Fahd University of Petroleum and Minerals, Saudi Arabia, E-mail: harunpirim@kfupm.edu.sa

RNA-seq further permits co-expression analysis to focus on splice variants and non-coding RNAs.

In this review, we provide an introduction and overview of what constitutes a co-expression network, followed by a guide of the different steps in co-expression analysis using RNA-seq data. We then describe commonly used and newly emerging methods and tools for co-expression analysis, with a focus on differential co-expression analysis to identify regulatory genes that underlie disease. We conclude with a discussion of the integration of co-expression networks with other types of data, to e.g. infer regulatory processes, and with future prospects and remaining challenges in the field.

This work is partly presented at 24th World Chemistry & Systems Biology Conference on October 03-04, 2018 in Los Angeles, USA