Biochimica et Biophysica Acta xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Review

Unraveling gene function in agricultural species using gene co-expression networks☆

Robert J. Schaefer [a], Jean-Michel Michno [a], Chad L. Myers [a,b,*]

[a] Biomedical Informatics and Computational Biology Graduate Program, University of Minnesota, Minneapolis, MN 55455, United States
[b] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Received 4 May 2016
Received in revised form 23 July 2016
Accepted 25 July 2016
Available online xxxx

Keywords:
Co-expression
Functional genomics
Agriculture
Crops
Network analysis
Guilt-by-association

ABSTRACT

Co-expression networks have been shown to be a powerful tool for inferring a gene's function when little is known about it. With the advent of next generation sequencing technologies, the construction and analysis of co-expression networks is now possible in non-model species, including those with agricultural importance. Here, we review fundamental concepts in the construction and application of co-expression networks with a focus on agricultural crops. We survey past and current applications of co-expression network analysis in several agricultural species and provide perspective on important considerations that arise when analyzing network relationships. We conclude with a perspective on future directions and potential challenges of utilizing this powerful approach in crops. This article is part of a Special Issue entitled: Plant Gene Regulatory Mechanisms and Networks, edited by Dr. Erich Grotewold and Dr. Nathan Springer.

1. Introduction

Despite extremely successful breeding programs substantially impacting agricultural yield, our understanding of gene function in most crop species is still limited. For instance, in maize, only ~1% of genes have functional annotations based on mutant analysis [1]. However, applications of high throughput, genome sequencing technologies are generating data that is beginning to expose the function of the rest of the genome. For example, standardized sequencing-based techniques, such as RNA-Seq, enable the measurement of gene expression for specific experimental conditions, developmental time points or different tissues. Surveying transcription across a large number of diverse experiments establishes an expression profile for each gene, which can be exhaustively compared to each another, building a network of putative co-regulatory relationships. Each *node* represents a gene and each *edge* shows the magnitude of co-expression between them, implying a probabilistic, functional relationship [2]. In addition to when and where it is expressed, these data can help establish a functional context for a gene, even where little other information exists.

In model species, where gene function can be established through reverse genetic approaches, co-expression networks have been shown to be a powerful tool for rapidly predicting potential functional links between genes. Furthermore, once represented as a network, topological and structural information shows that these biological networks share organization properties similar to other naturally occurring networks such as those seen in power grids, social interactions, and the world wide web [3]. Borrowing these systems biology based approaches developed in model species, systematic integration of large-scale whole genome expression datasets is now an active area of research in crops. Whole transcriptome sequencing technologies in crop species allow for a transfer of knowledge from decades of previous research in non-biological and model systems to build functional networks using co-expression. This shared domain knowledge allows for the direct application of many network based approaches in crop species allowing for rapid construction of robust, biologically coherent networks. However, there are unique characteristics of agricultural species that require special consideration in the application and interpretation of co-expression network approaches including high levels of nucleotide diversity, prevalent genotype by environment (GxE) interactions, and heterosis [4]. While not all agronomically important traits will be fully explained by variation in gene expression, the wealth of currently available gene expression information already available for many species coupled with the rate at which new expression studies are being performed makes co-expression analysis a powerful tool for unraveling gene function in crop species. The impact of these phenomena on interpreting functional relationships are just beginning to be explored in agricultural species and pose important considerations for applying co-expression based techniques.

☆ This article is part of a Special Issue entitled: Plant Gene Regulatory Mechanisms and Networks, edited by Dr. Erich Grotewold and Dr. Nathan Springer.

* Corresponding author at: Biomedical Informatics and Computational Biology Graduate Program, University of Minnesota, Minneapolis, MN 55455, United States.
E-mail address: cmyers@cs.umn.edu (C.L. Myers).

http://dx.doi.org/10.1016/j.bbagrm.2016.07.016
1874-9399/

In this review, we examine how gene co-expression networks have been used to unravel gene function, with a focus on agricultural species. First, we review generalized methods and techniques used to build and utilize co-expression networks. We then survey different strategies that previously have been used to examine gene function inferred from co-expression in several different crop species. Next, we focus on interpreting and understanding network structure that arises from co-expression networks built specifically in agricultural species. Finally, we provide our perspective on emerging applications of co-expression networks in helping to interpret the vast wealth of functional information that is now being generated for agricultural systems.

2. Building co-expression networks

2.1. Measuring co-expression among pairs of genes

Early methods in gene expression analysis discovered that when genes were hierarchically clustered, genes with similar biological function generally clustered near each other (Fig. 1A). A gene's pattern of expression across different samples is informative of its function [5]. The functional information originally observed in hierarchically clustered gene expression profiles can also be analyzed through the lens of pairwise relationships. Where the basic unit in a gene expression dataset is a gene's expression *profile* across a diverse set of experiments, the unit of interest in a gene co-expression network arises from quantifying the relationship among two genes' profiles. A typical approach is to systematically compare each pair of genes' profiles, representing similarity of two genes' profile with a single edge in a network, where the nodes represent genes and the edges reflect highly similar profiles (Fig. 1B).

The relationship between two genes' expression profiles can be measured several different ways [6]. The most widely used similarity metric used to calculate co-expression is the Pearson correlation coefficient (PCC), which measures the presence of a linear relationship between a pair of expression profiles. This similarity metric has an intuitive interpretation and is computationally inexpensive to compute, which may be factors that contribute to its widespread use for co-expression analyses[2,7–15]. Despite its extensive use, the Pearson correlation coefficient is not optimal for capturing nonlinear relationships, which possibly leaves meaningful relationships uncovered. Other similarity metrics have been proposed. Non-parametric measures such as the Spearman and Kendall correlation coefficient are alternatives to the PCC and can offer the advantage of robustness to outliers [16–18]. Additionally, more complex measures such as mutual Information (MI) attempt to quantify similarity more generally as a summary of statistical dependence, although this comes at the cost of computational efficiency [19]. Recent comparative studies on similarity metrics shows that in most cases, linear monotonic relationships capture instances where co-expression is informative, suggesting that similarity metrics designed for linear relationships such as PCC are suitable for most general purpose network analysis [20]. Highly similar co-expression patterns among sets of genes are often driven by a small subset of individuals or experiments that are collectively over- or under-expressed (Fig. 1A). While determining this subset of accessions is not explicitly handled by common co-expression measures, simultaneous clustering of accession profiles in an approach called bi-clustering can help determine which accessions are potentially 'driving' patterns of co-expression [21,22]. Additional reviews on uses and applications of similarity metrics can be found here [6,20,23,24]. For the remainder of this review, we focus on the Pearson correlation coefficient as the basis for measuring co-expression due to its widespread use. However, most of the concepts discussed apply more generally to networks derived from other metrics.

In addition to how co-expression is quantified between input genes, different approaches can be used to define the starting source when examining co-expression relationships among genes. The simplest

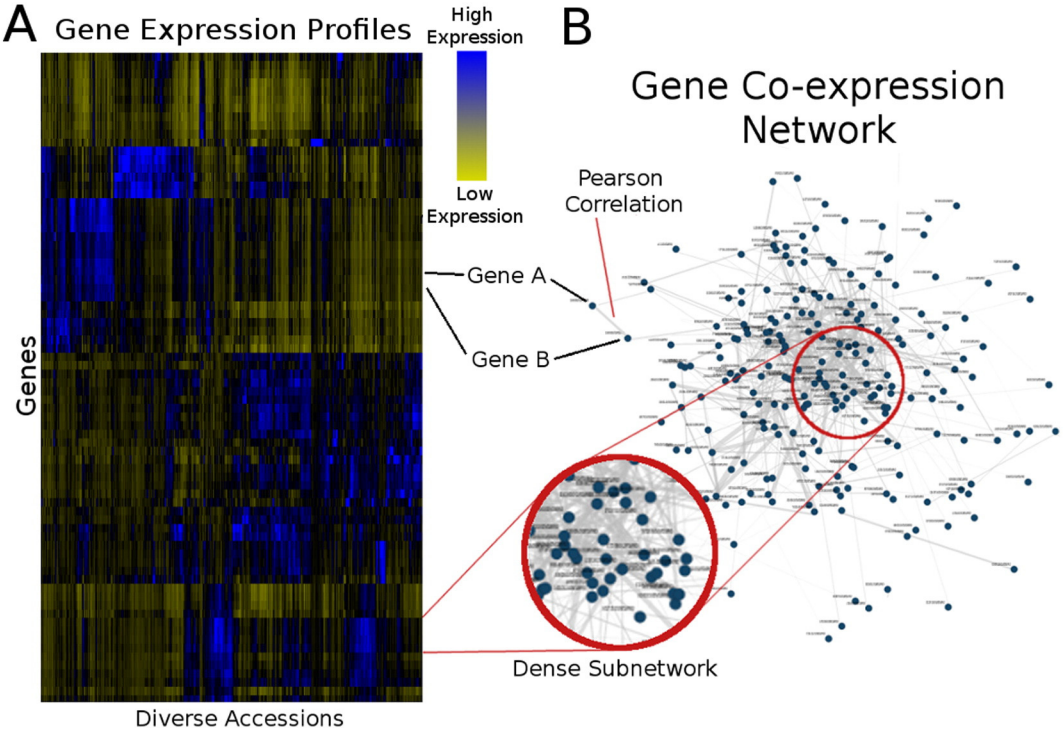

Fig. 1. Gene co-expression network construction and overview schematic. (A) Expression is measured for each gene within a genome to generate an expression profile across a diverse set of accessions (often tissues, developmental time-points, stress responses or genotypic variation). Genes are ordered, here, based on a hierarchical clustering algorithm to demonstrate the overall structure captured by the data. (B) Correlative structure can be quantified using a similarity metric (Pearson correlation coefficient here) applied to each pair of genes (Section 2.1). Interactions are normalized and thresholded to produce a network and displayed using a 'node' and 'edge' model (Section 2.2.1). Subnetwork structure, such as dense connectivity, relates back to the original correlation among gene expression profiles captured by variation among the diverse set of accessions. Biological interpretation then relies on understanding why profiles of genes are co-expressed (Section 2.2). Patterns here are potentially driven by a subset of accessions being over- or under-expressed, collectively.

Please cite this article as: R.J. Schaefer, et al., Unraveling gene function in agricultural species using gene co-expression networks, Biochim. Biophys. Acta (2016), http://dx.doi.org/10.1016/j.bbagrm.2016.07.016