Biochimica et Biophysica Acta 1860 (2017) 41–52

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

# Computational inference of gene regulatory networks: Approaches, limitations and opportunities☆

Michael Banf*, Seung Y. Rhee*

*Department of Plant Biology, Carnegie Institution for Science, 260 Panama Street, Stanford 93405, United States*

ARTICLE INFO

*Article history:*
Received 13 April 2016
Received in revised form 8 September 2016
Accepted 8 September 2016
Available online 16 September 2016

*Keywords:*
Computational systems biology
Gene regulatory network inference
Machine learning
Heterogeneous data integration
Gene network evaluation

ABSTRACT

Gene regulatory networks lie at the core of cell function control. In *E. coli* and *S. cerevisiae*, the study of gene regulatory networks has led to the discovery of regulatory mechanisms responsible for the control of cell growth, differentiation and responses to environmental stimuli. In plants, computational rendering of gene regulatory networks is gaining momentum, thanks to the recent availability of high-quality genomes and transcriptomes and development of computational network inference approaches.

Here, we review current techniques, challenges and trends in gene regulatory network inference and highlight challenges and opportunities for plant science. We provide plant-specific application examples to guide researchers in selecting methodologies that suit their particular research questions.

Given the interdisciplinary nature of gene regulatory network inference, we tried to cater to both biologists and computer scientists to help them engage in a dialogue about concepts and caveats in network inference. Specifically, we discuss problems and opportunities in heterogeneous data integration for eukaryotic organisms and common caveats to be considered during network model evaluation. 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

Gene regulatory networks are central to all biological processes of an organism. In their most basic form, they describe the complex web of transcription factor proteins that bind regulatory sequences of target genes in order to affect their spatial and temporal expression [1].

Understanding gene expression regulation has an immediate impact in biology and medicine as many traits and diseases are associated with mutations in regulatory sequences or dysfunctional transcriptional regulators [1]. In agriculture, changes in plant transcriptional regulation shaped modern crops such as maize, rice and wheat and enabled yield increases [2]. Elucidation of transcriptional regulatory systems could help improve metabolite production rates and resilience against environmental stresses [3,4].

Owing to their sessile nature, plants are subject to variations in their environment that drive adaptation. Yet, how gene-regulatory networks are rewired to control the adapted traits or drive adaptation is largely unknown. Despite extensive insights into the core components of the transcriptional machinery, how specificity is encoded during the highly dynamic process of transcriptional regulation remains an open question [5].

Given the interdisciplinary nature of the problem, biologists and computer scientists need to engage in a dialogue to solve the network inference problem.

In this review, we aim to (i) introduce the basic concepts and procedures used for gene expression based regulatory network inference, which are borrowed from several disciplines including statistics, information theory, graph theory and machine learning; (ii) discuss limitations of network inference based only on transcriptional datasets; (iii) present data types and strategies used for integrative network inference and causal link predictions; and (iv) describe caveats and solutions for the evaluation and selection of statistically and biologically relevant regulatory interactions.

## 2. Methods for gene regulatory network inference from expression data

Gene expression data generated by high-throughput technologies such as microarray or RNAseq still serve as one of the main sources for the development of gene regulatory networks. Therefore, we

☆ 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 authors.
*E-mail addresses:* michael.banf@gmx.net (M. Banf), srhee@carnegiescience.edu (S. Rhee).

http://dx.doi.org/10.1016/j.bbagrm.2016.09.003
1874-9399/© 2016 Elsevier B.V. All rights reserved.

start our review by highlighting some of the main concepts, methods and limitations in inferring regulatory networks from different types of gene expression data. For an in-depth comparison of the state-of-the-art network inference methods using expression data, we refer the reader to [6–9]. For an overview on gene expression normalization, filtering and pre-processing steps for network inference, see [10,11].

### 2.1. Correlation and information theoretic approaches

Approaches within this category employ statistical analyses of dependencies between expression patterns. The most basic models, called co-expression networks, exploit correlations between expression profiles of genes [12]. A popular correlation measure is Pearson's correlation coefficient $r$ [10], i.e.

$$r_{E_i,E_j} = \frac{cov(E_i, E_j)}{\sigma(E_i)\cdot\sigma(E_j)} \tag{1}$$

Here, $E_i, E_j$ denote gene expression profiles of two genes $i, j$ with covariance $cov(E_i, E_j)$ and standard deviations $\sigma(.)$. Other correlation measures include Spearman's correlation [10] or the more recently introduced weighted correlation coefficient [13]. More sophisticated correlation based approaches aim to distinguish direct from indirect, spurious correlations between genes by using partial correlations [14–16].

Information theoretic concepts [17,18] extend correlation to capture more complex statistical dependencies between expression patterns. This approach led to the development of a specific kind of association network called a relevance network [19]. Relevance networks define relationships between two genes $i, j$ based on an information theoretic property, called mutual information [19], based on their respective gene expression profiles $E_i$ and $E_j$. Mutual information is defined as

$$I(E_i, E_j) = \sum_{e_i \in E_i} \sum_{e_j \in E_j} p(e_i, e_j) log\left(\frac{p(e_i, e_j)}{p(e_i)p(e_j)}\right) \tag{2}$$

where $p(e_i, e_j)$ is the joint probability distribution of $e_i$ in $E_i$ and $e_j$ in $E_j$, and $p(e_i)$ and $p(e_j)$ denote the marginal probabilities.

Relevance networks are built by first constructing a fully connected graph for all gene pairs using mutual information to weight each link. Links whose associated weights lie below a certain threshold are removed from the network. The threshold, according to Butte and Kohane [19], is estimated by first randomizing the expression data, and then re-computing mutual information values to obtain a reference null distribution.

Various refinements of this idea have been proposed to discriminate direct from indirect effects [12,17,18,20,21]. The most prominent methods include ARACNE (algorithm for the reconstruction of accurate cellular networks) [22], CLR (context likelihood of relatedness) [21], MRNET (minimum redundancy, maximum relevance) [20] and C3NET (conservative causal core) [17].

The ARACNE algorithm [22] adjusts the construction of a relevance network by applying a constraint known as Data Processing Inequality to filter indirect interactions. The Data Processing Inequality states that, if gene $i$ interacts with gene $j$ via gene $k$, then the following inequality holds with respect to their corresponding mutual information values: $I(E_i, E_j) \leq min(I(E_i, E_k), I(E_k, E_j))$, i.e. the smallest of the mutual information scores $I(\cdot)$ within this inequality indicates an indirect regulatory interaction [22]. As a consequence, ARACNE evaluates all possible gene triplets and prunes individual (indirect) interactions within each triplet, if the Data Processing Inequality is violated beyond a certain tolerance threshold.

The CLR algorithm [21] first estimates the pair-wise mutual information values for all gene pairs. Then, it estimates the statistical likelihood of each mutual information value $I_{ij}$ for a particular pair of genes $(i,j)$ by comparing this mutual information value to a background distribution. For each gene pair $(i,j)$, two z-scores are obtained, one for gene $i$ and one for gene $j$, by comparing the mutual information value $I_{ij}$ with gene-specific distributions, $p_i$ and $p_j$. Here lies CLR's major advantage over the relevance network approach by Butte and Kohan [19] or ARACNE, as individual thresholds can be established by considering an individual background for each pair of genes. This is in contrast to relevance networks or the ARANCE approach, which use a global threshold for graph pruning.

The MRNET algorithm [20] incorporates a feature selection methodology, called Minimum Redundancy Maximum Relevance (MRMR), to infer interactions between genes. This algorithm first places each gene as a target gene with all remaining genes as its putative regulators. The MRMR method is then applied to select the best subset of regulators.

More recently, C3NET [17] has been proposed. This approach consists of two main steps. First, a relevance network is constructed and non-significant edges are pruned according to a chosen significance level. Then, only the most significant link for each gene, i.e. the highest mutual information value among the neighboring edges, is selected. This implies that the highest possible number of edges that can be inferred by C3NET is equal to the number of genes in the network.

In addition to aforementioned variations of the relevance network approach, concepts such as three-way and conditional mutual information have been proposed to directly address the problem of indirect interactions within gene triplets. For instance, the MI3 algorithm [23] uses three-way mutual information for inference, hypothesizing that gene regulation commonly involves more than one regulatory gene. Soranzo et al. [12] use only the conditional mutual information (CMI) to infer regulatory networks between gene triplets, pruning links that fall below a chosen threshold.

#### 2.1.1. Limitations

In general, co-expression and relevance networks are designed to help explore co-functionality of genes on a systems level [13]. In this context, the notion of separating direct from indirect regulatory effects between genes should not be confused with directionality or even causality. In general, correlation-based relationships are symmetric, i.e. bidirectional, although recent efforts to exploit time-delay effects using time-course expression data might allow for asymmetric relationships [24–26]. For an in-depth comparison of time-lagged information-theoretic approaches, see [25].

However, for most of the common approaches within this group of methodologies, directionalities between genes can only be assumed if regulator genes are known in advance. As a consequence, large-scale mining for biologically relevant graph-patterns, such as feedback or feed-forward loops, should be exercised with caution.

In general, correlations between gene expression profiles are used as an indicator of co-regulation [27], and a plethora of clustering approaches has been developed based on that premise [28–32]. Complementing clustering with various gene regulatory network inference approaches have been proposed to find a more robust method for identifying condition-specific regulators [31,33]. Advantages and limitations of these ideas are discussed in [32]

#### 2.1.2. Application examples in plant science

An in-depth tutorial on how to apply co-expression analysis for plant systems biology is given in [10]. The authors in [10] not only describe methodologies but also discuss statistical issues including how normalization of gene expression data can influence co-expression results. A more recent extensive overview on the construction and application of co-expression networks in plant species