Methods xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

ARMADA: Using motif activity dynamics to infer gene regulatory networks from gene expression data

Peter J. Pemberton-Ross, Mikhail Pachkov, Erik van Nimwegen *

Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form 22 June 2015
Accepted 23 June 2015
Available online xxxx

Keywords:
Transcription regulation
Gene regulatory network
Transcription factor
Network inference
Auto-regressive models
Regulatory motif
Motif activity

ABSTRACT

Analysis of gene expression data remains one of the most promising avenues toward reconstructing genome-wide gene regulatory networks. However, the large dimensionality of the problem prohibits the fitting of explicit dynamical models of gene regulatory networks, whereas machine learning methods for dimensionality reduction such as clustering or principal component analysis typically fail to provide mechanistic interpretations of the reduced descriptions. To address this, we recently developed a general methodology called motif activity response analysis (MARA) that, by modeling gene expression patterns in terms of the activities of concrete regulators, accomplishes dramatic dimensionality reduction while retaining mechanistic biological interpretations of its predictions (Balwierz, 2014).

Here we extend MARA by presenting ARMADA, which models the activity dynamics of regulators across a time course, and infers the causal interactions between the regulators that drive the dynamics of their activities across time. We have implemented ARMADA as part of our ISMARA webserver, ismara.unibas.ch, allowing any researcher to automatically apply it to any gene expression time course. To illustrate the method, we apply ARMADA to a time course of human umbilical vein endothelial cells treated with TNF. Remarkably, ARMADA is able to reproduce the complex observed motif activity dynamics using a relatively small set of interactions between the key regulators in this system. In addition, we show that ARMADA successfully infers many of the key regulatory interactions known to drive this inflammatory response and discuss several novel interactions that ARMADA predicts. In combination with ISMARA, ARMADA provides a powerful approach to generating plausible hypotheses for the key interactions between regulators that control gene expression in any system for which time course measurements are available.

1. Introduction

Understanding the structure, dynamics, and functioning of the genome-wide regulatory networks that control gene expression is one of the central challenges in systems biology. Gene regulatory networks allow individual cells to respond and adapt to changes in their environments, and allow multi-cellular eukaryotes to express a single underlying genotype, shared by all their cells, into a large variety of phenotypically and functionally distinct cell types. More than half a century has passed since the discovery of the basic biophysical mechanism underlying gene regulation [2], and during this time much has been learned about the molecular players involved in gene regulation, and the specific mechanisms through which they act. Very roughly speaking, in the cells of multi-cellular eukaryotes there are hundreds of regulatory proteins and RNAs expressed that bind in a sequence-specific manner to short sequence motifs within the DNA and RNA. The binding constellations of these regulatory proteins determine the rates at which genes are being transcribed, the stability of mRNAs, and the rates at which they are being translated.

We not only understand the basic molecular mechanisms, in well-studied model organisms most of the molecular players are known as well, i.e. comprehensive lists of transcription factors TFs [3] and regulatory RNAs such as miRNAs [4] are available, and for many of these there is also information about their targets and their functioning. However, knowing the molecular players and understanding the molecular mechanisms involved does not mean that we understand how gene regulatory networks function *as systems*. For example, how the actions of regulatory genes are coordinated to maintain and stabilize cell identity is not understood. Similarly, although it has recently become clear that, given an appropriate perturbation in the expression of regulatory proteins, cells can be driven from one cell type to another [5], what

* Corresponding author at: Biozentrum, University of Basel, Basel, Switzerland.
E-mail addresses: peter.pemberton-ross@unibas.ch (P.J. Pemberton-Ross), mikhail.pachkov@unibas.ch (M. Pachkov), erik.vannimwegen@unibas.ch (E. van Nimwegen).

http://dx.doi.org/10.1016/j.ymeth.2015.06.024
1046-2023/

perturbations would be needed to transdifferentiate cells from a particular cell type to a given desired target type is not understood. How to recognize the breakdown in control of gene expression, which may be associated with particular disease states, is another example of a systems-level question to which we currently have little insight.

To appreciate the magnitude of the challenge we face in answering such questions, it helps to recognize just how fragmentary our knowledge of genome-wide gene regulatory interactions still is in higher eukaryotes. For example, of the roughly 1500 TFs present in mammalian genomes [6], binding specificities are known for less than half, e.g. [7]. The ability of TFs to bind to their cognate sites depends on the local state of the chromatin which can be modified in a large number of ways, i.e. through chemical modification of the histone tails within nucleosomes. These epigenetic marks are both 'read' and 'written' by chromatin modifying enzymes which in turn may be recruited to specific loci by TFs bound to the DNA. This potentially complex feedback between chromatin state and TF binding is currently poorly understood. TFs may interact through direct protein–protein contacts with each other and with a large number of co-factors, and our knowledge of these interactions is very incomplete. Although regulation of transcription initiation is of crucial importance for the control of gene expression, expression is also regulated at the level of transcript processing (splicing, poly-adenylation), mRNA transport, transcript stability, translation initiation and elongation, and protein degradation. Although some aspects of this post-transcriptional regulation have been investigated in some detail, e.g. the role of micro-RNAs in regulating transcript stability and translation, by-and-large our knowledge of this post-transcriptional regulation is extremely limited. In addition, the 'activity' of regulatory factors is not only determined by their mRNA and protein expression level, but also by post-translational modification (e.g. phosphorylation at specific residues), by their localization within the cell, by their interaction with co-factors, and so on. In other words, although our knowledge of the individual players and interactions in gene regulatory networks has been steadily increasing, the things we *do not know* still outnumber the things we know by many-fold. Given this, it is clear that we are still very far removed from being able to meaningfully simulate detailed models of the functioning of gene regulatory networks. There is little point in taking all the information we happen to know, and pouring them into a mathematical model or computational simulation, without realistically dealing with the fact that there is much more we do not know.

1.1. Using gene expression data to infer regulatory networks

Instead of expecting to establish a detailed model of the functioning of the genome-wide gene regulatory network, much research focuses on more modest goals, such as identifying the key regulators operating in a particular model system. Since there are at least hundreds of potential regulators, it is generally unfeasible to experimentally investigate the role of all potential regulators. However, with the advent of high-throughput technologies such as next-generation sequencing, it has become relatively easy to measure gene expression and chromatin state genome-wide. Over the last decade, many researchers have thus turned to such methodologies with the aim of identifying the key regulatory interactions acting within their specific model systems.

From the point of view of computational methods, the question has thus become of how we can most efficiently use high-throughput data, such as genome-wide gene expression data, to learn about the key regulatory interactions acting in a given system. Indeed, a large number of methods for performing inference of regulatory networks from gene expression data has been proposed over the years, ranging from mostly descriptive methods that aim to summarize the structure of these high-dimensional datasets in terms of a relatively small number of statistical features, to highly specific methods that fit the data in terms of concrete models of the genome-wide gene expression dynamics, e.g. using coupled differential equations, Gaussian models, or Bayesian network models, see e.g. [8–10] for reviews.

On one end of this scale, methods that aim to fit the data using specific models of the underlying gene regulatory network generally suffer from the 'curse of dimensionality'. Put simply, because the number of possible regulatory network architectures is huge, the amount of data that would be necessary to reliably infer the regulatory network is many orders of magnitude larger than even the largest high-throughput datasets can provide. To uniquely predict a regulatory network from the data, these methods employ regularization schemes that aim to minimize either the total number of regulatory interactions, their magnitudes, or a combination of both. However, it is unclear to what extent we should expect such 'minimal' networks to match the true underlying biological network. Moreover, in order for the network inference to be computationally feasible, these methods are often forced to treat all genes as equivalent, thereby ignoring all kinds of relevant prior biological information. For example, many of such methods simply investigate the correlation or mutual information between all pairs of genes, and consider possible regulatory interactions between any pair of genes, even though prior biological knowledge indicates that most genes do not act as regulators.

On the other end of the scale, many methods focus simply on reducing the dimensionality of the data by identifying statistical descriptors that capture the main features of the data. These include well-known methods such a principal component analysis (PCA), which finds linear combinations of the variables (e.g. genes and conditions) which carry most of the variance in the data, as well as various clustering methods that divide the genes and or samples into a relatively small number of subsets that show similar expression profiles. Although such methods are very valuable in clarifying the structure of the data, it is generally difficult to relate the structures that they identify to underlying biological mechanisms. For example, when a particular subset of genes is predicted to form a 'co-regulated module', it is generally unclear what follow-up experiments could be done to further characterize or even validate this prediction.

1.2. Motif activity response analysis

In our view, the challenge facing methods for gene regulatory network reconstruction consist in reducing the dimensionality of the problem, so that models can be meaningfully fitted to the data, on the one hand, while at the same time incorporating relevant prior biological information, and formulating the models in terms of concrete biological mechanisms that are amenable to direct experimental follow-up, on the other hand. A few years ago we proposed an approach to regulatory network inference, called motif activity response analysis (MARA), which combines these desirable features [11]. First, recognizing that much of genome-wide mRNA expression levels are controlled by transcriptional and post-transcriptional regulators, MARA models gene expression levels explicitly in terms of the *activities* of TFs and miRNAs. To do this, MARA makes use of the fact that, both for miRNAs and for many TFs, targets genes can be computationally predicted based on DNA and RNA sequence analysis. That is, MARA first computationally predicts, for each of hundreds of TFs and miRNAs, which transcripts are regulated by each of these regulators. MARA then uses a very simple linear model to relate the observed expression levels of all transcripts in terms of the activities of the regulators. In this way, the very high-dimensional gene expression data, i.e. involving expression levels of tens of

Please cite this article in press as: P.J. Pemberton-Ross et al., Methods (2015), http://dx.doi.org/10.1016/j.ymeth.2015.06.024