Mathematical Biosciences 246 (2013) 326–334

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues

George Michailidis [a], Florence d'Alché-Buc [b,c,*]

[a] Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA
[b] INRIA-Saclay, AMIB, TAO, LRI umr CNRS 8326, Orsay, France
[c] IBISC, EA 4526, Université d'Evry, Evry, France

ARTICLE INFO

Article history:
Available online 28 October 2013

Keywords:
Gene regulatory network inference
Autoregressive models
Sparsity
Stability
Causality
Dynamic Bayesian networks

ABSTRACT

Reconstructing gene regulatory networks from high-throughput measurements represents a key problem in functional genomics. It also represents a canonical learning problem and thus has attracted a lot of attention in both the informatics and the statistical learning literature. Numerous approaches have been proposed, ranging from simple clustering to rather involved dynamic Bayesian network modeling, as well as hybrid ones that combine a number of modeling steps, such as employing ordinary differential equations coupled with genome annotation. These approaches are tailored to the type of data being employed. Available data sources include static steady state data and time course data obtained either for wild type phenotypes or from perturbation experiments.

This review focuses on the class of autoregressive models using time course data for inferring gene regulatory networks. The central themes of sparsity, stability and causality are discussed as well as the ability to integrate prior knowledge for successful use of these models for the learning task at hand.

1. Introduction

A number of technological advances, such as DNA microarrays, RNA-Seq [1], liquid chromatography tandem mass spectrometry [2], and similarly liquid or gaseous chromatography mass spectrometry [3], have enabled biomedical researchers to collect large amounts of transcriptomic, proteomic and metabolomic data. In addition, curated repositories containing both vast amounts of such data, as well as functional information, ontologies, gene and protein interactions, pathways, etc. are expanding at a fast pace (e.g. KEGG, IntegromeDB, BioGrid, GEO, NURSA, etc.).

The increasing availability of such high dimensional data and structured information have led to a number of novel learning problems, including that of *network inference*. Networks have become a key tool in computational biology due to their ability to capture at an appropriate level of abstraction biological processes. Overall, the study of biological networks including modeling, analysis, reconstruction and visualization aspects has become a key topic in bioinformatics and computational biology (for a review and recent trends see [4]).

A number of learning tasks have been studied in the literature, based on the type of biological network under consideration. For example, in metabolic reaction networks, the focus has been on learning enzyme kinetic parameters [5], stoichiometric analysis, as well as finding the operative modes of such networks subject to catalytic activity and steady state operational constraints. In protein interaction networks, predictions of interactions are based both on protein descriptors and labeled edges [6]. Information obtained from protein–protein interaction networks has proved useful in protein function prediction and in learning protein complexes [7], while predicting cellular responses using ontology information has been a key task involving signaling networks. In this review study, we focus on the problem of reconstructing (inferring) the structure of gene regulatory networks (GRN). Such networks involve interactions between DNA, RNA, proteins and other biomolecules, whose edges represent functional influences of one molecule on the other, rather than chemical interactions.

This learning task has become a central one in functional genomics, as the growing literature on the subject attests [8–10]. Two main types of data have been used to learn such networks: steady state data and time course data. steady state data are obtained from a long-term observation of gene expression, assuming the system reaches an equilibrium state. For instance, multiple biological replicates obtained at some late point in time provide such steady state data. Such data are usually obtained from microarray technologies, and provide a global view of the biological system under study in its natural state (wild type); however, their informational content for network reconstruction purposes is in general limited and accurate network inference usually requires a very

* Corresponding author at: IBISC, EA4526, 23, Bd de France, Université d'Evry et Genopole, 91037 cedex Evry, France. Tel.: +33 164853164.
E-mail address: florence.dalche@ibisc.fr (F. d'Alché-Buc).

0025-5564/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.10.003

large number of replicates [11]. On the other hand, time course data even for wild type measurements provide insights on the transitory behavior of the biological system which is induced by regulations, especially if the system is observed under different initial conditions due to perturbations, as discussed next.

A particularly informative source for the learning task at hand is data from perturbation experiments, involving specific gene knock-outs/downs or silencing. They may correspond to a single time observation point, selected so that the perturbation has manifested itself in the system, or could take the form of time series, as discussed above. The advantage of time course data obtained from perturbation experiments is that they contain significant information about the dynamics of the system and are shown to be a key component for network inference in the DREAM7 challenge on experimental design for parameter estimation in network models (more information regarding the DREAM challenge competition is provided in Section 9). The downside of perturbation data is that they are usually obtained from single gene knock-outs (downs). Hence, every replicate (time series or single time point observation) offers limited information about the overall system, especially when joint regulations are involved. Moreover, large scale perturbation experiments for most organisms are not readily available, due to technical complexities and cost considerations.

On the other hand, wild type time course data are still attractive for inferring relatively large scale GRNs, since they contain adequate information about regulatory interactions and are significantly less expensive to acquire compared to perturbation data.

For inferring GRNs, the majority of approaches in the literature belong to the class of *unsupervised methods*, although there has been work that assumes partial knowledge of the network which is either integrated as prior information in the model employed, or used in a *supervised approach* [12,13]. The class of unsupervised approaches can be divided in the following two categories: (i) *model-based* ones that aim to capture the dynamical behavior of the GRN by estimating the parameters of a chosen model [14–17,8,9,18–20], and (ii) *model-free* approaches that extract dependencies among state variables using information-theoretic criteria in the spirit of ARACNE [15,21,22].

This review primarily focuses on inferring GRNs from time course data and model-based approaches. Our goal is to emphasize the key elements that are common in the best off-the-shelf network inference algorithms and to outline the set of important features that such algorithms should possess to meet future challenges. A key feature is that of *sparsity*, due to the following facts. First, statistical analysis of known regulatory networks has shown that scale-free models are suitable to represent the topological structure of the network, thus reflecting their sparse nature. Second, most available data sets contain relatively few time points compared to the number of genes measured, thus making the use of sparse models obligatory. Another key element in network inference (and in learning complex structures in general) is that of *stability* of the algorithm. The concept of stability has been central for model selection in regularized regression [23] or as a construction principle in various randomized models, including bagging and random forests. Recent works explore the use of this concept in GRN inference [24,25]. Taking another angle, the ability to integrate prior knowledge into a model or in a learning method represents a valuable property in a field where partial knowledge coming from different sources may be available [26]. Finally a key question regarding network inference is the semantics associated with a direct edge in a regulation graph. Directed edges under certain conditions reflect causal relationships [27]. Even though estimating such relationships is known to be a very challenging task, *causality* nevertheless represents a central issue in network inference.

The remainder of the paper is organized as follows. Section 2 presents the problem of gene regulatory network inference from time course data and emphasizes desirable properties of a network structure inferred by a learning method. Section 3 gives an overview of the main Markov models used for network inference from time course data. In Sections 4 and 5, different works about Markov models and their associated network inference methods are reviewed and when it is possible, analyzed through the concepts of sparsity and causality. Section 4 focuses on linear autoregressive models for which sparse regression has been largely developed and from which Granger-causal networks can be inferred. Extensions of linear autoregressive models described in Section 5 consider generalized additive models and kernel-based methods. Section 6 gives a brief presentation of dynamic Bayesian networks that support, as a special case of autoregressive models, specific learning strategies. In Section 7, we highlight the notion of stability and describe how it has been recently used for model selection and to improve upon a base model. Section 8 addresses prior integration in the whole set of reviewed models, while Section 9 provides an overview of the performance of various methods in the DREAM computational challenges. Finally, Section 10 discusses recent trends and future challenges.

2. Gene regulatory network inference from time course data

In model-based approaches to network inference, a GRN is abstracted and considered as a dynamical system whose states correspond to different mRNA concentrations. The network structure is defined as a directed graph $\mathcal{G}$ whose nodes are associated to genes and whose directed edges represent the presence or the absence of regulations[1] from one regulating gene to a target gene. In the paper, $|\mathcal{G}| = p$ denotes the number of genes and A, a binary matrix of size $p \times p$, is the adjacency matrix of graph $\mathcal{G}$.

Assuming that we observe gene expression levels for wild type, we denote by $\mathbf{x}_T$ the p-dimensional vector of the gene expression levels measured at time T. Gene regulatory network inference consists in providing an estimate of A denoted by $\widehat{A}$, given the time course $\mathcal{S}_n = \{\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}\}$ of length n measured at equidistant time points $t_0, \ldots, t_{n-1}$, with $t_i = t_{i-1} + \tau,\ i = 1, \ldots, n-1$. In the case the time points are not regularly spaced, which happens rather frequently in biological experiments, the observations are smoothed by a nonparametric regression which is re-sampled subsequently. The sampling rate τ is in this case an additional hyperparameter of any discrete-time modeling. This estimation task is by definition unsupervised unless partial knowledge about the graph is available. The main part of the paper is devoted to the case when no edges are known. However, integration of prior knowledge will be discussed in detail in Section 8.

Model-based approaches mainly proceed in two steps: first, given a model of the dynamical system $\mathcal{M}$, they estimate its parameters from observed time course and second, they extract from its parameters an estimate $\widehat{A}$ of the target matrix A. In some cases, like in Dynamical Bayesian Learning, the network structure is included in the parameter set and the second step is straightforward.

2.1. Desirable properties for the estimated network

Let us discuss the properties for the network structure estimated from model $\mathcal{M}$ and time course data $\mathcal{S}_n$. Beyond structure consistency, which will not be discussed here per se, other properties related to what biologists expect from the automated inference process can be targeted. They include network sparsity and stability of the algorithm employed. Next, we provide a high level

[1] For sake of simplicity, we will only consider transcriptional regulations.