Chapter 3
Bayesian Networks for Modeling and Inferring Gene Regulatory Networks

Sebastian Bauer
Charité Universitätsmedizin Berlin, Germany

Peter Robinson
Charité Universitätsmedizin Berlin, Germany

ABSTRACT

Bayesian networks have become a commonly used tool for inferring structure of gene regulatory networks from gene expression data. In this framework, genes are mapped to nodes of a graph, and Bayesian techniques are used to determine a set of edges that best explain the data, that is, to infer the underlying structure of the network. This chapter begins with an explanation of the mathematical framework of Bayesian networks in the context of reverse engineering of genetic networks. The second part of this review discusses a number of variations upon the basic methodology, including analysis of discrete vs. continuous data or static vs. dynamic Bayesian networks, different methods of exploring the potentially huge search space of network structures, and the use of priors to improve the prediction performance. This review concludes with a discussion of methods for evaluating the performance of network structure inference algorithms.

INTRODUCTION

A multiplicity of mathematical tools has been developed to represent gene regulatory networks (GRNs) with different levels of detail. In the setting of network structure inference from microarray data, *Bayesian networks* (BNs) represent a commonly used tool to describe the network in a comparatively high level manner, in contrast, say, to ordinal differential equations. The purpose of this chapter is to provide necessary background knowledge of BNs.

The structure of this chapter is as follows: in the first section we provide a brief introduction into the biology of GRNs and the mathematical concepts on which the Bayesian networks are based. In

DOI: 10.4018/978-1-60566-685-3.ch003

the next section we present the theory of Bayesian networks and show how they can be adapted to *model* GRNs. We learn how we can use the model to *infer* or predict activity states of genes in terms of probability theory, which in general has been one of the classic uses of BNs. Yet, the probably most prominent application of Bayesian networks in computational biology has been for *reverse engineering* of gene regulatory networks, especially since the advent of high-throughput screening methods such as gene-expression microarrays. This is covered in the fourth section, in which we also discuss the issue of variable time lags in time-series data whereby the response time of one gene regulated by another varies greatly among the genes. We finish the chapter with conclusions, and provide directions which might be of interest for future research.

BACKGROUND

Biology

GRNs coordinate the changes in cellular behavior associated with development or response of the cell or organism to extracellular stimuli. Transcription factors are the molecules that activate or repress downstream genes by binding to promoter and other sequences (cis-regulatory modules) of genes, thereby modulating the rate of transcription of genes. Combinations of transcription factor binding events in any one promoter are one of the important factors determining the level of the corresponding mRNA in the cell. The regulatory state of the cell has been described as the total set of active transcription factors. However, a number of other molecules influence the activation state and concentration of transcription factors. For instance, signaling pathways consisting of ten or hundreds of proteins can transduce an extracellular event (such as the binding of a ligand to a receptor) into an intracellular biochemical signal by cascading protein modification events. For instance, a receptor-ligand binding event may induce phosphorylation (and activation) of an intracellular signaling molecule, which in turn phosphorylates other molecules, thereby propagating the signal through a cascade or network of proteins, some of which activate transcription factors and thereby influence the transcription of target genes. Other factors, such as non-coding RNAs, histone modifications, and CpG methylation, can also influence the level of mRNA of target genes. Therefore, measurement of mRNA levels can provide only a partial view of the regulatory state of a cell. At present, however, there remain major technical difficulties in obtaining large-scale measurements of protein levels or protein modifications, so that network structural inference has for the most part been attempted with mRNA data.

Graph Theory

Graphs are abstract entities of discrete mathematics which are used to encode relationships of interest between objects of the same domain. Formally, a *graph* is a pair $G=(V,E)$, in which V is finite set of *vertices*, representing the objects, and E a set of pairs of distinct elements of V, which is a binary relation over V. Elements of E are called *edges* (or arcs). The pairs may be ordered or not. An order implies a *direction*. If all edges of G are directed, the graph is *directed*. If at least one edge is directed we call the graph a *partially directed graph*. Otherwise the graph is an *undirected graph*.

A *path* with length n is a sequence of vertices $(v_1,...,v_n)$ which respects the edges, i.e., $(v_i, v_{i+1}) \in E$ for all i. A *cycle* is a special path whose start vertex v_1 equals to the end vertex v_n. A *directed path* is a