Chapter 7.31
Bayesian Network Approach to Estimate Gene Networks

Seiya Imoto
University of Tokyo, Japan

Satoru Miyano
University of Tokyo, Japan

ABSTRACT

In cells, genes interact with each other and this system can be viewed as directed graphs. A gene network is a graphical representation of transcriptional relations between genes and the problem of estimation of gene networks from genome-wide data, such as DNA microarray gene expression data, is one of the important issues in bioinformatics and systems biology. Here, we present a statistical method based on Bayesian networks to estimate gene networks from microarray data and other biological data. Because microarray data are measured as continuous variables and the relationship between genes are usually nonlinear, we combine Bayesian networks and nonparametric regression to handle continuous variables and nonlinear relations. Most parts of gene networks are still unknown, and we need to estimate them from observational data. This problem is equivalent to the structural learning of Bayesian networks, and we solve it from a Bayes approach. The main difficulty of gene network estimation is due to the number of genes involved in the network. Therefore, it leads to model overfitting to the observational data like microarray data. Hence, a combination of various kinds of biological data is a key technique to estimate accurate gene networks. We show a general framework to combine microarray data and other biological information to estimate gene networks.

INTRODUCTION

The microarray technology has produced a huge amount of gene expression data under various conditions such as gene knock-down, overexpression, experimental stressors, transformation, exposure to a chemical compound, and so forth. Along with this new data production, there have been considerable attempts to infer gene net-

works from such gene expression profile data, and several computational methods have been proposed together with gene network models such as Boolean networks, differential equation models, and Bayesian networks.

A Bayesian network is an effective method in modeling phenomena through the joint distribution of a large number of random variables. In recent years, some interesting works have been established in constructing gene networks from microarray gene expression data by using Bayesian networks. Friedman and Goldszmidt (1998) discretized the expression values and assumed multinomial distributions as the candidate statistical models. Pe'er, Regev, Elidan, and Friedman (2001) investigated the threshold value for discretizing. On the other hand, Friedman, Linial, Nachman, and Pe'er (2000) pointed out that the discretizing probably loses information of the data. In addition, the number of discretizing values and the thresholds are unknown parameters, which have to be estimated from the data. The resulted network strongly depends on their values. Then Friedman et al. (2000) considered fitting linear regression models, which analyze the data in the continuous variables (see also Heckerman & Geiger, 1995). However, the assumption that the parent genes depend linearly on the objective gene is not always guaranteed. Imoto et al. (2002) proposed the use of nonparametric additive regression models (see also Green & Silvermn, 1994; Hastie & Tibshirani, 1990) for capturing not only linear dependencies but also nonlinear structures between genes. In this chapter, we introduce a method for constructing the gene network by using Bayesian networks and the nonparametric regression, which is more suitable for estimating gene networks from microarray gene expression data than discrete type Bayesian networks.

Once we set the graph, we have to evaluate its goodness or closeness to the true graph, which is usually unknown. Hence, the construction of a suitable criterion becomes the center of attention of statistical genetic network modeling. Friedman and Goldszmidt (1998), used the BDe criterion, which was originally derived by Cooper and Herskovits (1992) for choosing a graph (see also Heckerman, Geiger, & Chickering, 1995). The BDe criterion only evaluates the Bayesian network model based on the multinomial distributions and Dirichlet priors. However, Friedman and Goldszmidt (1998) kept the unknown hyperparameters in Dirichlet priors and we only set up the values experimentally. We investigate the graph selection problem as a statistical model selection or evaluation problem and theoretically derive a new criterion for choosing a graph using the Bayes approach (see Berger, 1985). The proposed criterion automatically optimizes all parameters in the model and gives the optimal graph when we can score all candidate graphs.

The problem of finding an optimal Bayesian network is known to be NP-hard. The brute force method employing all computing resources in the world would even require time exceeding the lifetime of the solar system for finding an optimal Bayesian network of 30 genes from 100 microarray datasets. Our approach has made it possible to find optimal and near optimal Bayesian networks with respect to the score of the network in a reasonable time and has provided an evidence of the biological rationality in this computational approach (Ott & Miyano, 2003; Ott, Imoto, & Miyano, 2004; Ott, Hansen, Kim, & Miyano, 2005). For larger networks, we carefully employ the greedy hill-climbing algorithm for finding better gene networks.

The main drawback for the gene network construction from microarray data is that, while the gene network contains a large number of genes, the information contained in gene expression data is limited by the number of microarrays, their quality, the experimental design, noise, and measurement errors. Therefore, estimated gene networks contain some incorrect gene regulations, which cannot be evaluated from a biological viewpoint. In particular, it is difficult to determine the direction of gene regulation us-