Chapter 12
Structural Learning of Genetic Regulatory Networks Based on Prior Biological Knowledge and Microarray Gene Expression Measurements

Yang Dai
University of Illinois at Chicago, USA

Eyad Almasri
University of Illinois at Chicago, USA

Peter Larsen
University of Illinois at Chicago, USA

Guanrao Chen
University of Illinois at Chicago, USA

ABSTRACT

The reconstruction of genetic regulatory networks from microarray gene expression measurements has been a challenging problem in bioinformatics. Various methods have been proposed for this problem including the Bayesian Network (BN) approach. In this chapter, we provide a comprehensive survey of the current development of using structure priors derived from high-throughput experimental results such as protein-protein interactions, transcription factor binding location data, evolutionary relationships, and literature database in learning regulatory networks.

DOI: 10.4018/978-1-60566-685-3.ch012

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

The Bayesian Network (BN) has been proven to be a useful and important tool in biomedical applications such as clinical decision support systems (Beinlich, Suermondt, Chavez & Cooper, 1989), information retrieval (Baeza-Yates & Ribeiro, 1999), and discovery of gene regulatory networks (Friedman, Linial, Nachman & Pe'er, 2000). Automatic learning of BNs from observational data has been an area of intense research for more than a decade, yielding practical algorithms and tools (Spirtes, Glymour & Scheines, 1993). The ability of the BN approach to reconstruct genetic networks from microarray gene expression data has been extensively evaluated.

Consider a set of microarray experiments that measures the expression of a set of N genes over M different conditions. We denote the gene expression values by an $M \times N$ matrix $D = (d_1,...,d_n)$. The BN method discovers a directed acyclic graph (DAG) S such that the posterior probability $P(S|X=D)$ is maximized. Here $X = (X_1,...,X_N)$ denotes a set of random variables representing gene expression for genes $i = 1,...,N$. Let π_i be the set of parents of node i in an acyclic network S. Then, the probability $P(X = D|S)$ can be decomposed into the product of local probabilities of nodes specified by the network structure S:

$$P(X = D \mid S) = \prod_{i=1}^{N} P(X_i = d_i \mid X_{\pi_i} = D_{\pi_i}), \tag{1}$$

where X_{π_i} denotes the subset of variables corresponding to π_i and D_{π_i} the corresponding observations. For ease of notation, we will omit the symbol X but use D indicating that X takes an observation D. The nodes in the learned network correspond to genes or their products and the edges correspond to direct probabilistic dependencies, such as causality, mediation, activation, or inhibition between the genes. The posterior probability $P(S|D)$ is proportional to the product of the likelihood $P(D|S)$ and the prior probability $P(S)$ of network structure S based on prior knowledge, i.e.,

$$\mathrm{P(S|D)} \propto \mathrm{P(S)P(D|S)} \tag{2}$$

The main approach to learning BNs from data is based on the strategy of search-and-score, which attempts to identify the most probable network S given the data D. This network has the highest posterior probability. Depending on assumptions, maximizing this probability corresponds to maximizing a score function. There are several ways to define the score. A straightforward definition is the likelihood $P(D|S)$. For discrete data and multinomial distribution, the K2 score (Cooper & Herskovits, 1992) is often used to evaluate the networks generated. For a given network S, this score is defined as the likelihood:

$$P(D \mid S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!, \tag{3}$$

where N_{ijk} is the number of cases in D in which variable X_i has the kth value and the parent of i has the jth instantiation; q_i is the number of parents for i, and r_i the number of possible values of variable X_i. Thus,

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/structural-learning-genetic-regulatory-networks/38240

Related Content

The Social Spiders in the Clustering of Texts: Towards an Aspect of Visual Classification
Reda Mohamed Hamou, Abdelmalek Amineand Ahmed Chaouki Lokbani (2012). *International Journal of Artificial Life Research (pp. 1-14).*
www.irma-international.org/article/the-social-spiders-in-the-clustering-of-texts/81210

Usage of Comprehensive Learning Particle Swarm Optimization for Parameter Identification of Structural System
Hesheng Tang, Lijun Xieand Songtao Xue (2015). *International Journal of Natural Computing Research (pp. 1-15).*
www.irma-international.org/article/usage-of-comprehensive-learning-particle-swarm-optimization-for-parameter-identification-of-structural-system/126480

Parallel Hardware for Artificial Neural Networks Using Fixed Floating Point Representation
Nadia Nedjah, Rodrigo Martins da Silvaand Luiza de Macedo Mourelle (2011). *System and Circuit Design for Biologically-Inspired Intelligent Learning (pp. 295-308).*
www.irma-international.org/chapter/parallel-hardware-artificial-neural-networks/48900

Population Symbiotic Evolution in a Model of Industrial Districts
U. Merlone (2007). *Handbook of Research on Nature-Inspired Computing for Economics and Management (pp. 301-316).*
www.irma-international.org/chapter/population-symbiotic-evolution-model-industrial/21136

The Exposition of Fuzzy Decision Trees and Their Application in Biology
Malcolm J. Beynonand Kirsty Park (2009). *Advancing Artificial Intelligence through Biological Process Applications (pp. 375-394).*
www.irma-international.org/chapter/exposition-fuzzy-decision-trees-their/4988