Chapter 7.1
Problems for Structure Learning:
Aggregation and Computational Complexity

Frank Wimberly
Carnegie Mellon University (retired), USA

David Danks
Carnegie Mellon University, USA

Clark Glymour
Carnegie Mellon University, USA

Tianjiao Chu
University of Pittsburgh, USA

ABSTRACT

Machine learning methods to find graphical models of genetic regulatory networks from cDNA microarray data have become increasingly popular in recent years. We provide three reasons to question the reliability of such methods: (1) a major theoretical challenge to any method using conditional independence relations; (2) a simulation study using realistic data that confirms the importance of the theoretical challenge; and (3) an analysis of the computational complexity of algorithms that avoid this theoretical challenge. We have no proof that one cannot possibly learn the structure of a genetic regulatory network from microarray data alone, nor do we think that such a proof is likely. However, the combination of (i) fundamental challenges from theory, (ii) practical evidence that those challenges arise in realistic data, and (iii) the difficulty of avoiding those challenges leads us to conclude that it is unlikely that current microarray technology will ever be successfully applied to this structure learning problem.

DOI: 10.4018/978-1-60960-818-7.ch7.1

INTRODUCTION

An important goal of cell biology is to understand the network of dependencies through which genes in a tissue type regulate the synthesis and concentrations of protein species. A mediating step in such synthesis is the production of messenger RNA (mRNA). Protein products of one gene may help to regulate the rate of transcription into mRNA of the DNA reading frame of certain other genes. These dependencies among gene activities and their mRNA proxies have long been represented by directed graphs. Early in the 1990s, machine learning algorithms were developed for learning directed graphs representing causal relations from appropriate data samples. At about the same time, developments in microarray techniques made possible the simultaneous measurement of messenger RNA (mRNA) counts for thousands of distinct genes. This juxtaposition naturally led to a flood of studies in the computer science and biological literatures applying various search algorithms to gene expression data, with the aim of producing directed graphs that describe, for a tissue type, which genes regulate transcription rates of which other genes. Some of that work continues. We now know that the machine learning techniques are inappropriate and unsound in these applications, although they are potentially applicable to more recent measurements of RNA transcript concentrations in single cells. This chapter explains the statistical reasons why, as well as some of the relevant issues of computational complexity.

The short story is this: The goal of inference is the regulatory network within individual cells, but current microarray measurements are of mRNA counts extracted from large samples of cells. The machine learning algorithms exploit assumed symmetries between the network structure and a class of statistical properties of measurements. Assuming those symmetries hold for mRNA concentrations in individual cells and the regulatory network in the individual cellular level, and assuming all cells in the measured sample have the same regulatory network, it follows that the symmetry fails for measurements of concentrations aggregated from multiple cells. Experimental studies with real and simulated data confirm this failure.

THEORY: LEARNING FROM AGGREGATIONS

Microarrays are small chips a few square inches in size on which spots of DNA have been imbedded. A typical chip may contain thousands of spots, each spot composed of multiple copies of a small sequence of DNA. In the living cell nucleus, sections of DNA are copied ("transcribed") into a dual complementary molecule, RNA, which is the scaffolding for the synthesis, outside the cell nucleus, of cellular proteins. RNA can be extracted from tissue, and tiny luminescent beads can be chemically attached to RNA molecules obtained from tissue cells (e.g., from breast cancer cells). Each RNA molecule contains a sequence of bases that binds to a specific DNA sequence. When a suspension consisting of many RNA molecules from a tissue sample is applied to a microarray, the RNA molecules bind to the complementary DNA sites. By measuring the luminosity of each DNA spot, the relative concentration of each kind of RNA in the tissue sample can be estimated. From these concentrations, one can infer relative activity of genes—how much RNA is produced by various parts of the cell DNA in the tissues sampled.

Two fundamentally different strategies have been proposed to determine networks of regulatory relationships from microarray measurements. One strategy (Davidson, *et al*., 2002; Ideker, *et al*., 2001; Yuh, Bolouri, & Davidson, 1998) experimentally suppresses (or enhances) the expression of one or more genes, and measures the resulting increased or decreased expression of other genes. The method, while laborious, has proved fruitful in unraveling small pieces of the regulatory networks