A Bayesian Approach to Transforming Public Gene Expression Repositories into Disease Diagnosis Databases

Haiyan Huang
Department of Statistics
University of California at Berkeley
CA 94720, USA
hhuang@stat.berkeley.edu

Abstract

The rapid accumulation of gene expression data has offered unprecedented opportunities to study human diseases. The NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) is currently the largest database that systematically documents the genome-wide molecular basis of diseases. In this talk, I will introduce a our recent study on transforming public gene expression repositories into an automated disease diagnosis database. Relevant computational and statistical issues and difficulties will be discussed.

We have developed a systematic framework, including a two-stage Bayesian learning approach, to achieve the diagnosis of one or multiple diseases for a query expression profile along a hierarchical disease taxonomy. Our method, including standardizing cross-platform gene expression data and heterogeneous disease annotations, allows analyzing both sources of information in a unified probabilistic system. A high level of overall diagnostic accuracy was demonstrated by cross validation.

(This is a joint work with Dr. Jim Chun-Chih Liu and Dr. Xianghong Jasmine Zhou at USC.)