Datamining DNA microarray data: separating the wheat from the chaff

Sorin Draghici
Wayne State University / BioDiscovery
BioDiscovery Inc. 11150 W. Olympic Blvd., Suite 1150 Los Angeles, CA 90064
USA
(310) 966-9366
(310) 966-9346
sorin@biodiscovery.com
CAMDA00 Dataset 2: Leukemia

Draghici Sorin, Hoff Bruce, Shams Soheil

DNA microarray technology is gaining the spotlight through its ability to interrogate thousands of genes at the same time. However, this very ability is the source of the main problem in this field: the biological information is buried under a huge amount of numerical data. With the introduction of this technology, the emphasis has started to change from data generation to data mining and analysis.

Subsequent crucial issues are those of pre-processing and normalization, stages which are designed to separate the data carrying interesting biological information from noise and various other artifacts. The paper will discuss several data pre-processing and normalization techniques emphasizing their impact upon subsequent stages of analysis. Several normalization techniques are applied to the same data. Subsequently, the same gene selection criteria are applied and the results are shown to be different for different normalization techniques. This shows that the choice of the normalization technique has the potential to either emphasize or completely hide the valuable information buried in the data.

The paper addresses the following questions: 1. Is normalization important? How can normalization affect the subsequent data analysis? How do different normalization techniques determine different analysis results? 2. How can one assess the relevance of a particular sets of 'interesting' genes? 3. Is the success reported in [Golub,1999] to be expected in other microarray applications?

Keywords

Normalization, pre-processing, set of interesting genes, leukemia

Tools

Executable available

website

http://www.biodiscovery.com