Current Innovations in Microarray Analysis

A look at two-sided clustering and context-specific Bayesian clustering

Amit Kaushal
June 4, 2001
Overview

- To date, biologists have used (one-sided) clustering to analyze their data
- While clustering is informative, there is much more in the data that we can learn
- New techniques, such as two-sided clustering and context-specific Bayesian clustering attempt to make mine more information out of these microarrays
Cluster software released in 1998 by Michael Eisen, et al

- Implements a standard statistical algorithm to determine what genes show the most similar expression pattern over all data
How Cluster Works

- Compute a similarity score between every pair of genes
- Search for the highest score – this represents the most similar pair of genes
- Combine these genes into one node, and compute a similarity score between this node and every other gene/node
- Repeat recursively until there is one node, which is the whole tree
Benefits of Clustering

Cluster has become the industry standard for analyzing a biologist’s expression data. Why?

- One of the first programs able to process huge quantities of data in microarrays
- So easy to use, even a biologist can use it
- Acceptable results for a first-pass analysis
Drawbacks of Clustering

But there is a lot more information in the data, information that Cluster does not extract.

- Can not determine correlation between subsets of genes *and* experiments.
- The model is not flexible; it can not incorporate any prior knowledge we might have about the genes and their functions (ie promoter regions, clinical data).
Innovations in the Works

- Two-sided Clustering
  - Forms subsets of genes and experiments

- Context-specific Bayesian Clustering
  - Flexible, intuitive model for gene regulation
Two-sided Clustering

- More representational of biology in that genes that have common function will act together for the duration of the time that they carry out that function; two-sided clustering lets you see this subset

- Algorithm very similar to that of one-sided clustering

- Allows data to be clustered into subsets of genes and experiments
Bayesian Statistics $^{2,3,4}$

- Attempts to look at data and model gene relations based on causal relationships
- Cluster can only model data based on amount of transcription
Bayesian Statistics?

- Relies on a Bayesian Network
- For examples of basic concepts, see citation (2), pages 1-3
- BN are very good for describing processes that are locally dependent on each other
Advantages over Clustering

- Much richer than clustering, since Bayesian methods, given the same data set as a clustering program, can discovering “causal relationships, interactions between genes other than positive correlation, and finer intra-cluster structure”

- Can incorporate all kinds of information, not just mRNA output levels, and it makes biological sense

Friedman, N. et al. Using Bayesian Networks to Analyze Expression Data.

Friedman, N. and Y. Barash. Context-Specific Bayesian Clustering for Gene Expression Data. Recomb ’01.