A General Methodology for Integration of Microarray Data

Curtis Huttenhower
Olga Troyanskaya
Princeton University
chuttenh@princeton.edu

Microarray datasets tend to explore specific areas of biological function; integration of multiple microarrays allows construction of a more complete picture, but it can be difficult to combine independent datasets. Previous methods have attempted to integrate microarray data either purely statistically (Choi, 2003; Detours, 2003; Moreau, 2003) or for specific tasks (Ng, 2003; Pavlidis, 2003; Imoto, 2002; Hartemink, 2001). However, no general method for integration of microarray data with a focus on biological function has yet been proposed.

We present a method for the integration of microarray datasets employing a fixed structure Bayesian network. Rather than learning all interactions simultaneously, we focus on undirected functional interactions between pairs of genes. Using Expectation Maximization, we learn one set of network parameters per functional category of interest. This can be tuned to focus on particular functions or datasets, and the method can scale to include a variety of distance measures and analysis techniques. The outputs of these per-function networks are in turn integrated to produce a probability of functional relationship between any two genes.

We achieved 2-3 times enrichment of precision over random (at 5-15% recall) using evaluations against GO and KEGG annotations. In addition to predictions for individual gene pairs, the learned network parameters reveal information correlating entire microarray datasets with functional categories (e.g. heat shock with metabolism and organelle organization). As we integrate further processing methods and refine the network structure, we hope both to improve performance and to increase the ability of the technique to expose specific biological properties of microarrays.