

Proteome Profiling: Finding the Proverbial Needle in the Biological Haystack Expression Studies

Min Zhang, Ph.D. Candidate
Cornell University

Abstract:

High-throughput biotechnologies, such as microarray and mass spectrometry, simultaneously monitor the activities of thousands of genes at the RNA and protein level. Statistically, we are challenged by efficiently estimating high-dimensional parameters with noisy data. Furthermore, the signals in these large-scale analyses in genomics and proteomics are sparse and asymmetric. Here we propose a generalized shrinkage estimator based on empirical Bayesian thresholding, which is adaptive to the sparseness and possible asymmetry of the signals. The properties of this estimator have been investigated. Simulation study and application to microarray data demonstrate the performance of our approach.

Identifying polygenic effects on complex traits and profiling molecular features for clinical outcomes pose another challenging statistical issue as selecting variables with large p small n data. Likewise, the sparseness and possible asymmetry of the signals are the most important characteristics of the large p small n data, which should be exploited because of the limited sample size and/or the biological implication. We develop a Bayesian model selection approach to incorporate this *a priori* information. A heat-map is proposed to help researchers make informed decisions and control false discovery rate. This approach has been successfully applied in revealing sex-specific QTL underlying differences in glucose-6-phosphate dehydrogenase enzyme activity between two *Drosophila*