EXTRACTING MEANING FROM HIGH-DIMENSIONAL EXPRESSION DATA

J. Quackenbush

Dana-Farber Cancer Institute, Boston, MA, USA

† E-mail: johnq@jimmy.harvard.edu

The popular literature is rife with proclamations regarding the coming genomics revolution, but the application of genome scale techniques, including genome sequencing, transcript profiling, proteomics, and metabolomics, have largely been to cataloging response rather than developing a mechanistic interpretation of the patterns involved. While many balk at the challenges posed by "too much data," in fact the problem may be not enough data. The value of large datasets is that they can reveal features of the underlying biology provided they are filtered appropriately, and one approach that has proven successful is to integrate large datasets with other large datasets. In this way, genetic mapping and microarray data can be used to identify genes that are differentially expressed and genetically linked to a particular phenotype. Or changes in gene expression can be linked to changes in protein representation or metabolic flux. Similarly, prior knowledge about gene functional classes, or pathway membership, or interaction can be used as additional constraints on the data, as can constraints posed by evolutionary conservation of genes and pathways. The challenge moving forward is not one of collecting and analyzing data, but of integrating data to produce a comprehensive understanding of the biological systems under study.