Stochasticity and Networks in Genomic Data

John Quackenbush
Dana-Farber Cancer Institute and the Harvard School of Public Health

Two trends are driving innovation and discovery in biological sciences: technologies that allow holistic surveys of genes, proteins, and metabolites and a realization that biological processes are driven by complex networks of interacting biological molecules. However, there is a gap between the gene lists emerging from genome sequencing projects and the network diagrams that are essential if we are to understand the link between genotype and phenotype. â€˜Omic technologies such as DNA microarrays were once heralded as providing a window into those networks, but so far their success has been limited. Although many techniques have been developed to deal with microarray data, to date their ability to extract network relationships has been limited. We believed that by imposing constraints on the networks, based on associations reported through articles indexed in PubMed, we could more effectively extract biologically relevant results from microarray data and develop testable hypotheses that could then be validated in the laboratory. Using literature networks as constraints on a Bayesian network analysis of microarray data, we show that we are able to recover evidence for a wide range of known networks and pathways, even in experiments not explicitly designed to probe them.

With a putative gene-interaction network, the problem of producing viable models of the cell remains. While systems biology approaches that attempt to develop quantitative, predictive models of cellular processes have received great attention, it is surprising to note that the starting point for all cellular gene expression, the transcription of RNA, has not been described and measured in a population of living cells. To address this problem, we propose a simple (and obvious) model for transcript levels based on Poisson statistics and provide supporting experimental evidence for genes known to be expressed at high, moderate, and low levels. Although what we describe as a microscopic process, occurring at the level of an individual cell, the data we provide uses a small number of cells where the echoes of the underlying stochastic processes can be seen. Not only do these data confirm our model, but this general strategy opens up a potential new approach, Mesoscopic Biology, that can be used to assess the natural variability of processes occurring at the cellular level in biological systems.

Together these two approaches open new avenues of investigation that may help us in our eventual understanding of the function of biological systems, addressing many of the important questions that have arisen in the context of systems biology. Our ultimate goal will be to create predictive models that allow one to examine the current state of a biological system and to estimate the likelihood that, at some later time, the system will have evolved to a new state. Such an approach, if successful, could have a wide range of applications spanning laboratory, clinical, and translational biology.