

IEEE 7th BIBE Invited Plenary Keynote: Stochasticity and Networks in Genomic Data

October 14, 2007, Harvard Medical School Conference Center Rotunda, Boston, MA

John Quackenbush

Dana-Farber Cancer Institute and the Harvard School of Public Health
Harvard University
Boston, MA 02115
johnq@jimmy.harvard.edu

Abstract—

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. ‘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.



John Quackenbush received his PhD in 1990 in theoretical physics from UCLA working on string theory models. Following two years as a postdoctoral fellow in physics, Dr. Quackenbush applied for and received a Special Emphasis Research Career Award from the National Center for Human Genome Research to work on the Human Genome Project. He spent two years at the Salk Institute working on developing physical maps of human chromosome 11 and two years at Stanford University working on new laboratory and computational strategies for sequencing the Human Genome. In 1997 he joined the faculty of The Institute for Genomic Research (TIGR) where his focus began to shift to post-genomic applications with an emphasis on microarray analysis. Using a combination of laboratory and computational approaches, Dr. Quackenbush and his group developed analytical methods based on integration of data across domains to learn biological meaning from high-dimensional data. Since joining the faculty of Dana-Farber and the Harvard School of Public Health in 2005, his work has increasingly focused on the analysis of human cancer and expanded to embrace systems approaches to understanding and modeling biological problems.

Dr. Quackenbush is Professor of Biostatistics and Computational Biology and Professor of Cancer Biology at the Dana-Farber Cancer Institute and Professor of Computational Biology and Bioinformatics at the Harvard School of Public Health. He serves on a number of editorial boards, including *PLoS One*, *Bioinformatics*, *BMC Genomics*, and is Editor-in-Chief of *Genomics*. He is a member of numerous advisory boards, serves on a number of grant review panels, and has published more than 120 peer-reviewed scientific papers.