

## **Microarray Data Normalization and Transformation**

### **A proposal for an ISMB Tutorial**

**Instructors:** Catherine Ball, Stanford University  
John Quackenbush, The Institute for Genomic Research

Catherine Ball is a curator for the Stanford Microarray Database, which is used by hundred of researchers using microarray technology to address problems that include tumor biology, development, evolution, and basic cell biology. Her background in both bench biology and analysis and presentation of genome-scale data enables her to bridge the gap that frequently exists between researchers posing biological questions and the statisticians and computer scientists attempting to provide the means to answer them.

John Quackenbush is an Investigator at The Institute for Genomic Research, where he leads projects studying gene expression in human cancer, animal models of human disease, and in the model plant *Arabidopsis thaliana*. His group has developed a wide array of tools for microarray data analysis and they present regular courses on microarray analysis. Over the years, he has taught a large number of courses and has received many awards for his teaching.

Both instructors are active members of the Microarray Gene Expression Data society (MGED), which is working to establish standards for DNA microarray data and experimental descriptions.

#### **Goals and Overview:**

DNA microarray analysis is rapidly becoming the most widely used technique for the analysis of gene expression on a global scale. While laboratory protocols have become fairly robust, the techniques used for data analysis are still rapidly evolving. In general, the goal of using microarrays to examine gene expression is to identify patterns that display biologically relevant behavior. But before any downstream analysis is carried out, the crucial process of data normalization and transformation must occur to provide a sound basis for comparison between assays and genes. The goals of this workshop will be to provide an overview of the relevant issues associated with such analysis, a review of many of the techniques that are currently used, and an outline of many of the remaining challenges associated with both the analysis and the creation of an effective description of the process. Our hope is that the principles discussed here will be of use not only to those doing microarray analysis, but that they will be generally relevant to all those involved in large scale functional analysis.

#### **Audience:**

This workshop will attempt to provide a general overview of the process and will begin with a description of the biology underlying the analysis. At each step, an attempt will be made to both describe the relevant biological and relevant statistical assumptions so that

this presentation is accessible to biologists, statisticians, and computer scientists and will be of use to those starting to do microarray analysis as well as users experienced with the technique.

**Outline:**

- I. Introduction to Microarray Analysis
  - a. An overview of the technology
  - b. Expression measurements: The starting point
- II. Representing expression
- III. Expression from Fluorescence
  - a. What we measure
  - b. Signal-to-noise and saturation
  - c. Setting Thresholds
- IV. Why Normalize Data
  - a. Comparisons within and between datasets
  - b. Systematic variation in the data
  - c. Biological assumptions in normalization techniques
  - d. Strategies and algorithms for data normalization
  - e. Global versus Local Normalization
  - f. The use of exogenous controls
- V. Facilitating Comparisons
  - a. Replicate Filtering
  - b. Variance Stabilization
  - c. Estimating Errors
- VI. Experimental Design and Replication
  - a. Sources of error in microarray assays
  - b. Identifying and minimizing error through replication
- VII. Identifying Differential Expression
  - a. Error models and differential expression
  - b. Statistical approaches
  - c. Analysis of Variance (ANOVA)
- VIII. Toward Standards for Analysis
  - a. Are standards needed?
  - b. How do we describe analysis methods?
  - c. Is there a better way?