Statistical Issues in Microarray Data Analysis

Mark Reimers,
Center for Genomics and Bioinformatics
Karolinska Institute

Audience and Goals:

This course is for bioinformaticians who want to work with microarray data. Awareness is growing of the variability of this technology, and bioinformaticians find they need to work with the raw data, to ensure that their data mining algorithms work on valid numbers. The participants will obtain an overview of the main issues in data validation, pointers for addressing common problems, and introduction to software for specialized problems.

Pre-requisites are a good first course in statistics, and familiarity with microarray technology.

The instructor, Dr. Mark Reimers is director of the KISAC Bioinformatics Service Center, and co-director of the KI-chip Microarray facility, at the Karolinska Institute, Stockholm, Sweden. He has worked as a statistician in a variety of fields, has analyzed dozens of microarray data sets, and has developed his own model for multiple-probe data.

Outline:

Brief introduction to three leading technologies: printed cDNA array, spotted long oligonucleotide, and synthesized short oligonucleotide arrays by Affymetrix.

Overview of the analysis process:
Need for quality control of individual probes
Systematic variation between chips
Selecting significantly changed genes

Quality Control:
Wet-lab quality checks: RNA quality and incorporation
Detecting hybridization problems
Detecting uneven hybridization
Common patterns of uneven hybridization
Detecting background anomalies using negative controls
Using multiple positive controls to detect uneven hybridization

Normalization: How to ensure comparability
Sources of systematic bias: dye incorporation, time, temperature, buffer solutions
Detecting and compensating bias
Background:

- The controversy over background subtraction
- Other ways to deal with background

Dye-bias in two-color experiments
Intensity-dependent bias in one- and two-color experiments:
    Lowess normalization for two-color experiments
    Nonlinear normalization methods for one-color experiments
Detecting scanner or probe saturation
Combining scans at different settings to overcome saturation

Statistical Significance of Results
Corrections for multiple comparisons
False discovery rate
Correlation structure and randomization/permutation significance levels
Permutation methods to estimate false discovery rate
Normal technical and inter-individual variation: genes that show up frequently as false positives

(if time permits) Methods to estimate expression levels in multiple-probe systems
MAS 5.0
dChip
RMA