A NOVEL PROBABILISTIC FRAMEWORK FOR MICROARRAY DATA ANALYSIS: FROM FUNDAMENTAL PROBABILITY MODELS TO EXPERIMENTAL VALIDATION

by

Claudio A. Gelmi

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering

Fall 2006

ABSTRACT

Gene expression studies as currently done with printed arrays or pre-fabricated gene chips generate large quantities of data that are used to determine the relative expression levels of thousands of genes in a cell. The most compelling characteristic of these data sets is that the number of genes whose expression profiles are to be determined exceeds the number of replicates by several orders of magnitude. Standard spot-by-spot analysis seeks to extract useful information for each gene on the basis of the number of replicates available for the specific gene in question. As has become increasingly clear, this plays to the weakness rather than the strength of microarrays. On the other hand, by virtue of the sheer data volume alone, treating the entire data set as an ensemble, and developing fundamental theoretical distributions to represent these ensembles provides us with a framework for efficient extraction of gene expression information that plays to the strength of microarrays. Relatively little attention has been paid to studying distributions of complete microarray data sets; and virtually all of the published studies are empirical approximations fitted to observed data.

The primary objective of this dissertation is to present fundamental probability models for microarray data distributions that can be used for drawing rigorous statistical inference regarding differential gene expression. In this regard, we have departed from the standard gene-by-gene techniques that rely on *ad hoc* transformations needed to justify the use of classical statistics, since such techniques

play to the weakness of microarray technology. Instead, we consider the entire microarray data set as an ensemble and characterize it as such from first principles.

First, we present theoretical results that confirm what had previously been speculated, or assumed for convenience: that under very reasonable assumptions, the distribution of microarray intensities should follow the Gamma (not lognormal) distribution. It is subsequently established that a polar coordinate transformation of raw intensity data provides the basis for a technique in which each microarray data set is represented as a mixture of Beta densities for the fractional intensities (not intensity ratios), from which rigorous statistical inference may be drawn regarding differential gene expression. Using a Beta mixture model as its theoretical basis, a probabilistic framework for carrying out statistical inference was then developed. The final outcome of the inference is an ordered triplet of results for each gene: (i) a comparative fractional expression level, (ii) an associated probability that this number indicates lower, higher, or no differential expression, and (iii) a measure of confidence associated with the stated result (determined from the variability estimated from replicates, or else by propagation-of-error techniques when there are no replicates).

The application of the probabilistic framework is illustrated via a detailed treatment of experimental data from gene expression studies in *Deinococcus radiodurans* following DNA damage; the technique was also successfully tested on well known datasets that have been studied thoroughly in the bioinformatics literature using different statistical techniques. Additionally, the probabilistic framework was validated experimentally. The basic steps involved in the validation study were: i) the analysis of Affymetrix GeneChip array data and the selection of some candidate genes based on high probabilities of expression status and confidence associated with these

probabilities, and ii) the independent characterization of the real expression status (up-regulated, down-regulated or not differentially expressed) of the selected genes using a complementary high-precision, but not high-throughput polonies technology. The results of the probabilistic framework inference showed good agreement with the confirmatory results from the high precision, medium throughput polonies technique.

Finally, to assist potential users, the theoretical results have been reduced to software form and deployed as a toolbox for analyzing microarray data. The toolbox, designed in MATLAB®, is freely available from the Ogunnaike Research Group Web site: www.che.udel.edu/systems/people/cgelmi