Sequence Based Background Model for Affymetrix Arrays

Rafael A. Irizarry
Department of Biostatistics
Baltimore, MD 21209 Johns Hopkins University
Phone: 410-664-5946 Bloomberg School of Public Health
615 North Wolfe St. E3035
Baltimore, MD 21205

High density oligonucleotide expression arrays are widely used in many areas of biomedical research. These arrays use oligonucleotides with length of 25 base pairs that are used to probe genes. Typically each gene will be represented by 11-20 pairs of oligonucleotides referred to as probe sets. The first component of these pairs is referred to as a perfect match (PM) probe. Each PM probe is paired with a mismatch (MM) probe that is created by changing the middle (13th) base with the intention of measuring non-specific binding. The PM and MM are referred to as a probe pair.
After RNA samples are prepared, labeled and hybridized with arrays, these are scanned and images are produced and analyzed to obtain an intensity value for each probe. These intensities represent how much hybridization occurred for each oligonucleotide probe. However, part of the hybridization is non-specific and the intensities are affected by optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization.

A final step in the pre-processing of these arrays is to combine the 11-20 probe pair intensities, after background adjustment and normalization, for a given gene to define a measure of expression that represent the amount of the corresponding mRNA species. In this paper we illustrate the practical consequences of not adjusting appropriately for the presence of background noise and provide a solution. Our solution uses probe sequence information and an empirical bayes approach.