SAMPLE SIZE PLANNING FOR DEVELOPING CLASSIFIERS USING HIGH DIMENSIONAL DNA MICROARRAY DATA

K.K. Dobbin[†], R.M. Simon

National Cancer Institute, Bethesda, USA

[†] E-mail: *dobbinke@mail.nih.gov*

Many gene expression studies attempt to develop a predictor of pre-defined diagnostic or prognostic classes. If the classes are similar biologically, then the overall proportion of genes that are differentially expressed between the classes is likely to be small. This motivates a two-step process for predictor development, a subset of informative genes is selected for use in the predictor, and then the predictor constructed from these. Both of these steps will introduce variability into the resulting classifier, so both must be incorporated in sample size estimation. Previously reported methods provide sample sizes for identifying differentially expressed genes while controlling the false discovery rate (FDR), but such methods are not sufficient for planning studies to develop classifiers. We introduce a methodology for sample size determination that captures variability in both steps of predictor development. The method ensures that the classifier developed will achieve an expected misclassification rate within a specified tolerance of the best possible. As a corollary to our development we provide a novel method for calculating an optimal significance level cutoff for the gene selection step. The sample size methodology is shown to perform well on simulated data, and is applied to several real microarray datasets. We find that many prediction problems do not require a large training set of arrays for classifier development.