Sample Size Calculation for Microarray Studies with Survival Endpoints

Sin-Ho Jung*
Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
Samsung Cancer Research Institute, Samsung Medical Center, Seoul, Korea

Abstract

Oftentimes, we want to discover the genes whose expression levels are associated with a time-to-event endpoint, such as progression free survival or overall survival, through microarray studies. In this case, we need to adjust the false positivity in such discovery procedure for multiplicity of the genes using a multiple testing method. The most popular multiple testing methods used for gene discovery in microarray studies are to control the false discovery rate or the family wise error rate. In this paper, we review a FDR-control method to discover the genes associated with a time-to-event outcome and propose a sample size calculation method for microarray studies designed to discover genes whose expression levels are associated the chosen time-to-event outcome. These methods can be easily modified for other types of high throughput genome projects.

Keywords: Censoring; False discovery rate; Proportional hazards model

Introduction

Microarray has been a major high-throughput assay method to display DNA or RNA abundance for a large number of genes concurrently. One of the main objectives for conducting microarray studies is to discover the genes whose expression levels are associated with a time-to-event outcome (also called a survival outcome regardless the type of event), such as time to progression or survival. Discovery of the prognostic genes should be made taking multiplicity into account, but also with enough statistical power to discover important genes successfully. Biological samples for these high-throughput experiments are costly and usually difficult to obtain, so that having too many samples will waste valuable resources, while having too few samples may lead to no meaningful scientific conclusions. Therefore, it is necessary for us to develop a sample size calculation method for gene discovery.

One important aspect in gene discovery using a microarray study is to adjust the false positivity for multiplicity of the genes. The statistical procedures with such adjustment are called multiple testing. There are two types of false positivity we control in multiple testing for gene discovery with microarray data. One is called the family wise error rate (FWER), which is defined as the probability to discover any genes when all the genes under consideration are not associated with the clinical outcome. The other is called the false discovery rate (FDR), which is defined as the proportion of unassociated genes among the discovered genes. It is well known that the multiple testing procedures controlling the FWER are so strict that they select too few genes. So, some investigators prefer the multiple testing methods controlling the FDR.

A few methods have been developed for building genomic classifiers. Most methods are focused on sample size calculations for predicting classes and are not applicable to censored outcomes data directly. Some have considered sample size estimations for gene discovery with FDR control. For example, Jung derives a sample size formula for the Storey’s procedure under a weak dependence assumption [1,2]. Liu and Hwang propose similar sample size formulae that can be used for comparison of multiple independent samples. Others have also considered power and sample size calculations under FDR control [3]. Dobbin and Simon present analytical formulae for determining the number of biological replicates needed for developing a predictive classifier [4]. However, none of these publications provides a comprehensive investigation of sample size estimation for discovery of genes whose expression levels are associated with a survival outcome.

In this paper, we consider discovering genes whose expression levels are associated with a survival outcome of patients. At first, we review multiple testing with a FDR control and Cox’s regression method based on a proportional hazards model [5]. And, we propose sample size and power calculation methods for microarray studies to discover genes using Cox’s regression method while controlling the FDR and for sample size calculation for gene discovery with a FWER control [6,7].

Methods

False discovery rate

Suppose that we conduct m multiple tests, of which the null hypotheses are true for m0 tests and the alternative hypotheses are true for m1 = m - m0 tests. The tests declare that, of the m0 null hypotheses, A0 hypotheses are null (true negative) and R0 hypotheses are alternative (false rejection, false discovery or false positive). Among the m1 alternative hypotheses, A1 are declared null (false negative) and R1 are declared alternative (true rejection, true discovery or true positive). Table 1 summarizes the outcome of m hypothesis tests.

Benjamini and Hochberg defined the FDR as [8]

\[ FDR = \frac{E[R_0]}{R} \]

(1)

Note that this expression is undefined if Pr(R = 0) > 0. To avoid this issue, Benjamini and Hochberg (1995) redefined the FDR as [8]

\[ FDR = \Pr(R > 0)E\left(\frac{R_0}{R}\big| R > 0\right) \]

(2)

These two definitions are identical if Pr(R = 0) = 0, in which case we...
have \( FDR = E(R_i / R > 0) \) (= pFDR, which will be defined below).

If \( m = m_0 \), then FDR=1 by any critical value with \( Pr(R = 0) = 0 \). Pointing out this issue, Storey defines the second factor in the right hand side of (2) as pFDR, [9]

\[
pFDR = E \left( \frac{R_i}{R} \mid R > 0 \right)
\]

and proposes to control this quantity instead of FDR. Storey claims that \( Pr(R > 0) \approx 1 \) with a large \( m \), so that pFDR is equivalent to FDR [2].

We accept this argument in this chapter and do not distinguish between FDR and pFDR. Hence, definitions (1) and (2) are considered conservative and the conservativeness increases in \( m \).

\[ \text{Table 1: Outcomes of } m \text{ multiple tests.} \]

<table>
<thead>
<tr>
<th>Accepted hypothesis</th>
<th>Null</th>
<th>Alternative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>( A_j )</td>
<td>( A_i )</td>
<td>( m_i )</td>
</tr>
<tr>
<td>Alternative</td>
<td>( A_i )</td>
<td>( R_j )</td>
<td>( m_i )</td>
</tr>
<tr>
<td>Total</td>
<td>( A )</td>
<td>( R )</td>
<td>( m )</td>
</tr>
</tbody>
</table>

Benjamini and Hochberg (1995)[8] propose a multi-step procedure to control the FDR at a specified level. However this is known to be conservative and the conservativeness increases in \( m_0 \)[10]:

Suppose that, in the \( j \)-th testing, we reject the null hypothesis \( H_j \) if the p-value \( p_j \) is smaller than or equal to 2 (0; 1). Assuming independence of the \( m \) p-values, we have

\[
R_j = \sum_{j=1}^{m} I(H_j \text{true}, H_j \text{rejected})
= \sum_{j=1}^{m} Pr(H_j \text{true} | H_j \text{rejected}) Pr(H_j \text{rejected}) + o(m),
\]

Which equals \( m \alpha \), where \( m \cdot o(m) \rightarrow 0 \) in probability as \( m \rightarrow \infty \)[2].

Ignoring the error term, we have

\[
\text{FDR}(\alpha) = \frac{m \alpha}{R(\alpha)}, \tag{3}
\]

Where, \( R(\alpha) = \sum_{j=1}^{m} I(p_j \leq \alpha) \). Given \( \alpha \), estimation of FDR by (3) requires estimation of \( m \).

For the estimation of \( m \), Storey assumes that the histogram of \( m \) p-values is a mixture of \( m_0 \) p-values that are corresponding to the true null hypotheses and following U (0; 1) distribution, and \( m_1 \) p-values that are corresponding to the alternative hypotheses and expected to be close to 0 [2]. Consequently, for a chosen constant \( \lambda \) away from 0, none (or few, if any) of the latter \( m_1 \) p-values will fall above \( \lambda \), so that the number of p-values above \( \lambda \), \( \sum_{j=1}^{m} I(p_j > \lambda) \) can be approximated by the expected frequency among the \( m_0 \) p-values above \( \lambda \) from U (0, 1) distribution, i.e. \( m_0 / (1 - \lambda) \). Hence, given \( \lambda \), \( m_0 \) is estimated by

\[
\tilde{m}_0(\lambda) = \frac{\sum_{j=1}^{m} I(p_j > \lambda)}{1 - \lambda}
\]

By combining this \( m_0 \) estimator with (3), Storey (2002) obtains

\[
FDR(\alpha) = \frac{\alpha \times \tilde{m}_0(\lambda)}{R(\alpha)} = \frac{\alpha \sum_{j=1}^{m} I(p_j > \lambda)}{(1 - \lambda) \sum_{j=1}^{m} I(p_j \leq \alpha)}
\]

For an observed p-value \( p_j \), Storey defines the q-value, the minimum FDR level at which we reject \( H_j \), as [2]

\[
q_j = \inf_{\alpha \geq p_j} FDR(\alpha)
\]

This formula is reduced to

\[
q_j = FDR(p_j)
\]

if FDR(\( \alpha \)) is strictly increasing in \( \alpha \), see Theorem 2 of Storey (2003) [9]. The appendix in Jung (2005) shows that this assumption holds if the powers function of the individual tests is concave in \( \alpha \), which is the case when the test statistics follow the standard normal distribution under the null hypotheses [1]. We reject \( H_j \) (or, equivalently, discover gene \( j \)) if \( q_j \) is smaller than or equal to the prespecified FDR level.

The independence assumption among \( m \) test statistics is loosened to independence only among \( m \) test statistics corresponding to the null hypotheses [11], and to weak independence among all \( m \) test statistics [9,10]. These approaches are implemented in the statistical package called SAM [12].

Univariate Cox Regression

Let \( x_j \) denote the expression level of gene \( j = 1, ..., m \) for subject \( i = 1, ..., n \). In order to derive the asymptotic results for sample size calculation, we assume that the expression level is bounded in Section 2. For patient \( i \), we also observe \( Y_i \), denoting the minimum of survival time \( T_i \) and censoring time \( C_i \) together with event indicator \( \delta_i = I(T_i \leq C_i) \). We assume that \( T_i \) and \( C_i \) are independent.

In order to associate the survival time with the expression level of gene \( j \), we assume that, given \( x_j \), \( T \) has a hazard function of

\[
\lambda_j(t) = \lambda_0(t) \exp(\beta x_j),
\]

where \( \lambda_0(t) \) is an unknown baseline hazard function and \( \beta \) is the regression coefficient. By subtracting the sample mean and dividing by the standard deviation of gene \( j \), we assume that \( x_j \) has mean 0 and variance 1. In this case, \( \hat{\beta} \) denotes the log-hazard ratio between two groups of patients whose expression levels of gene \( j \) are one standard deviation apart. This regression model may not hold for all genes, but still provides a reasonable measure of relationship between survival time and each gene and the test statistic presented below controls the type I error [13,14].

By Cox (1972) the partial score and information functions are given by [5]

\[
U_j(\beta) = \sum_{i=1}^{n} \left[ z_i - \sum_{j=1}^{m} \frac{z_i \cdot x_i j (t) e^{\beta j}}{\sum_{j=1}^{m} Y_i (t) e^{\beta j}} \right] dN_i (t)
\]

and

\[
I_j(\beta) = -\frac{\partial U_j(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[ \frac{z_i \cdot x_i j (t) e^{\beta j}}{\sum_{j=1}^{m} Y_i (t) e^{\beta j}} - \frac{\sum_{j=1}^{m} z_i \cdot x_i j (t) e^{\beta j}}{\sum_{j=1}^{m} Y_i (t) e^{\beta j}} \right] dN_i (t)
\]

respectively, where \( N_i(t) = \delta_i I(X_i \leq t) \) and \( Y_i(t) = I(X_i > t) \) denote the event and at-risk events, respectively, and \( N(t) = \sum_{i=1}^{n} N_i (t) \). The partial maximum likelihood estimator \( \hat{\beta}_j \) is obtained by solving \( U_j(\beta) = 0 \) usually by using the Newton-Raphson methods.

By Anderson and Gill (1982), \( \hat{\beta}_j \) is a consistent estimator and
\( \sqrt{n}(\hat{\beta}_j - \beta_j) \) is asymptotically normal with mean 0 and its variance can be consistently estimated by \( nI_j(\beta_j) \). The expression level of gene \( j \) is not associated with the survival time if \( \beta_j = 0 \), and an over-expression (under-expression) of gene \( j \) shortens the survival time if \( \beta_j > 0 \) (if \( \beta_j < 0 \)). Hence, a two-sided test with marginal type I error rate \( \alpha \) rejects \( H_j: \beta_j = 0 \) if

\[
\sqrt{n} \frac{\hat{\beta}_j}{\hat{\sigma}_j} > z_{1-\alpha/2}
\]

Where \( \hat{\sigma}_j^2 = n/I_j(0) \) and \( z_{1-\alpha/2} \) is the upper 100\( \alpha \)-th percentile of \( N(0,1) \). Note that \( \sqrt{n}(\hat{\beta}_j/\hat{\sigma}_j) \) is asymptotically identical to the partial score test statistic \( U_j(0)/\sqrt{I_j(0)} \).

**Sample size and power calculation**

Let \( M_0 \) and \( M_1 \) denote the sets of genes for which the null and alternative hypotheses are true, respectively. Note that the cardinalities of \( M_0 \) and \( M_1 \) are \( m_0 \) and \( m_1 \), respectively. Since the estimated FDR is invariant to the order of the genes, we may rearrange the genes and set \( M_0 = \{1, \ldots, m_0\} \) and \( M_1 = \{m_0 + 1, \ldots, m\} \). For large \( m \) and under independence (or weak dependence) among the test statistics, we have [6,11].

By Storey and Tibshirani (2001) and Jung (2005), for large \( m \) and under independence (or weak dependence) among the test statistics, we have

\[
E(R(\alpha)) = \sum_{j=0}^{m} \Phi((z_{1-\alpha/2} - \sqrt{n} |\beta_j|/\sigma_j))
\]

by omitting the error term.

Let \( Z \) denote the random variable for the expression level of gene \( j \). From Appendix, we have

\[
\xi(\alpha) = \Phi((z_{1-\alpha/2} - \sqrt{n} \beta_j)/\sigma_j)
\]

where \( \Phi(*) = 1 - \Phi(*) \) and \( \Phi(*) \) denotes the cumulative distribution function of the standard normal distribution,

\[
\sigma_j^2 = \int_{\mathbb{R}} \left\{ E\left[Z_j^2F(t|Z_j)\right] \right\} dt - \left\{ E\left[Z_jF(t|Z_j)\right] \right\}^2
\]

and

\[
F(t|Z_j) = \exp\left(-\lambda(t) \exp(\beta_j Z_j)\right)
\]

are the survivor and probability density functions, respectively, of the survival time for a patient with gene expression level \( Z_j \) and \( G(t) = P(C \geq t) \). Here the expectations are taken with respect to the distribution of expression level of gene \( j, Z_j \). These integrals are usually obtained numerically. Hence, (4) is expressed as

\[
FDR(\alpha) = \frac{m_0 \alpha}{m_0 \alpha + \sum_{j=1}^{m_1} \Phi((z_{1-\alpha/2} - \sqrt{n} |\beta_j|/\sigma_j))}
\]

(5)

Note that FDR is decreasing in \( n \) and \( \beta_j \). Furthermore, by Jung (2005), FDR is increasing in \( \alpha \). If the effect sizes are equal among the prognostic genes, FDR is increasing in \( \beta_j/m \) and FDR increases from 0 to \( m_0/m \) as a increase from 0 to 1.

At the design stage of a study, \( m \) is determined when microarray chips are chosen for experiment and \( m_0 \) and \{\( \beta_j \), \( j \in M_1 \)\} are projected based on experience or from pilot data if any. The only variables undecided in (5) are \( \alpha \) and \( n \). With all other design parameters fixed, FDR is controlled at a certain level by the chosen \( \alpha \) level. So, we want to find the sample size \( n \) that will guarantee a certain number, say \( r_0 \), of true rejections with FDR controlled at a specified level \( \phi \).

In (5), the expected number of true rejections is

\[
E(R(\alpha)) = \sum_{j=0}^{m} \Phi((z_{1-\alpha/2} - \sqrt{n} |\beta_j|/\sigma_j))
\]

(6)

In multiple testing with FDR control, \( E(R(\alpha))/m_0 \) plays the role of the power of a conventional testing [15-17]. With \( E(R(\alpha)) \) and the FDR level set at \( r \) and \( \phi \), respectively, (5) is expressed as

\[
\phi = \frac{m_0 \alpha}{m_0 \alpha + r_0}
\]

By solving this equation with respect to \( \alpha \), we obtain

\[
\alpha^* = \frac{r_0}{m_0(1-\phi)}
\]

Given \( m_0, \alpha^* \) is the marginal type I error level for \( r_0 \) true rejections with the FDR controlled at \( \phi \). With a and \( E(R(\alpha)) \) replaced by \( \alpha^* \) and \( r_0 \), respectively, (6) yields an equation \( h(n) = 0 \), where

\[
h(n) = \sum_{j=0}^{m} \Phi((z_{1-\alpha^*/2} - \sqrt{n} |\beta_j|/\sigma_j)) - r_0
\]

(7)

We obtain the sample size by solving this equation using a numerical approach such as the bisection method. The final sample size may be chosen by rounding up the solution.

If we do not have prior information on the effect sizes, we may want to assume equal effect sizes \( \beta_j = \beta_0 \) for all \( j \in M_1 \). Let \( \sigma^2 \) and \( \sigma_0^2 \) denote \( \sigma_j^2 \) and \( \sigma_{0j}^2 \), respectively, obtained by replacing \( \beta_j \) with \( \beta_0 \). Then, (7) is reduced to

\[
h(n) = m_0 \Phi((z_{1-\alpha^*/2} - \sqrt{n} |\beta_0|/\sigma_0)) - r_0
\]

and, by solving \( h(n) = 0 \), we obtain a closed form formula:

\[
n = \left( \frac{\sigma_{0j}^2 + \sigma_j^2}{\beta_0^2} \right)^{1/2} + 1,
\]

(8)
where $\alpha' = r_0 \phi / \{m_j(1-\phi)\}$ and $1 - \beta' = r_1 / m_1$. Note that (8) is the conventional sample size formula when we want to detect an effect size of $\beta_j$ with power $1 - \beta'$ while controlling the type I error level at $\alpha'$.

Some practical distributional models for sample size calculation: Above sample size formula requires calculation of $\sigma_j^2$ and $\theta_j^2$, which depend of the distributions of survival time, censoring time and gene expression data. So, for a sample size calculation, we need to specify these distributions as accurately as possible. Suppose that there exist a database and a matching specimen bank to be used for a new microarray study. Once a sample size is calculated, we will randomly select the required number of patients from the database and specimen bank for the study. An exponential distribution can be assumed the truncated normal distribution with mean 0 and variance 1.

For bounded and standardized gene expression data, we may assume the truncated normal distribution truncated at ±2 (i.e. $\kappa = 2$), which can be obtained by truncating $\mathcal{N}(0, s^2)$ with $s = 1.90$, and the hazard rate of a patient with gene expression $z$ is specified as $\lambda(t|z) = \lambda_0 \exp(\beta z)$ with $\lambda_0 = 1$ and $\beta_0 = 0.5$ for all $m_j = 40$ prognostic genes. Under the assumptions on the survival and censoring distributions and the truncated expression data, we have $\sigma = 1.135$ and $\theta = 1.123$. Furthermore, we have

$$\alpha' = r_0 / \{m_j(1-\phi)\} = 0.842 \times 10^{-5}$$

and $1 - \beta = 1 - 30/40 = 0.75$, so that $z_{1-0.75} = 3.339$ and $z_{1-\beta} = 0.674$. Hence, from (8), the required sample size is given as

$$n = \left[\frac{(1.135 \times 3.339 + 1.123 \times 0.674)^2}{0.5^2}\right] + 1 = 83$$

Discussions

In this paper, we have considered discovery of genes whose expression levels are associated with a survival endpoint, and proposed sample size and power calculation methods for microarray studies that will be analyzed using the discovery method. We related the expression level of each gene with the survival outcome using a univariate Cox regression model and discovered genes by controlling the FDR to adjust for the multiplicity of genes. The proposed method is to calculate the sample size for a specified number of true discoveries while controlling the FDR at a given level. The input variables to be pre-specified are total number of genes for testing $m$, projected number of prognostic genes $m_j$, the baseline hazard rate $\lambda_0$, and the effect sizes of $\beta$ (the log-hazard rates between two patient groups whose gene expression levels are one standard deviation apart) for the prognostic genes. The method does not require any heavy computation, such as Monte Carlo simulations, so that we get a sample size in a second. Especially, if the effect sizes among the prognostic genes are identical, we have a closed form solution for the required sample size. The proposed method can be used to design a new study based on the parameter values estimated from pilot data.

The proposed methods can be used for other types of high-throughput biomarker study, such as genome wide association studies or gene sequencing studies, with minor modifications. The computer program for sample size calculation available from the author.
References


