PROJECT PROPOSAL

THE APPLICATION OF GROUP SEQUENTIAL METHOD IN THE DATA ANALYSIS OF MICROARRAY TECHNOLOGY

By

Yuntao Tian

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements for the degree in
Master of Science
In
APPLIED MATHEMATICS & COMPUTER SCIENCE
Indiana University South Bend

Spring 2006

Advisor
Dr. Yi Cheng

Committee Members:
Dr. Dana Vrajitoru
Dr. Morteza Shafii Mousavi

TABLE OF CONTENTS

1. Abstract

Currently, in the field of molecular biology, the application of microarray technology is widely used for various purposes. Because huge amounts of data can be generated from microarray experiments, proper statistical experiment design and data analysis are in high demand.

Group sequential experimental design and data processing method uses interim analysis for experiments with large sample sizes. These methods can suggest early termination of the experiment with overall the same significance and desired power as the fixed sample size design or even better. As a result, researchers will save the experimental cost with ethical and administrative benefits. This project will investigate the feasibility of the use of group sequential methods in microarray technology, which will deal with the data collection and test a set of comparable values to help the researcher to make an early termination for experiments. The R programming language will be used to simulate different group sequential algorithms. The successful codes will be used for real data from the cancer risk gene microarray study. It is expected that group sequential method will improve the experimental design strategy and data analysis for microarray technology compared with classical fixed sample size design.

2. Introduction

Microarray technology has great utility and provides more and more important findings in the areas of clinical diagnosis, drug discovery and biological or environmental testing. The data generated from microarray experiments can be so large that traditional methods of biological data analysis may have difficulty dealing with them properly. In addition, the current cost for microarray experiments is still very high, so the demand for a good statistical method for microarray data is in dire need in this area now.

The objective of this project is to develop a general procedure that can help the experimenter make a plan for the experiment from the analysis of existing data. In this project, a statistical method, group sequential method, will be used to evaluate whether the experiment can be terminated earlier based on the current achieved data, that is, if the evidence is strong enough to draw a conclusion. The computing programs of group sequential method will be used to analyze the data from a paper, which used microarray experiments to find the gene expression pattern for human breast cancer (Gruvberger 2001 [1]). It is expected that this project can develop the program tool using group sequential analysis to reduce the sample size from Gruvberger's paper and other further microarray research.

3. Literature Review

3.1 Group Sequential Method

Generally, in a complete random experimental design, the possible available sample resource will be assigned for some treatment. One or more certain observatory variables (for example, the blood pressure value from patients treated by certain medicine) will be obtained from each sample. In many cases, without treatment effect, the observatory variable in interest, denoted as X, follows a normal distribution with mean of μ_0 and variance of σ^2 ($\sigma^2 = E(X-\mu_0)^2$): $X \sim N(\mu_0, \sigma^2)$ [2]. When the treatment has effect on the samples, the new observatory variable might follow the new distribution denoted as $X \sim N(\mu_1, \sigma^2)$. Many statistical designs are performed as equivalence studies. That is, the study is to show the new treatment's effect being "the same as" or "at least no worse" than the effect of reference or "controlled" treatment. The study accepts the default hypothesis H_0: $\mu_0=\mu_1$ unless there is sufficient evidence indicating otherwise. In this case, it will reject the null hypothesis H_0: $\mu_0=\mu_1$ and admit the existence of the treatment effect. Accepting the alternative, H_A: $\mu_0 \neq \mu_1$, with a sufficient evidences means that the p-value of the any test (Pocock's test or O'Brien & Fleming's test) is smaller them the significance level α (usually it is 0.05 or 0.01).

In many experiments, the data accumulate steadily over a long period of time. Usually it is not practical to perform the experiment on all the available samples in the same trial, simply because of time constraints. Especially in the medical clinical and biological experiments, the sample source will be divided into several batches or groups for consecutive study because subjects may come sequentially. Each group will be studied and followed by next one. The collected data will accumulate in a timely manner.

It is reasonable to monitor and analyze the results, as they are available in order to take the action of early termination or some modification of the study. As long as the trial can be carried out in phases for a given significance level α for the whole sample space, it may test sequentially with interim analysis, at a carefully chosen modified α_i for each interim analysis such that overall significance level is α.

There are three major reasons to perform group sequential test: economic, ethical and administrative. Sequential interim analysis was originally proposed for economic reasons. Early termination can prompt or stop the development of the new product respectively. Either way will lead to savings on sample size, time and investment compared with the standard fixed sample size design. In medical and biological experiments, the total research fund and other resources are limited. By using the sequential method, it allows adjusted allocation of money and other resources for more studies. This is of utmost importance in microarray research, which usually requires much higher cost than classical biological approaches.

In trials dealing with human or animal subjects, it is ethical to minimize the possibility that the individuals are exposed to the uncertain treatment for too long with unexpected, unsafe, or inferior outcomes. For a potential positive treatment, it is desirable to terminate the trials sooner to provide the treatment to the future patients. For a potential negative, early termination means the subject can receive other more promising treatment sooner. Ethical consideration also provides the additional information for development of the treatment to be modified.

During the experiment, it is important to monitor whether the study is being executed as planned. The administrative purpose ensures that the samples are from the correct population and meet the criteria of the experiment as described in the protocol. The interim examination can detect a defect of the sample selection and suggest the revision before more resources are wasted and minimize systematic bias. For example in a gene microarray experiment, the examination of the data that comes from chips of the early groups can tell whether the selected chips with certain genes match the purpose of the study. If not, the researcher can change the chip type immediately.

The implementation of the group sequential method is actually dealing with the control of individual α_i for each interim test. Suppose overall α to be 0.05 as nominated level of type I error, obviously we have to control the α_i's smaller than α (α_i is the individual significant level for each interim test based on cumulative data, i=1,2,…,N. N is the total number of groups). For example, with 5 groups of α=0.05, we can use α_i ($\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$) for each cumulative test which controls the overall α. If the interim study using cumulative data with efficacy checking shows the test should stop before the 5th group, we should have a more "strict" α_i (should be less than 0.05) since we have not used all the possible samples.

There are several proposed algorithms for calculating critical values for group sequential analysis: Pocock (1977) [4], O'Brien & Fleming (1979) [5], Lan-Demets (α-Spending Function or Consumption Function, 1981) [6], Whitehead & Stratton (1983) [7], Wang & Tsiatis (1987) [8], Emerson & Fleming (1989) [9], Pampallona & Tsiatis (1994) [10], etc. Some of those will be discussed in the thesis. One of the specific aims of this project is to test an improved algorithm using R-programming simulation. On the other hand, existing algorithms are designed for equal group size testing, but in the real data from the paper I selected, the sample size is not equal for each experiment group. So it is necessary for me to modify the algorithm for my analysis. I will focus on the computation of critical values for the interim decision of H_0 rejection and acceptance in order to find exit probabilities in the sequential analysis using R-programming simulation.

3.2 Microarray Technology

Microarrays (used to be DNA arrays or gene chips) are the newly developed techniques to monitor biological gene expression for a large amount of genes in parallel [3]. This powerful tool allows the molecular characterization of a wide range of medical and biological problems, from the disease stages and response to stimuli, to the understandings of biodiversities and gene functions.

The microarray is typically a slide made of glass, polymer or metal. On the slide, hundreds or thousands types of DNA or other biological large molecules such as proteins are attached at different spots, called as "feature". In order to hybridize efficiently with samples, each feature contains at least millions of identical gene molecules. The features are usually printed on the microarray slide by a robot jet, or the DNA on the feature can also be synthesized in situ by photolithography, because the DNA is the repeat of four distinguish nucleotides: Adenine as "A", Guanine as "G", Cytosine as "C", and Thymine as "T". Out of these four types nucleotides, A and T can pair, so do G and C. The intact DNA molecule contains two complementary strands label as "+" and "-", for example, $\frac{+AGTCAGATCAGTA}{-TCAGTCTAGTCAT}$. The DNA strand can only be hybridized when sequences are complementary. The DNA sequence of a certain gene of any life form is unique. Therefore, when we want to study the level of that gene, we can put one strand of that DNA sequence on the microarray and label the complementary DNA strand in the sample with some detectable signal, usually fluorescence. Then we hybridize the labeled sample DNA with the microarray, and the target gene DNA will hybridize to its complimentary sequence. After that, we scan the microarray with a fluorescence reader. The gene expression level in the sample will be proportional to the fluorescence level detected on the microarray from the laser scanner.

Practically, the purpose of microarray is usually to evaluate the level of genes coming from the samples under a certain condition or treatment compared with the reference or control condition. The researchers can label with green fluorescence the sample from condition 1 and with a red fluorescence the sample from condition 2. If the gene expression level in sample 1 is in abundance, the feature will be greener, while if the gene expression in sample 2 is abundant, it will be red. If both are equal, the spot will be yellow, and if neither has high gene expression, it will appear black as expected. Thus, from the fluorescence intensities, the relative expression in both samples can be estimated and digitalized in numbers. The "distribution" of gene expression levels in the original scale is often not a normal distribution. One way to normalize the ratio data is to do the logarithmic transformation [11]. For example, genes expressed more in sample 1 will have ratio greater than 1 with logarithmic transformation of positive value, and genes expressed more in sample 2 will a have ratio less than 1 with logarithmic transformation of a negative value. Obviously, the equal expression will give a logarithmic value of 0. If sample 1 and sample 2 are identical from the same population, they will follow the normal distribution with mean of zero. By doing the statistical test with H_0: $\mu_1=\mu_2$, we can test the whether the logarithmic ratios are centered at zero.

One of the important utilizations of microarray technology is to compare different gene expression patterns of cancer patients to the patterns of normal people. This will construct the “fingerprint” gene map to predict the risk of bearing cancer for the pre-diagnostic patient, who might develop cancer in the future, so certain prevention treatments can be performed in advance to save the patient’s life. Many kinds of similes work have been done and much data is currently available. But for each study, a large amount of cancer samples have been collected with a great effort and each sample has to consume one microarray at least. Is this huge amount of experiments necessary for drawing the final conclusion in order to find the remarkable genes for cancer prediction? Answering this question is the other specific aim of this project. I will focus on the paper of “Estrogen Receptor Status in Breast Cancer Is Associated with Remarkably Distinct Gene Expression Patterns” by Sofia Gruvberger (2001, Cancer Research) [1] and utilize the group sequential test based the data and sample groups in Gruvberger’s paper to examed the possibility of early experiment termination with less cancer samples for that study. If the group sequential method can draw the similar conclusion as the paper suggests with less patients’ cancer samples and array chips, it would indicate that the group sequential method could be a promising analysis tools in microarray research.

4. Proposed Solution

For the first project objective, the strategy is to obtain a large amount of normal distributed random numbers (X_i) by R’s random number generation function. Those random numbers will be divided into five groups with equal amount of numbers. The formula for calculating the testing parameter Z: $z = (\sum_{i=1}^{mk} X_i) / \sqrt{mk\sigma^2}$ (m: group size, k the current groups for the interim test, k=1,2,3,4,5). First, I will use the group sequential test from Pocock’s algorithm and O’Brien & Fleming’s algorithm with type I error rates α. If the trial is stopped earlier with the same conclusion (reject or accept H_0) from a test of all five groups, this trial will be labeled as “succeed” with a sp%=1 for this trial and g=the current group number. If the conclusion from the sequential does match the conclusion of test of all five groups, (e.g. sequential method rejects H_0 at group 3, but when the test of all five groups finally accepts H_0.) the sp% for that trial will be zero and g=5. The whole process from the generation of random numbers will be repeated 1,000 to 100,000 times. The average of sp% will represent the success probability and the average of g will represent the expected group numbers by sequential method. After this step, the program with higher sp% and lower expected g, will be set up for the whole project.

After the achievement of first of objective, I will employ one of the different algorithms rather than Pocock's or O'Brien & Fleming's method by modifying the α calculation part in the program. In the project, the early termination also includes the acceptance of H_0 in the first several groups before finish all the samples. So we need to critical values (comparing values or boundaries) for each interim test. The program will compute two comparing values depending on α_i for each test, two values $Z_b<Z_a$. The boundary for accepting H_0 is Z_b: when the calculated Z-value is smaller than Z_b, we will accept H_0. The old algorithms never propose any solution for early acceptance of H_0, and this is a new idea in my project. The boundary for rejecting H_0 is Z_a: when the calculated Z-value is greater than Z_b, we will reject H_0. Then if the calculated Z-value falls between Z_a and Z_b, I will continue the experiment for next group. A similar simulation will be performed to calculate sp% and E(g) for each algorithm. The best algorithm will provide the highest sp% and smallest E(g) when given a large number of repeats of simulation, which are about 1000 to 10,000 depending on how soon the R can finish the simulation on my computer.

In the third step, I will utilize the best sequential method algorithm on the data from the Sofia Gruvberger's (2001, Cancer Research) paper [1]. Gruvberger's paper compares 58 breast cancer patients' samples with normal people's gene expression level. The 58 samples are divided into 5 groups/batches with various group sizes. The gene expression ratios (cancer/normal) of about 3000 distinct genes will be analyzed for each sample group. The sequential method criteria will be enforced for early termination test with the sp% and E(g) calculated.

Finally, the significant genes list from the above step and original conclusion from the paper will be compared to evaluate the validity and efficacy of employment of the group sequential in the microarray study.

5. Requirements

5.1 R Language

The main program-developing tool for this project is “The R Project for Statistical Computing”. R is a language and environment for statistical computing and graphics (http://www.r-project.org). R was developed from the S language at Bell Laboratories by John Chamber’s team, and both R and S are GNU based project. R provides an open source environment on different platforms (UNIX, LINUX, Windows and MacOS). Therefore, it is free to use R and the programs for R can be executed on different system with good portability and extensibility.

The advantages of R include an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, a well-developed, simple and effective programming language that includes conditionals, loops, and user-defined recursive functions and input and output facilities [12].

For the purpose of this project, R is sufficient to provide an adequate programming tool.

5.2 Data Description

Breast cancer is one of the most common forms of cancer in the United States. According to statistics from National Cancer Institute, the rate of breast caner is 129.1 cases per 100,000 populations with 134 for Caucasians and 118 for African American people. In 2002, female breast cancer was the third most common among the five major cancers, which was diagnosed at a late stage at the rate of 7.2 new cases per 100,000 women per year (Prostate: 7.6, Colon: 7.4, Rectum: 2.0, Cervix: 0.7). A lower rate of diagnosis at late stages is an early sign of the effectiveness of cancer screening efforts. It is very critical for the early diagnosis of breast cancer. As the current molecular biology technology develops, the microarray gene expression analysis is a very promising tool for the early diagnosis of cancers.

In order to set up the criteria for cancer microarray diagnosis, it is necessary to test the gene expression levels in the caner patients compared with healthy people and find the significantly different gene expression as a diagnostic marker. The DNA microarray technology allows for parallel analysis of the expression of thousands of genes to address complex questions in tumor biology. In Gruvberger's study, 58 grossly dissected primary tumors in 5 batches from node negative breast cancer patients, tumor sizes 20–50 mm, were collected. Human BT-474 breast cell line was used as a reference in all array hybridizations. For each gene, the fluorescent intensity of the most intense channel [red (Cy3) or green (Cy5)] for each sample was averaged over all samples. All genes for which the average exceeded 2,000 fluorescence units (scale 0–65,535 units) were included in the analysis. In addition, for all samples, the red and green intensities both exceeded 20 fluorescence units and that the union (of the two channels) spot area exceeded 30 pixels were required. These requirements left us with 3,389 of the original 6,728 genes. The original gene expression ratio data for each sample were stored in the public data file (http://research.nhgri.nih.gov/microarray/ER_data.txt) and is ready to be imported to an Excel file.

6. Expected Outcome

The development plan with expected outcome is designed as following:

I. The random number generation step with N(0,1) and N(μ ,1) will be set up. After that, the data will be equally divided into 5 groups. And the normal distribution property will be tested.

II. Pocock’s sequential method will be used to test the H_0: μ=0 with the generated 5 groups of data (α=0.05). The validity of sp% and E(g) calculation code will be examined.

III. After the program-testing step is finished, O'Brien & Fleming’s algorithm and my proposed algorithm will be used to modify Pocock’s method and their efficiency will be tested. The codes for Pocock’s function to calculate the boundaries will be changed for each other method.

IV. After achieving the best algorithm, the data from Gruvberger's paper will be studied. First the data (caner/normal gene expression level ratio) in the original Excel file will be formatted to a text file with comma as separations to fit the standards of R's data file. Then the data reading, alignment and transformation code will be written in the old program. Finally the transformed normal distributed data will be analyzed for all 3389 genes by the best group sequential method.

V. The comparison will be performed from the results of sequential method and the results of Gruvberger's paper. The conclusion will be given for the efficacy of the group sequential method.

VI. Documentations and Thesis Writing.

7.Conclusion

Most cancer microarray diagnosis depends on laboratory test. The application of this project will help doctors to set up the standards to estimate the level of cancer and treat patients easily. The combination of group sequential method and microarray technology is the advantage of this project. The completion of this project will hopefully provide a better way for cancer diagnosis.

8. Bibliography

Causton H, Quackenbush J, Brazma A,. (2003) *Microarray Gene Expression Data Analysis: A Beginner's Guide*. Blackwell Publishing. [3]

Emerson, S.S. and Fleming, T.R. (1989). Symmetric group sequential designs. *Biometrics*, 45, 905-923. [9]

Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS. (2001). *Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns*. Cancer Research. 2001 Aug 5;61(16):5979-84. [1]

Jennison C，Turnbull BW. (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC. [2]

O'Brien, P.C. and Fleming, T.R. (1977). A multiple testing procedure for clinical trials. *Biometrics*, 35, 549-556. [5]

Pampallona, S. and Tsianis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis treting with provision for early stopping in favor of the null hypothesis. *J.Statist. Planning and Inference*, 42, 19-35. [10]

Pieler R, Sanchez-Cabo F, Hackl H, Thallinger GG, Trajanoski Z，(2004). *ArrayNorm: comprehensive normalization and analysis of microarray data*. Bioinformatics. 2004 Aug 12;20(12):1971-3. [11]

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191-199 [4]

R project: http://www.r-project.org [12]

Reboussin DM, DeMets DL, Kim KM, Lan KKG, (1981). *Programs for Computing Group Sequential Boundaries Using the Lan-DeMets Method*. http://www.biostat.wisc.edu/landemets/ [6]

Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43, 193-200. [8]

Whitehead, J. and Stratton, I. (1983). *Group sequential clinical trials with triangular continuation regions*. Biometrics, 39, 227-236. [7]