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

Master of Science

in

APPLIED MATHEMATICS & COMPUTER SCIENCE
Indiana University South Bend

Yuntao Tian

2006

Advisor:
Dr. Yi Cheng

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

To my son: Frederick!

Acknowledgements

In the first place I want to acknowledge my advisor Dr. Cheng. During the work in this project, Dr. Yi Cheng gave me many supports and encouragements. Without Dr. Cheng's supervision and help, I could not choose this topic and finish my thesis.

I also would like to thank the help of Dr. Dana Vrajtoru and Dr. Morteza Shafii Mousavi for their support and guidance during the thesis writing. I would like to thank Dr. Hossein Hakimzadeh, Dr. Zhong Guan, and Dr. Dave Surma for the help with the experimental design and general advice.

Finally, I would like to dedicate this thesis to my parents, my husband Rui, and my son Frederick for their love, patience, and understanding! They allowed me to spend most of the time on this thesis for past couple of months. I love all of you!

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, with a significantly lower expected sample size, researchers will save the experimental cost with ethical and administrative benefits. This project will perform computerized simulation of different group sequential algorithms to find a desired one in order to investigate the feasibility of group sequential methods in microarray technology efficiently. 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.

Table of Contents

List of Tables

List of Figures

1. Introduction

The principle of statistics is applied in many fields, including medicine, government, education, agriculture, business and law. In the biological science field, statistics helps to develop and apply methodology for extracting useful knowledge from both experiments and data. This paper focuses on the application of statistical theory in the decision-making part during biological experiment design and data analysis.

Traditionally, when doctors performed clinical diagnosis for patients in the laboratory, the patients have to be taken many samples for testing in hospital. Now, this situation is changing. As new technologies advance, molecular gene expression profiling performed with microarray is applied in disease diagnosis in addition to the traditional immunohistological or pathological methods. 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. 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 (Causton, 2003). 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 data generated from microarray experiments can be so tremendous that traditional method of biological data analysis cannot deal with them properly; in addition, the cost for current microarray experiments is still at high level, so the demand for good statistical method for microarray is in dire need in this area now. On other hand, when we are looking for the important genes for cancer diagnosis, the large, whole genome microarrays are not needed for cancer classification. For diagnostics, the use of a few dozen genes, called cancer marker genes, is sufficient. When we perform the biological experiment to determine this marker gene list, how many groups of sample are needed is under the experiment consideration. By choosing too small amount of samples, we may not have enough information to draw the conclusion. On the other hand, if too many samples are chosen, it is a waste of resource and time.

The objective of this project is to develop a general procedure that can help the experimenters make plans for the experiment from the analysis of available 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. Group sequential methods are consistent comparing treatments in a series of interim analyses in a clinical trial. There are several algorithms for calculating critical values: Pocock (1977), O'Brien & Fleming (1979), which will be used in my program.

The computing programs of group sequential methods 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). In Gruvberger's paper, the goal is to find the breast cancer marker gene list. The author used 58 tumor samples in 5 batches to draw the conclusion. In this study, by using group sequential method programmed with R, the necessary sample size can be significantly reduced. It is expected that R program tool developed in this project can help researchers and doctors to reduce the sample size when using microarray as a diagnostic tool.

2. Literature Review

2.1 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, for the control group, the observatory variable in interest, denoted as X, follows a normal distribution with mean of μ_0 and variance of σ^2 ($\sigma^2 = E(X - \mu_0)^2$): $X \sim N(\mu_0, \sigma^2)$ (Jennison, 1999). 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 intends to show the new treatment's effect being "the same as" or "at least no worse" than the effect of the reference or "controlled" treatment. The study accepts the default hypothesis H_0: $\mu_0=\mu_1$ unless there is a 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 sufficient evidence means that the p-value of the test is smaller than 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.

There are three major reasons to perform the 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 allocation of money and other resources to be adjusted 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 a trial sooner to provide the treatment to the future patients. For a potential negative treatment, 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 in the sample selection and suggest the revision to minimize systematic bias before more resources are wasted. 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.

In theory, for group sequential test with two-treatment comparison, when a maximum number of groups K with size m for each group are chosen, the samples are allocated to two treatments followed by a randomization strategy which ensures that m patients (or subjects) are collocated to each treatment in each group. The accumulating data are analyzed after each group of 2m responses. For each k=1,…,K, a standardized statistic Z_k is computed from the first k groups of observations. The test stops with the rejection of H_0 if $|Z_k|$ is greater than a critical value c_k for k=1,…,K-1, otherwise, the trial will continue. If the test continues to the K^{th} analysis and $|Z_K| < c_K$, it stops at the final and H_0 is not rejected, otherwise, Ho is rejected. The series of critical values, $\{c_1,\ldots c_K\}$, are chosen to achieve a specified Type I error rate α_i at the certain stage of interim analysis such that the overall type I error rate is continued at α. Different types of sequential tests give different algorithms and different critical values.

When a trial is carried out in phases, given the significance level α for the whole sample space, it shall test sequentially with interim analysis, with carefully chosen α_i for the *i*th interim analysis *i*=1,…k, such that the overall significance level is α. The implementation of the group sequential method is actually dealing with the control of individual α_i for *i*th interim test. Assuming the overall α to be 0.05 as nominated level of Type I error, it can easily be entailed that α_i's should be smaller than α (α_i is the individual significant level for *i*th interim test based on cumulative data, i=1,2,…,K, where K is the total number of groups). For example, with 5 groups of α=0.05, we can use α_i ($\alpha_{1,}$ $\alpha_{2,}$ $\alpha_{3,}$ $\alpha_{4,}$ α_5) for each cumulative test that controls the overall α. If the interim study with calculated error rate smaller than α_i, it means we already have a more strict criteria (less than α) to stop the trial before the 5th group.

There are several proposed algorithms for calculating critical values for group sequential analysis: Pocock (197), O'Brien & Fleming (1979), Lan-Demets (*α*-Spending Function or Consumption Function (1981), Whitehead & Stratton (1983), Wang & Tsiatis (1987), Emerson & Fleming (1989), and Pampallona & Tsiatis (1994). 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 set that I selected, that sample size is not equal for each experiment group. So it is necessary for me to develop an algorithm with help form the existing methods, in order to carry out my analysis properly. I will focus on 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.

2.2 Microarray Technology

The modern development of molecular biology provides the possibility of accurate and reliable diagnosis of a disease in order to provide correct therapy decisions for the patients. Especially for cancer, the early diagnosis is crucial since for many types of cancers the late lethal stage is incontrollable, and cancer is the second frequent cause of death in the western world (Hoyert et al., 2005). Evaluating the molecular differences at the gene expression level in possible cancer patients is a promising direction in clinical practice for predicting cancer formation even before formation of cancer. The most notable tool in this direction of research is the microarray technology.

To understand the essence of gene expression data generated from microarray analysis, we need to understand the essential "central dogma" of molecular biology. The genetic information of most living organisms is stored in DNA, a long sequence (or polymerized organic molecules) of four different deoxyribonucleotides, which are sugar phosphate linked to one of the four types of nucleic bases: adenine (A), thymine (T), guanine (G), or cytosine (C). Out of these four types nucleotides, A and T can pair, so can G and C. The intact DNA molecule contains two complementary strands labeled as "+" and "-", for example, $\frac{+AGTCAGATCAGTA}{-TCAGTCTAGTCAT}$. The genome of the organism contains segments of DNA that encode genes, which are essentially the particular arrangements of DNA sequence. In the cells, the real functional materials are proteins, which determines how biological or pathological performances are achieved. However, the information for synthesizing proteins is stored in the genomic DNAs. In order to "transfer" the coded information from DNA to protein, DNA genes are transcribed into another type of polymerized organic molecules – messenger RNA (mRNA), which are similar to DNA but with different sugar phosphate and the nucleic base uridine replaces the appearance of thymine. The mRNAs are single strands and are translated to form protein. This process is called gene expression.

In the living organism, amount of DNA genes is constant. To control of protein's biological or pathological functions, the amount of mRNA present at a certain time is the important factor in the regulation of gene expressions, and represents a specific condition of the organism. Based on this theory, we can extract the mRNAs in the sample, and by some method called reverse-transcript PCR, we can make DNA copies of these mRNAs in the test tubes with the DNA amount proportional to their original biological value. These manually synthesized DNAs called as cDNA are used as the sample in the microarray technology.

Microarray is typically a slide made of glass, polymer or metal. On the slide, hundreds or thousands types of DNA are attached at different spots, called "features". 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. 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 cDNA sample as previous described with some detectable signal, usually fluorescence. Then we hybridize the labeled cDNA 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 certain condition or treatment compared with the reference or control condition. The researchers can label with green fluorescence for the sample from condition 1 and a red fluorescence for the sample from condition 2. If the gene expression level in sample 1 is in abundance, the feature will be greener, 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 (Pieler, 2004, Quackenbush, 2002). For example, genes expressed more in sample 1 will have a ratio greater than 1 with logarithmic transformation of positive value, and genes express more in sample 2 will have a 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: $\mu_1=\mu_2$, one can test the whether the logarithmic ratio, $\log(\mu_1/\mu_2)$, is centered at zero. Where μ_1 represents the population mean of the gene expression level in group 1; and μ_2 represents the population mean of the gene expression level in group 2.

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 a 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 this 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, published on Cancer Research (Gruvberger, 2001) and utilize the group sequential test based the data and sample groups in Gruvberger’s paper to exam 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 fewer patients’ cancer samples and array chips, it would indicate that the group sequential method can be a promising analysis tool in microarray research.

3. Simulation of Pocock’s test and O'Brien & Fleming Test

3.1 Pocock’s test

In the year of 1977, Pocock introduced an innovative algorithm for group sequential test, which followed the principle of “repeated significance test” (Pocock, 1977). In the clinical trials, the patient entry is usually sequential; as a result, the observations become available sequentially. On the ethical basis, the fixed sample size designs are less flexible. Before Pocock's method, the sequential methods (Armitage, 1975) had only been applied in a small fractions of actual trials, because Armitage's algorithm requires that the trials to meet the criteria of design with two treatments, patient entry in matched pairs, instantaneous patient evaluation, a normal or binary response and continuous surveillance of accumulating data. However, Armitage’s idea of repeated significance testing used in sequential method is still valuable. He proposed that, after each observation (for paired samples or the difference between two treatment arms), we could carry out a two-sided significance test, which can be very similar to fixed sample test. In this repeated significance testing, a "nominal" significance level α' could be used to stop the trial and declare the evidence of a treatment effect. When we define the overall significance level α, which is the Type I error rate and the power 1-β, which is the power of the test at a targeted treatment effect difference, we can deduce the value of α' and the total sample size N numerically.

Armitage's design is fully sequentially, not group sequential. By adapting Armitage's idea, Pocock suggested the GROUP sequential test instead of continuous assessment after every observation. The group sequential test can solve the disadvantage in Armitage's design by dividing the patient entry into a number of equal-sized groups in order to perform repeated significance test of accumulated data and make the decision of experiment termination. He contracted a constant nominal significance level (α_is) to analyze data at a relatively small number of times over the course of a study. The sample entry is divided into K equally sized groups containing m subjects on each treatment, and the data are analyzed after each new group of observations has been observed.

In Pocock's test, after the observations are obtained for each group, a cumulative standardized statistic Z_k (k=1,..,K) is calculated based on the difference treatment of arms,where $X_{Ai} \sim N(\mu_A, \sigma^2)$, $X_{Bi} \sim N(\mu_B, \sigma^2)$, i =1,2,…,and

$$Z_k = \frac{1}{\sqrt{2mk\sigma^2}}\left(\sum_{i=1}^{mk} X_{Ai} - \sum_{i=1}^{mk} X_{Bi}\right), \quad k = 1,\ldots \mathrm{K} \qquad (3.1)$$

The value of each Z_k suggests a individual significance level under current interim analysis. If the absolute value of Z_k is sufficiently larger than a preset limit, which is calculated to assume a Type I error rate α_i for *i*-th interim analysis, the study stops and we conclude rejection of H_0 at the *k*th analysis. This is in the form of repeated significance test: reject H_0 at stage k if $|Z_k| \geq C_p(K, \alpha)$, $k = 1, \ldots K$. The critical value $C_p(K, \alpha)$ is chosen to give overall Type I error α. $\Pr_{\mu A - \mu B = 0}$ {Reject H_0 at analysis k=1,k=2,…,or k=K}= α. If H_0 has not been rejected by the final analysis K, it is accepted.

Formally, the test is:

At *k*-th interim analysis k = 1 ,… , K-1 (3.2)

if $|Z_k| \geq C_p(K, \alpha)$ stop, reject H_0

otherwise continue to group k+1

At the end K-th, block,

if $|Z_k| \geq C_p(K, \alpha)$ stop, reject H_0

otherwise stop, accept H_0

The constants $C_p(K, \alpha)$ displayed in Table 1 for α = 0.01,0.05 and 0.1 were computed numerically using the joint distribution of the sequence of statistics $Z_1, \ldots, Z_k$. The computation of constants C_p is not a focus here and will not be further discussed in this thesis. I just used the existing constant values, shown in Table 1.

In each step of the interim analysis, the test rejects H_0 when the cumulative standardized statistic Z_k using the available data suggests a significance level below the two-side nominal significance $\alpha_i = 2[1- \Phi\{C_p(K, \alpha)\}]$, where Φ is the standard normal cumulative distribution function (cdf). So Pocock’s method is a repeated significance test with constant “nominal significance level” α_is since α_i does not change during each interim test. For example, if K=5 and α=0.05, the nominal significance level applied at each interim analysis is α'=2[1- Φ(2.413)]=0.0158 for *i* = 1,…K. It is essential that an interim significance level is below the nominated overall Type I error rate in order to avoid the "multiple-looks" problem mentioned earlier.

Table1 *Pocock tests: constants C_p (K, α)*
K groups of observations, Type I error probability α

	C_p (K, α)		
K	α = 0.01	α = 0.05	α = 0.1
1	2.576	1.960	1.645
2	2.772	2.178	1.875
3	2.873	2.289	1.992
4	2.939	2.361	2.067
5	2.986	***2.413***	2.122
6	3.023	2.453	2.164
7	3.053	2.485	2.197
8	3.078	2.512	2.225
9	3.099	2.535	2.249
10	3.117	2.555	2.270
11	3.133	2.572	2.288
12	3.147	2.588	2.304
15	3.182	2.626	2.344
20	3.225	2.672	2.392

3.1.1 Satisfying the Power Requirement

The power of a statistical test is the probably of detecting the treatment difference when the treatment effect exists at a certain value δ.

$$\Pr\ \{\text{Reject } H_0 \mid \mu_A - \mu_B = \pm\delta\} = 1 - \beta \tag{3.3}$$

To fulfill the satisfaction of a larger power requirement at a fixed δ, we need to increase the total sample size N. On the other hand, in order to detect a treatment effect more efficiently, we should not include too large samples in our trial. Too large sample size is not efficient. The issue here is to fulfill the power requirement with as fewer patients as possible.

The required sample size for each treatment arm in fixed sample test can be deduced as the following form:

$$n_f\ (\alpha,\ \beta,\ \delta,\ \sigma^2\ )\ =\ \{\Phi^{-1}(1\text{-}\alpha/2)\ +\Phi^{-1}(1\text{-}\beta)\}^2 2\sigma^2/\delta^2 \qquad (3.4)$$

with the null hypothesis of no treatment difference H_0: μ_A= μ_B against the two-sided alternative $\mu_A \neq \mu_B$ with significance of α and power 1-β at μ_A - μ_B = ±δ. This sample size calculation is proportional to σ^2/δ^2. When we consider the maximum possible sample size for group sequential test, n_f $(\alpha,\beta,\delta,\sigma^2)$ is serves as a starting point to multiply with a specific ratio of particular test. The ratio for Pocock's method is defined as $R_P(K, \alpha, \beta)$, which is a factor of maximum number of groups K, α, and β. Therefore, the maximum sample size per treatment arm of Pocock's test is given as $R_P(K, \alpha, \beta)\cdot n_f\ (\alpha,\beta,\delta,\sigma^2)$.

For example, when we perform a test at significance level α=0.05 and power 1-β=0.9 at μ_A - μ_B = ±1 given the population variance σ^2=4, the fixed sample size needed for each treatment can be calculated as n_f (0.05,0.1,1,4)=84.1. When we design the group sequential test for this trial, suppose that we want 5 groups in total (K=5), R_P(5, 0.05, 0.1) is 1.207. By multiplying 84.1 with 1.207, we need a total of 101.5 subjects per treatment to detect ±1 difference with power of 0.9. In each group then, we need 101.5/5=20.3, which rounds up to a integer as 21 per group. Following this plan, we study 21 subjects for each treatment A or B, calculate the standardized statistic statistic Z_k (k=1,..,K), and compare it with C_p (5, 0.05) to determine the early termination of the experiment.

The calculated values of R_P (K, α, β)s are given in Table 2 where the group number K ranges from 1 (actually, when K=1, it is identical to the fixed sample size test) to 20 at power 0.8 and 0.9 level. These values are required when I simulating the Pocock's test with R programs.

Table 2 *Pocock tests: constants R_p (K, α, β)*
K groups of observations, Type I error probability α and power 1-β

R_P (K, α, β)

		1 – β = 0.8			1 – β = 0.9	
K	α = 0.01	α = 0.05	α = 0.1	α = 0.01	α = 0.05	α = 0.1
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.092	1.110	1.121	1.084	1.100	1.110
3	1.137	1.166	1.184	1.125	1.151	1.166
4	1.166	1.202	1.224	1.152	1.183	1.202
5	1.187	1.229	1.254	1.170	***1.207***	1.228
6	1.203	1.249	1.277	1.185	1.225	1.249
7	1.216	1.265	1.296	1.197	1.239	1.266
8	1.226	1.279	1.311	1.206	1.252	1.280
9	1.236	1.291	1.325	1.215	1.262	1.292
10	1.243	1.301	1.337	1.222	1.271	1.302
11	1.250	1.310	1.348	1.228	1.279	1.312
12	1.257	1.318	1.357	1.234	1.287	1.320
15	1.272	1.338	1.381	1.248	1.305	1.341
20	1.291	1.363	1.411	1.264	1.327	1.367

3.1.2 Simulation of Pocock's test

In this section, I implemented an R program, labeled as R-1, to simulate the algorithm of the Pocock method. The user can change the parameters α, μ, σ^2, K, m, and how many times of simulation "r" to get the average values of early termination correct rate and group numbers. The idea of this program is to randomly generate m*K independent normally distributed observation "X"s from N(μ, σ^2), and then divided them into K groups, such that each group contains m values. This one set of observations represents the differences from the responses of two treatments (X_i=X_A-X_B). After the random arrangement of m*K observations, the accumulated Z-values will be computed from group 1 till group K according to the following formula (3.5):

$$Z_k = \frac{1}{\sqrt{mk\sigma^2}}\left(\sum_{i=1}^{mk} X_i\right) \quad , \ k \ = \ 1,\ldots \mathrm{K} \qquad (\ 3.5\ )$$

Then the Z-values will be compared with the constants Cp (K, α), under in a special case I chose Cp (5, 0.05) = 2.413. By following the strategy discussed above, the program will decide whether to reject H_0 and stop at current kth group (record the k to a parameter “h”), or continue to the next group. The program will compute up to the total of 5 groups are sampled and then make the final decision. If the early rejection matches the final decision, we set the correct rate as “1” and the stop-group number h=k for this trial, otherwise we set the correct rate as “0” for false group sequential rejection and use the total number of groups “5” for the value h. Then we start the next trial until we perform simulation r times (I used 1000 in my simulation). After the simulation, we will obtain the empirical correction rates and hs; by the percentage of it’s call it the expected efficiency of Pocock’s test. When we consider the power requirement, we need to set the sample size (m*K) first before simulation. For example, when I need to simulate the trial under the scenario where α=0.05, 1-β=0.9, μ=0.5 (actually setting δ=0.5 since we just generate one set of observations), σ^2=1, and K=5, we calculate n_f (α,β,δ,σ^2)=84.1. The sample size per group for Pocock’s test is m = R_P(K, α, β) n_f (α,β,δ,σ^2)=1.207*84.1/5=20.2. So I chose 21 as m for each group. Then I used these design parameters for the simulation of power consideration.

Simulation Results:

1. α=0.05, μ=0, σ^2=1, K=5, m=20, r=1000;

```
> pocock(0.05,0.5,1,5,20,1000):

Probability of Reject H0:    0.008
Average of Group Number:    4.925
Correct Rate of Program:    0.975
```

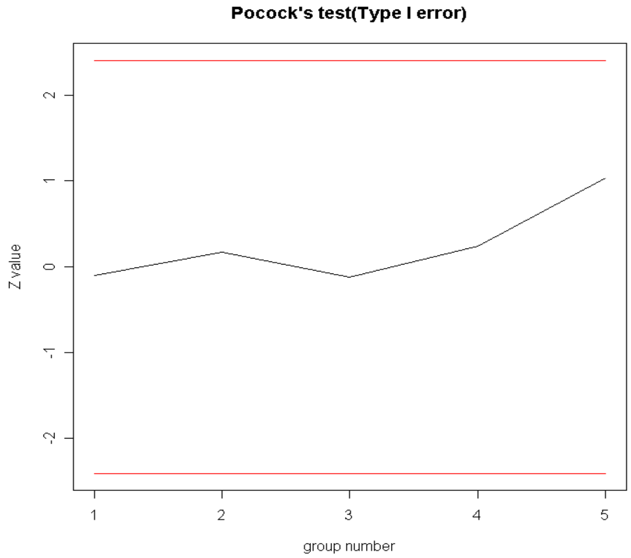

Figure 1
Simulation of Pocock's test with α=0.05, μ=0, σ^2=1, K=5, m=20, r=1000

The Figure-1 shows an example of a simulated trial for this simulation. The two red lines are the fixed boundaries Cp for Pocock's test. The black curve represents the change of calculated Z values. As anticipation, when μ=0, we expect the conclusion of no difference between treatments. So most of the test will continue to the last group and accept H_0. By saying the probability of Reject H0 is only 0.008, we mean that out of 1000 reptilians, only 8 trials reject H_0 at the 5th group. Since most of the terminations occur at the last group, the "Average of Group Number" is 4.925, very close to 5, although there are some early terminations to drop the values below 5. And most of the early rejection are incorrect which is indicated with the "Correct Rate of Program" = 97.5%.

As a result, the simulation is successful and the conclusions are reasonable with those assigned parameters. It also shows the Pocock's procedure is conservation since 0.008 is much smaller than the significance level α=0.05

2. α=0.05, μ=(0.1 to 1), σ^2=1, K=5, m=20, r=1000;

In the following expeniment, I simulated the situation when the treatment effect exists from a small 0.1 to a large 1.0 scale compared with the variance of 1.

Table 3 *The change of efficiency of Pocock's test with increasing treatment*

μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PRH	0.081	0.325	0.714	0.947	0.994	1	1	1	1	1
AGN	4.762	4.217	3.39	2.453	1.857	1.488	1.248	1.125	1.051	1.027
CRP	0.953	0.93	0.957	0.989	0.998	1	1	1	1	1

Probability of Reject H0: PRH
Average of Group Number: AGN
Correct Rate of Program: CRP
α=0.05, σ^2=1, K=5, m=20, r=1000

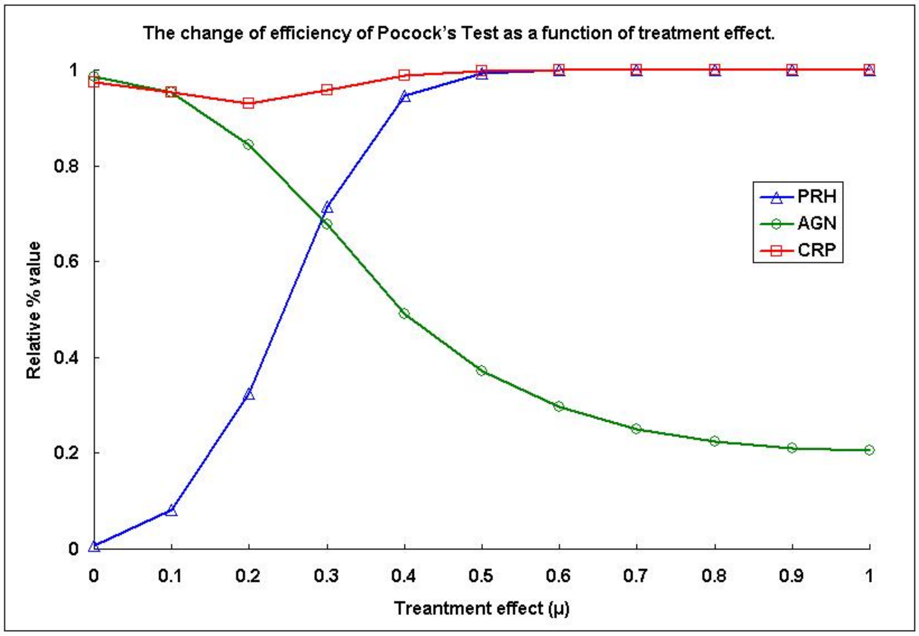

Figure 2
The change of efficiency of Pocock's test as a function of treatment

As the result shows, as treatment effect increases, the probabilities to reject H_0 (power of the test) increase rapidly. After the $\mu>0.5$, the PRH reaches 1.0, which is a great power value. The expected test group numbers AGN drop quickly as treatment effect increases. This indicates a good efficiency using Pocock's test when μ is considerably big and when $\sigma^2=1$. As show in figure-2, after $\mu>0.6$, the AGN is close to 1, which suggests most of the Pocock stops at the first interim analysis. The overall correct rate of Pocock's test is very close to 1 when μ is small since there are some false rejections. When $\mu>0.5$, CRP stays at 1, since Pocock's test made 100% right early rejection.

The following figure-3 shows the representative trials with boundaries and calculated Z values when (a) $\mu=0.1$, (b) $\mu=0.5$, and (c) $\mu=1.0$. The sequential process is notable when $\mu=0.5$ as Figure 2 (b) shows: At first two interim analyses, the Z values are between the boundaries Cp and the trials continue. However, for the 3rd analysis, H_0 was rejected and

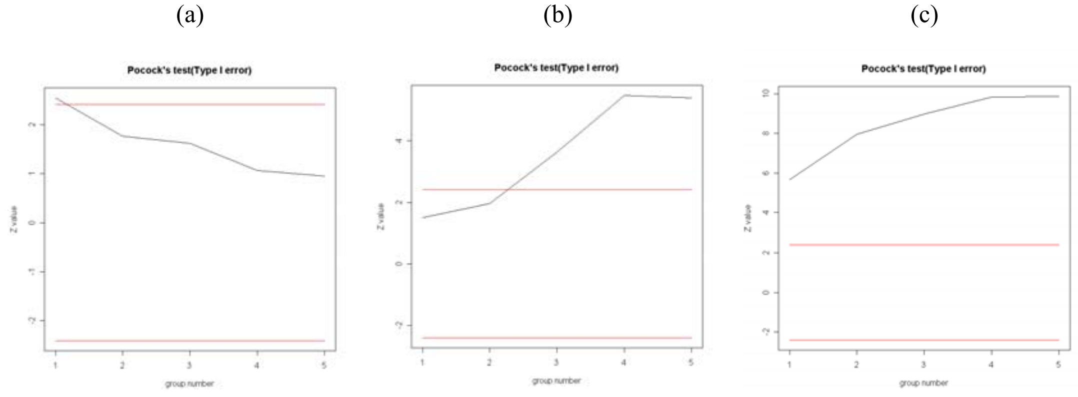

Figure 3
Sample Simulation Trials of Pocock's test with $\alpha=0.05$, $\sigma^2=1$, K=5, m=20, r=1000
(a) $\mu=0.1$, (b) $\mu=0.5$, (c) $\mu=1.0$

this decision was correct since the 4th and 5th Z are also beyond boundaries. Therefore, the real trial group number is 3 for this trial.

Figure 2(a) illustrates a false rejection example case when the treatment effect is small (μ=0.1). By Pocock's group sequential test, the H_0 was rejected at the 1st analysis. However, the final conclusion is accepting H_0. This early false termination is due to the small numbers of available samples size in the early stage of analysis. This is also the reason why the correct rate of group sequential test cannot always be 100%. So other reference (O'Brien, 1977) suggest we should set a wider boundary for the early steps of interim analysis, such as O'Brien & Fleming's test that I will discuss in section 3.2.

3. $\alpha=0.05$, $\mu=0.5$, $\sigma^2=1$, $K=5$, $r=1000$; $1-\beta=0.9$ with power consideration, $R_P(5, 0.05, 0.1)=1.207$.

I wrote the second program for simulation of Pocock's test with calculation of m by setting the power requirement. The group size m is calculated by $m = R_P(K, \alpha, \beta)\ n_f(\alpha,\beta,\delta,\sigma^2)$ when other parameters are given. The simulated results are:

```
> pocockIb(0.05,0.1,0.5,1,5,1000)
Group Size: 21
Probability of Reject H0 :    0.996
Average of Group Number:    1.762
Correct Rate of Program:    1
```

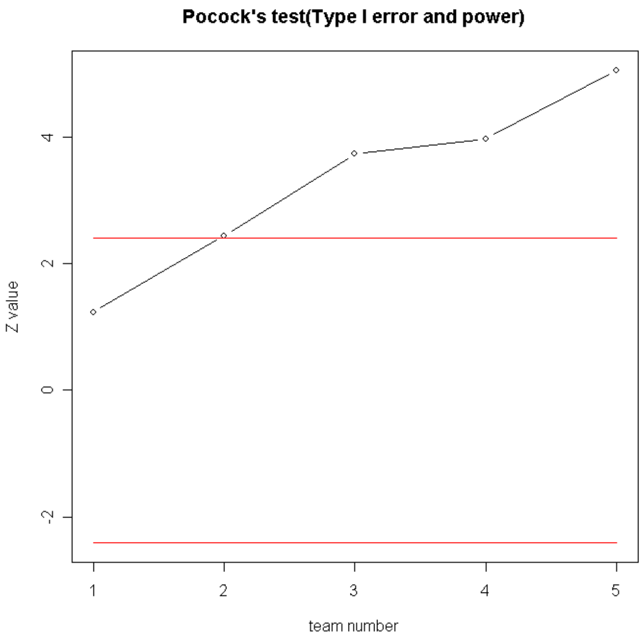

Figure 4

Simulation of Pocock's test with α=0.05, μ=0.5, β=0.1, σ^2=1, K=5, r=1000

As the results showed, the simulated Probability of Reject H_0 is 99.6%, which is much better than the designed power. And the figure-4 represents a trial sample in this simulation. As a conclusion, my program cans successfully the Pocock's algorithm with power requirements.

3.2 O'Brien & Fleming's test

O'Brien & Fleming was also developed based on the theory of repeated significance test in 1979 (O'Brien, 1979). The purpose of O'Brien & Fleming's paper was originally to propose a multiple testing procedure to improve the fixed sample test of clinical trials with dichotomous treatment response. Although O'Brien & Fleming's idea was based on Pearson chi-square test with the comparison with a critical value, the idea of normal standardized test statistics can still be adapted as Pocock's test.

When we define the same term of the standardized statistics Z_k, O'Brien & Fleming's test rejects H_0 after group k when $|Z_k| \geq C_k$ for a sequence of critical values. However, the C_k for O'Brien & Fleming's test is not a constant for individual k. The C_k decrease as we reach the later group of interim analysis. Formally, the test is written as:

At the k-th interim analysis, k = 1 ,… , K-1 (3.6)

if $|Z_k| \geq C_B(K, \alpha)\sqrt{K/k}$ stop, reject H_0

otherwise continue to group k+1

At the end of K block,

if $|Z_k| \geq C_B(K, \alpha)$ stop, reject H_0

otherwise stop, accept H_0

Thus, $C_k = C_B(K, \alpha)\sqrt{K/k}$, k = 1,…,K. Values of $C_B(K, \alpha)$ which ensure an overall Type I error probability α are provided in Table 4.

Table 4 *O'Brien & Fleming tests: constants C_B (K, α)*
K groups of observations, Type I error probability α

	C_B (K, α)		
K	α = 0.01	α = 0.05	α = 0.1
1	2.576	1.960	1.645
2	2.580	1.977	1.678
3	2.595	2.004	1.710
4	2.609	2.024	1.733
5	2.621	***2.040***	1.751
6	2.631	2.053	1.765
7	2.640	2.063	1.776
8	2.648	2.072	1.786
9	2.654	2.080	1.794
10	2.660	2.087	1.801
11	2.665	2.092	1.807
12	2.670	2.098	1.813
15	2.681	2.110	1.826
20	2.695	2.126	1.842

As we can calculated by setting $C_k=\Phi^{-1}(1-\alpha'/2)$, for K=5 and α=0.05, the nominal significance levels for analyses 1 to 5 are α_1’=0.000005, α_2’=0.0013, α_3’=0.0084, α_4’=0.0225, α_5’=0.0413. Obviously, the possibilities of rejecting H_0 at early analyses are very small and boundaries are very conservative. This ensures the strict examination with fewer observations in the beginning. As the experiment extends, more data will be evaluated under a relaxed condition, and overall Type I error rate is controlled as planed.

3.2.1 Satisfying the Power Requirement

Similar to Pocock’s test, O’Brien & Fleming’s test also uses a ratio $R_B(K, \alpha, \beta)$ to calculate the maximum sample size requirement for the group sequential analysis. Table 5 provides the constant $R_B(K, \alpha, \beta)$, which will be used in sample size calculations. The maximum sample size on each treatment arm is $R_B\ (K, \alpha, \beta)\cdot n_f(\alpha, \beta, \delta, \sigma^2)$.

Table 5 *O’Brien & Fleming tests: constants $R_B\ (K, \alpha, \beta)$*
K groups of observations, Type I error probability α and power 1-β

$R_B\ (K, \alpha, \beta)$

		1 – β = 0.8			1 – β = 0.9	
K	α = 0.01	α = 0.05	α = 0.1	α = 0.01	α = 0.05	α = 0.1
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.001	1.008	1.016	1.001	1.007	1.014
3	1.007	1.017	1.027	1.006	1.016	1.025
4	1.011	1.024	1.035	1.010	1.022	1.032
5	1.015	1.028	1.040	1.014	***1.026***	1.037
6	1.017	1.032	1.044	1.016	1.030	1.041
7	1.019	1.035	1.047	1.018	1.032	1.044
8	1.021	1.037	1.049	1.020	1.034	1.046
9	1.022	1.038	1.051	1.021	1.036	1.048
10	1.024	1.040	1.053	1.022	1.037	1.049
11	1.025	1.041	1.054	1.023	1.039	1.051
12	1.026	1.042	1.055	1.024	1.040	1.052
15	1.028	1.045	1.058	1.026	1.042	1.054
20	1.030	1.047	1.061	1.029	1.045	1.057

By comparing Table 2 and Table 5, we can see the maximum sample size requirement in O’Brien & Fleming’s test is smaller than Pocock’s test, because the effect of (μ_A - μ_B) on E(Z_k) increases with k and the power is gained primarily from the later analyses. Since nominal significance levels in the later analyses of O'Brien & Fleming's test are more relaxed than Pocock's test, they can gain more power than Pocock's boundaries with same group size (easier to reject H_0). Therefore, in the following simulation, we expected to see quicker termination by O'Brien & Fleming's test and lower expected group numbers of analysis.

3.2.2 Simulation O'Brien & Fleming Test and Comparison with Pocock's test

1. α=0.05, μ=(0 to 1), σ^2=1, K=5, m=20, r=1000;

We regard μ=0 no treatment difference, as a special case within the range of 0 to 1. I wrote a new R-program to simulate O'Brien & Fleming's test with α=0.05, μ=(0 to 1), σ^2=1, K=5 and m=20 with 1000 repetitions. The program only differentiates from previous one with O'Brien & Fleming's rejection critical values. The results are given as following:

Table 6

The change of efficiency of O'Brien & Fleming's test with increasing treatment

μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PRH	0.158	0.453	0.813	0.983	0.996	1	1	1	1	1
AGN	4.878	4.563	3.885	3.149	2.627	2.288	2.031	1.861	1.705	1.532
CRP	0.984	0.988	0.993	0.999	0.999	1	1	1	1	1

Probability of Reject H0: PRH
Average of Group Number: AGN
Correct Rate of Program: CRP
α=0.05, σ^2=1, K=5, m=20, r=1000

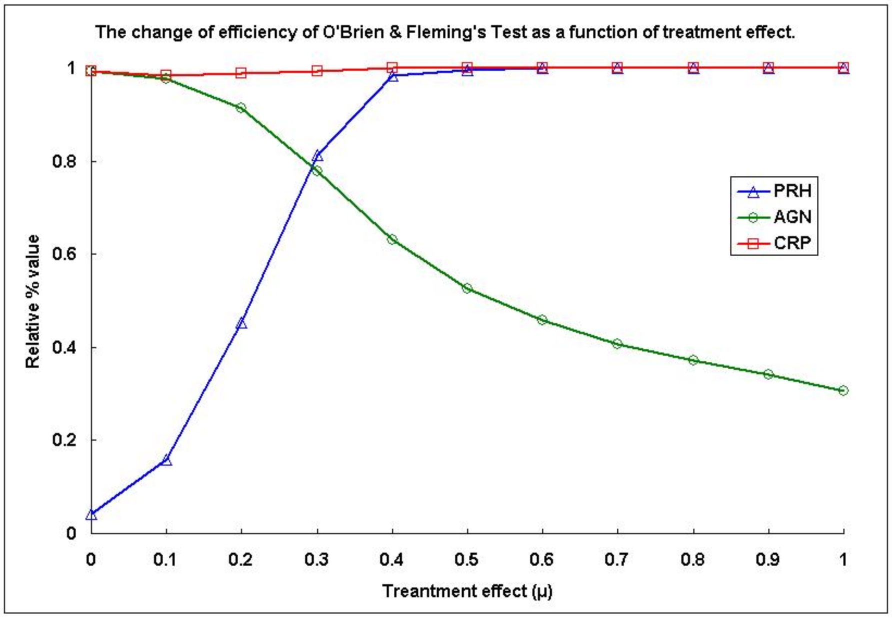

Figure 5

The change of efficiency of O'Brien & Fleming's test as a function of treatment

The above Figure 5 shows the trend of change of efficiency of O'Brien & Fleming's test treatment effect value increases. The conclusion is similar to Pocock's test but with some difference, which makes two tests favorable for different research goals.

The overall correct rate of O'Brien & Fleming's test is closer to 1 than the rate of Pocock's test, even when μ is small. This is due to the strict rejection critical value at lower μ range that the algorithm provides. So the false rejection is less likely to happen. The probabilities of rejecting H_0 (power of the test) is increased quicker than for Pocock's test. For μ>0.4, the PRH reaches 1.0, which indicates O'Brien & Fleming's test can provide better power. The expected test group numbers AGN of O'Brien & Fleming's test, however, drops slower than Pocock' test, which is also due to the strict rejection critical value at lower μ range. In addition, it is close to 1.5 when μ=1. We pay the price of larger sample size for higher correct rate, which may need more sample resource.

2. α=0.05, μ=0.5, σ^2=1, K=5, r=1000; 1-β=0.9 with power consideration, R_b(5, 0.05, 0.1)=1.026.

I wrote the second program for simulation of O'Brien & Fleming's test with calculation of m by setting the power requirement. The group size m is calculated by m = R_b(K, α, β) n_f (α,β,δ,σ^2) when other parameters are given. The simulated results are:

```
> OBFIb(0.05,0.1,0.5,1,5,1000)
Group Size:    18
Probability for Reject H0:    0.998
Average of Group Number:    2.742
Correct Rate of Program:    1
```

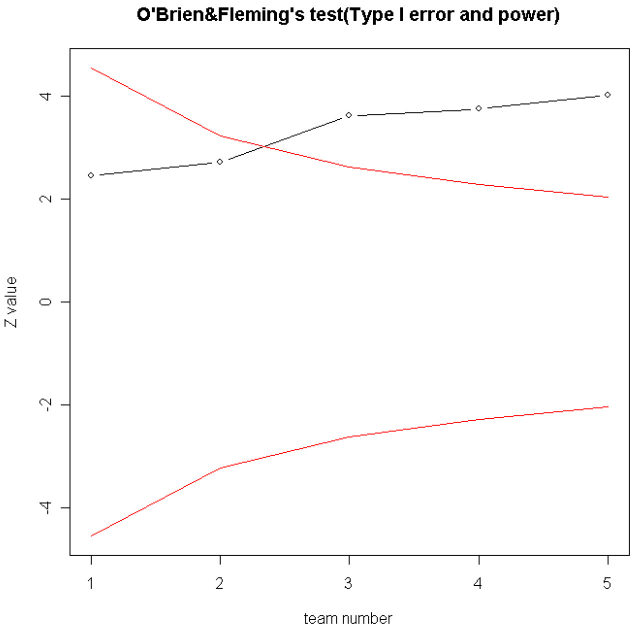

Figure 6

Simulation of O'Brien & Fleming's test

with α=0.05, μ=0.5, β=0.1, σ^2=1, K=5, r=1000

As the results showed, the group size m is calculated as 18. The simulated Probability of Reject H_0 is 99.8%, which is also much better than the designed power.

And the Figure 6 illustrates one of trials in the simulation. I also observed that the expected group number is 2.742, which is greater than Pocock's test.

4. Proposed Two Algorithms for Early Stop with Acceptance of H_0

In both Pocock and O'Brien & Fleming's tests, during the interim analysis, only the rejection of H_0 can be made, and there is no possibllity for early termination by accepting H_0. However, designs permitting early acceptance of H_0 are also appropriate because if it becomes in the interim analysis that H_0 is unlikely to hold, the trial should be ended to save the cost for the same ethical reasons that support the stops of early rejection of H_0.

4.1 Introduction for Early Termination with Acceptance of H_0

Five years after Pocock published the important algorithm of group sequential test, Gould and Pecore (1982) addressed the importance of early termination to accept the null hypothesis. They extended Pocock's procedure to allow acceptance of H_0 in at an interim stage. By simulation, they found that designs permitting both early acceptance and rejection of H_0 have a lower average cost than designs not permitting early acceptance even under both conditions when H_0 is true or false.

In order to allow early termination based on both acceptance and rejection of null hypothesis, we will introduce another critical value (comparing value or boundary) for each interim test. For the interim analysis at the end of k^{th} group of observation, the strategy defines two critical values a_k and b_k ($0 \leq a_k < b_k$). The boundary for accepting H_0 is a_k: when the calculated absolute value of standardized statistic Z_k is smaller than a_k, we will accept H_0. Same as other group sequential, the other critical value b_k is used to reject

H_0 when $|Z_k| > b_k$. As the idea of early termination by accepting H_0 has not been well studied yet and few algorithms have been proposed yet. I will define different strategies for setting the lower critical values a_k, and test the efficiency for each strategy. When the calculated Z_k falls between a_k and b_k, it continues the experiment for the next step of the interim analysis. A group sequential test with possible early stopping to accept H_0 is defined by pairs of constants (a_k , b_k) with $0 \leq a_k < b_k$ for k = 1,…,K-1 and $a_K = b_K$ is formalized as following

At k-th interim analysis k = 1 ,… , K-1 (4.1)

if $|Z_k| \geq b_k$ stop, reject H_0

if $|Z_k| < a_k$ stop, accept H_0

otherwise continue to group k+1

At the end of K, block,

if $|Z_k| \geq b_K$ stop, reject H_0

if $|Z_k| < a_K$ stop, accept H_0

Since $a_K= b_K$, termination at analysis K is ensured.

Sometimes, the early termination with acceptance of H_0 may not be appropriate at the first few analyses. It is possible to set a_k=0 to avoid early acceptance of H_0 when k is small.

4.2 Strategies for Calculating Critical Values

Here I propose two methods for early termination with acceptance of H_0.

4.2.1. Average Probability Method

Assumes there are K groups in each treatment, I spend overall α among K groups, and accumulated it as the trials go: k = 1,…,K.

$\alpha_a' = (1- k\,(1- \alpha)/K)/2 \quad a_k = \Phi^{-1}(\alpha_a')$ for early acceptance. (4.2)

$\alpha_b' = (1- k\alpha/K)/2 \quad b_k = \Phi^{-1}(\alpha_b')$ for early rejection. (4.3)

This form is based on same size group. In real world, the group size depends on the way that patients were recruited, and it might be unequal. So based on different members of groups, I spend α for each group to make the α's proportional to the group size. The following statement shows this modification. The parameter "pms" is the total number of sample, msg is the sum of sample numbers at the kth analysis for k=1,…K

$\alpha_a' = (1- (1- \alpha*msg/pms))/2$ for early acceptance. (4.4)

$\alpha_b' = (1-\alpha*msg/pms)/2$ for early acceptance. (4.5)

Simulation Results

Table 7

The change of efficiency of Average Probability Method with Early Acceptance of H_0

μ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PRH	0.016	0.123	0.401	0.773	0.943	0.975	0.994	0.998	0.999	1	1
AGN	2.652	2.79	2.953	2.668	2.271	1.843	1.556	1.308	1.172	1.088	1.031
CRP	0.911	0.922	0.878	0.892	0.956	0.976	0.994	0.998	0.999	1	1

Probability of Reject H0: PRH
Average of Group Number: AGN
Correct Rate of Program: CRP
$\alpha=0.05$, $\sigma^2=1$, K=5, m=20, r=1000

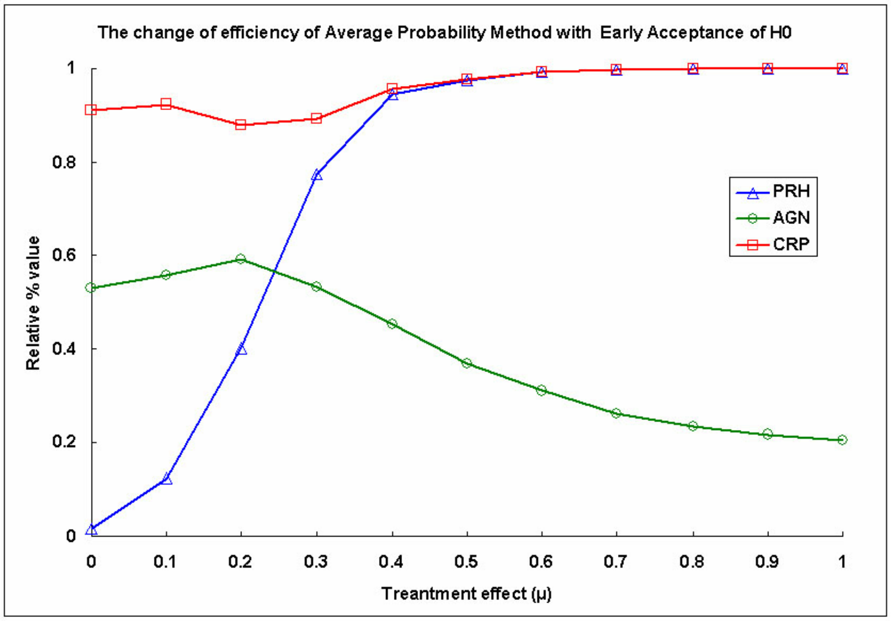

Figure 7

The change of efficiency of Average Probability Method with Early Acceptance of H_0

Table7 and Figure7 summarize the results of the simulation of “Average Probability Method” with early Acceptance of H_0. The effect of acceptance of H_0 can be observed from the peaks of Average of Group Number and Correct Rate of Program between $\mu=0$ to $\mu=0.4$. Since when μ is small, it is more likely to accept H_0 and stop the trial, so the AGN is relatively low in those cases. During the test, two types of false termination can be mad. The first is false acceptance, which means the early termination acceptance, but the final conclusion is rejection. The second is the false rejection, which is the same as for Pocock’s and O’Brien & Fleming’s tests. Since we increase the possibility of false early termination, we have lower CRPs compared with previous tests without early H_0 acceptance. This cannot be avoided because we take the risk of more early terminations. However, when $\mu=0.2$, the intermediate situation, the test provides the highest expected group number and worst correct rate. So when $\mu=0.2$ and $\sigma^2=1$, the test should be chosen more carefully.

4.2.2 Modified O’Brien & Fleming’s Method

As a consequence α’s vary, the changing in critical boundary values is not changed as quickly as α’s. In other word, using the standard normal cumulative distribution function (cdf) a_k or $b_k = \Phi^{-1}$(α’s), we can hardly achieve very large or very small boundary values for strict termination criteria. Those values will be beyond the allowable number range for the computer programs, and they will be calculated as infinite values. To neared this, I start from O”Brien & Fleming’s idea and introduce a variable constant Cm that can be defined by the user to meet special needs. This method I propose with early acceptance of H_0 is formalized as the following:

At k-th interim analysis,k = 1 ,… , K-1, Cm>0 (4.6)

if $|Z_k| \geq C_m \sqrt{K/k}$ stop, reject H_0

if $|Z_k| < C_m \sqrt{k/K}$ stop, accept H_0

otherwise continue to group k+1

At the end K block,

if $|Z_k| \geq C_m$ stop, reject H_0

otherwise stop, accept H_0

This algorithm offers more flexibility when we perform the experiment, where the outcomes show great treatment effects, however we need to chose the most important treatment effect. This is very important for the following real data analysis of microarray data.

4.3 Simulation Results

To compare with original O’Brien & Fleming’s test, I chose Cm=2.040, same as for the previous simulation. The following is the summary of results.

Table 8
The change of efficiency of Modified O'Brien & Fleming's test with Early Acceptance of H_0

μ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PRH	0.007	0.064	0.236	0.493	0.739	0.875	0.958	0.979	0.99	0.999	1
AGN	1.55	1.698	2.104	2.348	2.499	2.377	2.204	2.018	1.867	1.698	1.54
CRP	0.956	0.895	0.727	0.656	0.769	0.877	0.958	0.979	0.99	0.999	1

Probability of Reject H0: PRH
Average of Group Number: AGN
Correct Rate of Program: CRP
α=0.05, σ^2=1, K=5, m=20, r=1000

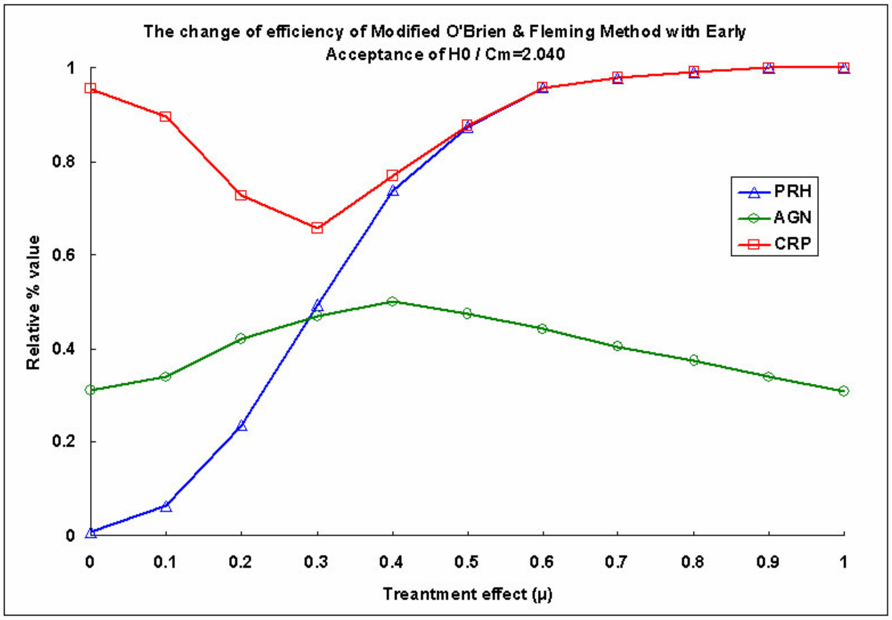

Figure 8
The change of efficiency of Modified O'Brien & Fleming's test of Early Acceptance of H_0. Cm=2.040

The overall efficiency of this method is not good when the treatment effect is small compared with original O'Brien & Fleming's test (Figure 5) and average probability test (Figure 7), because at small treatment effect range, the correct rate is low and the peak of expected sample group number locates at μ=0.4. This is reasonable because this test is designed for the large treatment effect. So we can still use it since the microarray data have $\mu>0.5/\sigma^2\approx0.1$, and we can easily change the boundary value to meet our goal. Further discussed will be provided in the next section.

If I increase the parameter Cm, which makes the boundaries larger, I obtain the following pattern. The detailed data are not listed since the figure gives adequate illustration.

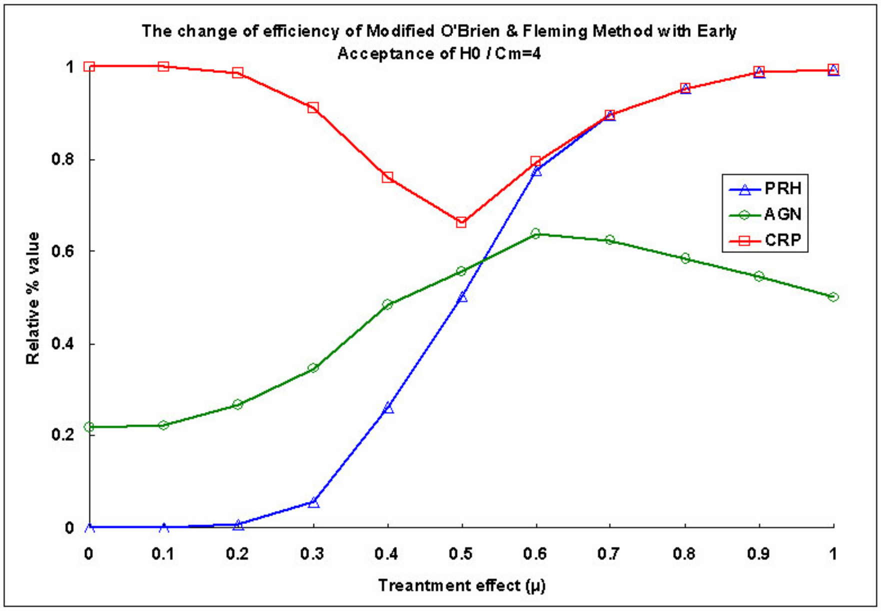

Figure 9

The change of efficiency of Modified O'Brien & Fleming's test of Early Acceptance of H_0. Cm=4.0

It is interesting that as Cm increases while the other parameters hold constant, the plot shifts to the right. This can be explained that increased boundary value can be used for larger treatment effect (the intermediate state shifts right for large μ values.

5. Method Application: Data Analysis of Microarray

In this section I used the above Modified O’Brien & Fleming’s test to analyze data from a microarray. The data are described in the following paragraph.

5.1 Analyze Data

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 this average exceeded 2,000 fluorescence units (scale 0–65,535 units) were included in the analysis. In addition, for all samples, it was enquired that 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 reason why I chose the Modified O’Brien & Fleming’s test is that if I used the algorithm of Average Probability Method with type I error rate α=0.05, I got this conclusion:

```
Probability of Reject H0: 1
Average of Group Number: 1
Correct Rate of Program: 1
Zmax: 28.52811
```

All of these genes are significant at 0.05 levels. I cannot find which one was more important, because even though I decrease α to 0.0001, the results were the same. If I use a very small $\alpha = e^{-20}$, the calculated a_k and b_k are “inf” which means they are beyond the ability of the R platform. We can see that the maximum calculated standard statistic Z is

28.52811, and if we use the computer program to calculate, the result $\Phi(28.5281)=1$. Actually any program will give the value $\Phi(Z>5)=1$. However, almost all the Z values for the 3389 genes are greater than 5 since the variance is very small compared with μ.

For these reasons, I chose the Modified O'Brien & Fleming's test with Cm=20, and unequal group size. I got the following results:

```
Reject H0 at Gene: 239       Z: -35.43014
Reject H0 at Gene: 252       Z: 27.1787
Reject H0 at Gene: 426       Z: 27.05163
Reject H0 at Gene: 902       Z: 28.52811
Reject H0 at Gene: 1355      Z: -27.40723
Reject H0 at Gene: 2463      Z: -31.40261
Reject H0 at Gene: 2473      Z: 24.73894
Reject H0 at Gene: 2617      Z: -28.20103
Reject H0 at Gene: 2628      Z: 25.31479
Reject H0 at Gene: 3154      Z: 24.07589
Reject H0 at Gene: 3345      Z: 27.75710
Probability of Reject H0: 0.00324
Average of Group Number: 1.012692
Correct Rate of Program: 0.9973436 Zmax: 28.52811
```

5.2 Microarray Data Results Discussion

1. The group sequential method with Modified O'Brien & Fleming's test can save the patient resource in great amount. The expected number of groups needed for this test is 1.01, which indicates that most of the trial can stop at the first interim analysis. Since the size for each group is 26,13,6,2,11, the trial stops after the first 26 patients samples. The small number of groups needed is also due to the first group containing the most samples in this study, and it can provide enough information for the statistical analysis.

2. Out of 3389 genes, eleven genes are selected as significant markers for breast cancer. The probability to reject H_0 is 0.324%. This gives us a relatively small set of genes we have to focus on. The names and functions of these genes obtained from

genbank (http://www.ncbi.nlm.nih.gov) are listed in the appendix 1. There are 7 genes found up regulated (positive Z value). They are HLA-DMB, HLA-B, FCGRT, COL4A2, RUFY3, FURIN, and DKFZp547O146. There are 4 genes found down regulated (negative Z value). They are RAE1, GSS, AURKA, and SET. Biological study with these genes associated with breast cancer might be important for future research.

6. Comparison Study of Different Group Sequential Test

6.1 Pocock and O'Brien & Fleming's Tests

As the following figure-10 shows, Pocock and O'Brien & Fleming's tests use different critical values for early rejection of H_0. In addition to this, Figures 2 and 5 show the difference between Pocock and O'Brien & Fleming's tests. For Pocock's test, the needed expected group number that will save our sample resource is low, but the correct rate at small treatment effect is not as good as O'Brien & Fleming's test. For the O'Brien & Fleming's test, the groups needed for analysis might be larger than Pocock's test, but we can get a better correct rate when the treatment effect is small.

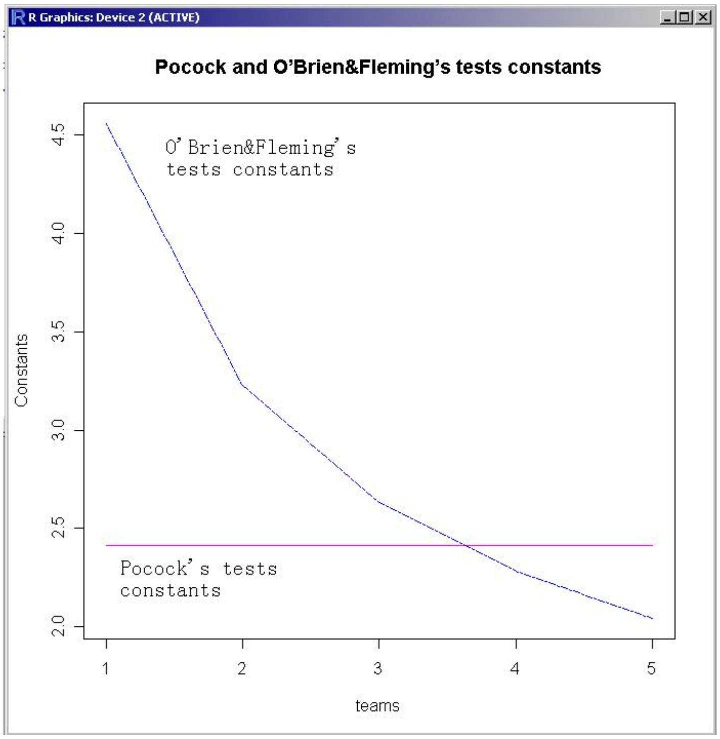

Figure 10

Comparison of Critical Values of Pocock and O'Brien & Fleming's tests

So based on the different purpose and property of our trials, we should select one of these two tests with caution.

6.2 Average Probability Method and Modified O'Brien & Fleming's Tests

By compare the Average Probability Method with Modified O'Brien & Fleming's Tests, I found that they are suitable for different data analysis purposes and they also apply better to different properties. The Average Probability Method is good for general analysis for study of treatment effect at a typical level (α=0.05, or α=0.01, etc.). When the data show great treatment effect and we need to focus on certain treatments, we can use the Modified O'Brien & Fleming's Tests with a carefully specified Cm values.

7. Conclusion

In this project, I successfully simulated the Pocock and O'Brien & Fleming's tests for group sequential methods with the R programming tool. Also I proposed two algorithms for group sequential tests with early acceptance of H_0 and simulated them with R. Comparisons were discussed for these tests based on the simulation results. The Modified O'Brien & Fleming's Test was used to research the data from a microarray paper with the purpose of finding the breast cancer marker genes. Eleven genes were found significantly important and the patient sample size was reduced to about 1 group of 26 patients with group sequential test from 5 groups of 58 patients. The completion of this project could 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.

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

Gould A, and Pecore V (1982). Group sequential methods for clinical trials allowing early acceptance of H0 and incorporating costs. Biometrika, 69(1) 75-80.

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.

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

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

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.

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.

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

Quackenbush J (2002), Microarray data normalization and transformation. Nat Genet. 2002 Dec;32 Suppl:496-501.

R project: http://www.r-project.org

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/

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

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

Appendix 1

List of Significant Genes from Modified O'Brien & Fleming's Test

No.239 RAE1 (RNA export 1, S.pombe) homolog - RAE1

Mutations in the Schizosaccharomyces pombe Rae1 and Saccharomyces cerevisiae Gle2 genes have been shown to result in accumulation of poly(A)-containing mRNA in the nucleus, suggesting that the encoded proteins are involved in RNA export. The protein encoded by this gene is a homolog of yeast Rae1. It contains four WD40 motifs, and has been shown to localize to distinct foci in the nucleoplasm, to the nuclear rim, and to meshwork-like structures throughout the cytoplasm. This gene is thought to be involved in nucleocytoplasmic transport, and in directly or indirectly attaching cytoplasmic mRNPs to the cytoskeleton. Alternatively spliced transcript variants encoding the same protein have been found for this gene.

No.252 major histocompatibility complex, class II, DM alpha - HLA-DMB

HLA-DMB belongs to the HLA class II beta chain paralogues. This class II molecule is a heterodimer consisting of an alpha (DMA) and a beta (DMB) chain, both anchored in the membrane. It is located in intracellular vesicles. DM plays a central role in the peptide loading of MHC class II molecules by helping to release the CLIP (class II-associated invariant chain peptide) molecule from the peptide binding site. Class II molecules are expressed in antigen presenting cells (APC: B lymphocytes, dendritic cells, macrophages). The beta chain is approximately 26-28 kDa and its gene contains 6 exons. Exon one encodes the leader peptide, exons 2 and 3 encode the two extracellular domains, exon 4 encodes the transmembrane domain and exon 5 encodes the cytoplasmic tail.

No.426. major histocompatibility complex, class II, DQ beta 1 - HLA-B

HLA-B belongs to the HLA class I heavy chain paralogues. This class I molecule is a heterodimer consisting of a heavy chain and a light chain (beta-2 microglobulin). The heavy chain is anchored in the membrane. Class I molecules play a central role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen. They are expressed in nearly all cells. The heavy chain is approximately 45 kDa and its gene contains 8 exons. Exon 1 encodes the leader peptide, exon 2 and 3 encode the alpha1 and alpha2 domains, which both bind the peptide, exon 4 encodes the alpha3 domain, exon 5 encodes the transmembrane region and exons 6 and 7 encode the cytoplasmic tail. Polymorphisms within exon 2 and exon 3 are responsible for the peptide binding specificity of each class one molecule. Typing for these polymorphisms is routinely done for bone marrow and kidney transplantation. Hundreds of HLA-B alleles have been described

No.902 Fc fragment of IgG, receptor, transporter, alpha - FCGRT

No.1355 glutathione synthetase - GSS

Glutathione is important for a variety of biological functions, including protection of cells from oxidative damage by free radicals, detoxification of xenobiotics, and membrane transport. The protein encoded by this gene functions as a homodimer to catalyze the second step of glutathione biosynthesis, which is the ATP-dependent conversion of gamma-L-glutamyl-L-cysteine to glutathione. Defects in this gene are a cause of glutathione synthetase deficiency.

No.2463 serine/threonine kinase 15 - AURKA

The protein encoded by this gene is a cell cycle-regulated kinase that appears to be involved in microtubule formation and/or stabilization at the spindle pole during chromosome segregation. The encoded protein is found at the centrosome in interphase cells and at the spindle poles in mitosis. This gene may play a role in tumor development and progression. A processed pseudogene of this gene has been found on chromosome 1, and an unprocessed pseudogene has been found on chromosome 10. Multiple transcript variants encoding the same protein have been found for this gene.

No.2473 collagen, type IV, alpha 2 - COL4A2

This gene encodes one of the six subunits of type IV collagen, the major structural component of basement membranes. The C-terminal portion of the protein, known as canstatin, is an inhibitor of angiogenesis and tumor growth. Like the other members of the type IV collagen gene family, this gene is organized in a head-to-head conformation with another type IV collagen gene so that each gene pair shares a common promoter.

No.2617 SET translocation (myeloid leukemia-associated) - SET

No.2628 KIAA0871 protein - RUFY3

No.3154 paired basic amino acid cleaving enzyme (furin, membrane associated receptor protein) - FURIN

The protein encoded by this gene belongs to the subtilisin-like proprotein convertase family. The members of this family are proprotein convertases that process latent precursor proteins into their biologically active products. This encoded protein is a calcium-dependent serine endoprotease that can efficiently cleave precursor proteins at their paired basic amino acid processing sites. Some of its substrates are: proparathyroid hormone, transforming growth factor beta 1 precursor, proalbumin, pro-beta-secretase, membrane type-1 matrix metalloproteinase, beta subunit of pro-nerve growth factor and von Willebrand factor. It is also thought to be one of the proteases responsible for the activation of HIV envelope glycoproteins gp160 and gp140. This gene is thought to play a role in tumor progression. The use of alternate polyadenylation sites has been found for this gene.

No.3345 hypothetical protein DKFZp547O146 - DKFZp547O146