Statistical Methodology 33 (2016) 71–82

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Symmetric directional false discovery rate control

Sarah E. Holte [a], Eva K. Lee [b], Yajun Mei [b,*]

[a] *Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA*
[b] *H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form
2 June 2016
Accepted 12 August 2016
Available online 24 August 2016

MSC:
62G10
62C25
62P10

Keywords:
Column permutation
Directional FDR
False discovery rate
Multiple testing
Symmetric decision
Three-decisions

ABSTRACT

This research is motivated from the analysis of a real gene expression data that aims to identify a subset of "interesting" or "significant" genes for further studies. When we blindly applied the standard false discovery rate (FDR) methods, our biology collaborators were suspicious or confused, as the selected list of significant genes was highly unbalanced: there were ten times more under-expressed genes than the over-expressed genes. Their concerns led us to realize that the observed two-sample t-statistics were highly skewed and asymmetric, and thus the standard FDR methods might be inappropriate. To tackle this case, we propose a symmetric directional FDR control method that categorizes the genes into "over-expressed" and "under-expressed" genes, pairs "over-expressed" and "under-expressed" genes, defines the p-values for gene pairs via column permutations, and then applies the standard FDR method to select "significant" gene pairs instead of "significant" individual genes. We compare our proposed symmetric directional FDR method with the standard FDR method by applying them to simulated data and several well-known real data sets.

* Corresponding author.
E-mail addresses: sholte@fredhutch.org (S.E. Holte), evakylee@isye.gatech.edu (E.K. Lee), ymei@isye.gatech.edu (Y. Mei).

http://dx.doi.org/10.1016/j.stamet.2016.08.002
1572-3127/

1. Introduction

This research is motivated from the analysis of a real gene expression data. As in the typical comparative genomics studies with high-throughput technologies, the data set we faced is from measuring the expression levels of $m = 54{,}675$ genes on $n = 16$ microarrays for two groups: $n_1 = 8$ healthy subjects and $n_2 = 8$ cancer subjects. The goal is to identify genes that are significantly differentially expressed between two groups with a potential of offering biomarker candidates.

Initially we thought this was a standard multiple hypothesis testing problem that often arises in modern biomedical applications such as genomic, proteomic, and metabolomic, and thus we blindly applied the standard false discovery rate (FDR) control method of Benjamini and Hochberg [2]: we calculated a two-sample t-statistic t_i for each gene i, permuted column data (randomly label cancer/normal subjects) to simulate the null distribution of the t-statistics, computed the corresponding two-sided p-values $p_i = \Pr_0\{|T| > |t_i|\}$'s for each gene i, and then used the standard Benjamini–Hochberg FDR method to select significant genes. However, when we reported the list of significant genes to our biology collaborators, they were suspicious, and felt the results did not make biology sense. We thought that this might be due to the simplicity of the Benjamini–Hochberg FDR method, and thus we re-analyzed data by applying more advanced FDR methods such as the robust FDR method of Benjamini and Yekutieli [3], the q-value of Storey [15,16], and the empirical Bayes estimate of the null distribution of Efron [4]. Unfortunately, our biology collaborators were still unsatisfactory to the results. After lengthy discussions, we realized that in our list of significant genes, we have selected ten times more negatively expressed genes than the positively expressed genes, but our biology collaborators preferred the list of significant genes to be balanced, since symmetry is common in many biology systems. More importantly, our biology collaborators did not use any specific biology knowledge to purposely choose negatively or positively expressed genes in the experiments.

It is natural to ask what happened to the data set we analyzed? Fig. 1 plots the histogram and QQ-norm plot of the observed t-statistics t_i's in our data set and both plots clearly suggest that the observed t_i's are highly skewed to negative and any normal distribution $N(\mu_0, \sigma^2)$ will likely be a poor approximation to the null distribution of t_i's. In other words, it is not clear how to estimate the null distribution $\Pr_0$ of t_i's for our data set. It is important to emphasize the role of the null distribution $\Pr_0$ of t_i's when genes are insignificantly differentiated expressed, since otherwise the corresponding p-values can be useless and thus the standard FDR methods are inappropriate. As mentioned in Efron [4], there are several methods to derive the null distribution of t_i's in the literature. The first one is the theoretical t-distribution under the assumption that the data x_{ij}'s are independent normally distributed, and this is often referred as the theoretical null distribution. The second method is data permutation methods by randomly labeling normal and cancer subjects and using the re-calculated t_i^* to simulate the null distribution. As pointed out in Efron [4], data permutation methods essentially approximate the null distribution of t_i's as $N(0, \sigma^2)$ after some suitable transformations, and do not help if the observed t_i's is not symmetric at 0. This view motivated Efron [4] to propose the third method that approximates the null distribution based on empirical Bayes: it is assumed that the null distribution is $N(\mu_0, \sigma^2)$ after transformations, where the null mean μ_0 is estimated from the observed t_i's that are likely from the null, say those between the first and third quartiles.

Unfortunately all these three existing approaches of estimating the null distribution of t_i's do not work in the case when the observed t_i's are asymmetric and highly skewed. One possible remedy is to extend the empirical Bayes method of Efron [4] by considering a mixture of normal or other distributions that can take into account the skewed or asymmetric properties of the observed t_i's. See, for instance, Zhao et al. [19] and Beana et al. [1], which applied the mixture distribution to address the skewness that is due to the non-null or significant genes. When the null distribution is skewed, one may still be able to use the mixture model to fit both null and non-null distribution of t_i's, but it is unclear how to classify the components of the mixture model between the null and non-null distribution. Moreover, such approach essentially assumes that the extreme behavior of the t_i's can be predicted based on the non-extreme values of t_i's, which is questionable or at least debatable.

In this article, we propose a novel FDR method that can circumvent the difficulty of estimating the null distribution of t_i's when they are highly skewed. Motivated by the rationale and remarks of our biologist collaborators, we note that the ultimate goal in FDR is not necessarily on estimating