Computers in Biology and Medicine 46 (2014) 1–10

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Empirical evaluation of consistency and accuracy of methods to detect differentially expressed genes based on microarray data

Dake Yang, Rudolph S. Parrish, Guy N. Brock*

Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40202, United States

ARTICLE INFO

Article history:
Received 3 June 2013
Accepted 3 December 2013

Keywords:
Microarrays
Differential gene expression
Moderated *t*-tests
Empirical study
Consistency of gene lists
Simulation study

ABSTRACT

Background: In this study, we empirically evaluated the consistency and accuracy of five different methods to detect differentially expressed genes (DEGs) based on microarray data.

Methods: Five different methods were compared, including the *t*-test, significance analysis of microarrays (SAM), the empirical Bayes *t*-test (eBayes), *t*-tests relative to a threshold (TREAT), and assumption adequacy averaging (AAA). The percentage of overlapping genes (POG) and the percentage of overlapping genes related (POGR) scores were used to rank the different methods on their ability to maintain a consistent list of DEGs both within the same data set and across two different data sets concerning the same disease. The power of each method was evaluated based on a simulation approach which mimics the multivariate distribution of the original microarray data.

Results: For smaller sample sizes (6 or less per group), moderated versions of the *t*-test (SAM, eBayes, and TREAT) were superior in terms of both power and consistency relative to the *t*-test and AAA, with TREAT having the highest consistency in each scenario. Differences in consistency were most pronounced for comparisons between two different data sets for the same disease. For larger sample sizes AAA had the highest power for detecting small effect sizes, while TREAT had the lowest.

Discussion: For smaller sample sizes moderated versions of the *t*-test can generally be recommended, while for larger sample sizes selection of a method to detect DEGs may involve a compromise between consistency and power.

1. Introduction

DNA microarrays have allowed investigators to compare gene expression values (measured as relative mRNA abundance) between two and more tissue samples for thousands of genes within the cell. However, because of the high dimensionality of the data with a relatively small number of replicates, microarrays have been referred to as 'An array of problems' [1]. Two of the main issues with microarray data are that they can be very noisy (both biological and technical noise), and that they contain a much larger number of mRNA expression measurements relative to the number of samples [2,3]. Hence, a central question concerning microarrays is the reproducibility of results from multiple studies of the same disease, in particular with regard to the lists of differentially expressed genes (DEGs) [4–6] and gene-sets for classification studies [7,8] that are found.

Recent studies have suggested that though the concordance between DEG lists from separate studies may be low, the false discovery rate of subsamples relative to the full data set tends to be low [9]. This suggests that each DEG list comprises mostly 'true' DEGs. This finding was reaffirmed when the correlation between DEGs was taken into consideration, in a study which investigated the consistency between DEG lists from two separate studies of the same disease [10]. Due to the high correlation between gene expression measurements, Zhang et al. [10] introduced a new measure for the concordance between two DEG lists, which took into account this correlation. Using the percentage of overlapping genes related (POGR) score, the authors demonstrated that while the DEG lists from two independent studies may not directly overlap with each other, each gene from one list is likely to be correlated with at least one gene from the second list.

An open problem not investigated by either of these two studies is the influence that the type of test statistic has on the reproducibility of the results. The *t*-test is a popular choice for detecting DEGs in microarray studies, and is well-known to have robust properties (e.g., to non-normality) when the sample size is sufficient (typically, $n \geq 25$). However, the traditional *t*-test has been documented to have problems in microarray studies, particularly for low expression levels when the sample size is small [11,12]. In this case, a gene with a low expression level but small variance can result in a large absolute *t*-statistic even when the

* Corresponding author.
E-mail addresses: d0yang03@louisville.edu (D. Yang), rudy.parrish@louisville.edu (R.S. Parrish), guy.brock@louisville.edu (G.N. Brock).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.12.002

Table 1
Characteristics of each of the five methods evaluated for determining DEGs in microarray data.

Method	Description
t-test	Ratio of difference in means divided by the standard error of the difference. Robust to violations of normality. May be sensitive to chance fluctuations in the estimate of variability.
SAM [11]	Moderated version of the *t*-test, which includes a positive constant in the denominator of the *t*-statistic as a stabilizing factor
eBayes [12,44]	Moderated version of the *t*-test similar to SAM, but based on an empirical Bayes approach which averages between the per-gene sample variance and a global (pooled) estimate of the variance
TREAT [16]	Extension to the eBayes method which tests whether differences in gene expression are above a given threshold. Essentially eliminates genes with low $\log_2$ ratios from the DEG list.
AAA [14]	Averages between a parametric (*t*-test) and a non-parametric (Wilcoxon rank-sum) test for differential expression.

mean difference in the expression level is small. These genes will be declared differentially expressed, even if the difference in expression is not biologically meaningful. Conversely, a gene with a high mean difference in the expression level may still result in a small *t*-statistic, if the estimate of the variance is unstable and unusually large (e.g., due to outliers).

Due to the huge data volume and inherent variation in microarrays, several statistical methods have been proposed to address these problems [11–13]. One of the earliest methods to appear was the significance analysis of microarrays (SAM) [11]. To solve the problem of unstable variances in gene expression measurements, SAM modifies the standard *t*-statistic by adding a small 'fudge factor' to the variance in the denominator. This modified test statistic is compared to an expected value under the null hypothesis, which is determined by permutations of the gene expression measurements. Differences between observed and expected values which are above a threshold are considered statistically significant, where the threshold is determined by the desired false discovery rate. Smyth [12] proposed a similarly derived empirical Bayes (eBayes) approach, which shrinks the estimated sample variances towards a pooled estimate. As an alternative to these two approaches, Pounds and Rai [14] proposed a method based on assumption adequacy averaging (AAA), which is robust to violations of normality. This approach incorporates an orthogonal test of normality of the gene expression measurements, and uses the resulting empirical Bayes posterior probability (EBP) estimate from this test to inversely weight the EBP values obtained from the *t*-test and the non-parametric Wilcoxon rank-sum test [15]. Lastly, as an extension to the empirical Bayes method of Smyth [12], McCarthy and Smyth [16] proposed to incorporate a biologically meaningful threshold into the test of differential expression (*t*-tests relative to a threshold, or TREAT). All of these approaches are designed to have robust performance, particularly for experiments with small numbers of arrays.

Although these methods have been proposed as improvements in detecting DEGs, few studies have empirically compared their performance [17]. Hence, in this paper we investigate both the consistency and power in determining DEGs between five different methods (traditional *t*-test, SAM, eBayes, AAA, and TREAT), based on three different empirical studies. In the first study we evaluate the effect of sample size reduction on maintaining a consistent list of DEGs using subsets from a single data set. In the second study, we evaluate the consistency for each method when comparing DEG lists obtained from two independent studies of the same disease. Lastly, we conduct a simulation study which evaluates the power and error rate of each method based on an approach which closely models the multivariate distribution of the original microarray data.

2. Methods

We performed three separate experiments to evaluate the consistency (reproducibility), sensitivity (power), and error rate (false discovery rate) of five different methods (*t*-test, SAM, eBayes, AAA, and TREAT) for determining DEGs in microarray data. The first experiment evaluated the self-consistency of each method based on subsets of the same microarray data set. The goal here was to evaluate how well each method performed at maintaining the same ordering of DEGs as the sample size is decreased. Secondly, we evaluated the consistency of DEG rankings for each method based on subsets of two different data sets concerning the same disease. The goal here was to evaluate whether certain methods outperformed others at maintaining a consistent list of DE genes across different studies of the same disease. Lastly, to ensure that the DEG lists returned by each method are appropriate and detecting truly DEGs, we conducted a simulation study to evaluate the power and false discovery rate of each method. Our simulation approach is based on a method for generating data which closely resembles the multivariate distribution of gene expression values observed in the original microarray data [18].

2.1. *Methods for detecting differentially expressed genes*

We selected five methods (*t*-test, SAM, eBayes, AAA, and TREAT) for determining DEGs in microarray data. In each case, the goal is to detect which genes are differentially expressed between two classes of samples (e.g., normal and diseased). A summary distinguishing characteristics of each of the methods is given in Table 1. Supplementary File A provides technical details concerning each of the methods, and references to software.

2.2. *POG and POGR scores*

To measure the consistency between two DEG lists, we use the percentage of overlapping genes (POG) metric and its extension to incorporate correlated gene expression changes, the POGR score [10]. The POG has previously been used to measure the reproducibility of DEG lists between different platforms [19], as well as from independent studies of the same disease [10]. The POGR score is a natural extension of the POG score which considers not only those genes which are shared between the two lists, but also those genes which are highly correlated with each other. The nPOG and nPOGR scores are normalized versions which account for the positive correlation of both scores with the length of the gene lists. They are analogous to the chance-corrected kappa coefficient [20]. A technical description of all the scores is given in Supplementary File A.

2.3. *Study design*

2.3.1. *Consistency of DE detection methods within a single data set*

Three data sets of different diseases were used to evaluate the consistency of methods to detect differential expression, based on subsets of the same data (Table 2). Before all the procedures we filtered the raw data to remove invariant transcripts, using the `nsFilter` function in the **genefilter** package [21]. Transcripts