International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

A Complete Summary of Non-Parametric Statistical Methods Used For Biological Microarray Data

Meenu Sharma, Rafat Parveen

Abstract*: Microarray technology is developed as a new powerful biotechnology tool, to analyze the expression profile of more than thousands of genes simultaneously. In recent times, Microarray is the most popular research topic. For extracting the differentially expressed genes from microarray data, numerous types of statistical tests are developed. The focus of microarray analysis is to predict genes that show different expression patterns under two different experimental conditions. The aim of this research paper is to explore various types of non-parametric methods proposed to analyze microarray expression data for predicting those genes which are differentially expressed, and a comparative analysis of various methods has been done. Besides, we also predicted the best condition for each method where they perform better and to investigate the disease development mechanism. Many types of statistical tests have been studied for identifying the differentially expressed genes, only very few studies have compared the performance of these methods. In our study, we extensively study and compare the different types of non-parametric methods.

Keywords*: About four key words or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

Microarray provides a way in biotechnology, to study the expression of thousands of gene expression profiles simultaneously [1]. Expression profiles of genes study are of relevance because it provides a way to identify or to predict disease markers that are having relevance in medical treatments. Many researchers want to identify genes which are differentially expressed in diseased condition [2]. The complete procedure used to analyze microarray to identify those genes which show differentially expression patterns across either between two different tissue samples or samples obtained under two different experimental conditions is shown in Fig. 1.The main aim of analyzing microarray data is to predict or identify which genes show differential expression across two different tissue samples or samples obtained under two different experimental conditions [1].

All statistical methods classified into two categories such as parametric methods and non-parametric methods. Parametric methods include t-statistics, Bayes t-test, B-statistics, ANNOVA-1 and Pearson's correlation coefficient, while non-parametric methods got very much attraction in this research field because of the availability of replicated data, this replicated data made difficult to get large samples. These non-parametric methods are significance analysis of microarray (SAM), SAMROC, this method uses very similar statistic to SAM (Significance analysis of microarray) is a non-parametric method uses similar statistic to t-statistic. SAMROC is a SAM with the use of receiver operating characteristics (ROC) curve, Zhao-Pan methods and SDEGRE [2]. Over the many years, numerous comparisons between statistical methods have been performed, to find which methods are most suitable to identify genes which show differential expression patterns. The main aim of this comparison is to screen out the best methods that exactly identify the highest proportion of differential expressed genes with having a small proportion of equivalently expressed (EE) genes which are falsely selected as differentially expressed [2].

Major problem found during the selection of genes are, we are only able to identify a less portion of genes which might have a significant effect in establishing any biological differences and less portion means less information and how many genes we have to select, who has any significant role in disease formation [3]. Analysis of differential expression has been done with the aim to identify those transcripts whose expression changes significantly between two tissue samples or different experimental conditions with respect to its mean and/or standard deviation. Under diseased conditions, many genetic factors (genes/miRNA) expressed abnormally or deregulated. This deregulation is influenced by many genetic and epigenetic factors. Experiments on microarray have been carried out to identify such gene expression disturbances under different experimental conditions. Differentially expressed genes are called significant or discriminative genes [4]. Identifying differentially expressed transcripts/genes in a small sample size from microarray data is very contradictory.

Manuscript published on November 30, 2019.
* Correspondence Author
Meenu Sharma*, Department of Computer Science, Jamia Millia Islamia, Jamia Nagar New Delhi 110025, India. Email: meenusharma2275@gmail.com
Dr. Rafat Parveen, Department of Computer Science, Jamia Millia Islamia, Jamia Nagar New Delhi 110025, India. Email: rparveen@jmi.ac.in

Any statistical test chosen inaccurately may result in wrong p-values. As a result, the identification of genes whose expression differs significantly under different experimental conditions may be incorrect and we also get a different list of differentially expressed genes by applying different statistical tests. Data distribution needs to take care of while during the selection of test statistics; this might have an effect on test performance [4].

The different statistical tests compared on the small sample size of microarray data are presented in Murie et al. [5]. Here, a comprehensive study of numerous parametric and non-parametric methods have been presented [1], [2], [3], [4], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], which are used for predicting genes which shows differentially expressed pattern from microarray expression profiles. A comparative study of the performance of several statistical methodologies have been provided here, based on simulated data sets to predict differentially expressed genes at cut off of p-values 0.05, and then a summary of all the statistical methods has been provided, based on their performance. A similar conclusion is drawn by Kim et al. [1] as of Jeffery et al. [14] such that the sample size, data distribution, variance assumptions, number of genes, number of samples, etc. have an impact on which test performs better.

In this paper, we compare non-parametric methods such as SAM, SAMROC, ZHAO-PAN METHOD, SDEGRE, Permutated t-test, Wilcoxon Rank Sum test, Modified Wilcoxon Rank Sum test, LIMMA, Shrink-t, and Soft threshold-t. In section 2, a review details of different statistical methods have been provided. A comparative study with an analysis of various statistical methods shown in Section 3. Finally, the conclusion presented in Section 4.

II. VARIOUS METHODS FOR PREDICTING DIFFERENTIALLY EXPRESSED GENES

In this section, we review various types of non-parametric statistical methods for identifying or predicting differentially expressed genes/ miRNAs from microarray experimental data for Homo sapiens. Non-parametric tests are also called as distribution-free methods because they do not assume any particular distribution data should follow. Most popular non-parametric tests used in a majority for predicting gene which shows differential expression patterns under different experimental conditions are given below.

A. SAM

To overcome the problem arises in t-statistics due to small variance, a similar statistical to t-test and for estimation of false discovery rate a permutation of repeated measurements have been utilized by SAM [1]. In case of less expressed genes, the computed t value may be high due to low s_i value (standard error). The shortcoming of t-statistic is that because of having a low value of both expression value and s_i, a high t-value is computed. Due to this reason, there is a high probability of counted this gene as the differentially expressed gene. Thus, to solve this problem SAM added a small positive constant called s_0 (also called "fudge factor"). SAM statistic is

$$t_{sam} = \frac{(\bar{x}_{i1} - \bar{y}_{i2})}{s_i + s_0} \quad (1)$$

Where, s_i stands for standard error found in groups mean such as $s_i = s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Where s stands for pooled estimate value for group standard deviation (i.e. $s = \sqrt{\frac{(n_1-1)*s_{i1}^2 + (n_2-1)*s_{i2}^2)}{(n_1+n_2-2)}}$. Here $(n_1 + n_2 - 2)$ is used to represent the degree of freedom (dof). With the assumption that the variance of the two groups is equal. The objective of setting the s_0 value is to make the variability of t_{sam} independent from s_i. It is achieved by calculating the t_{sam} variability as a function of s_i in the form of windows over the data. To evaluate the divergence MAD is used and data is divided into 100 windows of equal number size, by default. s_0 is selected for making the coefficient of variation $MAD_1, .. MAD_{100}$ small as small possible. The approach of estimating $s_0$is given in Algorithm 1.

Algorithm 1: A way of estimating s_0

1. calculate s_i values
2. calculate 100 quartiles of q_k of s_i values, where $k = 1,2, \ldots 100.$
3. for α = 0 to 1
4. compute $t_{sam}^{\alpha} = \frac{\bar{x}_{i1} - \bar{y}_{i2}}{s_i + s^{\alpha}}$, here s^{α} is the α-quantile of s_i values,
5. calculate $v_k^{\alpha} = 1.4826 * MAD\{t_{sam}^{\alpha} | s_i \in [q_{k-1}, q_k]\}$, here k=1,2,….100, where MAD is the median absolute distance.
6. now compute the coefficient of variance CV(α) of v_k^{α} values
7. $\alpha = \alpha + 0.05$
8. end for
9. select $\hat{\alpha} = argmin[CV(\alpha)]$ and $s_0 = s^{\hat{\alpha}}$. Argmin stands for an argument of the minimum. Returns value of α which minimizes CV(α) over the set of candidates for α as opposed to the minimum value itself.

B. SAMROC

A method has been proposed by Broberg [15] to provide a ranking to the genes according to their probability of being differentially expressed, referred to as SAMROC. This method is developed with the objective of estimating the false negative (FN) and false positive (FP) rates, and to reduce these errors. SAMROC statistics are very much identical with the SAM method, but only one difference is found instead of using a small positive constant this method uses different constant in the denominator. A used statistic is given below

$$t_{samroc} = \frac{(\bar{x}_{i1} - \bar{y}_{i2})}{s_i + s_{roc}} \quad (2)$$

The main attraction is to find the accurate value of s_{roc}constant for a given significance level of α. A criterion has been proposed by this method,

which is the distance of the curve from the origin, for selecting the best Receiver Operating Characteristic (ROC) curve. Users are able to compare the false positive error rate and a false negative error rate of various kinds of utilized statistical tests without using p-value. It further reduces the falsely declared positive or negative genes under a significance level of α and a small positive constant value s_{roc}.

C. Zhao-Pan method

A modified non-parametric approach has been adopted by Zhao and Pan [16], to identify the differentially expressed genes, from the expression data

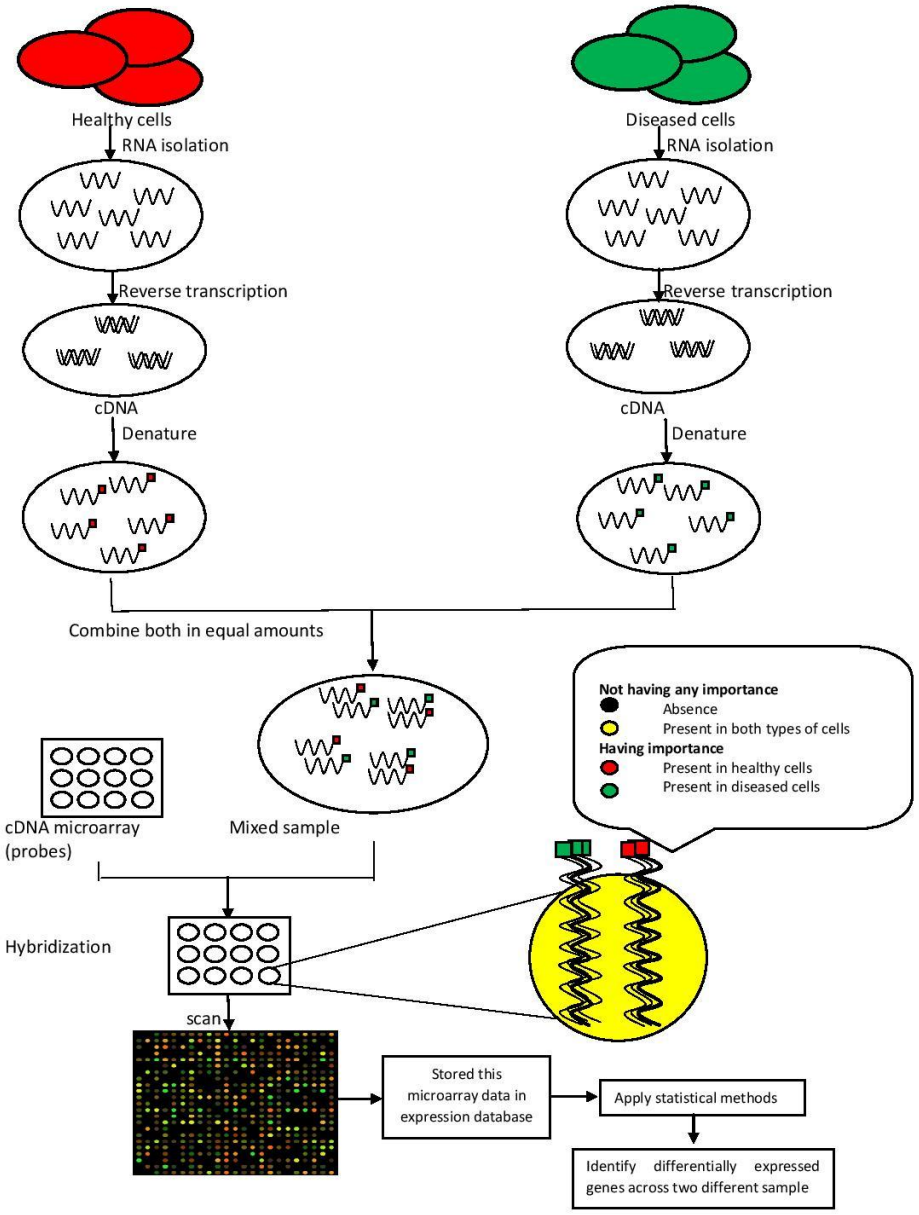

Figure 1: Methodology used for identifying differentially expressed genes.

developed from microarray experiments. The main motive of developing this method is in estimating the null distribution (it is the probability distribution of the test statistic when the null hypothesis is true) of test statistic like Z_i and via directly creating a null statistic, like z_i, in such a manner as the Z_i distribution is similar to the z_i distribution in the null hypothesis. This results in the removal of the solid assumption towards the null distribution of parametric methods. Through the previously described methods, such as empirical Bayes test, SAM method and MMM method, a common problem is found such as numerator and denominator of z_i and Z_i are considered as they are not dependent on each other. To overcome this problem, Zhao and Pan [16] use of z_i and Z_i which is given below

$$Z_i = \frac{\bar{y}_{i(a)} - \bar{y}_{i(b)}}{\sqrt{\frac{s^2_{i(a)}}{(k_1-k_2)} + \frac{s^2_{i(b)}}{k_2}}}, \quad z_i = \frac{\bar{y}_{i(a)} - \bar{y}_{i(c)}}{\sqrt{\frac{s^2_{i(a)}}{(k_1-k_2)} + \frac{s^2_{i(c)}}{k_2}}} \tag{3}$$

where

$$\bar{y}_{i(a)} = \frac{\sum_{p=1}^{k_1-k_2} y_{ip}}{k_1-k_2}, \qquad s^2_{i(a)} = \frac{\sum_{p=1}^{k_1-k_2} (y_{ip} - \bar{y}_{i(a)})^2}{(k_1-k_2-1)}$$

$$\bar{y}_{i(c)} = \frac{\sum_{p=k_1-k_2+1}^{k_1} y_{ip}}{k_2}, s^2_{i(b)} = \frac{\sum_{p=k_1-k_2+1}^{k_1} (y_{ip} - \bar{y}_{i(c)})^2}{k_2 - 1}$$

and

$$\bar{y}_{i(b)} = \frac{\sum_{p=k_1+1}^{k_1+k_2} y_{ip}}{k_2}, \quad s^2_{i(a)} = \frac{\sum_{p=k_1+1}^{k_1+k_2} (y_{ip} - \bar{y}_{i(b)})^2}{k_2 - 1}$$

Here, $k_1 and k_2$ represents the number of replicates under different experimental conditions. $\bar{y}_{i(a)}$ represents the mean of expression values under experimental condition a. $s^2_{i(a)}$ represents the variance of expression values under experimental condition a. At $k_1 > k_2$, $p = 1,2, \dots \dots k_1, k_1 + 1, \dots \dots, k_1 + k_2$.

D. SDEGRE

A new non-parametric method called SDEGRE [3] is developed based on relative entropy statistics and Kullback-Liebler distance calculated between two different groups of samples, for identifying differentially expressed genes. Relative entropy is the best way for measuring the difference between any two-distribution. Zero entropy value represents the two-identical distribution while large entropy value shows the more difference between two distributions. This method has the ability to identify each kind of distinction between the two groups of samples. This method does not need the data should follow any type of distribution or in other words, it is a distribution-free method. The noteworthiness of an individual gene, can be differentially expressed or not, it might be predicted using resampling based permutation. Suppose one gene expression in two experimental conditions, such as under the experimental condition 1 the observed expression levels are represented as $(s_1, s_2, \dots \dots s_m)$while under the experimental condition 2 the observed expression levels are represented by $(v_1, v_2, \dots \dots v_n)$ .To identify whether a specific gene is differentially expressed really or this is due to random fluctuation. The distribution density function of expression values under experimental condition 1 is denoted as $f(s)$ while under experimental condition 2 is denoted as $g(v)$. To take a decision whether this specific gene shows differential expression values or not, the hypothesis testing procedure is needed, such as

$H_0: f(s) = g$(v) and

$H_a: f(s) \neq g$(v)

The "Kullback-Liebler distance" between two distributions is given as,

$$H(f,g) = \frac{1}{2}\{\int f(t)\, log \frac{f(t)}{g(t)} dt + \int g(t) log \frac{g(t)}{f(t)} dt\} \quad (4)$$

Here, $f(s)$ and g (v) represents the density function. Kernel density estimation is used for computing the density function, described as

$$\hat{f}(s) = \frac{1}{nw_1} \sum_{j=1}^{n} \phi\left(\frac{s - s_i}{w_1}\right)$$

$$\hat{g}(v) = \frac{1}{nw_2} \sum_{j=1}^{n} \phi\left(\frac{v - v_i}{w_2}\right)$$

Here, ϕ(.) represents the density function of the standard normal distribution, m and n are the number of replicates under two experimental condition from $f(s)$ and $g(v)$ respectively, and w_1 and w_2 are the window width, both of these also control the quality of kernel estimation.

$$w_1 = C \sum_{i=1}^{n} \frac{d_i^h}{n},$$

$$w_2 = C \sum_{i=1}^{m} \frac{d_i^p}{m},$$

$$d_i^h = \min_{j \neq i} (|s_i - s_j|),$$

Where, and $d_i^p = \min_{j \neq i} (|v_i - v_j|)$ and C is a constant for control and we compute and then select it by comparing the results of the different choices.

E. Permutated T-test (Perm)

This test is also referred to as exact test, randomization and re-randomization test and comes under the category of t-test. This test involves the permutation or rearrangement of actual data points labels for every single gene. Usually, here 100 number of permutation is employed by default. Under the case of a high number of permutation than 1,000 then it needs to fix to 1,000. The next step is identical to t-statistics. To predict the confidence area, the inverse of this test is applied, and these confidences are also required further calculation. This permutation t-statistic [17] has more significance when we don't know about the distribution of data. This statistic has a similar problem as t-statistic has (called the "Behrens Fisher" problem).

F. Wilcoxon Rank Sum Test (RST)

The estimated result of t-statistic might not be reasonable for a non-normally distributed small size sample. Wilcoxon Rank Sum Test is another statistical method that can be used in a similar situation/condition. Because this method works on the rank transformed data, it is the most robust choice for computation on microarray expression data, which is commonly non-normal distributed or may have outliers. The first ranking of the combined sample has to be done, to perform the Rank Sum test. Then after this, a summation of ranks for the group a (such as $T_a = \sum ranks_{groupa}$) and summation of ranks for group b(such as T_b) is computed. If the number of samples in both groups are equal (e.g. n_a and n_b, where $n_a <= n_b$), then the RST statistic T is equal to the minimum value from T_a and T_b (such as $T = \min(T_a, T_b)$). If $n_a \neq n_b$ (unequal), then T_a has been calculated as the sum total value of the smallest sample ranks. Then, T_b is computed which is $T_b = n_a(n_a + n_b + 1) - T_a$. T is the minimum value from T_a and T_b. Small T value leads to the rejection of the null hypothesis, which means that the mean of the two samples is equal. The p-value for each and every gene has been calculated by $min(2 * min(T_1, T_2), 1)$ for a two-sided test. If n_a and n_b are large, then z-statistics is

$$z = \frac{(|T - mean_{w_a}| - 0.5)}{\sqrt{var_{w_a}}} \quad (5)$$

Where $var_{w_a} = n_b * \frac{mean_{w_a}}{6} = \frac{n_a * n_b * (n_a + n_b + 1)}{12}$ and $mean_{w_a} = \frac{n_a * (n_a + n_b + 1)}{2}$. Wilcoxon Rank Sum Test is the same as the Mann-Whitney test. T-test performs fast as compared to RST [4].

G. Modified Wilcoxon Rank Sum Test (RST)

In beginning, the ranks list has been created for the expression values of gene/miRNA for every gene/miRNA by modified RST throughout all the experiments in ascending order, and further do testing for the equality of means for the two ranked samples. In such a situation, where the number of samples under both the date sets/groups n_a and n_b are greater than 8, then normally approximated p-value used further (Walpole and Myers, 1993 [18]). As given in Troyanskaya et al. [19]: $z = \frac{(u_a - mean_{u_a})}{\sqrt{var_{u_a}}}$, where $var_{u_a} = n_a * n_b * \frac{(n_a | n_b | 1)}{12}$ and $mean_{u_a} = n_a + \frac{n_b}{2}$. Where, $u_a = T_a - n_a * \frac{(n_a | 1)}{2}$ and $T_a = \sum ranks_{group_a}$.

H. Linear Models for Microarray Data (LIMMA)

Under the situation of a few numbers of arrays and samples, result in variable standard deviation value for every single gene. While for large data, result in the problem of dimensionality. Limma test utilized empirical Bayes approach, to deal with such situation, implemented in Limma package of R [10], based on methods developed by Lonnstedt and Speed [1]. This empirical approach considers an assumption on the prior knowledge for the variance of unknown gene/miRNA data, such that variance of specific gene follows an inverse-gamma distribution (Γ^{-1} distribution).

Let, g represents the total gene number and k represents the total sample number. Then the linear model is as follows: $F(y_i) = X\alpha_i$, here y_i is used to represents the expression vector for every gene (i=1,….g) throughout the whole arrays or samples, X represents a full column rank design matrix and α_i is for coefficient vector. Few coefficients contrast are considered as relevant to biological interest for specific gene i, which is represented as $\beta_i = C^T \alpha_i$. This methodology is capable of comparing the number of properties without any limit, but the comparison is done only on two samples are compared so β_i is equal to the log fold change $\beta_i = (log\ \hat{x}_{1i} - log\ \hat{x}_{2i})$, denotes the mean of logged expression values of two groups. Contrast estimator $\hat{\beta}$ is assumed to follow a normal distribution; the residual sample variance (such as s_i^2) are to follow scaled chi-squared distribution (such as, χ^2_{dof}, and gene-specific variance follow an inverse-gamma distribution (Γ^{-1} distribution). Therefore, this statistic is written as: $\hat{\beta}_i | \sigma_i^2 \sim N(\beta_i \sigma_i^2)$, where N is used to represent it as normally distributed, $s_i^2 | \sigma_i^2 \sim \frac{\sigma_i^2}{d_i} \chi^2_{d_i}$ and $\sigma_i^2 \sim \Gamma^{-1}\left(\frac{d_0}{2}, \frac{d_0 s_0^2}{2}\right)$.

Posterior sample variance has been utilized by this model (such as, $\tilde{s}_i^2$) for gene i, it is $\tilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}$, here $d_0 (< \infty)$ and s_0^2 stands for the prior degree of freedom and variance respectively, and $d_i (< \infty)$ and s_i^2 represents the degree of freedom and the sample variance for specific gene i under an experimental condition respectively. In this case, two samples are used for comparing, so d_i is always equal to $n - 2$. The d_0 and s_0^2 are the two prior parameters used to represents the degrees of freedom and variance of a prior distribution respectively. Logged sample variance is fitted over the scaled F distribution for estimating the d_0 and s_0^2 parameters. Both of these parameters are calculated by making the empirical and expected values equal, under the two situations of two moments of $log(s_i^2)$ at the beginning. For any degree of freedom, the moments of $log(s_i^2)$ are finite, and distribution of $log(s_i^2)$ is more Gaussian than s_i^2, due to this reason $log(s_i^2)$ is most utilized than s_i^2.

Limma uses moderated t-statistic shown below

$$\tilde{t}_i = \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\hat{\beta}_i}{\tilde{s}_i} \quad (6)$$

The two-sample moderated t-statistic in the null hypothesis associated with a probability, which is estimated by reference to the t-distribution by using $d_0 + d_i$ as a degree of freedom.

Nonspecific filters might have an impact on the conditional distribution of test statistics. It is previously noted that the statistic used in Limma is based on the empirical Bayes approach which is used to model the gene-specific error variances (such as, $\sigma_1^2, \ldots . \sigma_m^2$) follow an inverse-gamma distribution. Inside Limma, the complete variance filtering is incorporated in the test statistics of Limma. Since correlation is found between the within-class variance estimator s_i^2 with overall variance, then the filtering on overall variance results in the depletion in the number of genes which have low s_i^2.

Limma includes both the moderated t-statistic and overall variance filtering, and it is utilized to identify the genes having a high variance from the microarray data. In this place, $\tilde{t}_i$ contracts the variance of the within-class estimate s_i^2 towards the common $\hat{s}_0^2$, which has its effect on genes that have large variance by decreasing the t-statistics denominator

value. Thus, it makes the t-statistics result bias by making it far from zero.

When genes have small variance, it leads to an increase in the t-statistics denominator, further it is equivalent t-statistics becomes bias almost close to zero. For this reason, variance filtering employed to remove low variance having genes.

Usually only two hyperparameters d_0 and s_0^2 are utilized, and both of these hyperparameters are calculated from the estimated standard gene-level variance s_i^2. The under dispersion of s_i^2 may result in the estimation of $\hat{d}_0^2$ which turn to be ∞. The filtering attached misfit of the conditional distribution of s_i^2 and Limma model, lead to the establishment of under dispersion. Main two reasons are there for estimating $\hat{d}_0 = \infty$. Firstly, it makes to removes completely the variance estimates of gene-level error, moderated t-statistics of every gene obtains the identical denominator, that leads to an investigation on the basis of fold change instead of t-statistics which is uncommon. Second, under the situation where both degree of freedom such as d_i and d_0 are less, but $\hat{d}_0 = \infty$ means that the correct null distribution to be heavy-tailed, yet inefficient comparison has been done between the moderated $\tilde{t}_i$ values and the standard normal/gaussian distribution. Due to this, the p-values are computed with the help of inefficient null distribution, identifying numerous true-null p-values which are approaches to zero (failure of control over the type-1 error). In this condition, the rejection of inverse gamma distribution found.

I. Shrink-t

Rein and Strimmer [18] created a moderated t-statistics on the basis of the James-Stein ensemble shrinkage estimation rule. After it is applied to the gene/miRNA variance estimators such as $v_1^2, \ldots\ldots, v_n^2$, the result we get from this rule in the form of the adjusted estimators which is represented as $\tilde{v}_i^2 = \hat{\lambda} v_0^2 + (1 - \hat{\lambda}) v_i^2$, here $\hat{\lambda}$ is referred to as the estimated pooling parameter, such as, $\hat{\lambda} = min \frac{(1, \sum_{i=1}^{n} var(v_i^2))}{\sum_{i=1}^{n} (v_i^2 - v_0^2)^2}$. The estimator v_0^2 is the median of $v_1^2, \ldots\ldots, v_n^2$, and $var(v_i^2)$ is calculated as

$$var(v_i^2) = \frac{(n_a + n_b)^3}{(n_a + n_b - 1)^3} \sum_{j=1}^{2} \sum_{k=1}^{n_j} \left(\frac{(x_{jki} - \bar{x}_{ki})}{n_a * n_b} - \frac{n_a + n_b - 2}{(n_a + n_b)^2} v_i^2 \right)^2 \quad (7)$$

Further, the shrink-t statistic can also have interpreted as: $\tilde{t}_i = \frac{D_i}{\tilde{v}_i^2}$, here D_i stands for the difference found between means of two groups.

J. Softthreshold-t (Soft-T)

Softthreshold test is also known as the L_1 penalized t-statistics which was devised by Wu [20]. Let us consider, for group a: number of treated samples are n_a, with mean $\bar{x}_{ai}$, having a standard deviation s_{ai} and its related degree of freedom $(dofa = (n_a - 1))$, and for group b: number of treated samples are n_b, with mean $\bar{x}_{bi}$, having a standard deviation s_{bi} and its related degree of freedom $(dofb = (n_b - 1))$. Now, for each gene/miRNA pooled standard deviation is $sab = \sqrt{\frac{(ssa + ssb)}{(dofa + dofb)} * \left(\frac{1}{(dofa+1)} + \frac{1}{(dofb+1)}\right)}$, where $ssa = rowsums((x_{ai} - \bar{x}_{ai})^2)$, and $ssb = rowsums((x_{bi} - \bar{x}_{bi})^2)$. Now, a shrinkage parameter denoted by Δ is computed as $\Delta = \bar{x}_{ai} - \bar{x}_{bi}$. The further numerator for this t-statistic is given below:

$$numerator = \begin{cases} \Delta - sgn(\Delta) * \lambda, & ifabs(\Delta) > \lambda, \\ 0, & ifabs(\Delta) <= \lambda, \end{cases}$$

Here, $sgn(.)$ is a sign function: $sgn(z) = 1$ if $z > 0$, $sgn(z) = -1$ if $z < 0$ and $sgn(z) = 0$ if $z = 0$. We also used to take $\lambda = 2 n_a m_i ab\lambda$ which represents a penalty factor, here $m_i = \sqrt{\frac{1}{n_a} + \frac{1}{(n_a + n_b)}}$. Now t-statistic is defined as:

$$t = \frac{numerator}{\sqrt{sab + \frac{\lambda^2}{(dofa + dofb)}}} \quad (8)$$

K. Other non-parametric tests

Kruskal and Wallis proposed a test which is named as Kruskal-Wallis test (KW test) [21]. This test comes under the category of non-parametric statistical methods, and its main objective is to identify whether the coming samples follow the same distribution or not. This test utilized to compare independent samples which are in more than two in numbers. Kruskal-Wallis test is similar in parameters as ANOVA-1. Ideal discriminator (also referred to as ID) method [19], it uses a resampling based approach. Ideal discriminator includes those genes/miRNAs which are maximally expressed in all the samples under experimental condition 1 and expressed minimally in all the samples under experimental condition 2. By this methodology, only those genes/miRNAs having the highest Pearson's correlation are being selected under the ID category. Then also validate their importance by comparing the Pearson correlation score of each and every gene/miRNA of real data with Pearson's correlation score computed from 50,000 data produced from data random permutation. Another useful nonparametric test is the Kolmogorov-Smirnov test (KS test) [22], this test employed for the continuous equality, one-dimensional probability distribution, which can be further employed to do a comparison between the samples of two populations. This Kolmogorov-Smirnov method computes the distance between sample empirical distribution function and the reference cumulative distribution function

or it can be in between the empirical distribution function of two samples. This method is too sensitive towards both the location and shape of two samples of cumulative distribution functions.

For large sample sizes, the p-value is very accurate, like $\frac{(n_a * n_b)}{(n_a + n_b)} >= 4$, here n_a and n_b are the sample sizes of two populations.

III. THEORETICAL COMPARISON

Remarkable improvement has taken place in the statistical tests used to study the microarray expression data with respect to the detection of differentially expressed genes. After analyzing all the statistical methods which are used to analyze microarray data, except non-parametric methods, a comparative Table I is prepared, which gives the summary of applied statistical methods, with their best performance condition, advantages, and limitations.

Table I: Comparison between different non-parametric statistical tests used to analyze microarray data.

Test/parameter	Sample size	Data distribution	Variance	Advantages	Limitations
SAM	Small	non-normally distributed	Equal variance	Avoids small variance problem. Uses permutation for correlations in genes and avoids parametric assumptions about the distribution of individual genes. Independence of genes correlates data with time.	Performs poorly when applied to the noisy dataset, uses permutation for correlations in genes and avoids parametric assumptions about the distribution of individual genes, not consistently performing well for small sample sizes, sample variance correction technique is not model-motivated
SAMROC	Small	Distribution-free	Equal variance	Enable to compare and minimize the FP and FN error rate of different test statistics without using P-values.	It performs poorly when data is skewed, SAM performs better than SAMROC when data follows a lognormal distribution (skewed data).
Zhao-Pan method	Large	NA	Equal variance	NA	Poor performance in the small sample size.
SDEGRE	Greater than 25	Distribution free	NA	This method is able to find some new differentially expressed genes which are not being found by any other methods.	It does not apply to the time series data. If a high number of permutation has been selected then it leads to an increase in computational cost.
Perm T-test	Both	non-normally distributed	NA	The goal is to get confidence intervals. Use it if 2 groups are distribution free. Simulate the null distribution repe:atedly randomly reassigning group labels.	Taking more time to compute test statistic than t-test, in case of normal distributions, it works poorly, especially for small sample sizes
Rank Sum	Large	non-normally distributed	NA	Robust for non-normal data having outliers. The difference between the 2 group's medians (or any other measure of location) is obtained by inverting the test.	Poor performance except for certain situations is likely to detect a location shift than two-sample t-stat
LIMMA	Both	Both distribution	Equal variance	Uses an empirical Bayes model for correcting sample variance. It can be used to compare two or more groups. It can be used for multifactorial designs (e.g. genotype and treatment).	It assumes that all the populations have the same standard deviation
Shrink-t	Small	Only work for non-normally distributions	Equal variance	Stabilizing t-test with very small replicates. Uses when the overfitting problem of 'large number of genes and a small number of samples' arises. Performs best for a lot of replicates of genes (except 2 replicates).	Its performance is not satisfactory in the case of 2 replicates of gene
Soft-Threshold-t aka L1 penalized t-statistic	Large (poor for all sample size as compared to others)	Poorer than others for both types of distributions	Estimators have low variance (main limitation)	Uses when the overfitting problem of 'large number of genes and a small number of samples arises.	Not useful for identifying differentially expressed gene/miRNA except certain situations, the weak consistency of penalized L1 estimators.

IV. CONCLUSION

An extensive review of numerous non-parametric statistical tests has been provided in this paper, which is used to predict genes having a different pattern of expression under two experimental conditions. Statistical test performance has been affected by a few conditions like the size of the sample, data distribution, variance of data, genes number, etc. So, for getting reliable and significant

results, we first require to get information about the real characteristics of the given dataset and then employ the correct appropriate statistical method on that particular dataset. The traditional statistical tests applied on microarray data to identify differentially expressed genes, or a clustering algorithm to identify the groups of genes that show similar behavior.

These methodologies are not successful for predicting the genes co-expressed differentially [23] which have significant biological roles. Statistical test which has better performance to identify genes which shows co-expressed differential patterns are not many, for this reason, further work is needed in this field.

ACKNOWLEDGMENT

I am thankful to Indian Council of Medical Research (ICMR), New Delhi, India for providing me funds for doing this research.

REFERENCES

1. S. Y. Kim, J. W. Lee, and I. S. Sohn, "Comparison of various statistical methods for identifying differential gene expression in replicated microarray data," Stat. Methods Med. Res., vol. 15, no. 1, pp. 3–20, 2006.
2. H. Andrew, G. Florence, and G. K. Bm, "Methods for Identifying Differentially Expressed Genes : An Empirical Comparison," vol. 6, no. 5, 2015.
3. X. Yan, M. Deng, W. K. Fung, and M. Qian, "Detecting differentially expressed genes by relative entropy," J. Theor. Biol., vol. 234, no. 3, pp. 395–402, 2005.
4. S. Bandyopadhyay, S. Mallik, and A. Mukhopadhyay, "A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data.," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 11, no. 1, pp. 95–115, 2013.
5. C. Murie, O. Woody, A. Y. Lee, and R. Nadon, "Comparison of small n statistical tests of differential expression applied to microarrays," BMC Bioinformatics, vol. 10, no. 1, p. 45, 2009.
6. W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," Bioinformatics, vol. 18, no. 4, pp. 546–554, 2002.
7. A. J. Vickers, "Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data," BMC Med. Res. Methodol., vol. 5, no. 1, p. 35, 2005.
8. J. Sreekumar and K. Jose, "Statistical tests for identification of differentially expressed genes in cDNA microarray experiments," Indian J. Biotechnol., vol. 7, no. October, pp. 423–436, 2008.
9. X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill, "Improved statistical tests for differential gene expression by shrinking variance component estimates," Biostatistics, vol. 6, pp. 59–75, 2005.
10. G. K. Smyth, "Statistical Applications in Genetics and Molecular Biology Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," vol. 3, pp. 2–3, 2011.
11. O. ElBakry, M. O. Ahmad, and M. N. S. Swamy, "Identification of differentially expressed genes for time-course microarray data based on modified RM ANOVA.," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 9, no. 2, pp. 451–66, 2012.
12. I. Lonnstedt and T. Speed, "Replicated microarray data," Stat. Sin., vol. 12, no. 12, pp. 31–46, 2002.
13. Y.-D. Tan, M. Fornage, and Y.-X. Fu, "Ranking analysis of microarray data: a powerful method for identifying differentially expressed genes.," Genomics, vol. 88, no. 6, pp. 846–54, 2006.
14. I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.," BMC Bioinformatics, vol. 7, p. 359, 2006.
15. P. Broberg, "Ranking genes with respect to differential expression," Genome Biol., vol. 3, no. 9, p. preprint0007, 2002.
16. Y. Zhao and W. Pan, "Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments," Bioinformatics, vol. 19, no. 9, pp. 1046–1054, 2003.
17. M. J. Anderson and C. J. F. T. E. R. Braak, "Permutation tests for multi-factorial analysis of variance," J. Stat. Comput. Simul., vol. 73, no. 2, pp. 85–113, 2003.
18. R. Opgen-rhein and K. Strimmer, "Statistical Applications in Genetics and Molecular Biology Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage," Stat. Appl. Genet. …, vol. 6, no. 1, 2007.
19. O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," Bioinformatics, vol. 18, no. 11, pp. 1454–1461, 2002.
20. B. Wu, "Differential gene expression detection and sample classification using penalized linear regression models," Bioinformatics, vol. 22, no. 4, pp. 472–476, 2006.
21. W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis Author (s): William H . Kruskal and W . Allen Wallis Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : http://www.jstor.org/stable/2280779 Accessed : 02," vol. 47, no. 260, pp. 583–621, 2016.
22. Kolmogorov, A. N., Sulla Determinazione Empirica di Una Legge di Distribuzione, Giornale dell'Istituto Italiano degli Attuari, 4. 83-91. 1933.
23. S. Bandyopadhyay and M. Bhattacharyya, "A biologically inspired measure for coexpression analysis," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 8, no. 4, pp. 929–942, 2011.

AUTHORS PROFILE

Meenu Sharma received her MSc and MPhil degrees in Bioinformatics from Jamia Millia Islamia, New Delhi, India in 2013 and 2016 respectively. She is currently pursuing a PhD from the same university. Her research interests include bioinformatics, data mining, computational biology, gene expression and systems biology.

Dr Rafat Parveen is an Associate Professor in the Department of Computer Science, Jamia Millia Islamia, New Delhi. She published 37 research articles and one book in national and international journals. Her research interests are computational biology and Bioinformatics, System Biology, cloud computing and cloud security.