Communications for Statistical Applications and Methods
2024, Vol. 31, No. 5, 571–584

https://doi.org/10.29220/CSAM.2024.31.5.571
Print ISSN 2287-7843 / Online ISSN 2383-4757

An approach based on clustering for detecting differentially expressed genes in microarray data analysis

Yuki Ando[1,a], Asanao Shimokawa[b]

[a]Department of Applied Mathematics, Tokyo University of Science, Japan;
[b]Department of Mathematics, Tokyo University of Science, Japan

Abstract

To identify differentially expressed genes (DEGs), researchers use a testing method for each gene. However, microarray data are often characterized by large dimensionality and a small sample size, which lead to problems such as reduced analytical power and increased number of tests. Therefore, we propose a clustering method. In this method, genes with similar expression patterns are clustered, and tests are conducted for each cluster. This method increased the sample size for each test and reduced the number of tests. In this case, we used a nonparametric permutation test in the proposed method because independence between samples cannot be assumed if there is a relationship between genes. We compared the accuracy of the proposed method with that of conventional methods. In the simulations, each method was applied to the data generated under a positive correlation between genes, and the area under the curve, power, and type-one error were calculated. The results show that the proposed method outperforms the conventional method in all cases under the simulated conditions. We also found that when independence between samples cannot be assumed, the non-parametric permutation test controls the type-one error better than the *t*-test.

Keywords: two group comparison, microarray data, DEGs, permutation test

1. Introduction

Genes are sometimes involved in diseases such as cancer. Therefore, it is medically essential to know which genes are involved in a disease, and this information can assist in the development of new drugs and treatment methods. However, there are tens of thousands of genes, and it is impossible to conduct experiments on each gene. Therefore, candidate genes involved in a disease are typically narrowed down using statistical analyses (Dudoit *et al.*, 2002). This method takes into consideration that the mere presence of a gene does not affect the organism. However, its expression reveals the information contained in the gene, and the information in the gene is predicted by measuring and analyzing the expression level, which is the numerical expression intensity. Differentially expressed genes (DEGs) are expressed in cells under various conditions (Pan, 2002). Because these DEGs can be regarded as candidates for genes involved in diseases, the detection of DEGs can narrow the list of genes. For example, DEGs whose expression levels differed between normal and cancer cells were likely to be cancer-related genes.

Genomic data were used to detect DEGs, and two data types are currently used. One was microarray data obtained from DNA microarray experiments, and the other was RNA-seq data obtained from

[1]Corresponding author: Department of Applied Mathematics, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan. Email:1422701@ed.tus.ac.jp

Published 30 September 2024 / journal homepage: http://csam.or.kr

next-generation sequencing experiments. Microarray data were used in this study. The expression level is defined as the fluorescence intensity of a fluorescent substance that shines more strongly when it is strongly expressed and is a continuous type of data (Draghici, 2012). These data roughly follow a logarithmic normal distribution and, when analyzed, a method that assumes a normal distribution with a logarithmic transformation (Holye *et al.*, 2002). However, the true distribution of microarray data is unknown because it has not been mathematically proven that the distribution of microarray data is lognormal. All data represent expression levels in matrices, with rows representing gene types and columns representing cell types, and differ only in the definition of expression levels. The data is characterized by a small sample size, whereas the dimensions of the data are very large because there are usually tens of thousands of gene types (Amaratunga *et al.*, 2014). This is because experiments to measure the expression levels require experimental animals such as rats and subjects that cannot be conducted in large quantities from an ethical point of view.

In this study, we focused on identifying the DEGs between the two groups. Examples of two-group comparisons include the cases mentioned above of normal cells and cancer cells, as well as cases of cells treated with drugs and cells not treated with drugs. Most of these are used to identify the causative gene of a particular disease or a gene affected by a specific substance. Outside of medicine, they are used to discover genes with different expression patterns between organisms, such as monkeys and chimpanzees. In addition, although not addressed in this study, it is sometimes used to compare expression levels in three or more groups of organisms (Churchill, 2004). This is the case when researchers want to compare the effects of multiple treatments or drugs for a particular disease or when they want to observe whether the expression levels change over time after the administration of a specific drug.

There are currently two primary methods for identifying DEGs using genomic data. One is based on fold change, and the other on testing. In a broad sense, fold change measures how much expression differs between two groups. The statistic that could be inferred to have a high probability of being a DEG with a large value was calculated, and the gene with a large value was designated as a DEG. Representative methods include the average difference (AD), weighted average difference (WAD) (Kadota *et al.*, 2008), and rank product (RP) methods (Breitling *et al.*, 2004). The AD and WAD methods utilize the difference between the mean expression levels of the two groups. The RP method focuses on the ratio of the expression levels of two groups and is known to be less affected by outliers than the AD and WAD methods. These methods are simple and easy to interpret, even without statistical knowledge; however, because they are not statistical methods, probabilities such as p values cannot be calculated, and it is difficult to determine which genes are DEGs. Another disadvantage of some methods is that they are vulnerable to outliers due to small sample sizes.

The test-based method performs statistical testing for each gene, and the gene with the most significant difference is designated as a DEG. Usually, genes are ranked by the p-value, and the genes with a small p-value are selected as DEG; and, because the p-value can be calculated, it is easier to determine the threshold to which a gene is designated as a DEG than the fold change-based method. As mentioned previously, because microarray data are often analyzed by assuming that they follow a normal distribution, the t-test (Welch, 1947) is often used. However, considering the sample size is too small to assume a distribution and the true distribution is unknown, the Wilcoxon rank-sum test (Wilcoxon, 1945) may be used. The disadvantages of test-based methods are their low accuracy, which is a result of the small sample size and multiplicity of tests. As for the multiplicity of tests, the type-one error can be reduced to some extent by adjusting the p-value using the Benjamini-Hochberg step-up method (Benjamini and Hochberg, 1995); however, this does not completely eliminate the type-one error.

To overcome these weaknesses, we propose a new clustering method. When a specific reaction or change occurs in a living organism, it may be accompanied by another change. In this case, when a gene controlling one of the changes was expressed, the gene that is controlling the other was also expressed. Therefore, many genes interact with each other. Because there are many genes whose functions are not yet known, even with current technology, this property may be used to predict the function of an unknown gene (Brown, 2002). Specifically, the expression levels of genes were measured under multiple conditions and clustered, and the genes showing similar expression patterns were identified. In this case, the genes classified in the same cluster can be interpreted as containing similar information. Based on this assumption, DEGs and non-DEGs were classified into different clusters when genes in the genome data were clustered. Therefore, DEGs can be detected by determining which cluster is a DEG cluster rather than by deciding whether each gene is a DEG. Consequently, we considered a method to determine whether each cluster was a DEG cluster using a test after clustering. By testing clusters instead of genes, the number of tests is reduced, and the sample size for each test is increased; thus, the accuracy is expected to be higher than that of the conventional method.

When a correlation exists between genes, independence between samples in the same cluster cannot be assumed. Because most of the tests currently used for two-group comparisons require the assumption of independence, it is necessary to consider how to deal with this problem. Instead of the t-test, we attempted to perform a nonparametric permutation test, which is considered suitable for the proposed method used for dealing with correlated data (Edgington and Onghena, 2014). We investigated the accuracy of the proposed method by comparing it with conventional methods through simulations. The methods compared were the t-test, the Wilcoxon rank-sum test, the proposed method using t-test, the Wilcoxon rank-sum test ignoring correlation, and the proposed method using a permutation test. For each of these methods, we calculated the accuracy when the data was generated by varying the correlation coefficients among the genes belonging to the same cluster, assuming that the genes formed several clusters. Other methods similar to the proposed method include gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005), significance analysis of microarray to gene-set analyses (SAM-GS) (Dinu *et al.*, 2007), and rotation gene set testing (ROAST) (Wu *et al.*, 2011). These are testing methods for gene sets. In the simulation, we compare the proposed method with these methods to see which one can detect DEGs with higher accuracy.

In the next section, we describe the proposed method in detail. Section 3 presents the setup and simulation results, and Section 4 summarizes the study.

2. Method using clustering

2.1. Procedure of the method

Let X_{ij} be the expression level of group 1 and $Y_{ij'}$ be the expression level of group 2 in gene i ($i = 1, \ldots, g$; $j = 1, \ldots, n_1$; $j' = 1, \ldots, n_2$). In the proposed method, g genes are first classified into K clusters. Let a_k be the total number of genes in the cluster $k (k = 1, \ldots, K)$. Then, the expression level of gene $l (l = 1, \ldots, a_k)$ belonging to cluster k in Group 1 is denoted by

$$X^k_{l1}, \ldots, X^k_{ln_1},$$

The expression levels in group 2 were

$$Y^k_{l1}, \ldots, Y^k_{ln_2}.$$

In this case, we use

$$X^k_{11}, \ldots, X^k_{1n_1}, \ldots, X^k_{a_k1}, \ldots, X^k_{a_kn_1},$$

and

$$Y^k_{11}, \ldots, Y^k_{1n_2}, \ldots, Y^k_{a_k1}, \ldots, Y^k_{a_kn_2}$$

as samples and a two-group comparison test was performed. If the expression levels of the two groups were regarded as significantly different, all the genes belonging to cluster k were identified as DEGs. This operation is performed for all k. In this study, the k-means method (Lloyd, 1982) and the Gaussian mixture model (GMM) (Banfield and Raftery, 1993) were used as clustering methods. The testing methods are discussed in Section 2.3. To implement the proposed method, it was necessary to specify the number of clusters at the time of clustering. Many methods have been proposed for estimating the number of clusters. Most are based on the information criterion, where the optimal number of clusters minimizes the information criterion, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). When using the proposed method, it is necessary to calculate these information criteria. If the actual number of clusters can be predicted, the proposed method can be executed using any number of clusters above the predicted number.

2.2. Clustering method

2.2.1. k-means method

The k-means method is one of the most representative nonhierarchical clustering methods. Because hierarchical clustering is known to be computationally expensive, this method was chosen for this study as it requires the clustering of a significant number of genes and is relatively computationally inexpensive. Let us consider assigning a binary indicator variable $u_{ik} \in \{0, 1\}$ to each data point $\boldsymbol{z}_i = (x_{i1}, \ldots, x_{in_1}, y_{i1}, \ldots, y_{in_2})$. This u_{ik} takes the value of 1 if gene i belongs to cluster k and 0 otherwise. The objective function J is defined as follows:

$$J = \sum_{i=1}^{g} \sum_{k=1}^{K} u_{ik} \|\boldsymbol{z}_i - \boldsymbol{c}_k\|^2,$$

where $\boldsymbol{c}_k = (c_{k1}, \ldots, c_{kn_1+n_2})$ is the center of cluster k, and $\|\cdot\|$ denotes the Euclidean norm. u_{ik} and $\boldsymbol{c}_k$ values that minimize J must be found. This can be achieved by repeating the following two steps: First, set the initial value of $\boldsymbol{c}_k$. In a typical k-means method, the initial value is determined randomly; however, another method is proposed to determine the initial value based on certain rules (Arthur and Vassilvitskii, 2007). Next, we minimize J for u_{ik} with $\boldsymbol{c}_k$ fixed. Then, u_{ik} is fixed, and J is minimized for $\boldsymbol{c}_k$. These two steps are repeated until convergence is achieved. Now, we consider the optimization of $\boldsymbol{c}_k$ with u_{ik} fixed. Because J is a function of $\boldsymbol{c}_k$, we find the minimum value by setting the partial derivative to zero as follows:

$$2 \sum_{i=1}^{g} u_{ik}(\boldsymbol{z}_i - \boldsymbol{c}_k) = 0.$$

To solve this, we obtain

$$\boldsymbol{c}_i = \frac{\sum_i u_{ik} \boldsymbol{z}_i}{\sum_i u_{ik}},$$

where $\boldsymbol{c}_k$ is interpreted as the average of all points belonging to cluster k.

2.2.2. GMM method

As mentioned above, microarray data are often tested by assuming a normal distribution. Therefore, we decided to use the normal distribution method for clustering. The GMM introduced in this section is a clustering method that uses a mixed normal distribution. The mixed normal distribution is a combination of several multivariate normal distributions and has the following probability density function:

$$f(z_i) = \sum_{k=1}^{K} \pi_k N\left(z_i \mid \boldsymbol{\mu}_k, \Sigma_k\right),$$

where $N(z_i|\boldsymbol{\mu}_k, \Sigma_k)$ is the probability density function of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and variance-covariance matrix Σ_k, π_k is the probability density of $0 \leq \pi_k \leq 1$, and $\sum_{k=1}^{K} \pi_k = 1$. In this case, the GMM method assumes that the data following a normal distribution with the same parameters belong to the same cluster. To find the clusters to which each z_i belongs, we assume that $P(u_{ik} = 1) = \pi_k$ for all i and calculate $P(u_{ik} = 1|z_i)$. From Bayes theorem, we obtain

$$\begin{aligned} P(u_{ik} = 1 \mid z_i) &= \frac{P(u_{ik})P(z_i \mid u_{ik} = 1)}{\sum_{m=1}^{K} P(u_{im})P(z_i \mid u_{im} = 1)} \\ &= \frac{\pi_k N(z_i \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m N(z_i \mid \boldsymbol{\mu}_m, \Sigma_m)}. \end{aligned}$$

The parameters $(\pi_k, \boldsymbol{\mu}_k, \Sigma_k; k = 1, and \cdots, K)$, which are necessary for the calculation were estimated using the maximum likelihood method. The log-likelihood function of the mixed normal distribution is as follows:

$$\log L\left(\pi_k, \boldsymbol{\mu}_k, \Sigma_k\right) = \sum_{i=1}^{g} \log \left\{ \sum_{k=1}^{K} \pi_k N(z_i \mid \boldsymbol{\mu}_k, \Sigma_k) \right\}.$$

When the partial differentiation is set to zero, we obtain

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{g} \gamma_{ik} z_i}{\sum_{i=1}^{g} \gamma_{ik}},$$

$$\Sigma_k = \frac{1}{\sum_{i=1}^{g} \gamma_{ik}} \sum_{i=1}^{g} \gamma_{ik} \left(z_i - \boldsymbol{\mu}_k\right)\left(z_i - \boldsymbol{\mu}_k\right)^T,$$

where

$$\gamma_{ik} = \frac{\pi_k N(z_i \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m N(z_i \mid \boldsymbol{\mu}_m, \Sigma_m)}.$$

This rate is known as the burden rate. As π_k is constrained to satisfy $\sum_{k=1}^{K} \pi_k = 1$ from the Lagrange multiplier, then

$$G = \log L\left(\pi_k, \boldsymbol{\mu}_k, \Sigma_k\right) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).$$

Differentiating G by π_k and λ and setting the derivative to zero, we obtain the following:

$$\pi_k = \frac{\sum_{i=1}^{g} \gamma_{ik}}{g}.$$

Because all the parameters calculated above are functions of the burden ratio, and the burden ratio also depends on the parameters, they cannot be calculated analytically. Therefore, they were computed using the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). In the EM algorithm, the initial values of the parameters are first determined. The burden ratio was calculated using the initial values. Finally, the parameters were recalculated using the burden rates. In this manner, the parameters and the burden ratio were alternately calculated, and the process was repeated until convergence was achieved. The calculation of the burden ratio is called the E-step, and that of the parameters is called the M-step. Unlike the k-means method, GMM does not strictly specify a single cluster to which each data point belongs because the output is the probability of belonging to each cluster. This clustering method is called soft clustering, whereas a clustering method that provides a binary output of belonging or not belonging to each cluster, such as the k-means method, is called hard clustering. However, due to the characteristics of the proposed method, when using GMM, the cluster with the highest probability of belonging was uniquely defined as the data cluster.

2.2.3. Clustering accuracy

If the data belonging to the same cluster are correlated, the higher the correlation coefficient, the better the clustering accuracy of the proposed method. The Euclidean distance between $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i'} (i \neq i')$ is considered. The expected value of the square of the difference in the first component of each vector is

$$\begin{aligned} E\left[(X_{i1} - X_{i'1})^2\right] &= E\left(X_{i1}^2 - 2X_{i1}X_{i'1} + X_{i'1}^2\right) \\ &= E\left[X_{i1}^2\right] + E\left[X_{i'1}^2\right] - 2\left\{\mathrm{Var}(X_{i1})\mathrm{Var}(X_{i'1})\rho(X_{i1}, X_{i'1}) + E[X_{i1}]E[X_{i'1}]\right\}. \end{aligned}$$

We can say that the larger the correlation coefficient, the smaller the expected distance between two points. In the k-means method, which performs clustering by Euclidean distance, and in GMM, data with similar numerical values are more likely to be classified into the same cluster, resulting in better clustering accuracy. Genes interact with one another to form complex networks. Cancer genes regulate the expression of other genes. Because several databases record which genes comprise the network, it may be possible to determine whether there are interactions between genes using these databases. This indicates that the proposed method can be applied to the data. DEG detection can be performed accurately if the proposed method can be applied to the data. This method can be used with both microarray and RNA-seq. However, it is the microarray that is easier to determine if it is operational. We now compare microarray and RNA-seq data. Microarray cannot detect unknown or rare transcripts. However, this means that many of the genes that can be detected by microarrays are genes that exist in the database. Therefore, it is often possible to know whether a gene forms a cluster by referring to the database. On the other hand, RNA-seq may contain genes that are difficult to determine if they form clusters or not.

2.3. Statistical testing

2.3.1. Choice of test

In the proposed method, because the expression levels of multiple genes obtained from the same cell were tested as samples, independence between the samples cannot be assumed if there is a relationship

between the genes. In most cases, a positive correlation was observed between the expression of these genes. One problem in this case is that the test statistic of the t-test becomes large because the sample variance decreases as the number of similar values increases. In other words, the difference between the groups was overestimated, which increased type-one errors. Therefore, it is necessary to transform the data or test the statistics to address these correlations. Currently, the following methods are used to test correlated samples: (1) Ignoring the correlation. (2) Summarizing the correlated data. (3) Correcting the test statistic of the t-test. (4) Performing a test that does not require random sampling.

When correlations are ignored, the t-test can be used to perform two-group comparisons (Galbraith *et al.*, 2010), which is often used in real analyses but cannot solve the problem. Data summarization is a method for generating independent data by determining a representative value, such as the sample mean for correlated data, and viewing it as a single sample (Virmani *et al.*,2006). However, because our motivation for proposing this method was to increase the sample size, we judged that a summary, which would reduce the sample size was not appropriate.

A correction for the t-test was proposed by Gonen *et al.* (2001). This method divides the test statistic by the number of correlated samples and the magnitude of the correlation. However, we believe that this method is not suitable for this study. This is because the corrected t-test assumes that the samples that are correlated with each other are known, such as when the samples are obtained from the same subject. Although it may be possible to predict microarray data by clustering or using genetic databases from previous studies, we decided not to use the corrected t-test to propose a method that can be used in a more general way.

Finally, we discuss the tests that do not require random sampling. As mentioned above, the problem of correlated data is caused by sample variance. Therefore, an increase in type-one errors can be prevented by conducting a test in which the test statistic is less affected by sample variance. Consequently, we consider a nonparametric permutation test that uses the difference in means. Nonparametric permutation tests do not require random sampling (Edgington and Onghena, 2014). However, some studies have shown that if the samples are correlated, the permutation changes the correlation structure, thus reducing the power of the test (Blair and Karniski, 1993). Therefore, we decided to compare the simulations in which the test is more suitable for microarray data, the permutation test controls the type-one error, and the t-test, which increases the type-one error but has no power problem.

2.3.2. Permutation test

The permutation test is a nonparametric test that calculates the p-value as the proportion of permuted data with a mean difference greater than that of the original data. Although the test statistic of the t-test may be used instead of the difference in means, we use the simple difference in means as an indicator to avoid the influence of sample variance. In the permutation test of the proposed method, the null hypothesis is that there is no difference in the population means, and we first compute the difference in means in cluster k as follows:

$$P = \frac{1}{a_k n_2} \sum_{l=1}^{a_k} \sum_{j'=1}^{n_2} Y_{lj'}^k - \frac{1}{a_k n_1} \sum_{l=1}^{a_k} \sum_{j=1}^{n_1} X_{lj}^k.$$

Then, $a_k n_1$ of the $a_k(n_1 + n_2)$ samples are selected as the expression values of Group 1, and those that are not selected are the expression values of Group 2. The difference in means was calculated again. This is repeated B times, and the calculated difference in the means is denoted by $P_1, \ldots, P_B$. B represents all combinations $\binom{a_k(n_1+n_2)}{a_k n_1}$ for a small sample size and any natural number for a large

Table 1: Results of calculating the AUC, detection power, and type-one error for each method for different correlation coefficients

Method	Evaluation indices	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
Welch t-test	AUC	0.703	0.704	0.703	0.701
	power	4.0×10^{-6}	8.3×10^{-6}	1.0×10^{-6}	1.1×10^{-6}
	type-one error	8.6×10^{-7}	8.6×10^{-7}	8.6×10^{-7}	8.5×10^{-7}
Wilcoxon rank-sum test	AUC	0.687	0.687	0.686	0.686
	power	0.000	0.000	0.000	0.000
	type-one error	0.000	0.000	0.000	0.000
k-means+Welch t-test	AUC	0.717	0.717	0.720	0.723
	power	0.865	0.867	0.872	0.879
	type-one error	0.672	0.673	0.674	0.680
GMM+Welch t-test	AUC	0.717	0.718	0.721	0.726
	power	0.867	0.871	0.878	0.883
	type-one error	0.673	0.675	0.680	0.686
k-means+Wilcoxon rank-sum test	AUC	0.714	0.714	0.716	0.720
	power	0.849	0.851	0.856	0.862
	type-one error	0.645	0.644	0.645	0.651
GMM+Wilcoxon rank-sum test	AUC	0.714	0.715	0.718	0.722
	power	0.853	0.857	0.862	0.867
	type-one error	0.646	0.648	0.652	0.658
k-means+Permutation test	AUC	0.773	0.775	0.782	0.785
	power	0.830	0.832	0.841	0.853
	type-one error	0.361	0.359	0.354	0.360
GMM+Permutation test	AUC	0.769	0.772	0.781	0.784
	power	0.830	0.834	0.846	0.856
	type-one error	0.362	0.361	0.356	0.363

sample size. In this case, the p-value of the permutation test is

$$\frac{\sum_{b=1}^{B} I(P_b > P)}{B}.$$

If the alternative hypothesis is that the population mean of Group 2 is larger, this p-value is used as is, whereas in the case of a two-tailed test, the p-value is doubled. The only assumption required to run the test is that the sample is permutable under the null hypothesis (Hayes, 1996).

3. Simulations

3.1. Simulation settings

In the simulations, we compared the proposed method with two conventional methods. As a common setup for all simulations, we used data with $g = 10000, n_1 = 3, n_2 = 3$. The number of DEGs was 3000; the genes formed several clusters, and the genes in the same cluster were positively correlated. The correlation coefficients ρ were assumed to be 0, 0.1, 0.5, or 0.9. The expression level of each gene followed a normal distribution, and the mean expression level was differentially expressed gene-by-gene, with values ranging from 3 to 15. In the DEG, the difference in the actual means was assumed to be 2. The number of clusters is arbitrarily set at 500 when we use the proposed method. This is the true number of clusters in the simulated data. When analyzing real data, estimation of the number of clusters is necessary. The number of simulations was set to 1000, and the area under the curve (AUC), detection power, and type-one error were used as accuracy evaluation indices. AUC is one of the evaluation criteria used to determine whether a data point is positive or negative. It represents the area under the curve plotted with the true positive rate on the vertical axis and the false

Table 2: Results of calculating the AUC, detection power, and type-one error for each method for different correlation coefficients when compared to the test for the gene set

Method	Evaluation indices	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
	AUC	0.648	0.645	0.637	0.632
k-means+GSEA	power	0.782	0.786	0.798	0.820
	type-one error	0.666	0.673	0.696	0.728
	AUC	0.654	0.648	0.636	0.632
GMM+GSEA	power	0.798	0.795	0.803	0.825
	type-one error	0.671	0.677	0.699	0.729
	AUC	0.691	0.692	0.698	0.705
k-means+SAM-GS	power	0.037	0.038	0.037	0.039
	type-one error	0.012	0.012	0.012	0.012
	AUC	0.697	0.699	0.706	0.711
GMM+SAM-GS	power	0.039	0.038	0.038	0.040
	type-one error	0.013	0.012	0.012	0.012
	AUC	0.691	0.692	0.698	0.711
k-means+ROAST	power	0.037	0.037	0.038	0.038
	type-one error	0.012	0.012	0.012	0.011
	AUC	0.697	0.699	0.707	0.711
GMM+ROAST	power	0.039	0.039	0.039	0.040
	type-one error	0.013	0.013	0.012	0.012
	AUC	0.773	0.775	0.782	0.785
k-means+Permutation test	power	0.830	0.832	0.841	0.853
	type-one error	0.361	0.359	0.354	0.360
	AUC	0.769	0.772	0.781	0.784
GMM+Permutation test	power	0.830	0.834	0.846	0.856
	type-one error	0.362	0.361	0.356	0.363

positive rate on the horizontal axis and takes values from 0 to 1. A value close to 1 indicates that the classification accuracy of the method is good, and the AUC approaching 0.5 indecates a completely random classification of positives or negatives. AUCs in this study are based on p-values. First, we make a ranking of genes by p-value. Then, the AUC is the area under the curve when the ROC curve is drawn with different thresholds for how many genes are determined to be DEGs. The value shown is the average of the AUCs for 1000 simulations. The significance level was set at 0.05 when used to obtain the power and type-one error.

3.2. Result and discussion

The simulation results are listed in Table 1. First, we calculated the AUC. The highest AUC is obtained by the method using the permutation test, followed by the method using the t-test after clustering, the method using the Wilcoxon rank-sum test after clustering, the method using only the t-test, and the method using only the Wilcoxon rank sum test. This indicates that the proposed method is more accurate than conventional methods when used to generate rankings. Specifically, we found that the t-test is more accurate than the Wilcoxon rank-sum test for both the proposed and conventional methods. This is because the true distribution is a normal distribution, and the parametric method fits better. The permutation test that outperforms the t-test is discussed later, but it appears that this result is due to the suppression of the type-one error. While the accuracy of the conventional method does not change significantly with a change in the correlation coefficient, the proposed method tends to improve its accuracy when the correlation coefficient is large. This is because the correlation improves the clustering accuracy. Among the clustering methods, k-means was slightly more accurate when the correlation was small, and GMM was slightly more accurate when the correlation was large.

Table 3: Results of calculating AUC, detection power, and type-one error for each method with different correlation coefficients when genes are correlated but do not follow the same distribution

Method	Evaluation indices	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
Welch t-test	AUC	0.705	0.703	0.698
	power	1.3×10^{-6}	1.0×10^{-6}	1.1×10^{-6}
	type-one error	1.57×10^{-7}	1.57×10^{-7}	2.57×10^{-7}
Wilcoxon rank-sum test	AUC	0.688	0.686	0.684
	power	0.000	0.000	0.000
	type-one error	0.000	0.00	0.000
k-means+Welch t-test	AUC	0.718	0.716	0.716
	power	0.866	0.870	0.874
	type-one error	0.673	0.675	0.680
GMM+Welch t-test	AUC	0.717	0.708	0.698
	power	0.870	0.875	0.874
	type-one error	0.674	0.683	0.691
k-means+Wilcoxon rank-sum test	AUC	0.715	0.713	0.714
	power	0.850	0.854	0.857
	type-one error	0.644	0.646	0.652
GMM+Wilcoxon rank-sum test	AUC	0.714	0.706	0.696
	power	0.856	0.859	0.857
	type-one error	0.647	0.656	0.665
k-means+Permutation test	AUC	0.774	0.779	0.780
	power	0.832	0.838	0.845
	type-one error	0.360	0.356	0.359
GMM+Permutation test	AUC	0.771	0.777	0.775
	power	0.834	0.841	0.846
	type-one error	0.363	0.360	0.366

Next, we consider detection power and type-one errors. First, both the detection power and type-one errors of the conventional method were minimal. This may be because the p values are generally large and there are almost no null hypotheses that can be rejected. This indicated that the conventional method did not function as a classifier when the significance level was 0.05 or lower. The proposed method using the t-test had the highest power, followed by the proposed method using the Wilcoxon rank-sum test, and finally, the proposed method using the permutation test. However, the type-one error showed the opposite trend, especially when the permutation test was much smaller than the other two. This indicates that the p values of the proposed methods using the t-test and Wilcoxon rank-sum test are generally smaller and that the proposed method detects an excessive number of DEGs. The large type-one error of the proposed method, even when the correlation coefficient was zero, was attributed to the loss of accuracy caused by clustering errors. However, the permutation test had reasonable accuracy. This tendency increased as the correlation coefficient increased, and it is recommended to use a permutation test to avoid excessive detection of DEGs. Overall, the GMM tends to have higher detection power and type-one errors than the k-means method.

Table 2 shows the results of comparing the proposed method with the test for gene sets. The clusters formed by k-means or GMM were considered as gene sets. First, the AUC values for all methods were above 0.6, indicating that all methods functioned as a method to determine the DEG. Among them, the proposed method had the highest AUC. Next to that, SAM-GS and ROAST had good accuracy. GSEA tended to have both high detection power and type-one error. On the other hand, both SAM-GS and ROAST were low. This indicates that the p-values calculated by GSEA often take small values, while SAM-GS and ROAST often take large values, and thus are not suitable for testing at a significance level of 0.05 under this sample size. The proposed method produces the

Table 4: Type-one error in the proposed method using k-means and the permutation test with increasing sample size under various correlation coefficients

Sample size	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
3 : 3	0.361	0.359	0.354	0.360
10 : 10	0.168	0.187	0.219	0.261
20 : 20	0.115	0.116	0.114	0.115
30 : 30	0.046	0.057	0.075	0.110

most stable accuracy.

Table 3 shows the simulation results when genes are correlated but do not follow the same distribution. More specifically, half of the genes in Cluster 1 followed Distribution 1, and the other half followed Distribution 2. Similarly, half of the genes in Cluster 2 followed Distribution 2, and the other half followed Distribution 3. Finally, half of the genes in Cluster 3 followed Distribution 3, whereas the other half followed Distribution 1. It can be seen that the accuracy of the conventional method does not change significantly. Because this method considers each gene individually, it is believed that there is no effect on the correlation structure. The results of the other methods were not significantly different; however, the AUC was generally lower. This may be because the correlation structure is no longer simple and the clustering accuracy decreases. In addition, unlike the previous simulation, an increase in the correlation coefficient did not necessarily improve the accuracy in terms of AUC, detection power, and type-one error. This is also due to changes in the correlation structure. However, the proposed method using a permutation test is still recommended for DEG detection.

Although the type-one error when using the permutation test was smaller than that of the t-test, it was still considerable. Therefore, to examine whether the use of the permutation test was appropriate, we checked whether the type-one error decreased when the sample size was increased. Table 4 lists the type-one errors for the proposed method using the permutation test as the sample size increases. The results show that type-one error decreases as the sample size increases for all correlation coefficients. It can also be seen that the type-one error tends to increase as the correlation coefficient increases, but it approaches 0.05 as the sample size increases.

4. Example

In the actual data analysis, we applied only the t-test and the method using k-means and permutation tests, which were determined to be the most useful in terms of the AUC in the simulation, to the observed microarray data in order to examine the effect of human X-box binding protein-1 (XBP1) on gene expression (Gomez *et al.*, 2007). Although the actual DEG number is unknown, by ranking the genes by p values, we calculated the number of genes commonly identified as DEGs when the top 1000, 3000, and 5000 genes out of the 22,283 genes were considered DEGs. The sample size was three for each group, and the number of clusters for the proposed method was 1000. The results are summarized in Table 5.

Less than half of all DEGs were determined to be DEGs using each method. Additionally, the same genes were more likely to be identified as DEGs when the threshold for identifying them as DEGs increased. We concluded that the number of genes detected by the proposed method significantly differed from that detected by the conventional method. This difference should be considered when performing detection. Another interpretation is that genes commonly identified as DEGs using these two methods have a very high probability of being DEGs. If the researcher wishes to narrow the number of DEG candidates to a more precise and smaller number, this may be accomplished by checking the results of multiple methods.

Table 5: Number and percentage of genes commonly identified as DEGs in each method

Number of DEGs	Number of common DEGs	Number of common DEGs/number of DEGs
1000	149	0.149
3000	1079	0.360
5000	2122	0.424

5. Conclusion

We proposed a new method to analyze microarray data to detect DEGs. The proposed method is an improvement over the test-based method. It can solve the drawbacks of the conventional method, such as low accuracy due to small sample size and multiplicity of tests. The simulation of the proposed method was performed under the assumptions that the expression levels followed a normal distribution and that there was a positive correlation among the genes. In addition, we conducted simulations for the case in which some genes were correlated but followed different distributions.

The results showed that all methods could detect DEGs in terms of AUC; however, it was necessary to use the proposed method because the conventional method did not work when the significance level was set at 0.05. Among the proposed methods, those using the t-test or the Wilcoxon rank-sum test have a high type-one error; therefore, the method using the permutation test is recommended. Because the method using the permutation test is less affected by sample variance, the p-value is less likely to become small, even when the assumption of independence is not valid, which is considered the reason why the type-one error can be controlled. When the accuracy was examined by changing the correlation coefficient, it was found that the stronger the correlation, the higher the accuracy of the proposed method, which can be attributed to the improvement in clustering accuracy.

The simulations showed that the method using the permutation test is the most recommended, even when genes are correlated but follow different distributions. However, unlike the aforementioned simulations, the accuracy did not increase as the correlation coefficient increased. This indicates that, even if the correlation becomes more robust, the accuracy of clustering is not likely to improve if the distribution of the samples is different. However, because such a situation may occur in actual microarray data, it is necessary to consider ways to cope with it. One way to address this issue is to change the clustering method. By clustering using correlation coefficients, it was possible to divide genes that influenced each other into the same cluster.

The actual data analysis showed that the genes detected by the proposed method significantly differed from those seen by the conventional method. This indicates that the genes likely to be detected using the proposed method differ from those detected using the conventional method. Unlike simulation data, real data do not always have normality, and the sample size is often small; therefore, it is difficult to determine whether the conventional t-test can be applied. However, the k-means method and permutation test can be considered in many cases because of their loose assumptions. Thus, the proposed method is easy to use.

The proposed method has room for improvement, as various testing and clustering methods can be applied. In this study, we applied tests that are generally available for a variety of data, but there are several tests for genes, such as the sequence kernel association test (SKAT) (Michael *et al.*, 2011) and the adaptive sum of powered score (aSUP) (Pan *et al.*, 2014). Further improvement in accuracy may be expected by incorporating these tests.

References

Amaratunga D, Cabrera J, and Shkedy Z (2014). *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*, Wiley, New Jersey.

Arthur D and Vassilvitskii S (2007). k-means++: the advantages of careful seeding, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027–1035.

Banfield J and Raftery A (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–321.

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series b-methodological*, **57**, 289–300.

Blair RC and Karniski W (1993). An alternative method for significance testing of waveform difference potentials, *Psychophysiology*, **30**, 518–524.

Breitling R, Armengaud P, Amtmann A, and Herzyk P (2004). Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS Letters*, **573**, 83–92.

Brown TA (2002). *Genomes*, Wiley, New Jersey.

Churchill GA (2004). Using ANOVA to analyze microarray data, *Biotechniques*, **37**, 173–175.

Dempster A, Laird N, and Rubin D (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series b-methodological*, **39**, 1–38.

Dinu I, Potter J, Mueller T *et al.* (2007). Improving gene set analysis of microarray data by SAM-GS, *BMC Bioinformatics*, **8**, 1–13.

Draghici S (2012). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*, CRC Press, New York.

Dudoit S, Yang YH, Callow MJ, and Speed TP (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111–139.

Edgington S and Onghena P (2014). *Randomization Tests*, CRC Press, Florida.

Galbraith S, Daniel JA, and Vissel B (2010). A study of clustered data and approaches to its analysis, *Journal of Neuroscience*, **30**, 10601–10608.

Gomez BP, Riggins RB, Shajahan AN *et al.* (2007). Human X-box binding protein-1 confers both estrogen independence and antiestrogen resistance in breast cancer cell lines, *The FASEB Journal*, **21**, 4013–4027.

Gonen M, Panageas KS, and Larson SM (2001). Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient, *Radiology*, **221**, 763–767.

Hayes AF (1996). Permutation test is not distribution-free: Testing H_0:$\rho = 0$, *Psychological Methods*, **1**, 184–198.

Holye D, Rattray M, Jupp R, and Brass A (2002). Making sense of microarray data distributions, *Bioinformatics*, **18**, 576–584.

Kadota K, Nakai Y, and Shimizu K (2008). A weighted average difference method for detecting differentially expressed genes from microarray data, *Algorithms for Molecular Biology*, **3**, 1–12.

Lloyd S (1982). Least squares quantization in PCM, *IEEE Transactions on Information Theory*, **28**, 129–137.

Pan W (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics*, **18**, 546–554.

Pan W, Kim J, Zhang Y, Shen X, and Wei P (2014). A powerful and adaptive association test for rare

variants, *Genetics*, **197**, 1081–1095.
Subramanian A, Tamayo P, Mootha V *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression, *PNAS*, **102**, 15545–15550.
Virmani T, Atasoy D, and Kavalali ET (2006). Synaptic vesicle recycling adapts to chronic changes in activity, *Journal of Neuroscience*, **26**, 2197–2206.
Welch BL (1947). The generalization of student's problem when several different population variances are involved, *Biometrika*, **34**, 28–35.
Wilcoxon F (1945). Individual comparisons by ranking methods, *Biometrics*, **1**, 80–83.
Wu D, Lim E, Vaillant F, Asselin-Labat M, and Visvader J, and Smyth GK (2010). ROAST: Rotation gene set tests for complex microarray experiments, *Bioinformatics*, **26**, 2176–2182.
Wu MC, Lee S, Cai T, Li Y, Boehnke M, and Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test, *The American Journal of Human Genetics*, **89**, 82–93.

Received March 14, 2024; Revised May 18, 2024; Accepted August 03, 2024