Indian Journal of Agricultural Sciences **90** (10): 1975–9, October 2020/Article

Mixture distribution approach for identifying differentially expressed genes in microarray data of *Arabidopsis thaliana*

ARFA ANJUM[1], SEEMA JAGGI[2], ELDHO VARGHESE[3], SHWETANK LALL[4], ANIL RAI[5], ARPAN BHOWMIK[6]*, DWIJESH CHANDRA MISHRA[7] and SARIKA[8]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

Received: 29 January 2020; Accepted: 30 January 2020

ABSTRACT

The basic aim of analyzing gene expression data is to identify genes whose expression patterns differ in the treatment samples, with respect to the control or healthy samples. Microarray technology is a tool for analyzing simultaneous relative expression of thousands of genes within a particular cell population or tissue in a single experiment through the hybridization of RNA. Present paper deals with mixture distribution approach to investigate differentially expressed genes for sequence data of *Arabidopsis thaliana* under two conditions, salt-stressed and control. Two-component mixture normal model was fitted to the normalized data and the parameters were estimated using EM algorithm. Likelihood Ratio Test (LRT) was performed for testing goodness-of-fit. Fitting of two-component mixture normal model was found to be capable of capturing more variability as compared to single component normal distribution and was able to identify the differentially expressed genes more accurately.

Key words: Differential gene expression, Microarray, Mixture distribution, Normal distribution

Differential Gene expression (DGE) provides the power to understand the biological variations between two different conditions or states like healthy or diseased, treated or control etc. Genes identified from DGE analysis are known as differentially expressed genes (DEGs) that are responsible for different expressions than rest of the genes in genome. In clinical research, DEGs are important to identify candidate biomarkers and therapeutic targets for drug designing.

Methods available for generating the expression data are DNA Microarray, RNA seq, Chip Seq etc. For RNA seq data, distributional approaches have been applied for expression analysis (Marioni *et al*. 2008, Mortazavi *et al*. 2008, Nagalakshmi *et al*. 2008, Anders and Huber 2010, Anjum *et al*. 2016). Other than RNA seq, microarray data is most commonly used for transcriptome/expression analysis. Microarray technique is a powerful technique that increases the speed at which differentially expressed genes are analysed and to determine its function. This technique is used for comparing the expression level of thousands of genes at a time.

The most common methods for microarray data analysis are clustering and heatmap approach (Brazma and Vilo 2000). Other than this, statistical techniques like t-test, multiple hypothesis testing, Baye's method have also been used on microarray data (Jeffery *et al*. 2006). Mixture distribution approach is another technique that can be applied on this data for expression analysis as within a whole data set there are different subsets that possess different properties that can be modelled separately. To know statistically significant evidence that any of the genes under study possesses a difference in expression across the groups/conditions/subpopulations is the main concern**.** The theory of mixture distribution model can be an effective tool in such situations.

A mixture distribution is a mixture of statistical distributions with a different probability density function in each component. This distribution is used in the situation when a population (complete set of genes) has subpopulations (like, up-regulated and down-regulated genes). Here components of mixture probability density are the densities of the subpopulations along with the weights as the proportion of each subpopulation in the overall population (Karim *et al*. 2011). Mixture model has become popular because they provide a simple mechanism to incorporate extra variation and correlation in the model along with model flexibility (McLachlan and Peel 2000, Yang *et al.* 2007).

[1]Ph D Scholar (anjum.arfa@gmail.com), [2]Head (DE) (Seema.Jaggi@icar.gov.in), [3]Scientist (Eldho.Varghese@icar.gov.in), [4]Ph D Scholar (shwetanklall@gmail.com), [5]Head (CABIN) and ADG (ICT) (Anil.Rai@icar.gov.in), [6]*Scientist and corresponding author (Arpan.Bhowmik@icar.gov.in), [7]Scientist (Dwijesh.Mishra@icar.gov.in), [8]Senior Scientist (sarika@icar.gov.in)

Pearson (1895) studied mixture distribution by mixing of different crab species and modelled mixture of two normal distributions and found that about 28% of genes appear to have an expression pattern that follows a mixture distribution. A mixture analysis approach was introduced by McLachlan *et al*. (2002) to the clustering of microarray expression data with respect to tissue samples on a very large number of genes.

In this article, mixture distribution approach is applied to microarray data of *Arabidopsis thaliana* for performing differential expression analysis. Joint likelihood density function is obtained and the parameters of the mixture model including the mixing weights (mixing proportions) are estimated. The performance of the mixture distribution model is compared with single distribution model. Further, R codes have been developed for fitting of mixture distribution and its testing.

MATERIALS AND METHODS

For this study, the data of *Arabidopsis thaliana* was used. It is known that *Arabidopsis thaliana* is a model organism for study because of its relative genetic simplicity, convenience and abundance, massive seed production, susceptibility to T-DNA insertions and basic life processes. The microarray data under two conditions, salt-stressed and control, was taken from Gene Expression Omnibus, with accession ID-GDS 3927 (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3927) and with platform GPL 198 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL198).

To model the variations with respect to expression levels, a class of mixture models are utilized that make use of a random threshold value for accommodating variations in the gene expression distribution. Distribution of expression scores/index (Z) can be considered as a mixture of two probability functions, representing the density function under two conditions as

$$g(Z) = pf_{\mathrm{i}}(Z) + (1 - p) f_2 (Z) \tag{1}$$

where, p is the proportion of subpopulation in the overall population, $f_i(Z)$ is the i^{th} component density which may be continuous or discrete for $i = 1,2$. An extension of this problem is to model genes that are under-expressed, expressed and over-expressed, leading to a three component mixture.

The mixture distribution for a random variable X that takes values in a sample space Θ, can be represented by a probability density function (or mass function in the case of discrete Θ) of the form

$$g(x) = \pi_1 f_i(x) + \ldots + \pi_k f_k(\mathrm{x}), \ (x \in \Theta) \tag{2}$$

where, $0 \leq \pi_i \leq 1$ for $i = 1,\ldots, k$ and $\pi_1 + \pi_2 + \ldots + \pi_k = 1$. The parameters $\pi_1, \pi_2, \ldots, \pi_k$ are the mixing weights or mixing proportions and $f_1(.),\ldots, f_k(.)$ are the component densities of the mixture respectively. The component densities $f_1(x), \ldots, f_k(x)$ can belong to the same or different parametric family. When there is a common functional form with different parameters, then

$$f_i(x) = f(x|\theta_{\mathrm{i}}) \tag{3}$$

where, θ_i denotes the parameters occurring in $f_i(x)$. The finite mixture density function will have the form

$$g(x \mid \Psi) = \sum_{i-1}^{k} \pi_i f(x \mid \theta_i), (x \in \Theta) \tag{4}$$

where, $\Psi = (\pi_1,\ldots,\pi_k, \theta_1,\ldots,\theta_k)$ is the complete collection of all distinct parameters occurring in the mixture model.

A random variable X has mixture normal distribution if $f_1(x)$ is normally distributed with mean μ_1 and variance σ_1^2 with mixing proportion $\pi_1 = p$ and $f_2(x)$ is normally distributed with mean μ_2 and variance σ_2^2 with mixing proportion $\pi_2 = (1-p)$. The mixture distribution of two normal distributions given above has five parameters, namely $p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. Let $f_1(x) = \phi_{\mu1, \sigma12}(x)$ and $f_2(x) = \phi_{\mu2, \sigma}{}^{2}{}_{2}(x)$ then

$$g(x) = \pi_1 \varphi_{\mu_1, \sigma_1^2}(x) + \pi_2 \varphi_{\mu_2, \sigma_2^2}(x)$$

$$= p \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu_1}{\sigma_1}\right]^2} + (1-p) \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu_2}{\sigma_2}\right]^2} \tag{5}$$

where $\pi_1 = p$ and $\pi_2 = 1\text{-}p$ is a mixture of two normal densities.

The mean and variance of the mixture distribution with k components are as follows:

$$\mu_m = E(\mathrm{x}) = \int x g(\mathrm{x}) \mathrm{dx} = \sum_{i=1}^{k} \pi_i \int x \, f_i(\mathrm{x}) \, dx = \sum_{i=1}^{k} \pi_i \mu_i$$

and

$$\sigma_m^2 = E(x - \mu_m)^2 =$$

$$\sum_{i-1}^{k} \pi_i E(x_i - \mu_i - \mu_m)^2 = \sum_{i-1}^{k} \pi_i \left[\sigma_i^2 + (\mu_i - \mu_m)^2\right] \tag{6}$$

The total number of parameters to be estimated depends on the distributions that are combined to form mixture distribution. The maximum likelihood estimation (MLE) method for parameter estimation is used in which the likelihood function is taken as the starting point. Incomplete data gives complicated likelihood functions, where MLE's usually have to be computed iteratively. The Expectation-Maximization algorithm, known as the EM algorithm, is a broadly applicable approach to the iterative computation of MLE's.

EM algorithm: Each iteration of the EM algorithm consists of two steps: the Expectation step (E-step) and the Maximization step (M-step). In the E-step, the algorithm finds the expected value of the log-likelihood, given the observed data and the initial parameter estimates. The M-step of the algorithm maximizes the expected log-likelihood obtained from the E-step and updates the parameter estimates. Unless it comes to some convergence criteria, the E- and M-steps are alternated repeatedly. After each, the log-likelihood is increased and thus the algorithm is guaranteed to converge to a local maximum.

The 'mixtools' (https://CRAN.R-project.org/

package=mixtools) package of R (Benaglia *et al.* 2009) provides a set of functions for analyzing a variety of finite mixture models. Many of the algorithms of the 'mixtools' package use EM algorithm. The function 'normalmixEM' implements the algorithm in mixtools. It returns the EM algorithm output for mixture of normal distributions. Other R packages are also available for Mixture of distributions like 'mixtNB' for the mixture of negative binomial distribution (Bonafede *et al.* 2016).

Likelihood Ratio Test (LRT): It is useful for comparing goodness-of-fit of two different distributions to the same set of data. The LRT statistic is given by following expression

$$LRT = -2\log_e \frac{l_s(\hat{\theta})}{l_m(\hat{\theta})} \tag{7}$$

LRT is the ratio of two likelihood functions; the single (s) distribution model has fewer parameters than the mixture (m) distribution model. Asymptotically, the test statistic is distributed as a χ^2 random variable, with degrees of freedom equal to the difference in the number of parameters between the two models. *LRT* compares two models provided the single model is a special case of the more complex model. *LRT* can be presented as a difference in the log-likelihoods as follows:

$$LRT = -2\left[\log_e l_s(\hat{\theta}) - \log_e l_m(\hat{\theta})\right] \tag{8}$$

Following steps are followed for analyzing differential gene expression:

i. For different genes, expression data for the two different conditions [control (C) and treated (T), i.e. salt-stressed] is taken**.** Fold change is then calculated as the ratio of the final value to the initial value, which describes how much a quantity changes from an initial to a final value.

ii. log2-fold change is calculated as follows:

$$log2FC = log2(T) - log2(C) \tag{9}$$

Fold changes greater than 1 (when T > C) become positive, while those lesser than 1 (C > T) become negative.

iii. The appropriate mixture distribution is fitted to the log2-fold change data and accordingly the parameters of the distribution are estimated.

iv. The goodness-of-fit of the model is tested and the fitted mixture distribution is compared with the single component distribution using likelihood ratio test.

R code has been developed for performing above steps of fitting mixture distribution.

RESULTS AND DISCUSSION

Gene expression data of two replicates under the two different conditions [Control (C) and Salt stressed (T)] for 22810 genes was taken and fold change was calculated. The log2-fold change was calculated. The one's with positive sign were up-regulated genes and the negative one's are down-regulated genes. Fig 1 and Fig 2 shows the histogram and plot of fold changes.

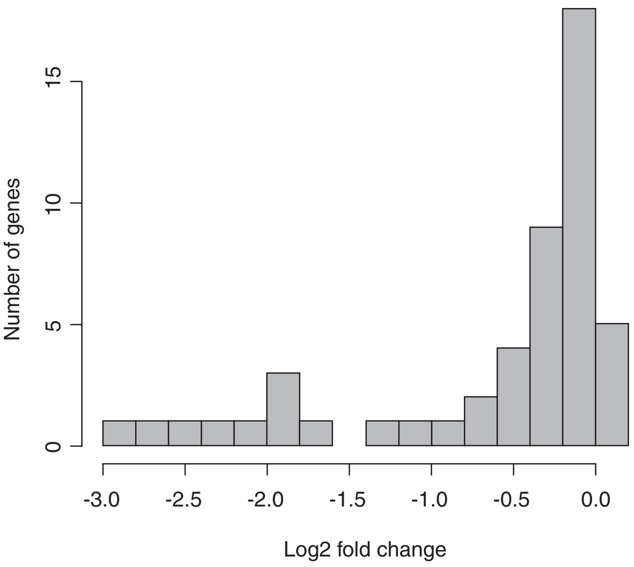

Fig 1 Histogram of log2-fold change values.

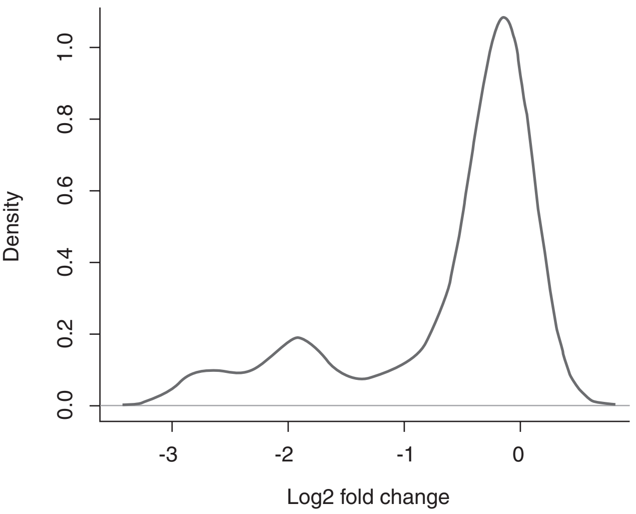

Fig 2 Plot of log2-fold change values.

Normal distribution was fitted to the data and the parameters were estimated as mean (μ) = 0.02297 and standard deviation (σ) = 0.53839. The fitted plot is shown in Fig 3.

The distribution of probabilities corresponding to different log2-fold change values is calculated and shown in Table 1.

It can be seen that the probability is ranging from 0 to 0.90. It clearly indicates that the distribution of probabilities cover a wide range of values and hence make it difficult to identify genes with differentially expressed value or reject genes which are not differentially expressed. If the genes are selected based on the probabilities, it will give a good selection only if the distribution covers probabilities in a smaller range. Therefore, the data was subjected to fitting of mixture normal distribution. The fitted two-component mixture plot is shown in Fig 4.

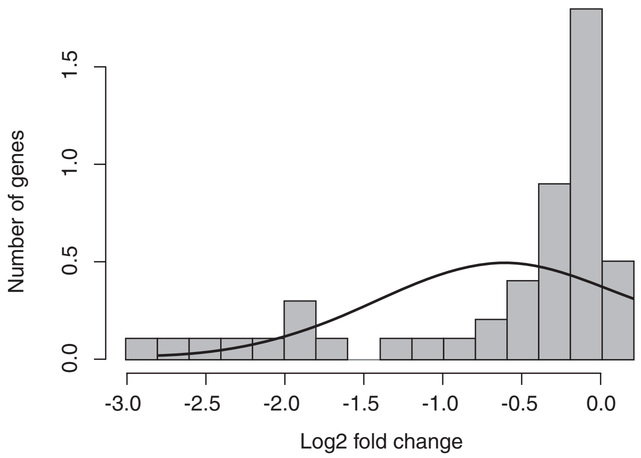

Fig 3 Fitted single normal plot.

Table 1 Distribution of the number of genes with respect to the probabilities

Probability interval	Frequency (No. of genes)
0.00 – 0.05	21
0.05 – 0.10	16230
0.10 – 0.15	4028
0.15 – 0.20	1358
0.20 – 0.25	595
0.25 – 0.30	286
0.30 – 0.35	132
0.35 – 0.40	68
0.40 – 0.45	41
0.45 – 0.50	21
0.50 – 0.55	11
0.55 – 0.60	12
0.60 – 0.65	2
0.65 – 0.70	3
0.70 – 0.75	1
0.75 – 0.80	0
0.80 – 0.85	0
0.85 – 0.90	1
Total	22810

The parameters of the two-component mixture model are estimated as follows:

	Component 1	Component 2
Mean	$_1$ = 0.258905	μ_2 = -0.0478084
Standard deviation	σ_1 = 0.972783	σ_2 = 0.2668608
Proportion (weights)	π_1 = 0.230776	π_2 = 0.7692240

Likelihood ratio test was performed to compare the goodness of fit of two models, one of which (the null model, single component normal model) is a special case of the other (the alternative model, two-component mixture normal model). The test based on the likelihood ratio expresses how likely the data can be fitted under single component normal model than the two-component mixture normal model. In case of significance, the likelihood ratio test indicates that fitting of single component normal model is less appropriate as compared to two-component mixture normal model. For the given data set, *LRT= 9176.395* which follows a chi-square with 4 degrees of freedom. Thus, it is found that the null model is rejected in favour of the alternative model as the calculated value of the *LRT* is much more higher than the tabulated value of χ^2 at 4 degrees of freedom and 5% level of significance which is 9.488. Hence, the alternative model of two-component mixture normal distribution fits the data more accurately as compared to a single component normal distribution.

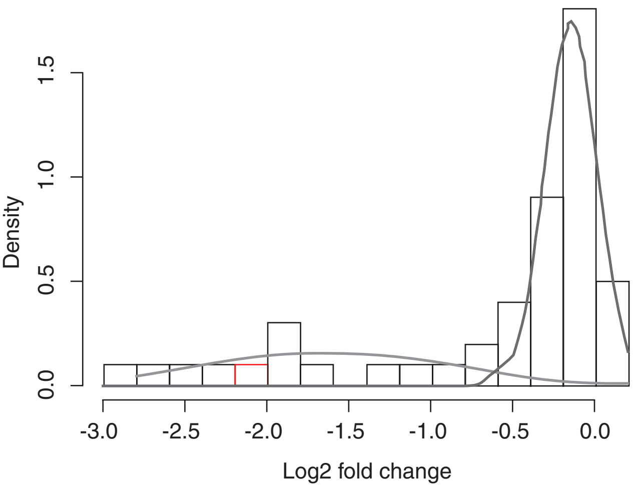

Fig 4 Fitted two-component mixture normal plot.

The distribution of probabilities corresponding to different log2-fold change values are calculated and shown in Table 2. It can be seen that the probability is ranging from 0 to 0.50 with maximum number of genes falling in the range of 0.05 to 0.35. It clearly indicates the distribution probabilities of genes covering a small range of values and hence enabling the selection of genes which are differentially expressed.

Table 2 Distribution of the number of genes with respect to the probabilities

Probability interval	Frequency (No. of genes)
0.00 – 0.05	12
0.05 – 0.10	15517
0.10 – 0.15	4875
0.15 – 0.20	1584
0.20 – 0.25	521
0.25 – 0.30	188
0.30 – 0.35	80
0.35 – 0.40	22
0.40 – 0.45	6
0.45 – 0.50	5
Total	22810

Table 3 Number of genes identified

Number of genes	Single normal ($\pm 3\sigma$)	Mixture normal ($_{m}\pm 3\sigma_{m}$)
Total genes identified	439	246
Number of down-regulated genes identified	97 (<-1.61465)	43 (<-1.98774)
Number of up-regulated genes identified	342 (>1.64007)	203 (>2.03255)

The total number of genes identified as differentially expressed are shown in Table 3 along with the cut-off values (within bracket) under the single normal and two-component mixture normal distributions. These cut-off values are obtained from ($\mu \pm 3\sigma$) and ($\mu_m \pm 3\sigma_m$) for single and mixture normal distribution respectively.

Out of a total of 439 genes identified with single normal distribution, 97 genes were down-regulated genes with cut-off value as -1.61465 in case of single normal i.e. these genes had difference values less than -1.61465 whereas 342 genes were up-regulated genes with cut-off value as 1.64007, i.e. these genes had difference values more than 1.64007. On the other hand, when a two-component mixture normal model was fitted to the data, 43 genes were identified as down-regulated out of a total of 246 genes as they had difference values less than the cut-off value -1.98774. The remaining 203 genes were up-regulated genes as they had difference values more than the cut-off value 2.03255. Therefore, the number of genes identified as down-regulated and up-regulated while fitting mixture normal distribution is less as compared to that of a single normal distribution. Thus, it can be concluded that mixture model is capable of capturing more variability and hence able to identify differentially expressed genes more accurately.

ACKNOWLEDGMENTS

The authors are grateful to the Editor and the referees for the constructive comments that have led to considerable improvement in the paper.

REFERENCES

Anders S and Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11(10)**: R106. DOI:10.1186/gb-2010-11-10-r106.

Anjum A, Jaggi S, Varghese E, Lall S, Bhowmik A and Rai A. 2016. Identification of differentially expressed genes in RNA-seq data of *Arabidopsis thaliana*: A compound distribution approach. *Journal of Computational Biology* **23(4)**: 239-47. DOI:10.1089/cmb.2015.0205.

Benaglia T, Chauveau D, Hunter D and Young D. 2009. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* **32(6)**: 1-29.DOI:10.18637/jss.v032.i06

Bonafede E, Picard F, Robin S and Viroli C. 2016. Modeling over dispersion heterogeneity in differential expression analysis using mixtures. *Biometrics* **72(3)**: 804-814.DOI: 10.1111/biom.12458

Brazma A and Vilo J. 2000. Gene expression data analysis. *FEBS Letters* ***480*(1)**: 17-24.

Jeffery I B, Higgins D G and Culhane A C. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* **7(1)**: 359.

Karim R, Hossain P, Begum S and Hossain F. 2011. Rayleigh mixture distribution. *Journal of Applied Mathematics*. Article ID 238290, DOI:10.1155/2011/238290.

Marioni J C, Mason C E, Mane S M, Stephens M and Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18(9)**: 1509-1517.DOI:10.1101/gr.079558.108.

McLachlan G J, Bean R W and Peel D. 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18(3)**: 413-22.DOI: 10.1093/bioinformatics/18.3.413.

McLachlan G and Peel D. 2000. *Finite Mixture Models*. New York: Wiley.

Mortazavi A, Williams B A, McCue K, Schaeffer L and Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5(7)**: 621-628. DOI:10.1038/nmeth.1226.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M and Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320(5881)**:1344-1349. DOI: 10.1126/science.1158441.

Pearson K. 1895. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London* A. **185**: 343-414.

Yang Y, Tashman AP, Lee JY, Yoon S, Mao W, Ahn K, Kim W, Mendell N R, Gordon D and Finch S J. 2007. Mixture modeling of microarray gene expression data. *BMC Proceedings* **1(1)**: S50.