Journal of Computer and Communications, 2025, 13(2), 223-235
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Gene Expression Data Analysis Based on Mixed Effects Model

Yuanbo Dai

College of Life Sciences, Wuhan University, Wuhan, China
Email: daiyuanbo2024@outlook.com

How to cite this paper: Dai, Y.B. (2025) Gene Expression Data Analysis Based on Mixed Effects Model. *Journal of Computer and Communications*, **13**, 223-235.
https://doi.org/10.4236/jcc.2025.132014

Received: December 27, 2024
Accepted: February 25, 2025
Published: February 28, 2025

Abstract

DNA microarray technology is an extremely effective technique for studying gene expression patterns in cells, and the main challenge currently faced by this technology is how to analyze the large amount of gene expression data generated. To address this, this paper employs a mixed-effects model to analyze gene expression data. In terms of data selection, 1176 genes from the white mouse gene expression dataset under two experimental conditions were chosen, setting up two conditions: pneumococcal infection and no infection, and constructing a mixed-effects model. After preprocessing the gene chip information, the data were imported into the model, preliminary results were calculated, and permutation tests were performed to biologically validate the preliminary results using GSEA. The final dataset consists of 20 groups of gene expression data from pneumococcal infection, which categorizes functionally related genes based on the similarity of their expression profiles, facilitating the study of genes with unknown functions.

Keywords

Mixed Effects Model, Gene Expression, Data Analysis, Gene Analysis, Gene Chip

1. Introduction

Gene expression data analysis, as an indispensable component of modern biological research, plays a pivotal role in unraveling the mysteries of life and driving medical advancements. This field of research delves deeply into the interior of cells, aiming to gain insights into regulatory mechanisms and functional characteristics of life by examining gene expression patterns under diverse conditions, such as disease states, environmental shifts, or drug treatments. Thanks to the rapid evolution of high-throughput sequencing technologies, including RNA-Seq

and single-cell sequencing, we can now obtain gene expression data with unprecedented depth and scope. However, the characteristics of these data—high dimensionality (thousands of genes), small sample sizes (constrained by experimental conditions and costs), and intricate association networks (interactions among genes, between genes and the environment)—pose unprecedented challenges for analysis [1].

Confronted with these challenges, statisticians and biologists are continually exploring novel analytical methodologies and tools to extract valuable information from vast datasets. Among them, the Mixed Effects Model has emerged as a versatile and potent statistical framework in gene expression data analysis, demonstrating its unique strengths and extensive application potential.

The essence of the mixed effects model resides in its capacity to consider both fixed effects and random effects simultaneously, thus capturing the sources of variation in the data more comprehensively. In the context of gene expression data analysis, fixed effects may embody known biological factors, such as gene functional categories, regulatory pathways, or specific experimental conditions. Random effects, on the other hand, might encompass genetic variations among individuals, experimental errors, or unknown environmental factors. By amalgamating these two effects within a single model, the mixed effects model can more precisely estimate gene expression levels and unveil underlying biological principles [2].

At the methodological level, gene expression data analysis utilizing mixed effects models encompasses a spectrum of approaches, ranging from linear models to nonlinear models, and from straightforward to intricate methodologies. The linear mixed effects model stands as one of the most fundamental and widely employed methods, presuming a linear relationship between gene expression levels and both fixed and random effects, with model parameters estimated through optimization algorithms like least squares. With the progression of machine learning technology, nonlinear mixed effects models have gradually been incorporated into gene expression data analysis, such as models grounded in neural networks, support vector machines, and other algorithms, which can more flexibly capture the nonlinear trends in gene expression levels [3].

At the application level, gene expression data analysis leveraging mixed effects models has found widespread use in various domains, including the identification of differentially expressed genes, the construction of gene regulatory networks, disease classification, and prognosis prediction. For instance, in screening for differentially expressed genes, the mixed effects model can simultaneously account for the impacts of multiple experimental conditions and individual differences on gene expression levels, thereby more accurately selecting genes with significant alterations in expression under diverse conditions [4] [5]. In constructing gene regulatory networks, mixed effects models can reveal interactions between genes and provide crucial clues for understanding the regulatory mechanisms of biological processes. In disease classification and prognosis prediction, mixed effects

models can harness information from gene expression data to enhance the accuracy and robustness of classification and prediction [6].

In conclusion, gene expression data analysis based on mixed effects models has emerged as a vital tool in modern biological research. It not only aids in gaining a deeper comprehension of life's regulatory mechanisms but also offers fresh perspectives and methodologies for disease diagnosis and treatment. With the ongoing advancement of technology and the continuous refinement of methods, it is conceivable that mixed effects models will play an even more pivotal role in the realm of gene expression data analysis [7].

2. Literature Review

A mixed effects model is a statistical model, either linear or nonlinear, that incorporates both fixed effects and random effects. In the analysis of gene expression data, mixed effects models can simultaneously account for fixed relationships among genes (such as gene function, regulatory networks, etc.) and random variations (like experimental errors, individual differences, etc.), thereby providing a more accurate depiction of changes in gene expression levels. Fixed effects typically pertain to factors that deterministically influence gene expression levels, such as gene sequences and mutation sites. Conversely, random effects pertain to factors that stochastically impact gene expression levels, such as experimental conditions and individual differences among samples. By introducing random effects, the mixed effects model can better capture data heterogeneity, enhancing the model's prediction accuracy and generalization capability [8] [9].

Currently, gene expression data can be analyzed at least at three progressively complex levels: firstly, analyzing the expression level of individual genes, focusing on whether the expression of each gene differs from the control under specific experimental conditions; secondly, categorizing genes into different classes and examining their shared functions, interactions, and collaborative regulation; and thirdly, attempting to infer potential regulatory regions and gene networks to elucidate observed patterns at the mechanistic level. Presently, research on gene expression data predominantly focuses on the second level, while the third level represents a more advanced research objective [10].

Various models have been employed in the analysis of gene expression data, including cluster analysis, multivariate statistics, pattern recognition, and neural networks. Genes with functional relevance are often co-expressed, and detecting gene clusters with similar expression profiles is an effective approach for studying gene function. Consequently, gene chip technology, which can simultaneously obtain numerous gene expression profiles, has been extensively utilized in biomedical fields such as disease diagnosis, genome sequencing, mutation and polymorphism detection, drug screening and development, novel gene discovery, and pathogen diagnosis. Mining the biological information contained within vast amounts of gene expression data is currently a focal topic in the field of biostatistics [11].

Gene Set Enrichment Analysis (GSEA) is a widely adopted method for analyzing gene expression data based on prior pathways. GSEA examines a group of genes with similar biological effects as a collective entity and has demonstrated superiority over single gene analysis in terms of stability, sensitivity, and biological relevance. The mixed effects model was developed by R.A. Fisher. Over the past two decades, this mixed effects model has garnered increasing attention from statisticians. In 2001, the mixed effects model was applied to analyze single gene expression data from lymphoma and yeast cells, yielding more accurate results than traditional methods. In 2003, we reanalyzed gene expression data from primates using a mixed effects model and identified novel differentially expressed genes across species. In 2008, we utilized the mixed effects model to reanalyze a previously analyzed gene expression dataset related to diabetes, revealing that the mixed effects model possessed superior testing ability compared to GSEA. In 2009, the mixed effects model was applied to the analysis of time-series gene expression data, demonstrating its advantages over other methods in terms of testing power, reduction of type I errors, and reliability. In 2013, the mixed effects model was employed in the study of heart disease-related genes, not only identifying known heart disease-related genes but also uncovering additional information. In contrast, there are no literature reports in China that analyze gene expression data using mixed effects models. Therefore, analyzing gene expression data based on mixed effects models is highly suitable for this study. As biology increasingly becomes a quantifiable discipline, future methods for analyzing gene expression data are poised to witness significant advancements.

3. Experimental Methods

3.1. Raw Data and Model Construction

1) Original experimental data

Pneumococcal infection is a very common pediatric disease to understand the pathogenesis of this disease identify the genes that cause pneumococcal infection and study their role in the disease is important research. The radiolabeled DNA microarray technology was applied to 1176 genes in mice infected with pneumococcal infection. The data contained 8 DNA microarray experimental samples 2 samples were obtained under experimental conditions without pneumococcal infection while the other 6 were obtained under conditions with infection. There were 90 genes whose expression levels changed significantly under the two experimental conditions among which 12 genes have been medically proven to be pathogenic (see Table 1).

2) Establish a mixed effect model

Our goal is to identify genes whose average expression levels change under the two experimental conditions, the basic idea is to first assume that the average expression levels of all genes do not change under the two experimental conditions, then test this assumption based on experimental data, it is clear that this can be reduced to a classic hypothesis testing problem. The mixed effects model

Table 1. Gene analysis list.

GenBank Accession No.	Gene/Protein Name	Function
M63122	Tumor necrosis factor receptor 1	Inflammatory reaction
X91810	Stat3, signal transducer and activator of transcription 3	Acute phase response factor
Z17223	Gax, growth-arrest-specific protein	Transcription factor, growth arrest
D10864	Id3, DNA-binding protein inhibitor	Cell cycle progression, growth
X74806	Von ebners gland protein	Middle ear gland protein
D30041	rac-beta serine/threonine kinase (rac-PK-beta); AKT2	Mitogenic signaling
D30040	rac-beta serine/threonine kinase (rac-PK-beta); protein kinase B	Mitogenic signaling
M86389	Heat shock 27-kDa protein (HSP27)	Celluar protection
Z27118	Heat shock 70-kDa protein (HSP70)	Celluar protection
D17695	water channel aquaporin 3 (AQP 3)	Water transportation
M63837	Platelet-derived growth factor alpha receptor (PDGFRa)	Proliferation
U03491	transforming growth factor beta 3 (TGF-beta3)	Anti-proliferation

of gene expression data may seem like a sample comparison problem in classical statistics, but microarray data have their own particularities, we can make the following assumption: for each gene i,i = 1, 2, ..., N, we have m samples of gene expression levels X obtained under the first experimental condition 1.....Xm, and the expression levels of n samples obtained under the second experimental condition Y1.....Y From the experimental data it can be seen that the number of genes N is very large (>1000) while the sample sizes m and n for microarray data are very small (typically <30). Therefore, traditional statistical tests such as t-tests and rank-based hypothesis tests are not applicable in this context. However, we can use gene expression levels X1.....Xm and Y1.....Y The test statistic Z is constructed as follows 1. In this way, the large number of genes N can be fully utilized. As shown in Figure 1, the mixed effect model analysis process is shown [12].

3.2. Genetic Data Processing and Basic Tasks

1) Information preprocessing of gene chips

To implement information mining on high-density gene chips, it is first necessary to read the chip data into a computer and form special format computer data files. This stage mainly involves using image processing techniques to read information from the chip based on the characteristics of the gene chip. This phase of

work can be referred to as pre-processing of gene chip information. The quality of pre-processing directly affects subsequent information mining. Therefore, researchers place great emphasis on pre-processing, designing various effective pre-processing methods tailored to the characteristics of high-density chips. The main task at this stage is to use specialized scanners to scan the gene chip, generating computer image files, determining hybridization site ranges (Segmentation) through grid division (Gridding), and obtaining the base sequence through steps such as signal intensity extraction [13].

a) The overlap of samples may affect the adjacent samples (Figure 2) due to the

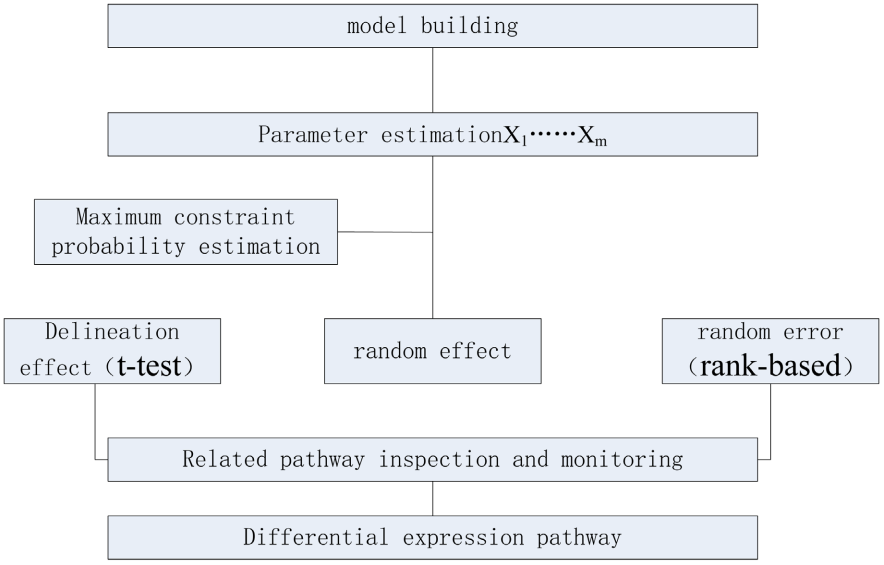

Figure 1. Analysis process of mixed effects model.

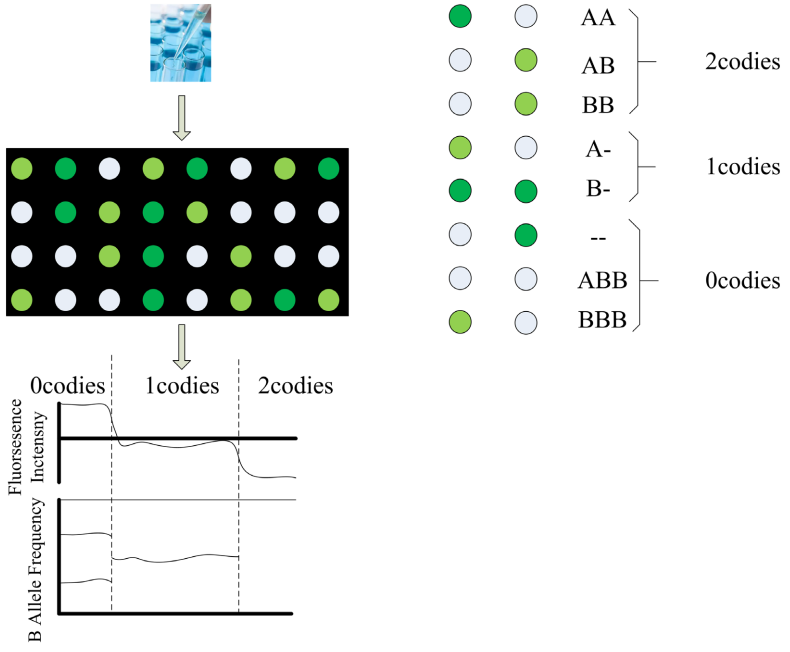

Figure 2. Overlap of sample points.

overlap of high intensity samples;

b) Due to the different types of experiments, different sample shapes may be produced (Figure 3);

c) Noise interference caused by contamination during the experimental process. In order to effectively solve the above main problems, information preprocessing is required for gene chips. Information preprocessing generally includes the following four steps.

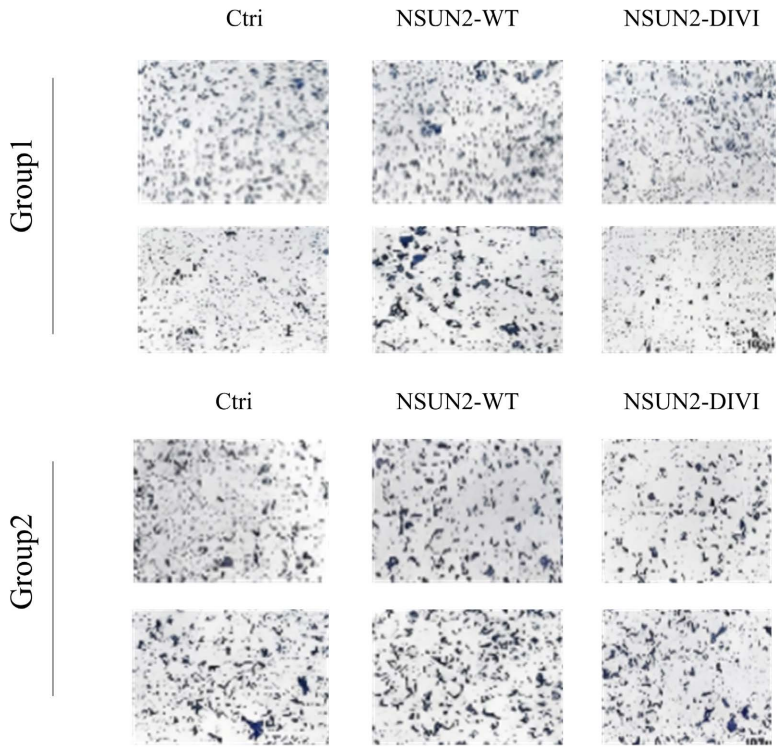

Figure 3. Different shapes of sample points.

a) Grid Division. To understand the number of samples in each row and column of the microarray image input into the computer, as well as the distance between adjacent samples, it is necessary to know the distances between samples. However, due to different chip designs and experimental conditions, it is impossible to have exact data. Therefore, the image needs to be divided into grids to understand this information [14].

b) Determination of sample point range. Identify the sample point area from the grid, each sample point is approximately circular due to the way the robot places cDNA on the slide and the method used to process the slide. Currently, there are many methods for fixed shape that can be used to segment microarray images, such as the fixed perimeter method and variable perimeter method. The better method is the variable shape segmentation method, mainly including Mann-Whitney test and SRG method.

c) Signal Strength Extraction. This step includes background intensity estimation saturation compensation and extraction of signal strength values. Due to the

issues of saturation and overlapping interference at high-value sample points it is not possible to simply use the intensity value of the current sample point as the signal strength. An accurate sample point theoretical model needs to be established and on this basis saturation compensation and interference correction should be performed to achieve precise signal strength extraction.

d) Standardization Processing. Due to the differences in samples, the imbalance in fluorescence labeling efficiency and detection rate, it is necessary to balance and correct the original extracted signals before further analysis of experimental data. The standardization processing (Normalization) mentioned above is precisely carried out for this purpose [15].

2) EM algorithm for hybrid models

For a given dataset D = {x(1) ..., x(n)}, when the potential model is a mixture model, there is usually no closed-form technique to directly maximize the likelihood score function. Listing the log-likelihoods of the mixture model makes this point clear: we get a sum of terms in a log form, which is a nonlinear optimization problem (not having a closed-form solution like in multivariate mixture models. The EM algorithm often increases the likelihood significantly in the initial few iterations and then converges slowly to the final value. However, the likelihood function is not necessarily concave with respect to the number of iterations. For many datasets and models, we can often obtain acceptable solutions with only 5 to 20 iterations. Of course, each solution provided by the EM algorithm is a function of the search starting point (since EM is a local search algorithm), so it is a good idea to restart the algorithm multiple times from a randomly selected starting point to avoid getting a poor local maximum. Note that whether K or P (or both) increases, the numerical value of the likelihood local maximum will increase significantly with changes in the dimension of the parameter space. When using maximum likelihood methods to estimate mixture distributions, some special cases need attention. For example, in a Gaussian Mixture Model, if the mean of a component equals a sample point and its standard deviation tends towards zero, the likelihood will increase infinitely. However, in this case, the maximum likelihood solution is likely to be finite. There are many methods to address this issue. The maximum finite value of the likelihood can be chosen to provide the estimated parameter value. Additionally, if the standard deviation is restricted to be equal, this problem does not occur. A more general approach is to use Bayesian methods to handle this issue, taking a prior distribution over parameters and no longer maximizing the likelihood but maximizing the MAP score function. This prior provides a framework that keeps the score function (MAP score function) away from problematic regions in the parameter space. Note that it is easy to extend the EM algorithm from maximizing the likelihood to maximizing the MAP (for example, replacing step M with step MAP, and so on). Another possible issue arises due to the lack of identifiability in the mixture distribution. A family of mixture distributions is said to be identifiable if and only if two members of the family are equal, meaning c = c, and for all k there exists some j. If a family of

distributions is unidentifiable, it cannot distinguish between its two different members, leading to estimation problems.

4. Experimental Results and Discussion

4.1. Conventional Results

We varied the number of Gaussian components in the model from 1 to 5 and performed model matching, listing the results of the model matching in Table 2. Using BIC as the selection criterion, we chose 8 = 2, meaning the Gaussian mixture model has 2 Gaussian components. Table 3 lists some of the model parameters when g = 2, with the model matching results being:

Table 2. Gaussian mixture model matching results.

G	Lob Like	AIC	BIC
1	1559.07	3122.15	3132.29
2	I-1253.98	2517.96	2543.31
3	I-1244.19	2504.37	2544.93
4	I-1239.70	2501.40	2557.17
5	I-1238.79	2505.59	2576.56

Table 3. Model parameters.

Number assigned to each component	79	1097
Estimate of mixing proportion for each component	0.107	0.893
Estimates of correct allocation rates for each component	0.560	0.992
Estimate of overall correct allocation rate		0.946
Estimated mean (asa row vector) for each component	−0.64832	0.07121
Estimated covariance matrix for component 1		5.09636
Estimated covariance matrix for component 2		0.263301
Criteria for this Clustering are AIC BIC	2517.96	2543.31

1097 genes fall into the first Gaussian component with an almost zero mean, indicating that for the vast majority of genes, their expression levels change little or not at all. Additionally, 79 genes fall into the Gaussian component with a mean of -0.64832, suggesting that these genes have undergone more significant changes in expression levels (experimental data has been standardized). The results show that the mixed effects model and gene set enrichment analysis (GSEA) methods jointly identified 12 differentially expressed pathways. Furthermore, 8 differentially expressed pathways were identified by the mixed effects model alone. Among these 10 differentially expressed pathways, the Wnt signaling pathway (Wnt signaling pathway) plays a crucial role in pneumococcal cell self-renewal, which has been confirmed in final pneumococcal infection data. The Hedgehog signaling pathway

(Hedgehog signaling pathway) is important in maintaining pneumococcal characteristics. During the expression process of pneumococcal infection, it reduces the self-renewal capability of pneumococci by interfering with the Hedgehog signaling pathway. Therefore, the mixed effects model has precise and high sensitivity in detecting expression processes, making it more effective in analyzing differentially expressed pathways compared to directly using GSEA.

4.2. Replacement Test

The hypothesis tests listed earlier are based on the assumption of drawing random samples from a certain distribution and the goal of the tests is to make a probabilistic statement about the parameters of the distribution. The ultimate goal is to draw inferences about the potential values of the underlying population based on the sample. Tests based on this principle are called permutation tests or randomization tests. Note that the above process does not provide any statistical inference from the sample to the entire population, but it does allow us to make conditional probability conclusions about treatment effects, with the condition being the observed data. Here is a simple example illustrating a permutation test: there are two sets of numbers:

Group1:	55	58	60
Group2:	12	22	34

In this example, the null hypothesis is that there is no difference between the two groups, and in the original data, the sum of Group1 is 173. In this case, if the null hypothesis is true, then no matter how the data are randomly assigned, the adjusted sum of Group1 will either be greater than or less than the original sum of Group1, but will not result in a significant skewness: the adjusted sum of Group1 will almost always be greater than or less than the original sum of Group1. If we regroup the data, we can derive 20 combinations and calculate the sum of Group1 for these 20 combinations:

Order	Group 1			Group 2			Sum
1	55	58	60	12	22	34	173
2	55	58	12	60	22	34	125
3	55	58	22	12	60	34	135
4	55	58	34	12	22	34	148
5	55	12	60	58	22	34	127
6	55	22	60	12	58	34	137
7	55	34	60	12	22	58	149
8	12	58	60	55	22	34	130
9	22	58	60	12	55	34	140

Continued

10	34	58	60	12	22	55	152
11	12	22	60	55	58	34	94
12	12	58	22	55	60	34	92
13	55	12	22	12	55	58	89
14	12	34	60	12	58	34	106
15	12	58	34	55	22	60	104
16	55	12	34	55	58	60	101
17	22	34	60	55	58	34	116
18	22	58	34	55	22	60	114
19	55	22	34	12	58	60	111
20	12	22	34	55	58	60	68

We can find that the sum of the remaining 19 combinations of Group1 is smaller than the sum of the original data Group1, which is the very skewed situation we mentioned earlier, so we reject the null hypothesis.

The permutation test requires that the data under the null hypothesis satisfy exchangeability. Specifically, in this article, the null hypothesis is that the gene expression levels have not changed significantly, allowing for the arbitrary permutation of the order of 8 DNA microarray experimental samples. If the gene data truly satisfies the null hypothesis, the test statistic Z calculated from the permuted data should be relatively evenly distributed on both sides of the test statistic derived from the original data (with roughly equal numbers of cases where it is greater or less than the original statistic), and will not exhibit significant skewness (almost all cases will be either much larger or much smaller than the original statistic). The purpose of using the permutation test here is to identify genes whose permutation data cause extreme skewness in the distribution of the test statistic Z, indicating that these genes expression levels may have changed significantly.

Here we define "extreme skewness" by arbitrarily swapping 8 samples under two experimental conditions to obtain 28 combinations of $C^2 = 28$, thus allowing us to calculate 28 test statistics Z. Observing the 28 Zs corresponding to each gene, if we define "extreme skewness" more broadly, it can significantly increase the false-negative rate (even to 100%), but it will greatly reduce the true-positive rate. We need to strike a balance between the two. We define: if only two of the 20 Zs are less than or equal to the Z calculated from the original data (and Z is greater than a selected threshold), then this gene is considered to be the one causing extreme skewness in the distribution of test statistics Z.

In this way, we can add genes that do not meet the data interchangeability under the null hypothesis (*i.e.*, no significant change in gene expression levels) to the initial experimental results, which is the first step of improvement. By processing this way, we further strengthen the experimental results by adding the results of

permutation tests to the previously single mixed model estimates, which clearly improves the recall rate. Of course, this sacrifices some accuracy (which should be higher), but ultimately yields more comprehensive experimental results. From the final experimental results, although the accuracy rate has not increased significantly, the recall rate has markedly improved.

The mixed effects model was used to analyze the same pneumococcal gene chip data Group 1 and Group 2 to observe the degree of expression pathways that conform to real biological effects. Among the 20 differentially expressed pathways detected by both the mixed effects model and GSEA, 19 of the differentially expressed signaling pathways in the mixed effects model have been proven to have a clear biological relationship with pneumococcus, while only 1 has been proven in GSEA. Therefore, using the mixed effects model to test for the biological validation of differentially expressed pathways is truly effective.

5. Conclusions

Based on an in-depth study of gene clustering processing methods, we propose a gene expression data method based on mixed-effect models and apply it to the gene expression data of 1176 genes in mice under two experimental conditions: with and without pneumococcal infection, achieving good results. This paper presents the following innovative viewpoints:

1) The gene expression data method based on mixed effects model is applied to the gene expression problem, which has its own advantages over other classical analysis methods: it can determine the number of categories, clearly define each category with a specific distribution (here it is Gaussian distribution), etc.

2) Gene expression data methods based on mixed-effect models, although intuitive and simple, are not without their limitations: the recall and precision rates are not very high; although they provide posterior probabilities as quantitative criteria for gene classification, these criteria are somewhat subjective and can lead to misclassification; the method is overly simplistic, resulting in somewhat rough classification outcomes. In this paper, we introduce permutation testing to improve this method, significantly enhancing both recall and precision rates, achieving excellent results.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Liu, X., Yang, Q., Liu, X., Tao, R. and Rao, W. (2025) A Visco-Hyperelastic Constitutive Model of Hydrogel Considering the Coupling Effect between Segment Motion and Interchain Slippage. *Journal of the Mechanics and Physics of Solids*, **196**, Article ID: 105996. https://doi.org/10.1016/j.jmps.2024.105996

[2] Ma, H., Geng, Y., Di Luzio, G., Li, G. and Wang, Y. (2025) Autogenous Shrinkage Model for Concrete Accounting for Compounding Effects of Mineral Admixtures. *Journal of Building Engineering*, **99**, Article ID: 111503.

https://doi.org/10.1016/j.jobe.2024.111503

[3] Recoules, C., Touvier, M., Pierre, F. and Audebert, M. (2025) Evaluation of the Toxic Effects of Food Additives, Alone or in Mixture, in Four Human Cell Models. *Food and Chemical Toxicology*, **196**, Article ID: 115198. https://doi.org/10.1016/j.fct.2024.115198

[4] Xie, S., Gan, C. and Lawniczak, A.T. (2024) Analyzing Decision-Making in Cognitive Agent Simulations Using Generalized Linear Mixed-Effects Models. *Mathematics*, **12**, Article No. 3768. https://doi.org/10.3390/math12233768

[5] Zhou, K., Zhang, S., Shang, J. and Lan, X. (2025) Exploring Immune Gene Expression and Potential Regulatory Mechanisms in Anaplastic Thyroid Carcinoma Using a Combination of Single-Cell and Bulk RNA Sequencing Data. *Computational Biology and Chemistry*, **115**, Article ID: 108311. https://doi.org/10.1016/j.compbiolchem.2024.108311

[6] Duan, J., Jiang, R., Shen, H., Xu, X. and Sun, D. (2024) Analysis of Nitrogen Metabolism-Related Gene Expression in Hepatocellular Carcinoma to Establish Relevant Indicators for Prediction of Prognosis and Guidance of Immunotherapy. *Computer Methods in Biomechanics and Biomedical Engineering*. https://doi.org/10.1080/10255842.2024.2438922

[7] Padala, S.K., Swamy, A.K. and Bhattacharjee, B. (2025) Air Temperature Prediction Models for Pavements Based on the Gene Expression Programming Approach. *Journal of Transportation Engineering, Part B: Pavements*, **151**, No. 3. https://doi.org/10.1061/jpeodx.pveng-1496

[8] Schunck, F., Kodritsch, B., Krauss, M., Busch, W. and Focks, A. (2024) Integrating Time-Resolved *nrf2* Gene-Expression Data into a Full GUTS Model as a Proxy for Toxicodynamic Damage in Zebrafish Embryo. *Environmental Science & Technology*, **58**, 21942-21953. https://doi.org/10.1021/acs.est.4c06267

[9] Qin, X., Zhang, S., Dong, X., Luo, T., Shi, H. and Yuan, L. (2024) An Improved Conditional Relevance and Weighted Redundancy Feature Selection Method for Gene Expression Data. *The Journal of Supercomputing*, **81**, Article No. 238. https://doi.org/10.1007/s11227-024-06714-5

[10] Zhang, Y., Zhang, W., Cao, W.J., *et al.* (2015) Mixed Effects Model Analysis of Tumor Expression Profile Gene Chip Data. *Advances in Modern Biomedical Science*, No. 3, 5.

[11] Zhu, B., Wu, Y., Qi, X., *et al.* (2014) Genome Wide Association Analysis of QTL-MAS2011 Public Dataset Using Mixed Linear Model and BayesCPi. *Journal of Animal Husbandry and Veterinary Medicine*, **45**, 692-698.

[12] Wang, Y.T. (2016) Tissue Specificity Analysis of APA Loci Based on Information Entropy and Generalized Linear Mixed Effects Model.

[13] Zhang, Y. and Zhai, S.D. (2004) Research Progress of Nonlinear Mixed Effects Model Method in Population Pharmacokinetics and Pharmacodynamics. *International Journal of Pharmaceutical Research*, **31**, 236-240.

[14] Fu, L.Y. (2013) Analysis of Parameter Estimation Methods for Nonlinear Mixed Effects Models. *Scientia Silvae Sinica*, **49**, 78-86.

[15] Qian, W.M. and Wang, J. (2003) Non Parametric Estimation of Random Effect Density in Linear Mixed Effects Models. *Proceedings of the* 11*th Annual Conference of the Chinese Society for Field Statistics*, *Part* 1.