Impact of Normalization Techniques in Microarray Data Analysis

Lwin May Thant
Cloud Computing Lab
University of Computer Studies, Yangon (UCSY)
Yangon, Myanmar
lwinmaythant@ucsy.edu.mm

Sabai Phyu
Cloud Computing Lab
University of Computer Studies, Yangon (UCSY)
Yangon, Myanmar
sabaiphyu@ucsy.edu.mm

***Abstract*— In a medical problem, investigation of powerful new tools is involved in essential role of high-throughput sequencing technologies. Microarray data sequencers produced a large and complex sets of data for the advancement of computational and statistical methods are appliance in the data analysis of medical area. In recent years, the researchers have processed on different methods that can accurately come out the gene expression on microarray data. In this paper, the several normalization algorithms for the analysis of RNA-seq differential data with differential species values are reviewed. Based on the studies, we present the reviews of normalization algorithms according their data nature to get the accurate data for building effective model.**

Keywords—microarray data, normalization, methods, techniques, features

I. Introduction

Microarray data is kinds of massive amount data with have various dimensions and the features amount are larger than their samples. It has many kinds of data such as RNA data, DNA data and various types of cancer. Microarray techniques have been used to analyst the genes of differential expression in tumors, drug treatments, virus infection and developing in animals. Due to the complexity of differential genes expression and high-dimensional throughput sequencing, there are many little-unknown useful correlations in the physical system for the saving of data. High-throughput sequencing technologies are very important sector in the analysis of gene expression.

Microarray data is occurring due to the variance of RNA-seq in gene expression. In microarray data, normalization is to identify and take out the sources of variance due to difference factors. Normalization is an extremely damage step but it has an effective result in microarray data. It aims to certify that expressions of genes are more as good as about the features (genes) and samples. After all, normalization methods of sequencing data analysis are still remaining a great deal of misgivings. The considerations of RNA-seq data normalization have been described in [1]. The selection of normalization process was affected on differential expression analysis. The analyst of [1, 2] described the procedures of normalization cause questions and this is necessary to add more useful procedures for researchers who is need to choose a normalization method.

The main goal of normalization is to discard the effect of technical impacts in the microarray sequencing. Exactly, effective normalization has the characteristic of the variance of gene, gene abundance and depth of sequencing. In addition, the considerations of genes variance should be the cell of sequence deeply or slowly. We reviewed the normalization methods that are widely used in differential expression of genes data. A powerful normalization method based on the analysis of microarray data sets with respect to their normalization methods to improve the outcome of differential gene expression analysis.

The part of the paper is organized as: II) provide a background knowledge for microarray data with detail explanation of what is microarray data set, characteristic of microarray data and theoretical analysis of microarray data and III) discussion and conclude the paper.

II. BACKGROUND

During the last period of 5 years, the different methods of normalization are emerged in research area differing from the types of approach to the microarray data. Microarray data sequencing and normalization techniques of microarray data are emerging with high-speed, 560 software tools are available [3], which of half are pre-processing the data like data-clustering, the reduction of data dimension and data-normalization. The sums of accessible devices are expanded within the improvement of sequencing advances and creating of maturing cells, qualities and cell populaces[4]. In the analysis of data, a huge amount of good software tools and methods have been carried out and can assist in method selection.

But, these good tools and methods are still need to updating and detailing of an analysis. Even if the previous recommendations are still valuable for processing of data, those are depending on facts of data processing (such as. [5]), it is not calculate how the selections of tools are impact on the analyzed study [6] or it is not implements all data processing aspects, such as doublet identification or filtering of cell (e.g. [7]). Normalization of the microarray sequencing number is a main role that select for cell-to-cell expression, depth of sequencing and another technical aspects.

Normalization is processed by scaling of different cells. For the different cells, the acceptance is that differential expression of genes between the sampled cells. Methods such as DESeq [8] and trimmed mean of M values (TMM) [9] normalization are essentially used here. The method of normalization could be used to characterize data sets. DESeq or TMM normalization is more compelling to differential expression but it is depending on the proportion of number between cells. This can be not clear in sequencing RNA information where the higher recurrence with durable normalization.

A. *Characteristic of Microarray Data*

For the computational methods, microarray data classification bearing an important challenge because of genes with smallest sizes of sample. There are five basic characters for microarray data:

- Small Sample Size
- Class Imbalance
- Data Complexity
- Dataset Shift
- Outliers

Small Sample Size

During past years, the rate of data selection is increased in microarray data, but there is a still remain the small sample size when choosing genes or features and modeling in a medical system. A big key to this problem is that small samples are extremely influenced by the calculation of the error rate, including the low statistical test rate, the mainly used t-test for between classes. If the estimation error rate is inappropriate, the result of classification methods is faulty. In order to overcome this difficulty, the validation method for the classification of error estimation must be chosen.

The efficient way to solve the estimating errors in microarray is to estimate cross-validation, since they are true errors on a general outcome. Nonetheless the difficulty with cross-validation is that it may be deeply overdone by the small sample size. Another solution is Bootstrap method that is increasing the performance corresponding to variance. Moreover, it may expensive the computational cost. Thus, this research should be removed into account when employing an estimation error method on the smallest size of microarray data.

Class Imbalance

Class awkwardness could be an issue when happen a dataset is impacted to a major lesson which has importantly more tests than other uncommon classes within the datasets. If the data is high-dimensional, example microarrays increase the bias to classify the majority class. To overcome these issues are under sampling and over sampling methods. In recent past, an ensemble method for classifiers is as a solution of this problem. Another arrangement to bargain with this issue ought to be one-class SVM qualified as it were with the select lesson which can cause to great a predictive performance.

Data Complexity

Data complexity represents the complexity of classification works, such as overlapping between classes, separability and determining the classification performance. In the microarray data area, several solutions already exist for this complexity measures. A novel approach to get the better of positive dependencies between complexity and error is K-NN classifiers.

Dataset Shift

The datasets are fundamentally apportioned to preparing set and test set is called the dataset shift. Dataset shift occurs in the sequencing of feature, composition of features or the boundaries of class lead to change a testing data happening. This problem solution is using normalization. The validation technique can be used in dataset shift. One of the foremost utilized procedures within the microarray zone is the K-fold cross-validation. Nevertheless, the cross-validation is solving the problem of dataset shift; a dangerous fact is that in accurate estimation. To overcome this dangerous fact, distribution optimally balanced stratified cross-validation (DOB-SCV) [10] is founded on that idea.

Outliers

In microarray datasets, some of the samples are labeled in incorrectly or defined which happened outliers are. These outliers are mostly occurred in the Colon dataset. In case of outlier exceptions are happening within the tests, it would be decided as a preprocessing degree within the classification of microarray spaces. Because of this exception can have a pretentious impact on the gene dataset choice and as a grouping on the prediction of model.

B. *Theoretical Analysis*

There are many normalization algorithms for the analysis of microarray data. However, traditional normalization methods are focused on the sequencing of cell. So, researchers are developing on the trend of variance of genes differential expression of genes and sequencing length. In a microarray data analysis, normalization of gene expression is to able more accurate between expression levels and within samples. A lot of normalization methods are developed such as Trimmed Mean of M values (TMM) AvgDiff, Total Count (TC) and so on. We studied the various normalization methods on differential gene expression data. Here, a lot of normalization algorithms are listed according their data expression as follows:

Normalization Based Gene Read Count

One of the facts is the read count normalization, also called the gene library size normalization. The library gene size means the total count of read gene produced by a sample. The main purpose of read count normalization is to fix the read gene by removing raw read gene in each sample. Many normalization methods have been developed that depend on the desire data. The second and fourth datasets in Table1 are used to apply in gene read count normalization. The gene library size is the total number of genes produced by a sample source and it has the various gene library sizes because of the different sample source. The normalization based gene library size is aim to make the gene library sizes similar by taking off the raw gene counts in each source. Among normalization methods based library size, there are recommended methods are below:

Trimmed Mean of M-values (TMM)

The number of RNA generation may not be regard as being directly, since we cannot estimate the expression depth and accurate lengths of the genes. Although, the correlation of RNA sequencing of two samples are more clearly be considered. In [11] propose a critical procedure that relate the general expression levels of genes between samples lower the acceptance that the mainly of them are not differential expressions. Trimmed mean of M values are expressed as the log expression ratios for weighted trimmed and it is yet robust.

To fully collect the detected M values, it is need to trim the values of both in M and A before counting the average weight. The variances of the gene expression weights are used to take for the case that log relative risks from genes with large amount of accounts have smaller variance on the algorithm. A trimmed is mean the average after cancelling the superior and interior of the data. Normalization of TMM factors between several samples can be computed by choosing the sample as the reference and the other sample is the test reference. The TMM normalization can be construct a statistical model and it is indicate the main role for microarray data.

scran's pooling-based Normalization

The TMM normalization is accurate to differential expression data but this method is depending on the computation of ratios between genes. In DESeq normalization, the mean values are calculated for each gene after cancelling all zeros. This calculation is need to escape the fact that a mainly gene is zero for means values, for example the ratio of gene to mean value cannot be determined. In [12] represent a scran's pooling-based normalization implemented by R package. In this normalization, the DESeq and TMM normalization is used by modifying the processing strategies.

In DESeq normalization, the size factors are calculated by using the estimateSizeFactorForMatrix function [13]. In this function, the zero ratios are cancelled in each gene by removing a factor of size equal to zero. The calcNormFactors function is used in TMM normalization implemented with the edgeR package [14]. For each gene, the calculation and trimming of M values are naturally deleted. The gene library size is denoted by the size factor that is the product of normalization and gene size factor for each gene. For each gene library normalization size, the total genes size was used exactly as the factor size of each gene.

In the DESeq and TMM normalization, the gene expression values are normalized by calculating the values that are dividing each gene expression count by the correlating gene factor size. The expression values are calculated directly on the variation of mean and coefficient of each gene. This approach is beneficially effect on the validity of microarray data.

Relative Log Expression (RLE)

The relative log expression (RLE) normalization is based on the gene hypothesis. In RLE normalization, the scaling factor of gene is calculated as the ratio of median. The median ratio of sample is used as the relation factor of all gene accounts to attain the gene hypothesis [15].

Upper Quantile (UQ)

The upper quantile normalization is deleting the gene with zero gene counts for all sample sources and the last gene counts are divided by the upper quantile of gene counts with no zero and multiplied by the mean upper quantile between the samples [16]. These upper quantile normalization methods are implemented by edgeR package [17].

Sample Based Normalization

Another aspect to analysis the microarray data is based on sample normalization. It is aim to get the accurate data quality and to become aware of biologically to the point of genes. Microarray data has the data variation challenge and sample based normalization is basically depend on the data variation. To solve those challenge, normalization methods such as singular value decomposition (SVD), remove unwanted variation (RUV) and principal component analysis (PCA) are main techniques to removing the unwanted variation from the data.

Remove Unwanted Variation (RUV) is performed the factor analysis on appropriate set of negative samples for removing the variation and it keeps the basic factor. RUV normalization is processed on three sectors in negative control genes, negative control samples and residual RUV. The amount of unwanted variation including the sample is the expression level of genes and EUV should be removed by thinking about the sample size [18, 19]. The singular value decomposition (SVD) method is upon the amount of surrogate variables using "BE" or "Leek" in [20, 21, and 22].

The method of "BE" is directly on a procedure of permutation and "Leek" is to provide an interface proposed in [23]. The PCA is performed by using residual scale matrix of SVD and it can determine the factors of variations. After using one of these three normalization method, a suitable statistical approach such as linear model or ComBat should be used to get the normalized data.

Gene Length Normalization

The microarray data with biological bias nit discovered but discovered in the RNA sequencing data learning is the effect of gene distance on the approximation of abundances. In general, the huge amount of gene count has the larger gene contrast to smaller genes have the difference size of genes. RPKM/FPKM (reads/fragments per kilo-base per million mapped reads) [24, 25 and 26] is used to correct the estimation of bias. Other technique to alter the length of gene is TPM (transcripts per million) method, which draw into accounts of gene length and the sequencing length read.

However, it still remain the suffering of biases that gene sequencing depth and technical effect. The normalization methods based on the total gene count or effective gene count and tend to carry out the slowly when gene samples are transcript distributions. Moreover, RPKM/FPKM and TPM methods are more suitable to apply when the result is to contrast the gene expression level across the gene and the differential gene expression investigate that the levels of gene expressions must be compared between the samples.

Normalization Based Gene Distribution

DBNorm normalized the gene distributed data implemented R package [27]. In a DBNorm R package contains the fitting functions such Polynomial, Fourier and Gaussian distributions. The performance of this method achieves the better results and can also be apply the bioinformatics analysis. The gene distribution normalization change the values from one to another scale but it keep the original scale value. Assume that there are two sources of microarray data, these two sources of data are merged at the same time before normalization. It is aim to obtain the two clusters with one source at one cluster. The distribution of two sources are mixed well based on their source nature and minimize the gap between two sources. The DBNorm normalization is the strongest gene expression normalization methods in current age.

Sctransform's variance stabilizing transformation

In [28] act with the procedure with three steps. Firstly, utilize a linear model using sequencing depth to set the model facts for each gene. In the second procedure, kernel regression is used to estimate the out coming facts or parameters that are directly upon a differential gene expression and this are removing the noisy data. The kernel regression using the ksmooth function to estimate the relationship of model parameter and gene mean values on global trend with R packages. The kernel bandwidth is firstly selected using R function.

The adjustment of factor and independent regularizations are performed using R package. In the final step, negative binominal is performed second time to fix the parameters of the model and it is reduce unique molecular identifier counts into Pearson residuals using affine transformation function. Calculation of the Pearson residuals representing a variance stabilization transformation and that is removing the dependencies of genes expression and variation [28]. These approach describe the procedure of normalization is effectively remove the gene variation and without damaging the biological information.

In theoretical section, we are only focused on the gene read count, gene distribution, variance-stabilizing transformation, sample based and gene length normalization. In addition, other strategies for normalization are also have according to their associated data. Housekeeping gene or other associated gene normalizations are not described.

III. DISCUSSION AND CONCLUSION

During the last period, microarray data normalization has been applied to get an effective and accurate data in the analysis of microarray data research environment. Microarray data analysis is more and more concentrate in sample base normalization to get the effective information with the development of biological gene data. In microarray data, the differential genes expression level is very complex and compact between gene conditions. In addition, several types of noisy are involved because of the various factors are altering during the production of microarray data which can control to the accuracy and precision of data. In order to investigate these improper biological effects, these effects are needed to be collected and removed.

As a purpose to collect and remove biases, data normalization must be performed. In a theoretical analysis section, Trimmed Mean of M-values (TMM), scran's pooling-based normalization method, Relative Log Expression (RLE) and Upper Quantile (UQ) are effective methods for normalization-based gene read count. For sample-based normalization, Singular Value Decomposition (SVD), Remove Unwanted Variation (RUV) and Principal Component Analysis (PCA) are powerful methods. The useful methods in normalization-based gene distribution are DBNorm and Sctransform's variance stabilizing transformation.

Variance present in microarray data, trend in current big data research area is still key challenge. With the development of new technologies and effective methods of normalization, the microarray data will become a valuable for research work, diagnosis analysis and medical data analysis. After normalization, we get the data with greater accuracy and precision and the result data can help to build the better model. This paper reviews the algorithms from many sides of views. As future work, the algorithm with more accurate and to get the better quality of data are need to develop for efficient and effective. The dataset sizes are more abundant and distribute, the researchers are need to compare the accuracy and performance of the model on spark platform.

REFERENCES

[1] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evalu-ation of statistical methods for normalization and differential expression in mRNA-Seq experiments,"BMC Bioinformatics, vol. 11, article 94, 2010..

[2] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," Genome Biology, vol. 11, no. 3, article r25, 2010.

[3] Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA- database.

[4] Svensson V, Beltrame E. d. V., Pachter L. A curated database reveals trends in single cell transcriptomics. bioRxiv.

[5] Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. 2019; 20(269):1–21.https://doi.org/10.1186/s13059-019-1898-6.

[6] Tsuyuzaki K, Sato H, Sato K, Nikaido I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. Genome Biol. 2020;21 (9):1–17.https://doi.org/10.1186/s13059-019-1900-3.

[7] Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun. 2019;10(1):1–11.https://doi.org/10.1038/s41467-019-12266-7

[8] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106.

[9] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010

[10] Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation. DOI: 10.1080/095281300146272

[11] Mark D Robinson, Alicia Oshlack "A scaling normalization method for differential expression analysis of RNA-seq data" Robinson and OshlackGenome Biology2010,11:R25. http://genomebiology.com/2010/11/3/R25

[12] AaronT.L.Lun, Karsten Bach and John C. Marioni "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts" Lunet al. Genome Biology (2016) 17:75 DOI 10.1186/s13059-016-0947-7

[13] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550

[14] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40

[15] Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome biology 11: 1.

[16] Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics 11: 1. https://doi.org/10. 1186/1471-2105-11-1.

[17] Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. https://doi.org/10.1093/ bioinformatics/btp616 PMID: 19910308

[18] Risso D, Ngai J, Speed TP, Dudoit S (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology 32: 896–902. https://doi.org/10.1038/nbt.2931 PMID: 25150836Leek JT (2014) svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic acids research 42: e161–e161

[19] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28: 882–883. https://doi.org/10.1093/bioinformatics/bts034 PMID: 22257669

[20] Leek JT (2011) Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data. Biometrics 67: 344–352. https://doi.org/10.1111/j.1541-0420.2010.01455.x PMID: 20560929

[21] Buja A, Eyuboglu N (1992) Remarks on parallel analysis. Multivariate behavioral research 27: 509– 540. https://doi.org/10.1207/s15327906mbr2704_2 PMID: 26811132

[22] Leek JT (2011) Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data. Biometrics 67: 344–352. https://doi.org/10.1111/j.1541-0420.2010.01455.x PMID: 20560929

[23] Leek JT (2011) Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data. Biometrics 67: 344–352. https://doi.org/10.1111/j.1541-0420.2010.01455.x PMID: 20560929

[24] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5: 621–628. https://doi.org/10.1038/nmeth.1226 PMID: 18516045

[25] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28: 511–515. https://doi.org/10.1038/nbt.1621 PMID: 20436464

[26] Data B A survey of best practices for RNA-seq data analysis.

[27] Qinxue Meng1, Daniel Catchpoole, David Skillicorn and Paul J. Kennedy “DBNorm: normalizing high-density oligonucleotide microarray data based on distributions” Meng et al. BMC Bioinformatics (2017) 18:527 DOI 10.1186/s12859-017-1912-5

[28] Christoph Hafemeister, Rahul Satija “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression “Hafemeister and Satija Genome Biology (2019) 20:296 https://doi.org/10.1186/s13059-019-1874-1.