# Analysis of Normalization Method for DNA Microarray Data

Omar Salem Baans[1], Asral Bahari Jambek[2] and Khairul Anuar Mat Said[3]
[1, 2and 3]*School of Microelectronic Engineering, Universiti Malaysia Perlis, Perlis, Malaysia*
Email: [1]omersalim4901@gmail.com, [2]asral@unimap.edu.my, [3]anuarsaid91@gmail.com

**Abstract - Normalization is a process of removing systematic variation that affects measured gene expression levels in the microarray experiment. The purpose is to get more accurate DNA microarray result by deleting the systematic errors that may have occurred during the making of DNA microarray Image. In this paper, five normalization methods of Global, Lowess, House-keeping, Quantile and Print-tip are discussed. The Print Tip normalization was chosen for its high accuracy (32.89 dB and its final MA graph shape was well normalized. Print tip normalization with PSNR value of 33.15dB has been chosen as a new normalization method. The results were validated using four images from the formal database for DNA microarray data. The new proposed method showed more accurate results than the existing methods in term of four parameters: MSE, PSNR, RMSE and MAE.**

***Keywords- Normalization, Global, Lowess, House-keeping, Quantile, Print-tip, microarray, Background correction, M-A plot, DNA.***

## I. INTRODUCTION

Gene expression measurements provide clues about the regulatory mechanism, biochemical pathways and broader cellular function. By gene expression, it can be understood as the transformation process of gene's information into proteins. The formal transformational pathway of protein begins from DNA (deoxyribonucleic acid) which is copied to the mRNA (messenger ribonucleic acid) and, finally, this molecule passes from nucleus to cytoplasm carrying the information to build up proteins [1].

There are many microarray analysis software packages available on the market whether commercial or freeware. Basically, each software program can be separated into three main tasks: (1) gridding or addressing, which is the process of specifying coordinate to every spot on the slide. (2) the segmentation which decides the classification of each pixel either as foreground which corresponds to be an interesting spot or as background which acts as an error or noise. (3) the Intensity Extraction which is the step to calculate green and red for foreground fluorescence intensity for each spot on the array [2, 3, 4].

Subsequently, there are many processes to inspect the results and also to correct the errors that have occurred. The background correction method which ignores the effect of intensity of the background. This can be achieved by subtracting the value of the background intensity from the value of foreground intensity or any other suitable method to neglect the effect of background intensity. Another process to increase the accuracy is the normalization method which we are going to discuss in this paper [5, 6].

Normalization is a process of removing systematic variations that affect measured gene expression levels in microarray experiments. The purpose of normalization is to adjust for effects which arise from variations in the microarray technology rather than from biological differences between the RNA samples or between the printed probes. Imbalances between the red and green dyes may arise from differences between the labeling efficiencies or scanning properties of the two flours complications perhaps by the use of different scanner settings [7, 8]. The aim of the paper is to review and make some comparison between various methods in microarray data normalization.

Section II discusses the purpose of normalization. Several normalization algorithms are discussed in section III, while section IV discusses the comparison of the different methods. Section V and VI represent methodology and results of each method and section VII conclude this paper.

## II. LITERATURE REVIEW

### *A. Purpose of normalization and normalization expression graphs*

The purpose of normalization is to adjust for effects which occur from variations in the microarray technology rather than from the biological differences between the RNA samples or between the printed probes. It can be regarded as a sort of calibration process that improves the comparability among microarrays treated alike. Imbalances between the red and green intensities may arise from differences between the labeling efficiencies or scanning properties of the two slides which may due to the use of different scanner settings. If the imbalance is more complicated than a simple scaling of one channel relative to the other, then a function of normalization will need to be performed. As an example of the importance of the normalization process, by comparing Fig. 1 with Fig. 2, a different was observed once the background correction is ignored. Fig. 1 represents M-A plot for a red and green intensity before correcting the background values, thus, it shows irregular distributions of the spot around the plot. However, there is a spot regulation in Fig. 2 for normalized intensities. For more details refer to [9, 10].

Published Paper Link: https://doi.org/10.35118/apjmbb.2019.027.4.04

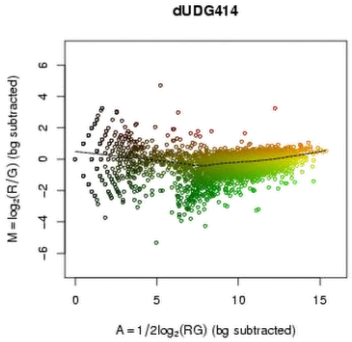


Fig. 1: M-A Plot for No-Background Corrected Slide [9]

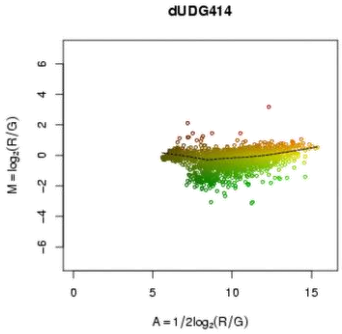


Fig. 2: M-A Plot for Background Corrected Slide [9]

### B. *Normalization Graph Expression*

Normalization can be expressed in two types of graphs. First one is the logarithm of the red intensity versus the logarithm of the green intensity (log R vs. log G) as shown in Fig. 3. The second one is M-A plot, it is 45° rotation of standard scatter plot as shown in Fig. 4. Write R (Red intensity) and G (Green intensity) for the background-corrected red and green intensities for each spot, normalization is usually applied to the log-ratios of expression, which will be written as in Equation (1). The mean of log-intensity of each spot will be written as in Equation (2), a measure of the overall brightness of the spot. (The letter M is a mnemonic for minus while A is a mnemonic for addition) [11].

$$M = \log R - \log G \quad (1)$$

$$A = (\log R + \log G)/2 \quad (2)$$

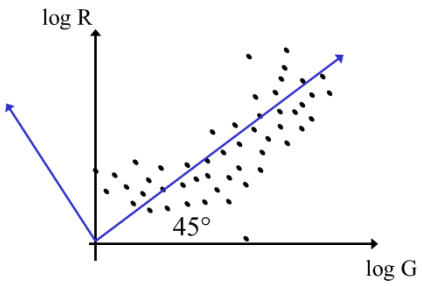


Fig. 3: Log R vs. Log G

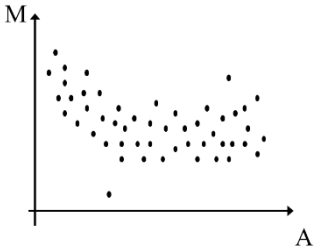


Fig. 4: M-A Plot

### C. *The latest trend in microarray normalization*

There are many studies on the DNA microarray normalization. As a result, many methods were created and their results were drawn as M-A plot or any other type of plots representation. This section will discuss and elaborate these methods in order to choose the most suitable one and develop it for further microarray analysis.

The first method is Global normalization, the underlying assumption of this approach is that the total of mRNA labeled with either R-value (sum of red intensities) or G value (sum of green intensities) is equal. While the intensity for any spot may be higher in one channel than the other, when averaged over thousands of spots in the array, these fluctuations should average out. Consequently, in this method, it takes the value of c out of a log (R/G). The c value is equal to the main assumption that equal to the log of the total R (Red intensity) over total G (Green intensity) which can be expressed by the variable K, Equation (3) and (4) explain this method [12].

$$log_2(R/G) \rightarrow log_2(R/G) - c = log_2 R/(I \quad (3)$$

$$K = \sum (R/(G) \quad (4)$$

House-keeping method is a similar method that uses a fixed value to subtract or add to the (M) value. However, this method requires a specific gene call house-keeping gene. The expression of the house-keeping gene is assumed to be constant. Therefore, after hybridization, the intensity of these genes is identified and the difference should be calculated which would be used later for normalizing the other genes. [13]

The intensity-dependent normalization (Lowess) runs a line through the middle of the MA plot, shifting the M value of the pair (A, M) by m = mean (M), as shown in Equation (5). One estimate of m is made using the Lowess function (Locally Weighted Scatterplot Smoothing). As in Fig. 5 and Fig. 6, the difference between Global and Lowess normalization can be noticed in M-A plot form [14].

(5)

Published Paper Link: https://doi.org/10.35118/apjmbb.2019.027.4.04

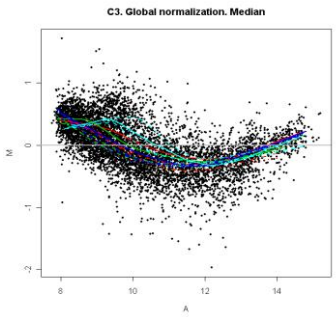


Fig. 5: Global normalization [12]

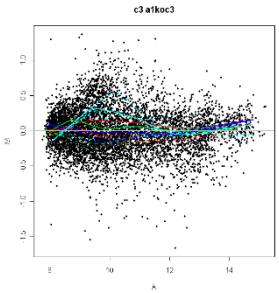


Fig. 6: Lowess normalization [12]

The Print-tip normalization is similar to Lowess normalization that repeating itself in groups, where each group is separated than the others. Thus, Print Tip normalization starts by dividing A value ([Log R + Log G]/2) into tip groups. Then, each group is normalized by subtracting its M-value (Log R – Log G) from its corresponding value (lowess(A)) of the tip group as in Equation (2). This value (lowess(A)) is equal to the mean of M value inside each tip group. The normalized log-ratios (N) will replace the M values to restore back the red and green intensities. A simpler form of Print-tip is shown in Equation (6) where lowess (A) is the global loess curve plotted in Fig. 7. Refer to Fig. 8 for the final figure of the Print-tip normalization [15].

$$N = M - \text{lowess}(A) \quad (6)$$

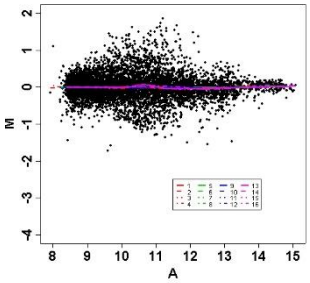


Fig. 7: After Print-tip normalization [12]

.

Lastly is the Quantile normalization method which is also one of the most favorable approaches used especially in normalization between arrays. First, rearrange the genes in each column as in the second table in Fig. 8. Then, take the mean in each row and replace the whole row by the mean value as in the third table in Fig. 8. Finally, reorder each gene in its original place with its new value [5]

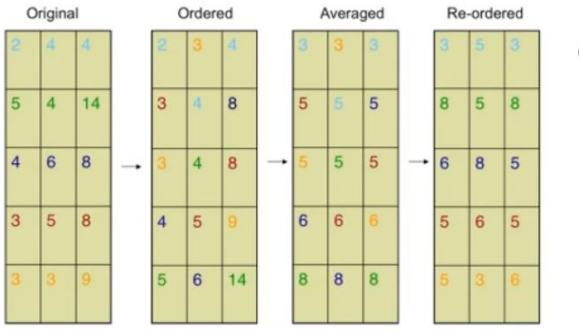


Fig. 8: Quantile normalization

### *D. Comparison of different Normalization approaches*

In this section, the existing system algorithm as discussed in section III will be analyzed and discussed to find out the similarities and variations among the different normalization methods. Table 1 summarized the comparison of these algorithms.

TABLE 1 COMPARISON BETWEEN DIFFERENT SYSTEM ALGORITHMS

| No. | (Yang Et Al., 2012) | (Berger Et Al., 2004) | (Martin Et Al., 2002) | (Smyth & Speed, 2013) | (Adriaens Et Al., 2012) |
|---|---|---|---|---|---|
| Year | 2012 | 2004 | 2002 | 2013 | 2012 |
| Method | Global | Lowess | House Keeping | Print Tip | Quantile |
| Function | Log (R/ KG) | Log ( R/ G) – C(A) | N = M – Housekeeping Value | N= M-Loess (A) | Mean Of Rows after Reorder |
| Variable | K = sum(R) / sum(G) | LOWESS Function | House Keeping | Global Loess | NA |

From table 1, it can be seen that all methods used are mainly the value of M which equal to log of red intensity minus log of green intensity. However, three methods have different value to subtract from M. To illustrate, Global normalization use the log of the addition of each of red and green intensity while the other two methods are using median and global median.

In term of the final shape of the normalization on M-A graph, there are similarities between Lowess and Print-tip methods because both have a straight median line in the value of (M = 0) due to their similarities on subtracting the mean or median from M. However, in Global normalization, there is a curve around the value of (M= 0) due to the subtraction of the total R and G.

House-keeping and Quantile normalization methods do not use M-A plot, consequently, their final graphs do not always take a straight line of the mean on the (M=0). In addition, house-keeping requires knowing the expected intensity value of its genes to compare it with the final intensity value; therefore, it is

Published Paper Link: https://doi.org/10.35118/apjmbb.2019.027.4.04

difficult to examine it in this project. The main reason for testing the house-keeping method because it is one of the main and most commonly use fixed normalization type methods.

According to this review, we suggest Print-tip normalization method to be used because when comparing to the global normalization, its final figure is simpler and easier to read, and can also be compared easily to various plots. A straight line (M=0) is easier to read than the Global normalization curve. However, when it is compared to loess normalization's final figure, there was not much different in the value of M after the normalization and thus, in the end, the M value is noticeable [15].

*E. Results Validation Parameters:*

In order to choose the most accurate method, as discussed by Chaurasia et al [16], four parameters were used to compare each method results with the Princeton results. These parameters are:

*1. MSE (Mean square Error)*: is defined as some sort of average or sum (or integral) of the square of the error between two intensities as In Equation 3.

$$MSE=1/(N*M)\times \sum (x(i,j)-y(i,j)) \wedge 2 \quad (7)$$

Where: x(i, j) is the original intensity from Princeton, Y(i, j) is the intensity for a specific method, M and N are the dimensions of the image.

*2. Peak Signal to Noise Ratio (PSNR):* defined as the ratio between signal variance and reconstruction error variance as in Equation 4.

$$PSNR=20 * \log (max)-10\times \log (MSE) \quad (8)$$

Where mean squared error (MSE) and max is the maximum possible pixel value of the intensity

*3. RMSE (root mean square error):* is defined as the square root of mean square error as in Equation 5.

$$RMSE = \sqrt{MSE} \quad (9)$$

*4. MAE (maximum absolute error):* is defined as the maximum absolute value, the difference between Princeton intensity and one of the reviewed methods as in Equation 6.

$$MAE=Max (|x(i,j) - y(i,j)|) \quad (10)$$

In this work, Matlab version R2013a 9.0 and its Image Processing Toolbox which supports an extensive range of image processing operations are used for data analysis and technical computing due to its high performance and powerful language. This work is implemented using a personal computer with a processor: Intel (R) Core (TM) i3 −1.80 GHz.

## III. Methodology

Following the previous steps and in order to examine the suitable method which would provide more accurate algorithm among normalization algorithms that was reviewed in Section II, four DNA microarray images were used as shown in Fig. 9, these images are from Princeton University microarray database. The formula codes were applied according to normalization methods that have been discussed in Section two. These normalization methods are Global normalization, Lowess normalization, Print Tip normalization and Quantile normalization. However, housekeeping normalization will not be examined because it requires house-keeping gene from the manufacturer. Princeton University microarray database provide the measured intensity information for each image. Thus, this information was used as a reference to compare and validate this research results according to four parameters, these parameters are MSE, PSNR, RMSE and MAE. For an overview of the parameters, the reader should be referred to Section II.

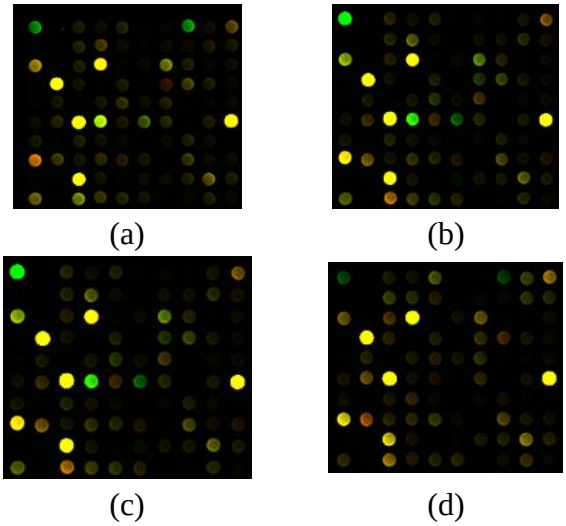


Fig. 9: Princeton DNA Microarray Images

*A. The Proposed Method*

The Lowess normalization and Print Tip showed the best result for normalization as proved by Baans et al [17]. Therefore, in this research, a new mixed algorithm of these two was proposed and its result was compared with all other existing methods. This new algorithm starts by calculating the Lowess value (m) according to Equation (5) and the Print Tip values (PT) for each Print Tip period as they were explained in Section II. Then, the Print Tip values (PT) for each Print Tip period would be replaced by a new value. This new value equals the mean of Lowess value (m) and Print Tip values (PT) for each Print Tip period as in Equation (5). After that, the PT value was subtracted from M in each PT Group as in Equation (11). Finally, the normalized red and green intensities are calculated using A and PTnew values according to Equations 12 and 13.

$$PTnew= (m+PT(i))/2 \quad (11)$$

$$Rn= \surd(2\wedge(2.a+m)) \quad (12)$$

$$Gn= \surd(2\wedge(2.a-m)) \quad (13)$$

*B. Result's Validation*

In Term of four parameters, MSE, PSNR, RMSE and MAE, the new normalization algorithm validated by comparing its results with the existing algorithm using Princeton results as a reference for the four images in Fig. 9, these images are real pictures obtained from a public database of the Princeton University microarray database. It is important to notice that the databases calculations are in a form of 16 bits while the algorithm that used in this work is in 8 bits. Therefore, it is compulsory to change the databases to 8 bits forms before comparing it with this project's result. In this section, it is also

Published Paper Link: https://doi.org/10.35118/apjmbb.2019.027.4.04

important to notice that the red and green intensity will not be considered. Therefore, the error will be calculated for 200 spots regardless of the colors whether it is red or green.

## IV. Result and Discussion

This work discusses five methods for normalization applied on four DNA microarray images in Fig. 9. The spots intensity results for these methods were compared with the Princeton databases. The comparison was done depending on four basic parameters. These parameters are PSNR (Peak Signal to Noise Ratio), MSE (Mean Square Error), RMSE (Root Mean Square Error) and MAE (Maximum Absolute Error) as they were discussed in Section 2.

From these tables, Table 2, Table 3, Table 4 and Table 5, it can be seen that the new proposed algorithm gave more accurate results especially in term of MAE in all the tables. However, in term of MSE, PSNR and RMSE, the new algorithm was the best algorithm in Table 2 and Table 3 only. To illustrate, Print Tip showed the best results regarding these three parameters in Table 3 while Lowess normalization was the best in Table 1. However, the Quantile normalization gave huge amount of error.

PSNR (Peak Signal to Noise Ratio) for normalization presents much lower results than the PSNR results for intensity extraction as in II by around 10 dB units for each image. This is because the normalization comes after intensity extraction, so the noises that were appeared during the intensity extraction, it also would be included for normalization.

Table 2: Accuracy of Normalization Methods Applies on Princeton Image (A) in Fig.

| Method | MSE | PSNR | RMSE | MAE |
|---|---|---|---|---|
| Global | 39.24 | 32.23 | 6.26 | 13.23 |
| Lowess | 43.05 | 31.83 | 6.56 | 14.00 |
| Quantile | 1444.03 | 16.57 | 38.00 | 168.37 |
| Print Tip | 38.36 | 32.33 | 6.19 | 15.84 |
| New | 36.40 | 32.55 | 6.03 | 12.21 |

Table 3 Accuracy of Normalization Methods Applies on Princeton Image (B) in Fig.

| Method | MSE | PSNR | RMSE | MAE |
|---|---|---|---|---|
| Global | 37.82 | 32.39 | 6.15 | 17.02 |
| Lowess | 60.28 | 30.36 | 7.76 | 24.74 |
| Quantile | 1670.65 | 15.94 | 40.87 | 171.00 |
| Print Tip | 36.61 | 32.53 | 6.05 | 16.44 |
| New | 41.61 | 31.97 | 6.45 | 13.57 |

Table 4 Accuracy of Accuracy of Normalization Methods Applies on Princeton Image (C) in Fig.

| Method | MSE | PSNR | RMSE | MAE |
|---|---|---|---|---|
| Global | 31.15 | 33.23 | 5.58 | 14.67 |
| Lowess | 33.41 | 32.93 | 5.78 | 13.67 |
| Quantile | 1219.27 | 17.30 | 34.92 | 135.21 |
| Print Tip | 32.86 | 33.00 | 5.73 | 15.55 |
| New | 30.75 | 33.29 | 5.55 | 10.80 |

Table 5 Accuracy of Normalization Methods Applies on Princeton Image (D) in Fig.

| Method | MSE | PSNR | RMSE | MAE |
|---|---|---|---|---|
| Global | 22.01 | 34.74 | 4.69 | 14.25 |
| Lowess | 21.35 | 34.87 | 4.62 | 15.25 |
| Quantile | 2527.36 | 14.14 | 50.27 | 254.97 |
| Print Tip | 27.85 | 33.72 | 5.28 | 16.18 |
| New | 21.76 | 34.79 | 4.67 | 11.18 |

These findings support the finding of Smyth et al [15] as he mentioned that the “print-tip loess normalization provides a well-tested general purpose normalization method which gives good results on a wide variety of arrays”. It is best combined with diagnostic plots of the data. When the diagnostic plots show that biases still remain in the data after normalization, further normalization steps such as house-keeping or quantile normalization between the arrays may be undertaken. Besides that, the new algorithm represented the most accurate results than all other existing methods.

## V. Conclusion

In this paper, Normalization is defined as a process to delete systematic error which is why it is important and necessary. Since there are many normalization methods that exist, five most commonly used normalization algorithms such as Global, Lowess, House-keeping, Quantile, and Print-tip have been tested and compared to find the most suitable approach in a general normalization process. For that purpose, a Matlab code was built for each method for two slides; the ideal and real microarray slides. The results were shown in two forms, Table of red and green intensities and M-A graph. The results show that Global, Lowess, and Print-tip have more accurate result once compared with an ideal image result while Print-tip has the advantages than the other two especially in term of final graph shape. By combining Lowess and Print Tip normalization, a new algorithm for normalization was proposed and applied on four DNA microarray image from Princeton website. Using Princeton results, this new algorithm was compared with the existing algorithms; the results validate this algorithm as one of the best algorithms for DNA microarray normalizations.

## Acknowledgment

The author would like to acknowledge the support from the Sciencefund under a grant number of 03-01-15-SF0229 from the Ministry of Science, Technology & Innovation, Malaysia.

## Reference

[1] Mark Schena, “Micropuce Biochip Technology,” Oxford University Press, 1999.

[2] Borda, Monica, et al. "FPGA based SoC for automated cDNA microarray image processing." E-Health and Bioengineering Conference (EHB), 2011. IEEE, 2011..

[3] Rao, Youlan, et al. "A comparison of normalization techniques for microRNA microarray data." *Statistical*

Published Paper Link: https://doi.org/10.35118/apjmbb.2019.027.4.04

*applications in genetics and molecular biology* 7.1 (2008).

[4] Hovatta, I., Kimppa, K., Lehmussola, A., Pasanen, T., Saarela, J., Saarikko, I., ... & Vihinen, M. (2005). DNA microarray data analysis. CSC, *2nd edn., Scientific Computing Ltd*..

[5] Yang, Yee Hwa, Michael J. Buckley, and Terence P. Speed. "Analysis of cDNA microarray images." Briefings in bioinformatics 2.4 (2001): 341-349.

[6] Seidel, Chris. "Introduction to DNA microarrays." *Analysis of micro array data: a network-based approach 1 (2008): 1*.

[7] Geeleher MP, Morris D, Golden A, Hinde J. BioconductorBuntu User's Manual.

[8] Karakach, Tobias K., et al. "An introduction to DNA microarrays for gene expression analysis." *Chemometrics and Intelligent Laboratory Systems104.1* (2010): 28-52.

[9] Adriaens, Michiel E., et al. "An evaluation of two-channel ChIP-on-chip and DNA methylation microarray normalization strategies." BMC genomics 13.1 (2012): 1.

[10] Babu, M. Madan. "Introduction to microarray data analysis." *Computational genomics: Theory and application* (2004): 225-249.

[11] Dudoit, Sandrine, et al. "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." *Statistica sinica* (2002): 111-139..

[12] Yang, Yee Hwa, et al. "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." *Nucleic acids research* 30.4 (2002): e15-e15.

[13] Bilban, Martin, et al. "Normalizing DNA microarray data." Current Issues in Molecular Biology 4 (2002): 57-64.

[14] Berger, John A., et al. "Optimized LOWESS normalization parameter selection for DNA microarray data." BMC bioinformatics 5.1 (2004): 1.

[15] Smyth, Gordon K., and Terry Speed. "Normalization of cDNA microarray data." Methods 31.4 (2003): 265-273.

[16] Chaurasia, K., & Sharma, N. (2015). Performance Evaluation and Comparison of Different Noise, apply on PNGImage Format used in Deconvolution Wiener filter ( FFT ) Algorithm Kalpana Chaurasia and Mrs. Nidhi Sharma *. Evolving Trends in Engineering and Technology, 4, 8–14. http://doi.org/10.18052/www.scipress.com/ETET.4.8.

[17] Baans, O., Jambek, A., Hashim, U., & Azah, N. (2016). Performance Comparison of Image Normalization Method for DNA Microarray Data. Pertanika J. Sci. & Technol., 25, 59–68. Retrieved from http://www.journals-jd.upm.edu.my/Pertanika PAPERS/JST Vol. 25 (S) Jan. 2017/7-JTS(S)-0083-2016-4thProof.pdf.

[18] Exploration and analysis of DNA microarray and protein array data, 2004. Dhammika Amaratunga & Javier Cabrera.Wiley & Sons, Inc .

Published Paper Link: https://doi.org/10.35118/apjmbb.2019.027.4.04