The 2017 IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017), September 12-14, 2017

Software Profiling Analysis for DNA Microarray Image Processing Algorithm

Omar Salem Baans
School of Microelectronic Engineering
University Malaysia Perlis
Perlis, Malaysia
omersalim4901@gmail.com

Asral Bahari Jambek
School of Microelectronic Engineering
University Malaysia Perlis
Perlis, Malaysia
asral@unimap.edu.my

***Abstract*— Microarray analysis is one of the most suitable tools available for scientists concerned with DNA sequences to study and examine gene expression. Through microarray analysis, the gene expression sequence can be obtained and biological information on many diseases can be acquired. The gene expression information contained in the microarray can be extracted using image-processing techniques. Microarray image processing consists of three main steps: gridding, segmentation and intensity extraction. This paper analyses the computational time for this microarray image processing. The results show that the intensity extraction consumes majority of the overall computational time. More detail analysis reveals that this high computational time is due to the background correction part of the process, as discussed in the second part of this paper.**

Keywords*— DNA; Microarray; software profiling; image processing; gridding, normalization; intensity extraction.

I. Introduction

Gene expression regulates the production of proteins which control all cellular processes in the biological system of the human being. Understanding gene expression and the mechanism of protein production is important for many applications in terms of diagnosis and finding suitable treatments for diseases. Using the cDNA microarray, it is possible to diagnose rapidly and efficiently the level of gene expression in the sample [1], [2].

The single-stranded molecule mRNA, obtained from the original DNA degrades easily. Therefore it is transformed into a stable double-stranded molecule called complementary DNA for further examination. Microarray technology is based on forming cDNA on a matrix, where each matrix element represents a spot that carrying gene expression probes. Each of these spots undergoes hybridization with the two samples: normal as the reference and cancer as the target. The samples have been labeled using green (reference) and red (target) dyes. The array is then scanned by two light sources to analyze the hybridization process. If the spot is indicating green dye, it shows a high expression of the normal sample while the red dye shows that a high expression of the target is found on that gene. In addition, black dye shows there is no expression from either sample. Yellow dye shows an equal expression of both samples on that spot. When the hybridization is completed, the gene expression is scanned and the microarray image is printed out [3].

There are many microarray analysis software packages available in the market, both commercial and freeware. Each software program can be separated into three main tasks. The first is gridding or addressing, which is the process of specifying coordinates for every spot on the slide. The second is segmentation which classifies each pixel as either foreground, corresponding to a spot of interest, or as background, which is regarded as error or noise. The third and final task is intensity extraction which is the step where green and red fluorescence intensity is calculated for each foreground spot on the array [4], [5].

The intensity extraction step can be calculated using more than one step to obtain the most accurate results. In this paper, it was calculated in two steps. The first step is background correction to remove the effect of the intensity of the background. This is done by subtracting the value of the background intensity from the value of foreground intensity. The second step is normalization to remove any systematic variation that may arise during the scanning process. This may be due to the use of different scanner settings and may affect the measured gene expression levels in the microarray experiment [6].

The aim of this paper is to examine and analyze the profiling of DNA microarray image-processing steps. Section II reviews existing work on DNA microarray image processing. The methodology of this project will be elaborated in Section III, while Section IV explores and discusses the results of this study. Section VII concludes the paper

II. LITERATURE REVIEW

Various researchers have studied the development of microarray image processing. The first step in processing microarray image is gridding. The existing tools for allocating the grid structure in a microarray image often require human intervention which causes variations in the gene expression results. Figure 1 shows an example of a gridded microarray image. Gridding can be achieved as follows. First, the image is converted into a grayscale image before autocorrelation profile is obtained for the image. Then, the horizontal and vertical profile is calculated before the central region and the gap

Citation: O. S. Baans and A. B. Jambek, "Software profiling analysis for DNA microarray image processing algorithm," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, 2017, pp. 129-132.

Published manuscript: DOI: 10.1109/ICSIPA.2017.8120592

between each pair of neighbor spots is defined. Another way of gridding is to perform the gridding manually according to the known size of each spot. Eventhough this method require lesser time, this method tend to gives an incorrect result, especially when the spot sizes are unequal [7, [8], [9].

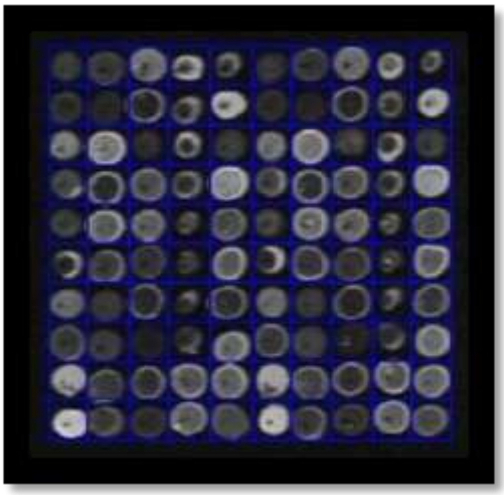


Figure 1: Gridded microarray image.

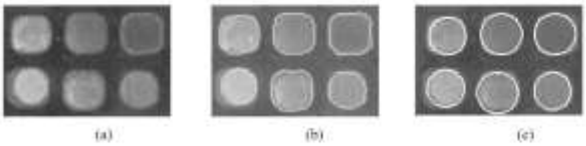


Figure 2: Comparison of segmentation results: (a) shows the original spot image, (b) is the segmented spot image using morphological segmentation and (c) is the image segmented by GenePix [10].

The second stage of microarray image processing is segmentation. Various segmentation methods has been developed by researchers such as algorithm based on k-means clustering, mathematical morphological and histogram analysis. Figure 2 shows the difference result between GenePix and mathematical morphological segmentation [10]. In the third step, intensity extraction of the spot is calculated by taking the median of the pixels inside the spot. Each spot consists of many pixels and the intensities of these pixels are not the same. However, the intensity of the spot consists of both background intensity and foreground intensity. Thus, the background correction must be performed before taking the median value of the spot.

The standard background correction method is usually applied by taking the mean of the foreground values and the median of the background values, and subtracting the latter from the former to define the true value of the spot intensity. Sometimes the background value is greater than the foreground intensity value. Thus the final result will be negative for that spot. Therefore, many researchers have attempted to develop a more logical and accurate method. A method to avoid negative intensity values was presented by Edwards using the local median of the background value [12]. This method is similar to the standard method but when the resultant value is less than a specific small threshold value, a monotonic function is applied, whereas when it is larger than the threshold value, it is considered to be the true intensity of the spot.

After background correction, the final step in microarray image processing is normalization [11]. One of the method used in this step is Print-tip normalization [13]. In this method, each M-value (Log R – Log G) is normalized by subtracting from it the corresponding value of the tip group loess curve that is dependent on the A value ([Log R + Log G]/2) whose value should be fixed. The normalized log-ratios (N) are the residuals from the tip group loess regressions. A simpler form of Print-tip is shown in Equations 1 and 2 where loess (A) is the global loess curve plotted in Figure 3. Figure 4 shows the final figure of the Print-tip normalization [13].

$$(\text{Log } R/G) - C\ (A) = \log R/\ (k\ (A)\ G) \ \ldots\ldots\ldots\ldots \quad (1)$$

$$N = M - \text{loess}\ (A) \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

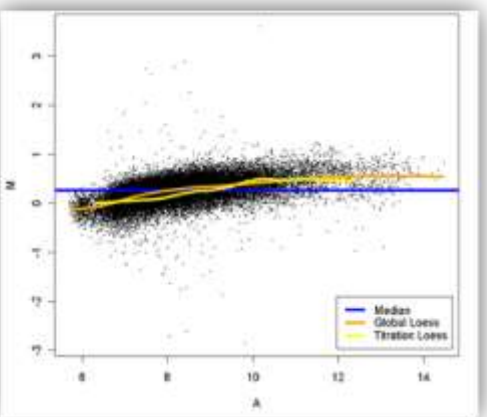


Figure 3: Before Normalization

Citation: O. S. Baans and A. B. Jambek, "Software profiling analysis for DNA microarray image processing algorithm," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, 2017, pp. 129-132.

Published manuscript: DOI: 10.1109/ICSIPA.2017.8120592

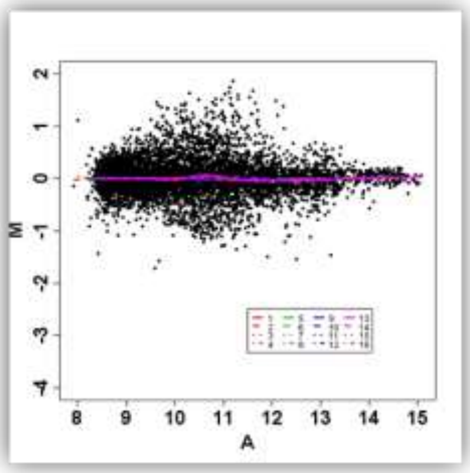
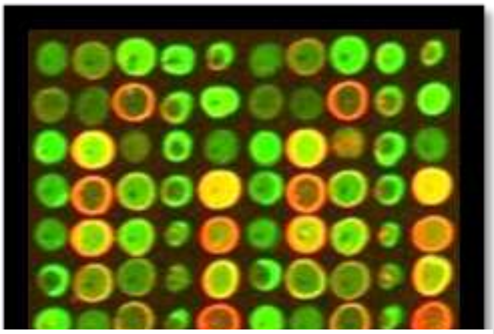


Figure 4: After Normalization

III. METHODOLOGY

The microarray image processing algorithm as provided in [16] is used as the initial case study in this work. For this experiments, the computational profiling of the code is analysed in two part. The first part was for profiling the overall DNA microarray analysis steps, consisting of gridding, segmentation and intensity extraction. The second part was for profiling the intensity extraction steps. This includes background correction and normalization.

In the first part, we import a real DNA microarray slide image and crop it to 100 spots. Cropping part of the image instead of using the whole microarray slide makes the process easier and simpler for the software to handle. The real microarray slide is shown in Figure 5 was used [16]. An ideal microarray image with 100 spots was created manually as a control image is shown in Figure 6. The control image allows us to monitor the accuracy of the algorithm during the whole experiments.



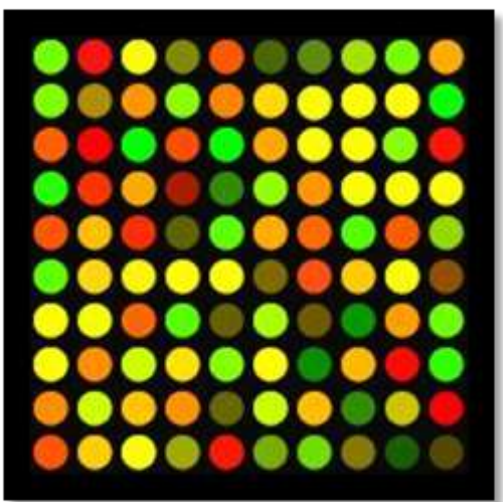


Figure 5: Microarray Image real slide with 100 spots

Figure 6: Microarray Image Ideal slide with 100 spots.

The code is divided into three main functions according to the microarray steps: gridding, segmentation and intensity extraction. The algorithm starts by importing the whole microarray image slide using the 'imread' command. Then, the image is cropped to 100 spots as shown in Figures 5 and 6 using the command 'imcrop'. The image intensity is read as a matrix of pixels. In this work, this image with size of 220 x 227 pixels is cropped. Thus, each spot extends to around 20 pixel diameters in average. The second part of the experiment focuses in profiling the intensity extraction part. This part was divided into two steps: background correction and normalization.

During the computational time profiling, the variation in the results could occurred depending on the state of the PC. Therefore, the measurement of the computation time is repeated five times, and the average and percentage of each step for these five times is taken. The programs are simulated using a laptop computer (processor: Intel(R) Core (TM) i3-3217U CPU @ 1.80GHz, 5.00 GB RAM) and using MATLAB software version 2013a. .

IV. RESULTS AND DISCUSSION

As mentioned previously in the methodology section, the experiment is performed in two part. The first part profiles the overall DNA microarray image processing, while the second part shows the results for profiling the intensity extraction step only.

A. *Profiling DNA Microarray Image Processing:*

The profiling for overall microarray image processing algorithm is showed in Table 1. From the table, it can be shown that the intensity extraction took 1.75s which represents around 77.6% of the DNA microarray image processing time. At the same time, it can be seen that gridding took very little time, i.e., 0.16s which is only 7% of the whole process. Similarly, segmentation took only a small percentage (15%), though somewhat more than gridding.

The gridding and segmentation steps are performed after the image is converted into grayscale. Therefore, the size of the image when these two functions are performed is equal to one third of the image size used during intensity extraction. The latter case uses RGB image which consists of three colors: red, green and blue. The intensity extraction step also includes many repeated loops and functions that also require much time such as 'imtophat' and 'imcrop'. Segmentation requires only the grayscale image and does not take much time other than during mathematical morphological computation.

TABLE I. PROFILING DNA MICROARRAY IMAGE PROCESSING

step	**1st (s)**	**2nd (s)**	**3rd (s)**	**Average (s)**	**Percentage (%)**
gridding	0.16	0.156	0.157	0.16	6.93
segmentation	0.344	0.343	0.36	0.349	15.47
intens_extract	1.749	1.792	1.71	1.75	77.60

B. Profiling DNA Microarray Intensity Extraction:

The profiling result for the second part of the task is shown in Table 2. From the table, it can be shown that background correction consumed 0.35s, which is equivalent to 92% of the intensity extraction computational time. Background correction crops the image 100 times, once for each spot, and takes the median value 200 times, twice for each spot, i.e., once for red and again for green. These huge amounts of repetition for an image with three components (red, green and blue) are the main reason for the high time consumption for background correction. This directly contributed to the total computational time for intensity extraction as discussed in the previous section. On the other hand, normalization consumes only 0.029s (7.7%) of the intensity extraction time due to its simple mathematical and logical operations involved. These results support the work in [14], which states that gridding does not require an RGB image because grayscale is enough to define each spot location [11]. However, gridding could consumes more time than segmentation when using an RGB image [15].

TABLE II. PROFILING DNA MICROARRAY INTENSITY EXTRACTION

step	1st (s)	2nd (s)	3rd (s)	Average (s)	Percentage (%)
background	0.347	0.338	0.363	0.35	92.25
normalization	0.029	0.029	0.03	0.029	7.7

IV. CONCLUSION

In this paper, the profiling of the DNA microarray image-processing algorithm is discussed. During the experiment, the task is divided into two part. The first part of the experiment divides the algorithm into three functions according to the microarray image-processing steps: gridding, segmentation and intensity extraction. The second part of the experiment concentrated on the intensity extraction step which consist of background correction and normalization. In general, the intensity extraction consumed majority of the time for the microarray process. This was mainly due to the background correction part, as has been proven in the second part of this work. Gridding and segmentation consume very little time because grayscale image was used as compare the RGB format during intensity extraction.

ACKNOWLEDGMENT

This research was funded by ScienceFund, Ministry of Science, Technology and Innovation (MOSTI), Malaysia (2015).

REFERENCES

[1] Kooperberg, Charles, et al. "Improved background correction for spotted DNA microarrays." Journal of Computational Biology 9.1 (2002): 55-66.

[2] Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences. 2000 Aug 29;97(18):9834-9.

[3] Rao, Youlan, et al. "A comparison of normalization techniques for microRNA microarray data." Statistical 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] Ahmad MM, Jambek AB, bin Mashor MY. A study on microarray image gridding techniques for DNA analysis. InElectronic Design (ICED), 2014 2nd International Conference on 2014 Aug 19 (pp. 171-175). IEEE.

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

[7] Adwan, Omar, Amjad Hudaib, Hussam Nawwaf Fakhouri, and Loai M. Alnemer. "Microarray Gene Expression Extraction and Archiving Based on Local Spot Locations." (2013)..

[8] Mabrouk, Mai, Islam Fouad, and Amr Sharawy. "Automated statistical and morphological based gridding methods for noisy microarray image processing." Journal of Bioinformatics and Intelligent Control 2.4 (2013): 265-275..

[9] P.Thamaraimanalan et al, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, February- 2014, pg. 353-360.

[10] Fielden, M. R., et al. "GP3: GenePix post-processing program for automated analysis of raw microarray data." Bioinformatics 18.5 (2002): 771-773.

[11] Yang, Yee Hwa, et al. "Comparison of methods for image analysis on cDNA microarray data." Journal of computational and graphical statistics 11.1 (2002): 108-136.

[12] Edwards, David. "Non-linear normalization and background correction in one-channel cDNA microarray studies." Bioinformatics 19.7 (2003): 825-833

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

[14] Alhadidi, Basim, Hussam Nawwaf Fakhouri, and Omar S. AlMousa. "cDNA microarray genome image processing using fixed spot position." American Journal of Applied Sciences 3.2 (2006): 1730-1734.

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

[16] Robert Bemis, "DNA MicroArray Image Processing Case Study", https://www.mathworks.com/matlabcentral/fileexchange/2573-dna-microarray-image-processing-case-study

Citation: O. S. Baans and A. B. Jambek, "Software profiling analysis for DNA microarray image processing algorithm," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, 2017, pp. 129-132.

Published manuscript: DOI: 10.1109/ICSIPA.2017.8120592