INTENSITY EXTRACTION AND NORMALIZATION ALGORITHM DEVELOPMENT FOR DNA MICROARRAY IMAGE PROCESSING

by

OMAR SALEM NASSER BAANS
(1630312155)

A thesis submitted in fulfillment of the requirements for the degree of
Master of Science in Electronic Engineering

School of Microelectronic Engineering
UNIVERSITI MALAYSIA PERLIS

2018

UNIVERSITI MALAYSIA PERLIS

DECLARATION OF THESIS

Author's Full Name	:	OMAR SALEM NASSER BAANS
Title	:	INTENSITY EXTRACTION AND NORMALIZATION ALGORITHM DEVELOPMENT FOR DNA MICROARRAY IMAGE PROCESSING
Date of Birth	:	26 MAY 1988
Academic Session	:	2016/2017

I hereby declare that this thesis becomes the property of Universiti Malaysia Perlis (UniMAP) and to be placed at the library of UniMAP. This thesis is classified as:

☐	**CONFIDENTIAL**	(Contains confidential information under the Official Secret Act 1997)*
☐	**RESTRICTED**	(Contains restricted information as specified by the organization where research was done)*
☑	**OPEN ACCESS**	I agree that my thesis to be published as online open access (Full Text)

I, the author, give permission to reproduce this thesis in whole or in part for the purpose of research or academic exchange only (except during the period of 2 years, if so requested above)

Certified by:

SIGNATURE	**SIGNATURE OF SUPERVISOR**
07250878	Assoc. Prof. Dr. Asral Bahari Jambek
(PASSPORT NO.)	**NAME OF SUPERVISOR**
Date:	Date:

ACKNOWLEDGMENT

In the name of Almighty ALLAH, the Most Gracious, the Most Merciful, who enabled me to understand, execute and finish this research project, without his help, I would not have been able to come this far. I am forever grateful, alhamdulillah.

I would like to express my deepest appreciation to my main supervisor Associate Professor Doctor Asral Bahari Jambek for his generous support, outstanding assistance, guidance, and the countless of the facilities that have been provided during the implementation of this project in the school of Microelectronic Engineering at University Malaysia Perlis. His supervision and exceptionally caring nature on both the personal and academic level has been essential to the progress of the project.

Special thanks also to all postgraduates and staff in the School of Microelectronic for their time and making life in the school interesting and entertaining. It has been a pleasant learning experience working with all of you. I could never express my thankfulness and gratefulness feeling in the most significant way other than mentioning the people who are involved directly or indirectly in completing my project and all my tasks. It is such a great honor for me to be able to work with them.

I also extend my special thanks and appreciation to those who have been tirelessly giving me moral support and have assisted me in various capacities, my lovely family, and friends who are really comforting and showing me their support through the whole time of this project.

TABLE OF CONTENTS

LIST OF TABLES

NO. **PAGE**

LIST OF FIGURES

NO. **PAGE**

LIST OF ABBREVIATIONS

cDNA	Complementary Deoxyribonucleic Acid
DNA	Deoxyribonucleic acid
G	Green intensity
Gb	Green Background intensity
Gf	Green Foreground intensity
Gn	Normalized green intensities
Intens_Extract	Intensity Extraction
Lowess	Locally Weighted Scatterplot Smoothing
MAE	Maximum Absolute Error
mRNA	Messenger Ribonucleic acid
ms	milliseconds
MSE	Mean Square Error
norm	Normalization
PCR	Polymeric Chain Reaction
PSNR	Peak Signal To Noise Ratio
PT	Print Tip Normalization
QRN	Quantile normalization
R	Red intensity
Rb	Red Background intensity
Rf	Red Foreground intensity
RGB	Red, Green, Blue
RMSE	Root Mean Square Error
Rn	Normalized red intensities
RNA	Ribonucleic acid
SRG	Seeded Region Growing

Pembanguanan Algoritma Untuk Pemprosesan Imej Mikrotatasusunan DNA

ABSTRAK

Mikrotatasusunan mempunyai beberapa ribu bintik yang mewakili pelbagai jenis gen manusia pada slaid. Setiap bintik terdiri daripada dua sampel (sampel biasa sebagai rujukan dan sesuatu kanser sebagai sasaran). Sampel dilabelkan dengan warna hijau (rujukan) dan warna merah (sasaran). Jika bintik berwarna hijau dilihat pada gen, ia menunjukkan ungkapan tinggi oleh sampel biasa manakala warna merah untuk sampel sasaran. Untuk menunjukkan peratusan keamatan merah dan hijau bagi setiap bintik, mikrotatasusunan menjalani pemprosesan imej di mana terdapat sejumlah besar data yang meningkatkan kebarangkalian kesilapan dan mengambil banyak masa. Mengaplikasikan pemprosesan imej membersihkan sisa-sisa yang tak dikehendaki pada imej mikrotatasusunan dan menyelesaikan masalah carian bintik dengan kejituan yang tinggi dan masa penggunaan yang singkat. Pemprosesan imej melibatkan penggridan, segmentasi, pengekstrakan keamatan dan penormalan. Penggridan mengenal pasti bintik pada imej mikrotatasusunan. Kemudian segmentasi boleh melakukan pemisahan antara piksel latar depan dan latar belakang. Ketiga, purata keamatan latar depan dan latar belakang bagi setiap bintik dikira. Keempat, keseimbangan baki warna yang tak dikehendaki untuk mengurangkan hingar. Tujuan kerja ini adalah untuk meningkatkan pengekstrakan keamatan dan langkah penormalan untuk algoritma pemprosesan imej mikrotatasusunan DNA menggunakan MATLAB. Tiga kaedah untuk memperuntukkan dan mengira nilai keamatan latar belakang dibincangkan dan dibandingkan. Kaedah-kaedah tersebut adalah GenePix, ScanAlyze, dan QuantArry. Selain itu, lima alternatif untuk pengekstrakan keamatan digunakan pada imej slaid mikrotatasusunan bagi mencari nilai keamatan yang paling jitu bagi setiap titik dalam mikrotatasusunan dua-warna. Alternatif berkenaan adalah Standard, Kooperberg, Edward, Morph dan No-background. Berdasarkan keputusannya, kaedah Edward menunjukkan hasil yang paling jitu untuk mengekstrak keamatan latar depan dan latar belakang dan mengira keamatan muktamad untuk setiap titik sebanyak 39.7 dB dari segi PSNR. Kaedah yang lebih baik telah dicadangkan untuk pengekstrakan keamatan dengan meningkatkan lokasi latar belakang, di mana kaedah ini menunjukkan keputusan yang sangat jitu sebanyak 41.36 dB dari segi PSNR dan 2.2 dari segi RMSE. Selain itu, MAE adalah sekitar 9 dengan menggunakan kaedah yang dicadangkan manakala sangat tinggi bagi algoritma pengekstrakan keamatan yang sedia ada. Selain itu, lima algoritma penormalan, Global, Lowess, House-keeping, Quantile, dan Print-tip telah diuji dan dibandingkan untuk mencari pendekatan yang paling sesuai untuk proses penormalan. Penormalan Print-tip telah dipilih untuk penormalan kerana kejituan yang tinggi iaitu sekitar 32.89 dB dari segi PSNR dan bentuk graf MA akhirnya sangat ternormal. Sehubungan dengan perkara ini, kaedah yang dicadangkan untuk penormalan digunakan. Ia meningkatkan kejituan sebanyak 33.15 dB dari segi PSNR, 32.63 dari segi MSE dan ralat kejadian menjadi sangat kecil dengan sekitar 12 dari segi MAE. Akhirnya, keperibadian algoritma telah dilakukan, ia membuktikan bahawa algoritma yang dicadangkan menggunakan masa yang kurang berbanding projek Bemis sekitar 347.7 milisaat.

Algorithm Development for DNA Microarray Image Processing

ABSTRACT

Microarray has several thousands of spots that represent various parts of human genes on a slide. Each of the spot consists of two samples (normal as a reference and cancer as a target). The samples are labeled into green (reference) and red (target) dyes. If the spot is indicating green dye, it shows a high expression of the normal sample whereas red dye shows a high expression of the target spotted on that gene. In order to indicate the percentage of red and green intensity for every spot, microarray undergoes image processing where there are huge amount of data that increase the probability of error and consume much time. Applying the image processing clears unwanted residues on the microarray image and solves the spot finding problem with high accuracy and short time consumption. The image processing involves gridding, segmentation, intensity extraction and normalization. Gridding addresses the spots on the microarray image. Then segmentation can perform separation between the foreground and background pixels. Thirdly, the averages of the foreground and background intensity for each spot are computed. Fourthly, unwanted balance of the colors is balanced to cut back the noises. The aim of this work is to improve the intensity extraction and normalization step for DNA microarray image processing algorithm using MATLAB. Three methods for allocating and calculating the background intensity values were discussed and compared. These methods were GenePix, ScanAlyze, and QuantArry. Besides that, five alternatives for intensity extraction were applied to a microarray slide image in order to find the most accurate intensity value for each spot in the two-color microarray. These alternatives were Standard, Kooperberg, Edward, Morph and No-background. Based on the results, Edward method shows the most accurate results to extract foreground and background intensity and to calculate the ultimate intensity for each spot by 39.7 dB in term of PSNR. An improved method was proposed for intensity extraction by increasing background locations, where this method showed very accurate results by 41.36 dB in term of PSNR and 2.2 in term of RMSE. Besides that, using the proposed method the MAE is around 9 while it is very high for the other intensity extraction existing algorithms. On the other hand, five normalization algorithms, Global, Lowess, House-keeping, Quantile, and Print-tip, have been tested and compared to find the most suitable approach for normalization process. Print Tip normalization was chosen for normalization because of its high accuracy which was around 32.89 dB in term of PSNR and its final MA graph shape was well normalized. In relation to this matter, a proposed method for normalization was applied. It increases the accuracy by 33.15 dB in term of PSNR, 32.63 in term of MSE and the occurrence of errors become very small by around 12 in term of MAE. Finally, algorithm profiling has been done, it proved that the proposed algorithm consumes less time than the Bemis project by around 347.7 milliseconds.

CHAPTER 1: INTRODUCTION

1.1 Overview

Gene expression regulates the production of proteins which control all cellular processes in the human biological system. The understanding of gene expression and the mechanism of protein production has many applications in terms of diagnosis, staging and finding suitable treatments for diseases. Using the cDNA (Complementary Deoxyribonucleic Acid) microarray, it is possible to diagnose easily and efficiently the level of gene expression in the sample (Kooperberg et al., 2002).

DNA microarray has thousands of spots that represent different parts of human genes on the slide. Each of the spots undergoes the hybridization with the two samples (normal as reference and cancer as a target). The samples have been labeled into green (reference) and red (target) dyes. After that, the array is 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 a high expression of the target is spotted on that gene. Further, the black dye shows there is no expression of both samples and a yellow dye having an equal expression of both samples on that spot.

When the hybridization is completed, the gene expression is scanned and microarray image is printed out. During the scanning process, the information of the gene expression might not be properly labeled. These problems could affect the analysis of the microarray image.

In order to recover the microarray image, an improvement and enhancement of the image must be done before interpreting the image information. The algorithms of DNA microarray image processing might help to reduce the noises on the image and at once the image quality can be improved. Apply the image processing can clear the unwanted residues on the microarray image and solve the spot finding problem (J, S, & Pradeep, 2002).

Image processing involves gridding, segmentation, intensity extraction and normalization. Gridding is to separate and address spots on the microarray image. Then, segmentation will do the separation between the foreground and background pixels. Third, densities of the foreground and background intensity for every spot are computed. Finally, normalization is performed to remove unbalancing intensities between the red and green color of the image (Buhler, Ideker, & Haynor, 2001).

In this chapter, problem statement of this study is presented. Next, objectives of the research are explained followed by the scope of the thesis and its expected outcomes. Lastly, the structure of the thesis is reviewed.

1.2 Problem Statement

DNA microarray is a technique that observes thousands of genes simultaneously in order to identify the gene expression patterns. This technique improves the efficiency of the gene expression analysis (Nepomuceno, Troncoso, & Aguilar-Ruiz, 2011) and (Devi Arockia Vanitha, Devaraj, & Venkatesulu, 2014). Moreover, typical DNA microarray image consists of nearly 6700 array spots (Jain et al., 2002). Each of the

gene expression consists of a nucleotide (probe) that represents specific gene which indicates the condition of each part of the human body. Two samples (normal cDNA and cancer cDNA) are being labeled into two different fluorescent dyes. Then, the samples are be hybridized on the DNA microarray and the result of the hybridization are printed out as DNA microarray image (Belean, Terebes, & Bot, 2014 and Nagaraja, Pradeep, Manjunath, & Karthik, n.d, 2009).

During the microarray image acquisition process, the image might contain a lot of noises. Some of the spots cannot be detected, this is because of poor visual quality of the microarray image. This happens if there are dust on the slide or other noises that interrupts during print imaging process (Jain et al., 2002). Moreover, these noises may affect the analysis result of microarray image. Thus, digital image processing can analyze the information and determine each of the spot locations on the DNA microarray image. Therefore, microarray image quality can be improved and the image analysis can be interpreted using digital image processing (Buhler et al., 2001), (J et al., 2002) and (Belean et al., 2014). Furthermore, computation of microarray image required a larger number of data storing and consume a great amount of energy and time to the huge number of spots on DNA microarray image. In order to analyze and identify each of the spots, it might require expensive operation due to the complex processes (Sterpone, 2009).

This work improves intensity extraction and normalization processes and implements the improved algorithms with the other DNA microarray image processing steps onto one code that will provide a huge improvement in the DNA microarray analysis.

1.3 Research Objectives

The objectives of this study are as follows:

i. To study DNA microarray image processing.

ii. To propose an improved method for DNA microarray image processing especially for intensity extraction and normalization steps.

iii. To assess the performance of the proposed algorithms in term of accuracy and time with the existing methods.

1.4 Research Scope

This research is embarked based on the following scopes; firstly to study and understand DNA microarray image processing that consists of three main steps namely gridding, segmentation, intensity extraction and normalization. This project mainly concentrates on the last two steps which are the intensity extraction and normalization. Secondly, implement the intensity extraction algorithms to discuss and verify the results and to choose the most accurate method. The algorithm will be based on Standard, Edward, Kooperberg and No Background. This step will be limited to an image with only 100 spots because it is easier to discuss and compare the results. Thirdly, based on global, Lowess, Quantile and Print Tip normalization, implement normalization algorithms and discuss the results and to choose the most accurate method. Fourth, propose improved methods of intensity extraction and normalization. These methods are simulated and their results would be validated to an authorized database. In this project, four parameters have been used to validate the results, these parameters are MSE,

PSNR, RMSE and MAE. Finally, the algorithm profiling has been done to prove that the improved algorithms help to reduce the time consumed to each step compared to the other existing algorithms.

1.5 Thesis Organization

This thesis consists of five chapters. Chapter One gives a general idea of the research, where it gives an overview of the gene expression and DNA microarrays. In this chapter also, the problem statement, project objectives, and scopes were addressed and discussed.

Chapter Two presents the literature review, which explains the basic fundamental theories of the DNA microarrays, and the theoretical procedure of the DNA microarray slide preparation and DNA microarray image processing steps. Furthermore, the available algorithms of the intensity extraction step and also normalization various algorithms.

Chapter Three demonstrates the methodology and the procedure of the project. It illustrated the image, software and tools for this research. Besides that, the preintensity extraction steps namely gridding and segmentation are also discussed. Furthermore, the existing intensity extraction algorithms for background correction, foreground calculation, and normalization were elaborated. Finally, it explains the improvement that has been done to the existing methods and the profiling comparison classifications.

Chapter Four contains the results obtained, data analysis and explanation of the project. The result obtained from the experiments were tabulated and plotted into graphs for analysis. Then, all the discussion and explanation of the obtained results were presented in this chapter.

Last but not least, Chapter Five summarizes the output of the thesis based on the objectives of this research. The recommendations for future works that might improve the finding of this project are also discussed.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter begins with a short introduction about the gene expression and the DNA Microarray. Then, the explanation expands to DNA microarray steps, hybridization, and scanning microarray image. After that, this chapter illustrates the main steps of microarray image processing namely gridding, segmentation and intensity extraction. Furthermore, it explores five existing methods that performs intensity extraction. Finally, it reviews the existing algorithms of normalization.

2.2 Gene Expression

Gene expression regulates the production of proteins which control all cellular processes in the human biological system. The understanding of DNA and the mechanism of the building of the proteins has many applications in terms of the identification of the nature of an illness or other problem, staging and finding suitable drug for diseases. Using the cDNA microarray, it is possible to diagnose rapidly and efficiently the level of genes in the sample (Venkat, Rao, G, & Raj, 2012).

The informational pathway in gene expression is as follows starting by DNA and ending by protein through by mRNA (DNA → mRNA → protein). The protein coding information is transmitted by an intermediate molecule called messenger ribonucleic acid mRNA. This molecule passes from nucleus to cytoplasm carrying the information

to build up proteins (Lockhart & Winzeler, 2000). Figure 2.1 illustrates the transformation from DNA to mRNA then to protein.

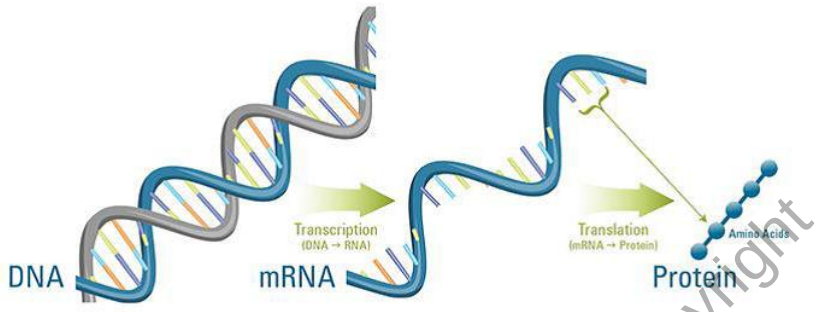

Figure 2.1: Gene Expression From DNA to RNA to Protein (Xiang & Chen, 2000)

2.3 Preparing the Slide

The mRNA acid is a single-stranded molecule from the original DNA and is subject to degradation, so it is transformed into stable complementary DNA for further examination. Microarray technology is based on creating DNA microarrays which represent gene-specific probes arrayed on a table such as a glass slide or microchip (Campbell, Hatfield, & Heyer, 2007).

Usually, samples from two sources are labeled with two different fluorescent markers and hybridized on the same array (glass slide). The hybridization process represents the tendency of two single-stranded DNA molecules to bind together. After hybridization, the array is scanned using two light sources with different lengths (red and green) to determine the amount of labeled sample bound to each spot through hybridization process. The light sources induce fluorescence in the spots which are

captured by a scanner and a composite image is produced, which is further on processed to determine spot characteristics in order to estimate gene expression levels. The most common use for DNA microarrays is to measure, simultaneously, the level of gene expression for every gene in a genome. In this way, the microarray compares genes from normal cells with abnormal or treated cells, determining and understanding the genes involved in different diseases. The microarray technology is used also in toxicological research and monitoring environmental effects on different genomes (P. Bajcsy, 2005).

2.4 Microarray Image Processing

Classical genomic microarray experiment involves complex steps including slide production and scanning. Brief steps of a microarray experiment can be summarized as follows (Borda, Belean, Terebes, & Malutan, 2011):

i. Generation of array ready cDNA (selecting cell material) and PCR (Polymerase Chain Reaction) for DNA amplification, to read more about PCR (Giordano, Ferrance, Swedberg, Hu, & Landers, 2001).
ii. cDNA selection and microarray slide printing.
iii. Selection of specific cell material to be tested from target tissues and fluorescent labeling.
iv. Hybridization of the target material on the microarray slide
v. Microarray image scanning.
vi. Image filtering and spot detection.
vii. Intensity extraction in order to evaluate gene expression.

viii. High order processing (Clustering and interpretation, gene regulatory network estimation).

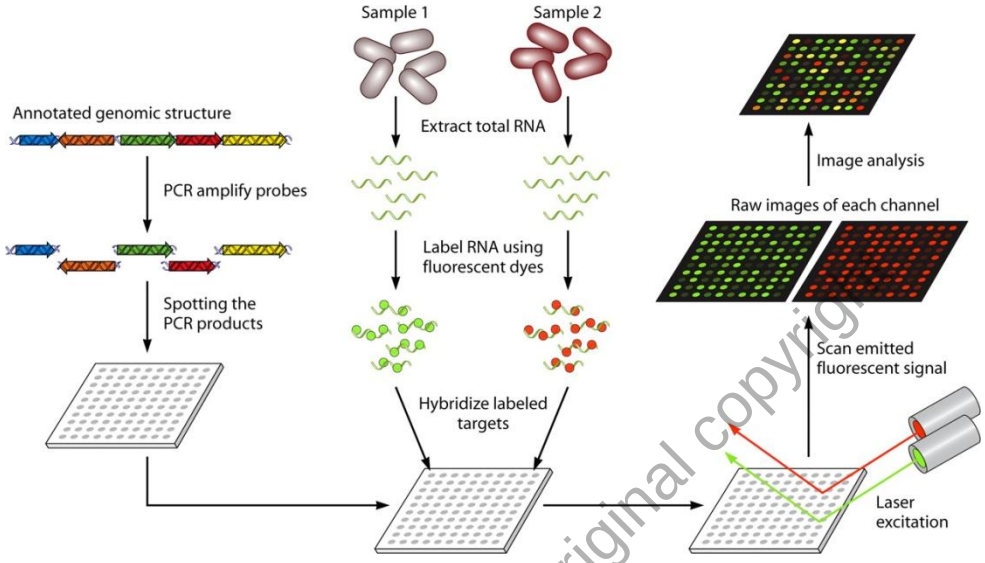

Figure 2.2: cDNA Microarray Slide Ppreparation (Borda et al., 2011)

Steps i, ii, iii, iv and v, as shown in Figure 2.2, are carried out by companies producing microarray slides, where special laboratory conditions need to be met in order to accomplish the process. Regarding steps vi and vii, it represents a chain of image processing techniques that this thesis focuses on as in Figure 2.3. The classical flow of processing a microarray image is generally separated in the following steps (Xiang & Chen, 2000): gridding, segmentation, intensity extraction and normalization. The first step is gridding and it means to assign coordinates to every element of the spot array. The second step namely, segmentation, is to classify a group of pixels as spot pixels by which each individual cell in the grid must be selected to determine the spot signal and to estimate the background hybridization. The third step is quantification and it is the step that deals with measuring the intensity of the spot signal and the background (Mabrouk, Fouad, & Sharawy, 2009). Finally, for more accuracy and to

remove the rest of the noises, normalization should be applied for the extracted intensity. The next sections will explore each step separately in detail.

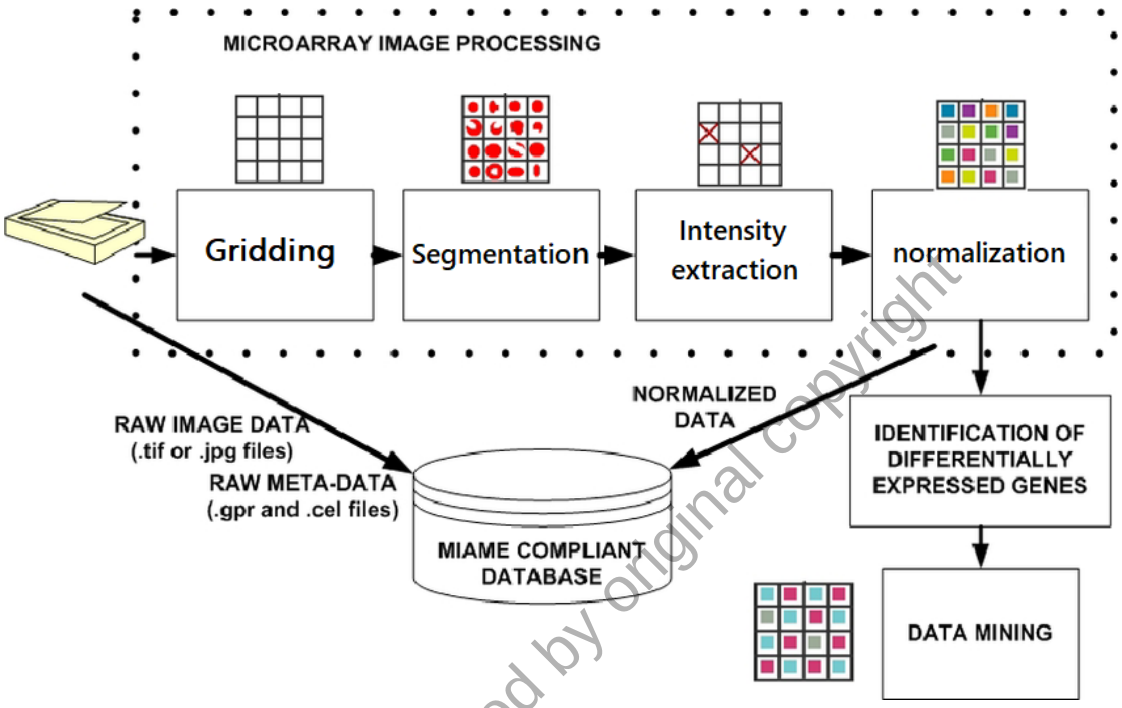

Figure 2.3: Microarray Data Processing Workflow (Peter Bajcsy, Liu, & Band, 2014)

2.5 Gridding Review

Gridding is the process of assigning coordinates to each of the spots as in Figure 2.4. Automating this part of the procedure permits high-throughput analysis. The basic structure of microarray image is determined by the manufacture of the chip and is therefore known. For example in (Bemis, 2010), image as in Figure 5.1 of Appendix E, there are eight rows and four columns of grids and within each grid, there are around 10 rows and 10 columns of spots depending on the slide manufacturer. However, to address the spots in any microarray image, a number of parameters need to be estimated (Yang, Buckley, Dudoit, & Speed, 2002b). These parameters include:

i. The separation between rows and columns of grids.

ii. Individual height and width of grids.

iii. The separation between rows and columns of spots within each grid.

iv. Small individual height and width of spots, and

v. The overall position of the array in the image.

Within a batch of microarray images produced together, the last of these parameters are usually shows huge differences between microarrays manufacture companies. Other parameters that may in some cases need to be estimated include misregistration of the red and green channels, rotation of the array in the image and skewness in the array. The last two parameters are important issues for automated gridding algorithms, but a lesser problem if manual grid placement is used. In addition, with the improvement of printing and scanning technologies, some of these parameters such as misregistration between the two channels and small individual translations of spots are likely to decrease in importance (Matthew E. Ritchie et al., 2007).

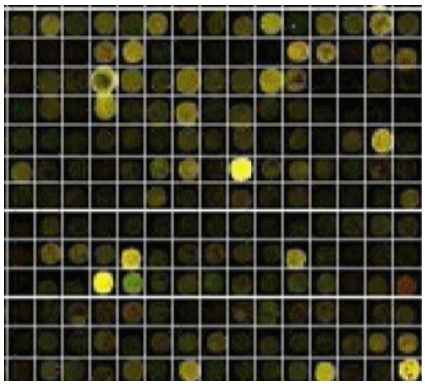

Figure 2.4: An example for Gridding DNA Microarray Image (Mukhtar, Jambek, & Mashor, 2017)

In order to achieve higher levels of accuracy in the measurement process, it is desirable for the gridding stage to be enhanced by allowing user intervention. However,