

**DEVELOPMENT OF A FULLY AUTOMATED IMAGE  
ANALYSIS METHOD FOR HIGH DENSITY cDNA AND  
array CGH MICROARRAY BASED GENOMIC STUDIES**

*Submitted to the  
Cochin University of Science and Technology  
in partial fulfillment of the requirements for the award of the degree of  
**Doctor of Philosophy**  
In the Faculty of Technology*

*By*  
**Deepa J**

*Under the guidance of*  
**Dr. Tessamma Thomas**



**DEPARTMENT OF ELECTRONICS  
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY  
KOCHI- KERALA, INDIA 682022**

*March 2013*

*Development of a Fully Automated Image Analysis Method for High Density cDNA  
and array CGH Microarray Based Genomic Studies*

***Ph.D Thesis in the field of Image Processing***

***Author***

---

*Deepa J.  
Research scholar  
Department of Electronics  
Cochin University of Science and Technology  
Kochi-682022  
Kerala, India  
email:deepaj@ceconline.edu*

***Research Advisor***

---

*Dr. Tessamma Thomas  
Professor  
Department of Electronics  
Cochin University of Science and Technology  
Kochi-682022  
Kerala, India  
email:tess@cusat.ac.in*

*March 2013*

*Dedicated to .....*

*All, who....  
Guided,  
Helped,  
Supported  
&  
Blessed me to complete this work*



**DEPARTMENT OF ELECTRONICS  
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY  
KOCHI-22**

**CERTIFICATE**

*This is to certify that this thesis entitled **Development of a Fully Automated Image Analysis Method for High Density cDNA and array CGH Microarray Based Genomic Studies** is a bonafide record of the research work carried out by **Mrs. Deepa J** under my supervision in the Department of Electronics, Cochin University of Science and Technology. The results presented in this thesis or part of it has not been presented for the award of any other degree.*

**Dr. Tessamma Thomas**  
(Supervising Guide)  
Professor  
Department of Electronics  
Cochin University of Science and Technology

Kochi-22  
5-3-2013



## **DECLARATION**

*I hereby declare that this thesis entitled **Development of a Fully Automated Image Analysis Method for High Density cDNA and array CGH Microarray Based Genomic Studies** is based on the original research work carried out by me under the supervision of **Dr. Tessamma Thomas** in the Department of Electronics, Cochin university of Science and Technology. The results presented in this thesis or parts of it have not been presented for the award of any other degree.*

Kochi-22  
5-3-2013

**Deepa J**





## **Acknowledgement**

I thank God Almighty for his immense blessings all throughout my journey.

I express my deepest sense of gratitude to Prof. (Dr.) Tessamma Thomas my research guide, for her faithful guidance, valuable advice, immense patience, continuous supervision, encouragement and support throughout.

I express my heartfelt gratitude to Professor (Dr.) C. K. Aanandan, Head of the Department of Electronics, and Cochin University of Science and Technology for extending facilities in the department for my research work.

I am indebted to Prof. (Dr.). K. Vasudevan, Dean of Technology for his whole-hearted support during the tough period of my work. I deeply thank Professors, Dr. P. R. S Pillai, Dr.P. Mohanan, Dr.James Kurian, Dr. Supriya M for their continuous encouragement. I remember Prof. (Dr.) K.G Balakrishnan for his affection and support.

I also acknowledge the help rendered by my co researchers Deepa Sankar, Reji,A.P, Praveen.N, Anatharesmi, Sethunadh, Anusabareesh Tina P.G and Nobert Thomas of Department of Electronics, CUSAT. I express my deep sense of gratitude to all the technical and office staff of Dept. of Electronics, CUSAT, for the help they have rendered to me.

I also acknowledge with thanks for the support of my teachers Prof. (Dr.) V.P Devassia, Prof (Dr.) Mini M.G and Prof (Dr.) R. Gopikakumari. I also express my gratitude for the support from my Head of the Institution Prof. (Dr.) Jyothiraj V.P. My sincere thanks to all my colleagues in College of Engineering Chengannur for their support. I am indebted to Prof. Balachandran, Head of the Department of History, NSS College, and Changanacherry for his help and encouragement.

It is beyond words to express my gratitude to my parents and my mother in law for their love and support extended to me. I remember the support and love from my sister and brother in law in my tough times. I am indebted to my husband and my kids for their sacrifice, during the period of my study.

**Deepa J**

## Table of Contents

<b>I. Introduction.....</b>	<b>1</b>
1.1 Digital Image Processing .....	2
1.2 Digital Image Processing Steps .....	3
1.3 Image Processing in Molecular Biology Research- A Review .....	5
1.4 Significance of the Study .....	7
1.5 Objectives.....	<b>8</b>
1.6 Contributions of the Thesis .....	8
1.7 Thesis Outline.....	10
<b>II. Microarray Technology .....</b>	<b>13</b>
2.1 Introduction .....	<b>14</b>
2.2 Biological Background .....	14
2.2.1 Deoxyribonucleic acid.....	15
2.2.2 Chromosomes .....	16
2.2.3 Genes.....	17
2.2.4 Proteins .....	17
2.3 Central Dogma of Molecular Biology .....	18
2.4 Need for microarrays .....	21
2.5 Fabrication of cDNA Micoarrays .....	21
2.6 Microarray Experiment .....	23
2.7 Applications of Microarrays.....	25
2.8 Challenges in Using Microarrays .....	26
2.9 Types of Microarray .....	27
2.9.1arrayCGH Microarrays.....	27

2.9.2 Protein Microarrays .....	28
2.9.3 Tissue Microarrays .....	28
<b>III. Overview of Microarray Image Analysis .....</b>	<b>29</b>
3.1 Introduction .....	30
3.2 Microarray Image Processing .....	32
3.2.1 Gridding .....	33
3.2.2. Segmentation .....	34
3.2.3 Identifying Background pixels .....	35
3.2.4 Quantification of Spot Intensity .....	36
3.2.5 Normalization .....	37
3.2.6 Spot Quality Assessment.....	38
3.3 Image Processing Challenges .....	39
3.4 Microarray Data Base .....	42
<b>1V. Development of a Fully Automatic Gridding Technique for High Density Microarray Images .....</b>	<b>43</b>
4.1 Introduction .....	44
4.2 Literature Survey .....	44
4.2.1 Manual Gridding .....	45
4.2.2 Semiautomatic Grid Alignment Technique .....	46
4.2.3 Fully Automatic Grid Alignment Techniques .....	48
4.3 Automatic Gridding of microarray images using intensity projection profile of best subimage.....	48
4.3.1. Pre-Processing .....	49
4.3.2 Global Parameter Estimation (Global gridding.....	57
4.3.3 Local Gridding .....	60
4.3.3.1 Identification of the best subimage.....	61

4.3.3.2 <i>Parameter Estimation</i> .....	63
4.4 Implementation .....	64
4.5 Experimental Results and Performance Analysis .....	65
4.6. Conclusions .....	76
<b>V. Development of Automatic Adaptive Seed Region Growing Technique for Microarray Spot Segmentation .....</b>	<b>77</b>
5.1 Introduction .....	78
5.2 Image Segmentation Techniques .....	78
5.3 Microarray Image Segmentation: A Review .....	80
5.4 Region Growing Based Segmentation .....	82
5.5 Development of Automatic Adaptive Seed Region Growing Method .....	84
5.6 Background Extraction .....	89
5.7 Intensity Calculations .....	91
5.7.1 Intensity Ratio .....	92
5.8 Performance Analysis of segmentation method using Monte- Carlo simulations .....	95
5.9 Implementation .....	96
5.10 Result and Discussions .....	97
5.11 Conclusions .....	107
<b>VI. Microarray Spot Intensity Quantification and Normalization .....</b>	<b>109</b>
6.1 Introduction .....	110
6.2 Log transform .....	110
6.3 Visualization of Microarray data using various Representations .....	114

6.3.1 Scatter Plots: Diagonal and MA Plots.....	114
6.3.2 Box plots.....	117
6.4 Normalization Techniques .....	118
6.4.1 Within Array Normalization .....	119
6.4.1.1. <i>Linear Regression of Cy5 against Cy3</i> .....	120
6.4.1.2 <i>Linear Regression of Log Ratio against Average Intensity</i> .....	118
6.4.1.3 <i>Lowess Regression of Log Ratio against Average Intensity</i> .....	122
6.5 Implementation .....	123
6.6 Experimental Results .....	123
6.7 Conclusions .....	130
<b>VII. Spot Quality Evaluation .....</b>	<b>131</b>
7.1 Introduction .....	132
7.2 Literature Review .....	132
7.3. Factors Affecting the Spot Quality.....	134
7.4 Quality measures .....	134
7.4.1 Signal Level .....	135
7.4.2 Coefficient of variation .....	137
7.4.3 Coefficient of Determination (CD).....	138
7.4.4 Spot Area .....	140
7.4.5 Coefficient of variation in the Local background .....	144
7.4.6 Deviation from Global background .....	146
7.5 Saturated Spots .....	147
7.6 Composite Quality Factor .....	148
7.7 Implementation of Spot Quality Evaluation Method .....	148
7.8 Experimental Results .....	149

7.9 Conclusions .....	162
<b>VIII. Implementation of Image Analysis Method on arrayCGH Images .....</b>	<b>163</b>
8.1 Introduction .....	164
8.2 Comparison between array CGH and cDNA arrays .....	165
8.3 array CGH Image Analysis - Case Study.....	167
8.4 Experimental Results .....	170
8.5 Conclusions .....	173
<b>IX. Conclusions and Future work .....</b>	<b>183</b>
9.1 Conclusions .....	184
9.2 Scope for Further Investigations .....	186
<b>References .....</b>	<b>189</b>
<b>List of Publications .....</b>	<b>203</b>
<b>Resume</b>	





## List of Figures

Figure 2.1	Cell structure.....	14
Figure 2.2	DNA structure .....	15
Figure 2.3	Chromosome structure .....	16
Figure 2.4	Genetic code and amino acids corresponding to codons.....	18
Figure 2.5	Central Dogma of molecular biology .....	19
Figure 2.6	Flow of information from DNA to protein .....	20
Figure 2.7	Robotic Spotting.....	22
Figure 2.8	Experimental procedures for microarray experiment .....	24
Figure 3.1	Microarray Data Processing Steps.....	30
Figure 3.2	Composite Microarray Image .....	31
Figure 3.3	Structure of an Ideal Microarray Image.....	32
Figure 3.4	Gridding .....	33
Figure 3.5	Different segmentation schemes.....	35
Figure 3.6	Local background regions used by different software .....	36
Figure.3.7	Examples of microarray defects (A) .....	39
Figure 3.8	Examples of Microarray defects (B) .....	40
Figure 3.9	Different morphological deviations in microarray spots.....	41
Figure.4.1	Preprocessing steps-1and 2 .....	52
Figure 4.2	Preprocessing steps-3 and 4 .....	54
Figure 4.3	Different morphological operations .....	56
Figure 4.4	A Microarray image .....	58
Figure4.5	Intensity Projection profiles .....	59
Figure 4.6	Microarray image after global gridding.....	60
Figure 4.7	Different Preprocrssing steps.....	60
Figure 4.8	Binary reference image and Identified optimum sub image .....	62
Figure 4.9	Intensity projection profile of optimum subimage .....	63
Figure 4.10	Gridded Image .....	64

Figure 4.11	Implementation of new gridding method on two subarrays .....	66
Figure 4.12	Different gridded subarrays with contaminations .....	67
Figure 4.13	A gridded subarray with more than 50%contamination and Microarray image .....	68
Figure 4.14	Results of applying Gridding using the intensity projection profile of whole subarray .....	69
Figure4.15	Gridding accuracy vs Coefficient of variation of subarray images. ....	71
Figure 4.16	Results of applying gridding algorithm on Gaussian Noise added Microarray image .....	73
Figure 4.17	Results of implementing gridding method on Salt and pepper noise added microarray image.....	74
Figure 4.18	A Salt-and-pepper noise added microarray image (b) Preprocessed image(c) Optimum subimage.....	74
Figure 4.19	A gridded noisy image using the projection profile of whole image .....	75
Figure 5.1	Different segmentation schemes .....	78
Figure 5.2	Basic seed region growing approach of segmentation .....	84
Figure 5.3	AASRG applied to a good spot .....	88
Figure 5.4	ASSRG applied to a black hole.....	89
Figure 5.5	Global back ground regions and their intensity Distribution (histogram).....	91
Figure 5.6	Different real microarray spots from SMD with their biological information .....	94
Figure 5.7	microarray spot after adding noise at different variance level after segmentation .....	96
Figure 5.8	Spots with irregular shape and the segmented foreground regions and ratio estimates (AASRG method) .....	97
Figure 5.9	Different spots (with artifact and the segmented foreground regions and ratio estimates (AASRG method) .....	98
Figure 5.10	Comparison of segmentation methods applied on different spots–ASRG and conventional SRG method used in MAGIC tool .....	100

Figure 5.11	Segmentation of a high background spot using AASRG.....	100
Figure 5.12	Mean Square error vs. SNR (dB).....	101
Figure 5.13	Classification errors vs. SNR .....	102
Figure 5.14	Different existing microarray segmentation methods applied to two spots.....	103
Figure 5.15	Intensity ratio evaluated using AASRG method on a microarray subarray .....	104
Figure 5.16	Gene expression ratio of first 15 genes from <i>Saccharomyces cerevisiae</i> microarray. ....	105
Figure 5.17	Comparison between true ratio evaluated using AASRG and ScanAlyze (SMD) for gene YAL054C (Spot-15) .....	106
Figure 6.1	cDNA Microarray image with 756 spots and their expression ratio .....	113
Figure 6.2	Scatter plot of Log intensities .....	115
Figure 6.3	MA plot .....	116
Figure 6.4	Box plot for intensity levels of two channels .....	117
Figure 6.5	Scatter plot of the log intensities of image .....	120
Figure 6.6	A Microarray image .....	121
Figure 6.7	MA plot -linear regression fit through the data .....	121
Figure 6.8	MA plot and the Lowess fit through the microarray data.....	122
Figure 6.9	A Microarray Image (Yeast genome) .....	123
Figure 6.10	MA plot for four subarrays before and after linear regression normalization.....	124
Figure 6.11	Box plot of log ratios of sub arrays.....	125
Figure 6.12	Lowess regression (curve in red colour) on log ratio.....	125
Figure 6.13	MA plot for Normalized data .....	126
Figure 6.14	MA plot for the whole array .....	126
Figure 6.15	Two subarrays with different spatial bias .....	127

Figure 6.16	MA plot before and after applying linear regression (1) first subarray (2) second subarray .....	128
Figure 6.17	MA plot data before and after applying lowess regression (1) first subarray (2) second subarray .....	129
Figure 7.1	Variation of $q(s)$ with signal intensity .....	136
Figure 7.2	Five different spots and their $q(s)$ values .....	136
Figure 7.3	Spots with different quality measures $q$ ( $CV$ ) indicating the homogeneity of intensities .....	138
Figure 7.4	Scatter plot of pixel intensities of two channels and the linear regression fit (green line) for a good quality spot (left) .....	139
Figure 7.5	Scatter plot of pixel intensities of two channels and the linear regression fit (green line) for spot with contamination (left) .....	140
Figure 7.6	Spots with different $q$ ( $CD$ ) values .....	140
Figure 7.7	Gallery of spots from cDNA based Microarray and their red green intensities graphed three dimensionally .....	141
Figure 7.8	Relation between ratio of area ( $A/A_0$ ) and quality .....	142
Figure 7.9	Spots with different area and the corresponding quality measures $q$ (a) and 3D intensity plot. ....	143
Figure 7.10	Background variations (a) High intensity background (b) Black holes .....	142
Figure 7.11	Two spots with the quality measures $q$ ( $CV_b$ ) .....	145
Figure 7.12	A spot with high background compared to global background .....	147
Figure 7.13	A array with 195 spots .....	150
Figure 7.14	Results of applying quality measure for signal (a) Quality value (b) Rejected spots .....	151
Figure 7.15	Results of applying quality measure for coefficient of variation (a) Quality value (b) Rejected spots .....	152
Figure 7.16	Results of applying quality measure for coefficient of determination (a) Quality value (b) Rejected spots .....	153
Figure 7.17	Results of applying quality measure for area $q$ (a) (a) Quality value (b) Rejected spots .....	154

Figure 7.18	Results of applying quality measure for area q (CVb) (a) Quality value (b) Rejected spots .....	155
Figure 7.19	Results of applying quality measure for q (gb) (a) Quality value (b) Rejected spots .....	156
Figure 7.20	Results of applying $q_{com}$ and rejected spots .....	157
Figure .7.21	Flagged spots using SMD tool .....	158
Figure .7.22	Flagged spots using new quality control algorithm .....	158
Figure.7.23	ROC plot for different thresholds .....	159
Figure 8.1	Log ratio vs. chromosome plot .....	166
Figure 8.2	Microarray image with 2463 BAC clones in triplicate .....	167
Figure 8.3	Image after global gridding.....	168
Figure 8.4	Image after local gridding .....	169
Figure 8.5	Log <sub>2</sub> ratio vs. genomic position plot for chromosome2 (left) and chromosome 8 .....	170
Figure 8.6	(a)log <sub>2</sub> ratio vs. genomic positions plot for chromomsome2(b) (b) Plot after smoothening.....	171
Figure 8.7	log <sub>2</sub> ratio vs. genomic positions plot for chromosome 8 . b) Plot after smoothening .....	172



## List of Tables

Table 4.1	Comparison of Gridding accuracy between two methods with varying spot size and intensity levels.....	70
Table 4.2	Comparison of Gridding accuracy between two methods for subarrays with same number of spots but different CV values .....	70
Table 4.3	Gridding accuracy for noisy microarray images .....	72
Table 4.4	Execution with different number of spots time for gridding sub arrays.....	76
Table 5.1	Classification error at different noise levels .....	102
Table 5.2	Computational efficiency of AASRG .....	107
Table 6.1	Conversion from fold ratios to log (Base 2) ratios .....	112
Table 6.2	Expression ratio (Log 2 ratio) of spots .....	113
Table 7.1	Spots with Quality measures .....	150
Table 7.2	Cut off values of quality measures .....	150
Table 8.1	Chromosome 2 Data.....	174
Table 8.2	Chromosome 8 Data.....	177





## **ABBREVIATIONS**

DNA	-	Deoxyribonucleic acid
RNA	-	Ribonucleic acid
PCR	-	Polymerase chain reaction
FISH	-	Fluorescence in situ hybridization
CNV	-	Copy number variation
cDNA	-	Complementary DNA
SNP_	-	Single nucleotide polymorphism
mRNA	-	Messenger RNA (ribo nucleic acid)
tRNA	-	Transfer RNA
ssDNA	-	single stranded DNA's
PCR	-	Polymerase Chain Reaction
PMT	-	Photon Multiplier Tube
TIFF	-	tagged image file format
array CGH	-	array based comparative genomic hybridization
SMD	-	Stanford Microarray Database
GEO	-	Gene Expression Omnibus
BAC	-	Bacterial Artificial Chromosome



# CHAPTER 1

## Introduction

---

*Recent advances in molecular imaging technologies open new roads for the disease diagnosis and treatment. These imaging techniques are used to represent, characterize, and quantify biological processes at the cellular and sub cellular levels within intact living organisms. Microarray technology is one such molecular imaging technique that enables the acquisition of genomic data on a scale that was previously unimaginable. Image processing techniques are used to quantify the information from microarrays. A brief introduction of the current developments in molecular biology is presented in this chapter. The fundamental digital image processing steps are explained. The significance of the present study and major contributions of the present research work are highlighted.*

---

## **1.1 Digital Image Processing**

The importance of digital images in this technological era is increasing tremendously and image processing is becoming vital component in our life. It has become an important tool in different fields like communication, remote sensing, robotics, industry and medicine. The rapid developments in image acquisition systems and computer aided analysis methods provide a great help in the diagnosis and analysis of complicated diseases and now medical imaging has emerged into one of the most important sub-fields in scientific imaging. Digital image processing enables to enhance features of interest and at the same time attenuate irrelevant details in the given context. A combination of modern microscopy and digital image processing techniques has radically changed biological research (Kherlopian, A. R. et al, 2008). Biologists study cells and generate 3D confocal microscopic data sets; radiologists identify and quantify tumors from MRI and CT scans (Doi, K., 2006). Today molecular interactions and structural dynamics are visualized as digital images to study the biological system. Automated image analysis of tissue samples now plays an important role in speeding up the drug discovery process. Analysis of these diverse types of images requires sophisticated computerized quantification and visualization tools. With the advent of internet, it is now possible to search and retrieve images from a large database of digital images and medical image databases are key components in diagnosis and preventive medicines.

Microarray Technology is one of the fastest-growing new technologies in the field of genetic research in which digital image processing techniques are applied for feature extraction and analysis. Scientists are using DNA microarrays to investigate everything from gene discovery, disease diagnosis to drug design. The image processing steps used in microarray experiments have major impact on the quality of microarray data.

## 1.2 Digital Image Processing Steps

A digital image can be represented as a two dimensional function  $f(x, y)$ , where  $x$  and  $y$  are spatial coordinates and the amplitude of  $f$  at pair of coordinates  $(x, y)$  is the intensity or gray level of the image at that point (Gonzalez R.C et al., 2002). Images can be classified according to the source of illumination as Visual, X-ray, UV, IR, Acoustics microwave images and so on. Different sensors are used to acquire these images depending on the application.

Digital image processing refers to the processing of digital images using computers which can be classified as low level, middle level and high level. The low level processing involves primitive operations like enhancement and noise reduction where both the input and output are images. Middle level processing involves extracting features, classification of objects, description of objects etc. In middle level processing inputs are images, while the outputs are attributes such as edges, contours etc. High level processing refers to image understanding from the available knowledge. This includes pattern recognition and computer vision applications. The following sections describe different image processing steps.

**Image Enhancement:** This is the preprocessing step that is applied to improve the quality of images. This can be performed both in spatial as well as frequency domain. Different algorithms are applied to emphasize, sharpen or smoothen the images before further processing. Contrast enhancement, histogram equalization, spatial and frequency filtering are some of the commonly used image enhancement techniques.

**Image Segmentation** refers to the process of partitioning an image into different regions or objects based on some criterion. The level of detail to which the partitioning is carried depends on the problem being solved. Segmentation techniques are classified into three categories such as region based, boundary based and edge based. Autonomous segmentation is one of the most difficult tasks in image processing.

**Image Restoration:** This technique is used to recover an image that has been degraded by noise, using a prior knowledge of the degradation phenomenon such as blur, random noise, nonlinearities due to sensors and geometric distortions. Most of the restoration techniques model the degradation process and attempt to apply the inverse function to obtain an approximation of the original image. Blind restoration techniques attempt to solve the problem without prior knowledge about the degradation.

**Representation and Description** Segmentation results in groups of pixels representing different regions. For further analysis, these regions should be represented and described in suitable form. Basically representation is based on either internal characteristics (region) or external characteristics (boundary) of the image. The represented region should be described by descriptors. For example a boundary can be described by features such as length, number of concavities etc.

**Morphological processing** deals with the tools used for extracting image components that are useful for representation and description of shapes, such as boundaries, skeletons, convex hull. Morphological techniques are used for pre or post processing of images. Some of the morphological operations are filtering, filling, and thinning. Erosion and dilation are two primitive morphological operations.

**Pattern Recognition** is the process that assigns labels to the objects based on descriptors. Pattern is an arrangement of descriptors. Pattern class is a family of patterns that share some common properties. Pattern recognition by machine involves techniques for assigning patterns to their respective classes.

**Compression** techniques are used to reduce the storage required to save an image, or the bandwidth required to transmit it without any appreciable loss of information. Various compression standards have been developed for this purpose.

### **1.3 Image Processing in Molecular Biology Research- A Review**

Molecular biology concerns with understanding the interactions between the various systems of a cell, including the interactions between the different types of Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA), Protein biosynthesis as well as learning how these interactions are regulated. Much of the work in molecular biology is quantitative, and recently much work has been done at the interface of molecular biology and computer science, in bioinformatics and computational biology. As of the early 2000s, the study of gene structure and function, and molecular genetics has been among the most prominent sub-fields of molecular biology. Expression cloning, Polymerase chain reaction (PCR), Gel electrophoresis, Macromolecule blotting and probing arrays are some of the major technologies used in molecular biology to characterize, isolate, and manipulate the molecular components of cells and organisms. Molecular Imaging emerged in the early twenty-first century as a discipline at the intersection of molecular biology and in vivo imaging. It enables the visualization of the cellular function and the follow-up of the molecular process in living organisms without perturbing them (Massoud et al., 2003; Betzig E. et al., 2006). Many areas of research are being conducted in the field of molecular imaging. Advancements arising from this research can enhance our knowledge of disease, lead to earlier disease detection and accelerate drug discovery. The present pace of advancements in biotechnology and functional genomics is making parallel progress in molecular imaging innovations and applications (Subramanian et al., 2001). The development, validation, and application of these novel imaging techniques in living subjects should further enhance our understanding of disease mechanisms and go hand in hand with the development of molecular medicine. The various existing imaging

technologies differ in five main aspects: spatial and temporal resolution, depth of penetration, energy expended for image generation (ionizing or non ionizing, depending on which component of the electromagnetic radiation spectrum is exploited for image generation), availability of injectable/biocompatible molecular probes, and the respective detection threshold of probes for a given technology.

Inventions of fluorescent molecule and developments in image processing techniques have great impact on the way research is being conducted in molecular biology (Zhang, J et al., 2002). Fluorescence is a quantum mechanical property of molecules and atoms whereby a photon of one energy level (typically the higher energy) is absorbed, and a photon of another energy level (typically the lower energy) is emitted. With precisely designed probes and instruments it is now possible to monitor the behavior of hundreds to thousands of single molecules within a living cell, at spatial resolutions that approach molecular-length scales (Bruchez M. P. et al., 2009).

Fluorescence in situ hybridization (FISH) is a cytogenetic technique developed by biomedical researchers in the early 1980's used to detect and localize the presence or absence of specific DNA sequences on chromosomes. FISH tests provide promising molecular imaging biomarkers to accurately and reliably detect and diagnose cancers and genetic disorders (Moter, A et al., 2000). FISH uses fluorescent probes that bind to only those parts of the chromosome with which they show a high degree of sequence complementarity. Fluorescence microscopy can be used to find out where the fluorescent probe is bound to the chromosomes. FISH is often used for finding specific features in DNA for use in genetic counseling, study of copy number variation (CNV), and species identification.

DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning



of the human body. Microarray technology evolved from Southern blotting, (E. M Southern, 1975), where fragmented DNA is attached to a substrate and then probed with a known DNA sequence. The first DNA microarray produced in Patrick Brown's laboratory in Stanford by Schena et al in 1995. They utilized gridding robots to print DNA from purified cDNA (complementary DNA) clones on glass microscope slides. The slides were interrogated with fluorescently labeled RNA samples and the specific hybridization between a cDNA clone on the slide and labeled RNA in the sample used to infer the expression levels gene corresponding to each clone. This technology enables the analysis of expression of thousands of genes in a single experiment. Now microarrays have wide range of applications. Types of microarrays include: DNA microarrays, Protein microarrays, Peptide microarrays, Tissue microarray, Cellular microarrays, Chemical compound microarrays, Antibody microarrays, Carbohydrate arrays. Data drawn from the microarray chip are mainly fluorescent images organized into matrix of spots, whose intensity is proportional to specific, site dependent, DNA hybridization. One of the major advantages of this technique is the parallelism of the process. With just one experiment it is possible to collect large number of relevant data necessary for genomic analysis. One of the major challenges in the data extraction from microarray is the image processing phase. The accuracy of this phase has substantial impact on the accuracy and effectiveness of the subsequent steps. Development of advanced intelligent image processing technique is a major requirement for speeding up the real time diagnosis and implementation procedures of microarray based analysis.

#### **1.4 Significance of the Study**

The rapid advancement of technologies for fabrication of high quality microarray makes image analysis and quantification of microarray data become a major task. Fabrication inconsistency, irregularities of spot morphology and varying surface intensity distribution are common problems with these high density

arrays. Moreover, the experimental conditions and image acquisition stages introduce noise and other sources of variability. In real microarray experiments, the exact position and size of the spots may vary due to the several reasons, mainly related to mechanical constraints and hybridization inconsistencies. In addition, spot intensity levels are highly variable and weak spots are often difficult to be detected. During the last few years, different image analysis methods have been developed; most of them require input parameters and at times manual intervention for accurately extracting information from the array. This imposes big burden for the biologists who use microarrays in their research. Image analysis has large impact on downstream analyses such as clustering or the identification of differentially expressed genes. An automatic image analysis method capable of handling high density microarrays is essential for the high throughput analysis. The method should be robust against noise and contaminations that commonly occur in different stages of microarray development.

## **1.5 Objectives**

The main objectives of the research are:

1. To develop a novel method for automatic gridding of high density microarray images.
2. To develop a novel method for image adaptive segmentation of microarray spots.
3. Perform intensity quantification and develop a novel method for spot quality assessment

## **1.6 Contributions of the Thesis**

The major contributions of the thesis are summarized as follows:

### **1.6.1 Development of a Novel Gridding Method for High Density Microarray Images**

A novel method for locating subarrays and individual spots within the microarray image has been developed using the intensity projection profile of the best subimage. The method is capable of processing the image without any user

intervention, does not demand any input parameters, and is found to be suitable for gridding microarray images with irregular spots and varying surface intensity distribution. Performance analysis indicates that the new method is robust against various noises and contaminations that are found common in microarray images. On comparison with an existing intensity projection profile method, the new scheme shows superior performance while gridding images with large coefficient variations. The new method has been implemented using MATLAB software.

### **1.6.2 Development of Automatic Adaptive Seed Region Growing (AASRG) Method for Microarray Images**

A novel segmentation method called image adaptive seed region growing has been developed. The seed and threshold value are selected automatically depending on the characteristics of the spot. Local background intensity was calculated by considering the both local and global background intensity characteristics which is found to be accurate for high density microarray images. Block processing method is used for reducing the computation time required for implementing the algorithm on high density microarray images. Monte-Carlo simulations were conducted to study the segmentation accuracy and classification error of the AASRG method. The new segmentation algorithm has been implemented on different real microarray from Stanford Microarray Database (SMD). The performance of the new algorithm is compared with MAGIC 2.2 which uses conventional SRG method for feature extraction, and it is found that better segmentation accuracy is attained with AASRG while segmenting spots with low intensity, irregular shape and size.

### **1.6.3 Development of New a Spot Quality Assessment Method**

Various quality measures are used during image processing step to evaluate the quality of individual spots so that bad spots can be excluded from further analysis. A new scheme for automatic filtering of low quality spots has been developed. Different quality measures are defined for this purpose. A

Composite quality score was assigned to each spot which has been used to check the quality of microarray spots.

The new image analysis tool has been tested on various real cDNA and array CGH microarray images available at Stanford Microarray Database (SMD). Quantification of image intensities were carried out and expression ratio was computed. Normalization techniques were used to reduce the systematic errors and bias introduced during the microarray experiment.

## **1.7 Thesis Outline**

**Chapter 1** presents an introduction to digital image processing and its applications in biological research. Significance of the present study, objectives and contributions of this research work are also summarized.

**Chapter 2** provides a biological background and discusses the need for microarray based analysis. Basic principle of microarray technology and different steps in developing a microarray are explained. Types of microarrays and their applications are also presented.

**Chapter 3** deals with various stages of the microarray data analysis. Image processing steps are also explained in detail. The challenges in applying the image processing methods on real microarray images are described. An overview of relevant image analysis methods for microarray images is described.

In **Chapter 4**, a novel method for gridding high density microarray images has been proposed. The state of the art gridding methodologies for microarray images are explained. Implementation, performance measures such as accuracy, robustness against noise and computation time is narrated. The suitability of proposed algorithm for gridding high density microarray images is described.

In **Chapter 5**, a novel segmentation algorithm using automatic adaptive seed region growing method is explained. A review of literature on various

segmentation techniques used for extracting the foreground and background intensities from the microarray is presented. The implementation and performance analysis are also illustrated. Comparison of the new method with existing method and advantages of using the new method for high density microarray images are explained.

**Chapter 6** explains the quantification of spot intensity and normalization methods used in this research. The log transformation methods and graphical tools used for the examining the results are described. The implementations of normalization on various microarray images are narrated.

In **Chapter 7** a new scheme for spot quality assessment is explained. Various factors that effect the quality of microarray spots are described. State of the art quality assessment techniques for automatic filtering out of low quality spots in the high throughput analysis is presented. Performance analyses of the new method on various microarray spots are presented.

**Chapter 8** describes implementation of the new fully automated image analysis method on arrayCGH images. HT29 cell line based array CGH image has been used for testing and biological validation. The implementation and the results are highlighted in this chapter.

A brief summary of the research work and important conclusions are highlighted in **Chapter 9**. Suggestions for future research are also provided. Remaining sessions of the thesis include the bibliography followed by list of publications.



## **CHAPTER2**

### **Microarray Technology**

---

*Life sciences are currently at the center of information revolution. Dramatic changes are being registered as a consequence of the development of techniques and tools that allow collection of biological information in extremely large quantities. Development of microarray technology empowered the biological researches to monitor the bioactivity of populations of the cells on a high throughput basis. This chapter give an introduction to the microarray technology and its applications.*

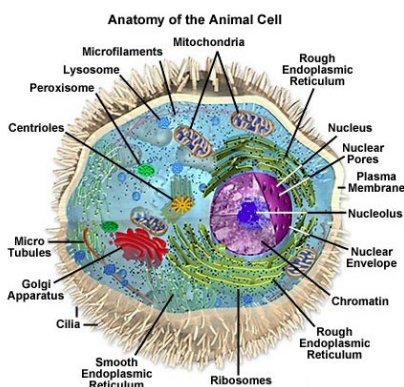
---

## 2.1 Introduction

There has been a dramatic change in the field of biological technologies over the last decade, which resulted in complete sequencing of many important model organisms including human genome. The fundamental strategy of the era of functional genomics is to expand the scale of biologic research from studying single genes or proteins to studying all genes or proteins simultaneously using a systematic approach. Microarray technology has made it possible to monitor the expression levels of thousands of genes in parallel and become a standard tool in molecular biology. A brief review of the basic biology is presented in the following session.

## 2.2 Biological Background

All living organisms consist of cells, which contain nucleic acids and proteins. Within cells there is an intricate network of organelles that all have unique functions. Figure 2.1 shows the internal structure of an animal cell. Nucleus is the largest organelle within the cell. The nucleus of the cell contains most of the cell's genetic material, organized as multiple long linear DNA molecules in complex with a large variety of proteins, such as histones, to form chromosomes.



**Figure 2.1** *Cell structure*



### 2.2.1 Deoxyribonucleic acid (DNA)

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism. DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Thus, a base, sugar, and phosphate together are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell. DNA contains the instructions needed for an organism to develop, survive and reproduce. To carry out these functions, DNA sequences must be converted into messages that can be used to produce proteins, which are the complex molecules that do most of the work in our bodies. Figure 2.2 shows the DNA double helix structure and its building blocks.

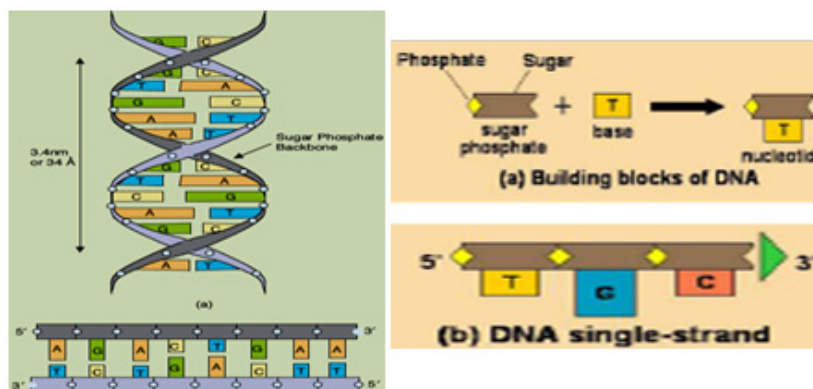
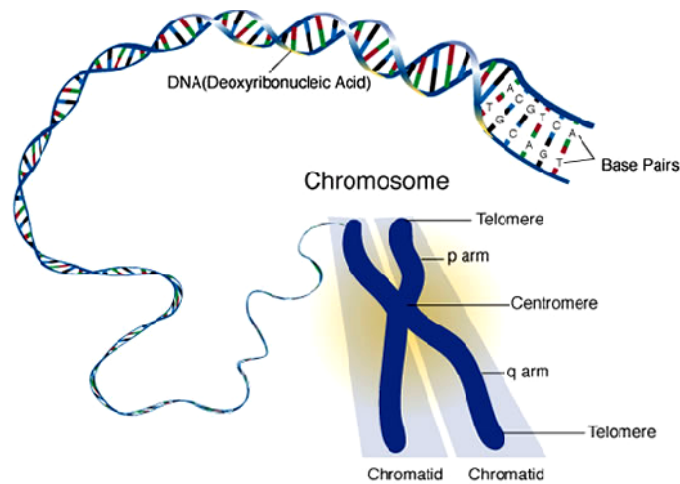


Figure 2.2 DNA structure

### 2.2.2 Chromosomes

The DNA is coiled and super coiled to form chromosomes. Each chromosome has around 50 to 250 million bases. Human cells contain two sets of chromosomes, one set inherited from the mother and one from the father. The egg from the mother contains half of the 46 (23) and the sperm from the father carries the other half of 46 chromosomes. Together the baby has all 23 pairs of chromosome in which 22 pairs are autosomes and 1 pair of sex chromosomes. Figure 2.3 shows the structure of chromosome. Each chromosome has a constriction point called the centromere, which divides the chromosome into two sections, or “arms.” The short arm of the chromosome is labeled the “p arm.” The long arm of the chromosome is labeled the “q arm.” The location of the centromere on each chromosome gives the chromosome its characteristic shape, and can be used to help describe the location of specific genes.



**Figure 2.3** *Chromosome structure*

### 2.2.3 Genes

Genes are the working subunits of DNA. Each gene contains a particular set of instructions, usually coding, for a particular protein or for a particular function. There are nearly 50,000 to 100,000 genes, each being made up of hundreds of thousands of chemical bases. The genes contain the information to make the necessary proteins. *Gene expression* is the process by which genomic information at DNA level is converted into functional proteins. Each cell expresses, or turns on, only a fraction of its genes. The rest of the genes are repressed, or turned off.

The genetic code is the set of rules by which information encoded within genetic material (DNA) is translated into proteins (amino acid sequences) by living cells. The genetic code can be expressed in a simple table with 64 entries. The Genetic code comprises of 64 triplets of nucleotides which are called as the codons. Except few exceptions, each codon encodes for one of the 20 amino acids which produces redundancy in the code that is most of the amino acids is encoded by more than one codon. This is known as the degenerative property of the codon.

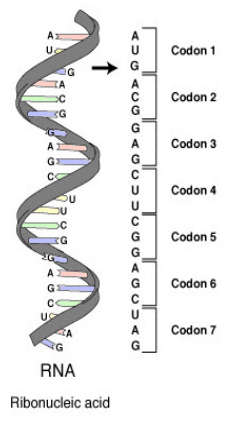
### 2.2.4 Proteins

They are involved in virtually in all cell functions. Each protein within the body has a specific function. Some proteins are involved in structural support, while others are involved in defense against germs. Proteins are important building blocks for all body parts, including muscles, bones, hair, and nails. Proteins circulate throughout the body in the blood and are normally harmless. Proteins vary in structure as well as function. They are constructed from a set of 20 amino acids and have distinct three-dimensional shapes. Occasionally, cells produce abnormal proteins that can settle in body tissue, forming deposits and causing disease.

## 2.3 Central Dogma of Molecular Biology

The central dogma of molecular biology describes the flow of genetic information within a biological system. This flow is from DNA to RNA to proteins. In order to make proteins, the gene from the DNA is copied by each of the chemical bases into the messenger RNA (ribonucleic acid) or mRNA. The composition of mRNA is similar to DNA except for a few characteristic differences.

The sugar molecule present in mRNA is ribose, and among the four nitrogenous bases, thymine (T) is replaced by uracil (U). The mRNA moves out of the nucleus and uses cell organelles in the cytoplasm called ribosomes to form the polypeptide or amino acid that finally folds and configures to form the protein. Biological decoding is accomplished by the ribosome, which links amino acids in an order specified by mRNA, using transfer RNA (tRNA) molecules to carry amino acids and to read the three mRNA nucleotides at a time. Figure 2.4 shows genetic code and formation of amino acids by triplets of nucleotides.



		2nd base			
		U	C	A	G
1st base	U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
		UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
		UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)
	C	UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan
		CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
		CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
		CUA (Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine
A	CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine	
	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine	
	AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine	
	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine	
G	AUG <sup>[A]</sup> (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine	
	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine	
	GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine	
	GUU (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine	
	GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine	

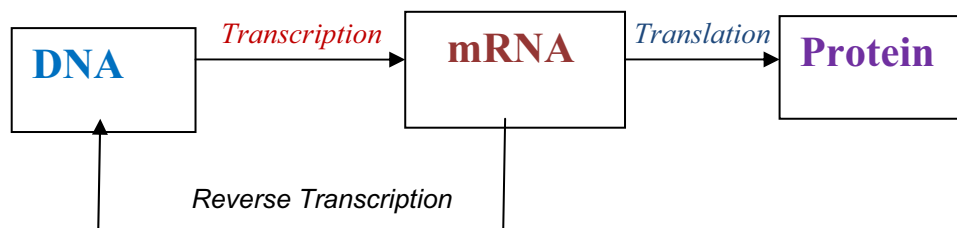
**Figure 2.4** Genetic Code and Amino acids corresponding to Codons

Three different processes responsible for these transformations are shown in Figure 2.5. They are:

**Replication:** A double stranded nucleic acid is duplicated to give identical copies. This process perpetuates the genetic information.

**Transcription:** A DNA segment that constitutes a gene is read and transcribed into a single stranded sequence of mRNA. The mRNA moves from the nucleus into the cytoplasm.

**Translation:** the mRNA sequence is translated into a sequence of amino acids as the proteins are formed. During translation, the ribosome reads three bases (a codon) at a time from the mRNA and translates them into one amino acid.



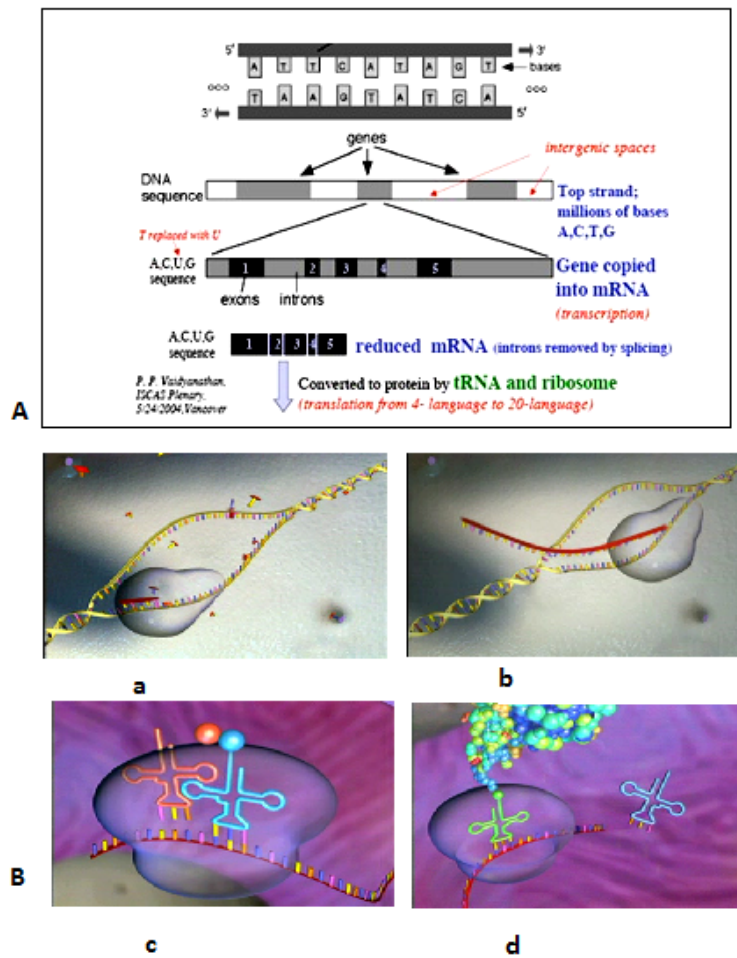
**Figure 2.5** *Central Dogma of molecular biology*

The presence of mRNA in a cell indicates that the gene is active. Every cell of an individual organism contains same DNA, carrying the same information. Each cell expresses only a fraction of its genes and produce different set of proteins that defines the function of the cell.

### **Reverse transcriptase**

Reverse transcription is the process by which a reverse transcriptase enzyme mediates the creation of a DNA complement (complementary DNA or cDNA) from an RNA strand. The discovery and use of reverse transcriptase has greatly

improved knowledge in the area of molecular biology. Reverse transcriptase is used for gene expression analysis, to create cDNA libraries from mRNA and, along with other enzymes, allow cloning, sequencing, and characterization of RNA. Figure 2.6 show the different steps in the generation of protein from gene.



**Figure 2.6. (A.)** Flow of information from DNA to protein. **(B).** Protein synthesis (a) Replication (b) Transcription (c) Translation (d) Protein formation

## 2.4 Need for Microarrays

The amount of protein generated from each gene determines both the morphology and the function of a given cell. Small changes in the expression levels can change at organism level resulting in various diseases like cancer. Therefore comparing the expression levels of the genes in different conditions such as differing environments, treatments, time points, phenotypes, or clinical outcomes are of extreme interest for scientists. This need stimulated the development of high throughput techniques such as microarrays. Microarray allows the interrogation of thousands of genes at the same time. The first microarrays were developed at Stanford University by Schena et al. in 1995. Development of better surface technologies, more powerful robots for arraying, labeling techniques and improved computational power and automated analyzers have vastly improved the power and efficiency of microarray, while also lowering the cost of these analyses. Microarray is currently used to analyze different systems, including the classification of microbes and human microbial pathogens, cellular responses to pathogens, drug and toxic exposures, tumor classification, detection of gene fusions, comparative genomic hybridization, alternative splicing detection (exon junction array/exon arrays) and gene expression profiling via analyzing global mRNA levels.

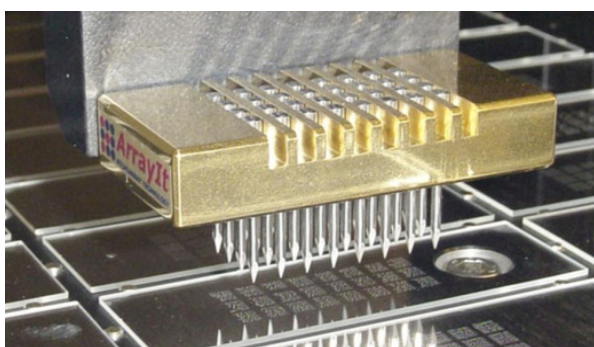
## 2.5 Fabrication of cDNA Microarrays

A DNA microarray consists of a solid substrate (glass, nylon or plastic) on which known single stranded DNA's (ssDNA) corresponding to known genes are deposited. Researchers have a database of over 40,000 gene sequences that can be used for this purpose. Single stranded DNA fragments are prepared by either Polymerase Chain Reaction (PCR) or by using already synthesized oligonucleotides. PCR technique creates billions of copies of specified DNA fragments and oligonucleotides have to be presynthesized. These ssDNA fragments (called *probes*) are spotted at fixed locations that are arranged in regular grid like

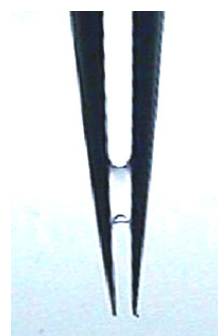
pattern. The attachment with the substrate is either covalent or electrostatic attraction.

There are two technologies for making microarrays: *robotic spotting and in-situ synthesis*. Manufacture by robotic deposition may proceed through the deposition of PCR amplified or oligo nucleotide single stranded DNA segments at specified locations using spotting robots. *In-situ synthesized arrays* are fundamentally different from robotic spotted array. Instead of presynthesising oligo nucleotides, oligos are built up base by base on the surface of the array. *In-situ* can be divided into photolithography, ink jet printing and electrochemical synthesis.

Fig.2.7 shows a robotic spotting method where a robotic arm moves the cassette containing the pins over the microtiter plates containing probes and dips pins into the wells to collect the first batch of ssDNA. The robotic arm is then moved over the array and the pins touch the surface of the array at specified locations to deposit ssDNAs. If more than one array is being synthesized the cassette is moved to subsequent arrays. Before collecting the next ssDNA batch to be spotted, the pins are washed to ensure no contamination. The final step of array production is fixing, in which surface of substrate is modified so that no additional DNA can stick to it.



**Figure 2.7(a)** *Robotic Spotting*



**(b)** *Pin structure*



## 2.6 Microarray Experiment

A cDNA microarray works by exploiting the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. A typical two channel or two colour microarray experiment involves two samples such as tumor tissue (test sample) and healthy tissue (reference sample). Once the known ssDNA segments corresponding to different genes are fabricated on the slides, there are four steps in the microarray experiment to measure the gene expression.

### Step 1: Sample Preparation and Labeling

Sample preparation refers to extracting and purifying the mRNA from the tissues of interest (eg: normal as well as tumorous tissues). Extracted mRNA is reverse transcribed into cDNA. To allow detection of which cDNA that will bind to the complementary part in glass substrate, a labeling process is carried out. The tumor and the normal cDNA samples are labeled with different fluorescent dyes. Now most laboratories use fluorescent labeling by using two dyes Cy5 (excited by red laser) and Cy3 (excited with green laser). For Cy5 the excitation wavelength is 550nm and emission wavelength is 581nm and for Cy3 the excitation and emission wavelengths are 649nm and 670nm respectively (Dov Stekel, 2003).

### Step2: Hybridization

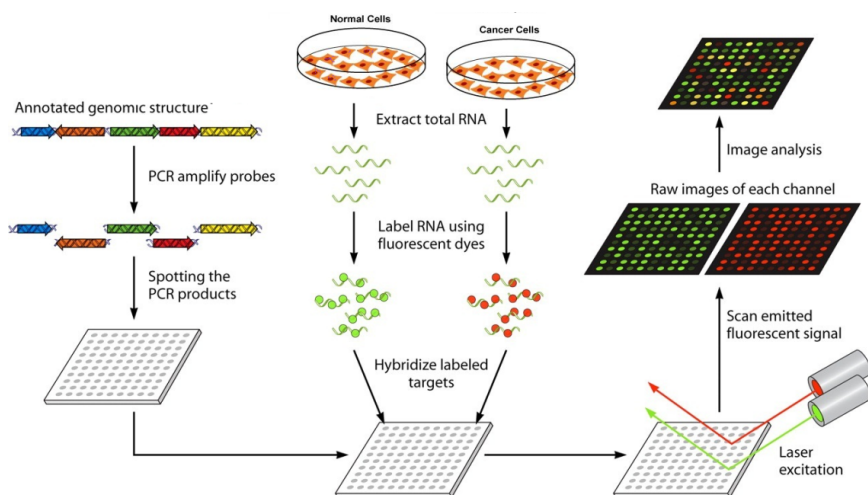
Hybridization is the step in which the ssDNA sequences on microarray slide (called *probes*) and the labeled cDNA(called *targets*) forms hetroduplexes according to Watson –Crick base pairing rule known as cross hybridization (Sorin D ,2003).The fluorescently labeled targets are pooled and allowed to hybridize under stringent conditions. The chip is then incubated for 12 to 24 hours in temperature. At this temperature, DNA strands in the slide encounter the complementary strands and match together to create a double stranded DNA.

### Step3: Washing

After hybridization the slides are washed to remove excess labeled samples (which are not hybridized) from the array and the array is dried using a centrifuge or blowing clean compressed air.

**Step 4: Scanning**

The final step of the laboratory process is to produce an image of the surface of the hybridized array. The slide is placed in a scanner. It is essentially a fluorescent microscope that is specialized for acquiring microarray fluorescent image on the standard microscopic slide format. The scanner contains lasers that are focused onto the array. Lasers excite dyes in the labeled targets and emit photons. The emitted photons are amplified by using photon multiplier tube (PMT). The fluorescence of the dye measured by a PMT is converted to digital image. With two colour arrays, the output of the scanner is two monochrome images; one for each of the two lasers in the scanner. These monochrome images are imported into a software in which these are combined to create the red-green composite image of the microarray. Both the monochrome and the composite image are usually stored in tagged image file format (TIFF) with a resolution of either 8 or 16 bits. The amount of dye on the slide at each spot position is detected and this signal indicates which genes are active and how much mRNA was produced. Fig 2.8 shows the typical two channel microarray experiment.



**Figure 2.8** Experimental procedures for microarray experiment.

## 2.7 Applications of Microarrays

Microarrays have been extensively used for biological research such as sequencing, single nucleotide polymorphism (SNP) detection, investigation of genetic mechanism in living cell such as comparing healthy and malignant tissues, studying cell phenomena over time, study the effect of various factors such as interferons, oncogene transfection. Other applications include:

**Gene Discovery:** DNA Microarray technology helps in the identification of new genes, know about their functioning and expression levels under different conditions.

**Genotyping:** It is the process of determining differences in the genetic make-up (genotype) of an individual by examining the individual's DNA sequence using biological assays and comparing it to another individual's sequence or a reference sequence.

**Disease Diagnosis:** DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancers have been classified on the basis of the organs in which the tumors develop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer.

**Drug Discovery:** Microarray technology has extensive application in Pharmacogenomics. Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help in the identification of the biochemical constitution of the proteins synthesized by the diseased genes. The researchers can use this information to synthesize drugs which combat with these proteins and reduce their effect. They can also be used to

monitor changes in gene expression in response to drug treatments. Gene expression microarray analysis can be valuable at all stages of the drug discovery process, including target identification and validation, mechanism of action studies and the identification of pharmacodynamic endpoints.

**Toxicological Research:**

Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants. They are widely used for ecotoxicology to understand the mechanism of action of toxicants on living organisms. Such knowledge would help to develop predictive simulation models of toxic effects, to link molecular biomarkers with population-level effects, and then to anticipate ecologic risk assessment issues for new chemicals. Gene expression profiles represent the primary level of integration between environmental factors and the genome, providing the basis for protein synthesis, which ultimately guides the response of organisms to external changes.

**2.8 Challenges in Using Microarrays**

The advantages and possibilities of the microarray technology are numerous, but their users are challenged by many issues. Even if the experiment is performed several times with exactly same material and preparations in exactly the same experimental conditions, after scanning and image analysis, they show variation in the quantified values. Noise is introduced at each step of the various procedures such as mRNA extraction, transcription, labeling, non-specific background hybridization, and other artifacts. Scanning issues such as dynamic range limitation and inter channel alignment, type of the image processing techniques used, characteristics of individual pin tips, properties of specific probe source plates, are some of the sources variation. Due to the mechanical constraints, it is common to have irregular-shaped spots in microarray slides such as doughnut

shaped spots and spots with irregular edges. One of the major problems involved with the technique of DNA microarrays is concerned with the amount of data that is produced. Large-scale, high-throughput experimental methods require a great deal of information processing and data analysis.

## **2.9 Types of Microarrays**

Microarrays are not limited to gene expression analysis. Different microarrays are developed for studying the interaction between various biomolecules such as protein microarrays, tissue microarrays, antibody microarray and chemical compound microarray .

### **2.9.1 arrayCGH Microarrays**

Copy-number variations (CNV) are form of structural variations, which are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. Most CNVs are benign variants that will not directly cause disease. However, there are several instances where CNVs that affect critical developmental genes do cause disease. These gene amplifications and deletions cause various genetic disorders and even cancer.

DNA microarray based comparative genomic hybridization (array CGH) is a technique that allows simultaneous monitoring of copy number of thousands of genes throughout the genome. In this technique, DNA fragments or "clones" from a test sample and a reference sample differentially labeled with dyes (typically, Cy3 and Cy5) are hybridized to mapped DNA microarrays and imaged. Copy number alterations are related to the Cy3 and Cy5 fluorescence intensity ratio of the targets hybridized to each probe on a microarray. Clones with normalized test intensities significantly greater than reference intensities indicate copy number gains in the test sample at those positions. Similarly, significantly lower intensities

in the test sample are signs of copy number loss. BAC (bacterial artificial chromosome) clone based CGH arrays have a resolution of the order of one million base pairs (1Mb).

### **2.9.2 Protein Microarrays**

They are used to track the interactions and activities of proteins, and determining function on a large scale. Its main advantage lies in the fact that large numbers of proteins can be monitored in parallel. Protein microarrays are rapid, automated, economical, and highly sensitive, consuming small quantities of samples and reagents. The high-throughput technology behind the protein microarray is relatively easy to develop since it is based on the previously-developed DNA microarray technology.

### **2.9.3 Tissue Microarrays**

Tissue microarrays enable the high throughput analysis of a large number of tissue samples that have been collected and archived through the use of paraffin blocks or formalin. Tissue microarrays are different from DNA microarrays where each spot on an array represents a cloned cDNA or oligonucleotide that binds to the target sequence. With tissue microarrays, each array has patient specific histological samples from cancer infected tissues. The tissue microarray technique is best suited for screening one genetic marker or protein across thousands of samples where as DNA microarrays are best suited to study gene expression across thousands of genes.

## CHAPTER 3

### Overview of Microarray Image Analysis

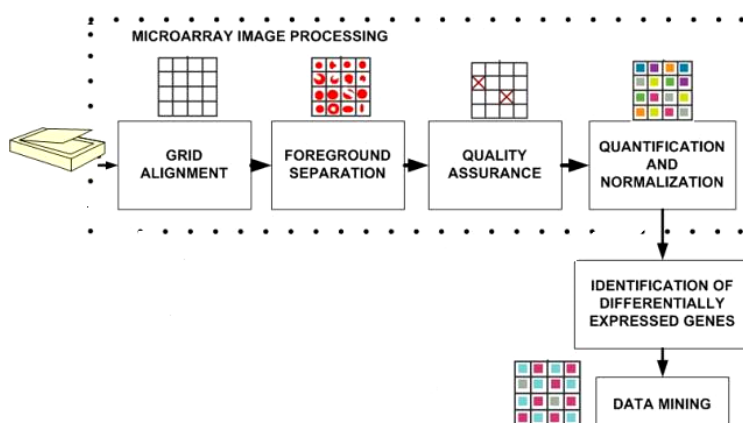
---

*Image processing is the first step in knowledge discovery from the microarray and has a potentially significant impact on downstream analyses. Microarray image processing consists of three major stages such as gridding, segmentation and quantification. This Chapter explores the different image processing steps and discusses the challenges in microarray image analysis.*

---

### 3.1 Introduction

The microarray data processing starts with image acquisition using laser scanners and ends with the results of data mining that have to be interpreted by biologists. The major steps of data handling in a microarray processing are shown in Figure 3.1. The microarray data processing workflow includes image processing (grid alignment, foreground separation, spot quality evaluation, data quantification and normalization) (2) data analysis (identification of differentially expressed genes, data mining, integration with other knowledge sources, and quality evaluation and repeatability of results, and (3) biological interpretation



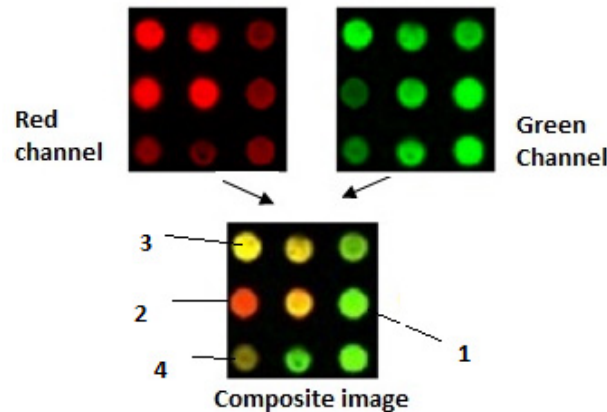
**Figure 3.1** *Microarray Data Processing Steps*

The raw microarray image from the scanner needs to be preprocessed so that fluorescent intensity associated with each arrayed spot can be determined accurately. The scanner resolution is an important factor that determines the quality of the microarray image.

The composite image provides a convenient way of identifying genes or gene transcripts present in greater abundance in the test sample compared to the reference sample. Usually reference sample is labeled with green dye (Cy3) and



test sample is labeled with red dye (Cy5). A green spot in the composite microarray image indicates that the gene present in greater concentration in the reference sample compared to test sample. A gene expressed more in test sample will produce a red spot and a gene that equally expressed in both test and reference samples will appear as yellow. A gene that is not expressed in both samples will appear as a black spot. Figure. 3.2 shows an example of spots in a microarray with different fluorescent intensities. The red and green channels are shown separately by adding colours to each channel. The composite image is obtained by overlapping these channels. Consider a gene that is expressed abundantly in tumor tissue compared to normal tissue (spot2). The spot corresponding to this gene will yield an intense spot on red channel compared to green channel. Spot 1 with green intensity indicates that the corresponding gene is more expressed in normal tissue compared to tumor tissue, while spot 3 corresponds to a gene that expressed equally in both tissues. Spot 4 represents a gene that has not expressed in either tissue.

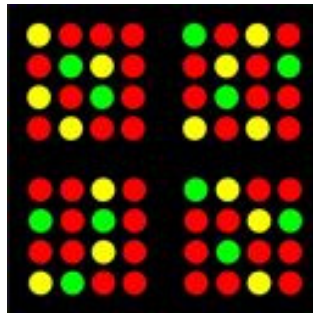


**Figure 3.2** *Composite Microarray Image*

### 3.2 Microarray Image Processing

In microarray experiments the data extraction from the scanned slides has large impact on subsequent analysis. Fluorescent intensities correspond to levels of hybridization of targets to probes spotted on the slide. Measured intensity of the spot also includes contribution of nonspecific hybridization and fluorescence emission from the chemicals on the slide. It is important in microarray image to adjust these background intensities for accurate estimation of the spot intensity. Image processing is the first processing step in the analysis of microarray data to quantify the intensity values of the spot and its local background. Fig. 3.3 is the structure of an ideal microarray image. Spots within the array are aligned horizontally and vertically in blocks called subarrays. The ideal microarray image has the following properties:

- All the subarrays are of the same size.
- The spacing between subarrays is regular.
- The size and shape of the spots is the same for all the spots.
- No dust or scorches and other contamination are on the slide.
- There is minimal and uniform background intensity across the image.
- A perfect image should only reflect measures of the fluorescence intensities for the dye of interest.



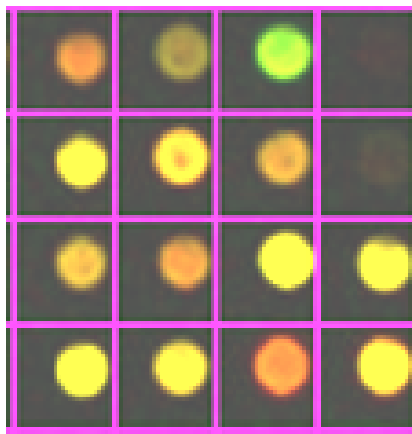
**Figure 3.3** Structure of an Ideal Microarray Image

Real microarray images deviates from perfect regularity in the positions of the subarrays and positions of individual spots within the subarray. Also, there are deviations from the ideal uniform shape, size and background intensity levels.

### 3.2.1 Gridding

The first step of the image processing is the identification of the position of subarrays and then that of each spot within a subarray. This addressing procedure is called “gridding” in microarray literature. To address each spot a number of parameters must be estimated, including separation between rows and columns of subarrays, row and column spacing between spots and average diameter of the spots. Figure 3.4 shows an exmple of gridded spots.

Several academic and commercial packages are available for gridding; most of them require prior knowledge of the image specific parameters or direct user intervention to find the position of the spots. It is important that the addressing procedure be accurate, to ensure precision of the subsequent steps of image analysis.



**Figure 3.4** *Gridding*

### 3.2.2. Segmentation

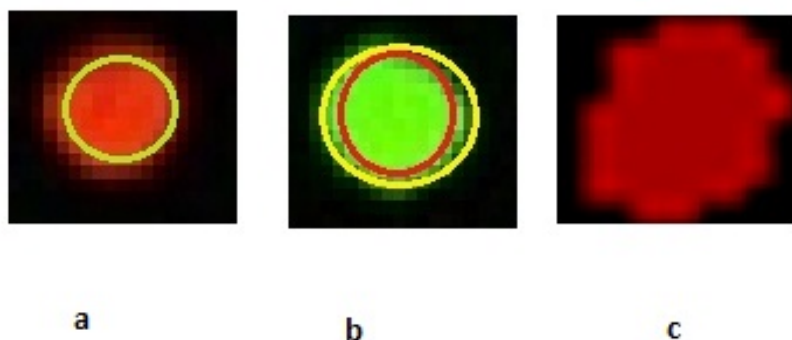
Segmentation of the image is generally defined as the process of partitioning the image into different regions, each having certain properties (Soille, 1999). Identifying the spots and separating the background from the foreground is a fundamental problem in DNA microarray data analysis. Existing segmentation techniques for microarray images can be categorized into four groups. They are:

1. Fixed circle segmentation
2. Adaptive circle segmentation
3. Adaptive shape segmentation
4. Histogram segmentation

**Fixed circle segmentation** fits a circle with constant diameter to all the spots in the image. All the pixels inside the circle are collected and used for foreground calculation. The problem with fixed circle segmentation is that it gives inaccurate results if the spots are of different size, which is a common case in microarray images. Figure 3.5 (a) shows fixed circle implementation.

**Adaptive circle segmentation** method fits a circle with variable diameter onto the region containing the spots. This method is able to resolve spots of different size, but performs less well on irregular shaped spots. One approach is to fit concentric circles over the spot. The pixels inside the inner circle are used for calculating the signal intensity. Pixels outside the outer circle are used for calculating the background intensity as shown in Figure 3.5(b).

**Adaptive Shape segmentation** method uses two common approaches, Seed region growing (SRG) (Adams et al., 1994) and Watershed (Vincent et al., 1991). Both these methods require the specification of starting points, or seeds. These methods have the advantage of being able to cope with the irregular shaped spots. Figure.3.5(c) shows adaptive shape segmentation.



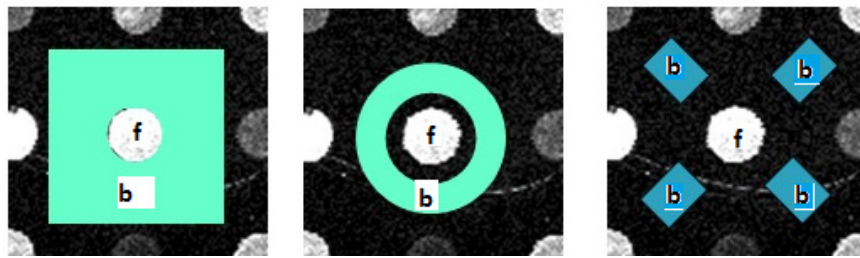
**Figure 3.5** *Different Segmentation Schemes (a) Fixed circle (b) Adaptive circle (c) Adaptive shape*

**Histogram segmentation method** does not explicitly classify pixels into foreground and background. Instead, this method analyzes the pixels within designated region and estimate the foreground intensity from the intensity distribution of pixels. This method uses a target mask that is chosen to be larger than any given spot in the microarray. For each spot foreground and background intensity estimates are determined from the histogram of pixel intensities within the masked area. Histogram based method give reliable results for irregular shaped features. But the method is found unstable if the spot size is small compared to the circular mask chosen.

### 3.2.3 Identifying Background pixels

Signal intensity of the spot includes contributions from the nonspecific hybridization and other fluorescence from the substrate. Background intensity is subtracted from the signal intensity to provide more reliable estimate of hybridization intensity of each spot. It is a common practice to identify the background pixels from the local background pixels that are not part of the foreground region within each grid after segmentation. Local background calculation is a difficult task for high density microarray images. Different image analysis software defined the local background differently. Some of the commonly used local

background regions are shown in Figure 3.6. For a spot with foreground region marked as *f*, back ground intensity is calculated using the pixels within region marked as ‘*b*’.



**Figure 3.6** Local background regions used by different software (a) Scanalyze (b) ImaGene (c) Spot and GenePix

### 3.2.4 Quantification of Spot Intensity

After the pixels belonging to foreground and background have been separated, the next step is calculation of intensity of the spot. Image analysis software usually calculates the following intensity measures for each channel.

1. *Signal Mean (Foreground Mean)*: Mean of the pixel intensities of the foreground region.
2. *Background Mean*: Mean of the pixel intensities of the back ground region.
3. *Signal Median*: Median of the pixel intensities of the foreground region.
4. *Background Median*: Median of the pixel intensities of the back ground region.
5. *Signal Standard deviation*: The standard deviation of pixel intensities of foreground region.
6. *Background Standard deviation*: The standard deviation of pixel intensities of background region
7. *Number of pixels*: Number of pixels in the foreground region.

The purpose of spot intensity quantification is to combine the pixel intensity values into a unique quantitative measure that can be used to represent the expression level of a gene deposited in a given spot (for cDNA arrays) or copy number variation in array CGH microarrays. The true signal intensity of each channel is measured as difference between foreground and background intensity.

### **Data Transformation**

It is common practice to transform microarray data before proceeding with the further analysis in order to improve comparability and signal to noise ratio. Usually the raw intensities are transformed into log intensities. There are several advantages for this transformation:

- The variability should be constant at all intensity levels.
- The distribution of experimental errors should be approximately normal.
- The distribution of intensities should be approximately normal.

In microarray data analysis logarithm to base 2 is common. The reason is that the ratio of intensities of the two channels (Cy5/Cy3) is transformed into the difference between log intensities of the Cy5 and Cy3 channels. Therefore a 2 fold up regulated genes correspond to a log ratio of +1 and 2 fold down regulated genes correspond to a log ratio of -1. Genes that are not differentially expressed have a log ratio of 0. Log ratio is defined as  $\log_2 \frac{R}{G}$ , where R (red) is the intensity value of the Cy5 channel and G (green) is the intensity value for the Cy3 channel.

### **3.2.5 Normalization**

The complexity of the microarray experimentation process often introduces systematic bias into intensity measurements. Systematic biases can be caused by concentration and amount of DNA placed on the microarrays, wear of arraying equipment such as spotting pins, the quantities of mRNA extracted from samples, reverse transcription bias, lack of spatial homogeneity of the slides,

scanner settings, saturation effects, background fluorescence, linearity of detection response and ambient conditions (D. Amaratunga et al., 2003). Normalization is a general term for collection of methods that are used for solving these systematic biases. In addition dye bias is a common problem in almost all multichannel experiments. Generally, Cy5 (red) intensities tend to be higher than Cy3 (green) intensities. The reasons of imbalance between the channels occurs due to following

- The Cy3 and Cy5 labels may be differentially incorporated into DNA of different abundance.
- The Cy3 and Cy5 dyes may have different emission response to excitation at different abundance
- The Cy3 and Cy5 emissions may be differentially measured by the photomultiplier at different intensities.
- The Cy3 and Cy5 intensities measured at various areas on the array may differ due to tilt in the array.

The normalization can be performed on *within array level* correcting for technical bias inside each array and on *between array level* correcting for distributional differences between different arrays from the experiment.

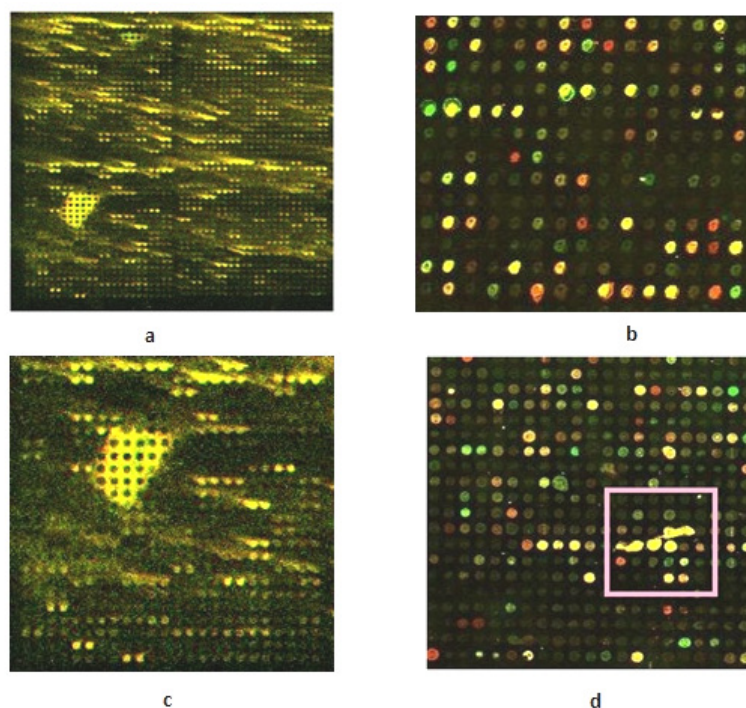
### 3.2.6 Spot Quality Assessment

Noise, irregularities of spot, shape and position are common problems, especially in large-scale high density microarrays. Quality control is a major requirement to flag out low quality spots. Without a good scheme to produce high quality data, any data mining tools can lead to misleading results. Most of the image analysis methods include procedures for flagging out spot on the basis of one or more quality measures, so that flagged spots are rejected from further analysis. Area of the spot, signal to noise ratio, spot regularity, variation in the intensity of foreground and background are some of the commonly used quality measures.



### 3.3 Image Processing- Challenges

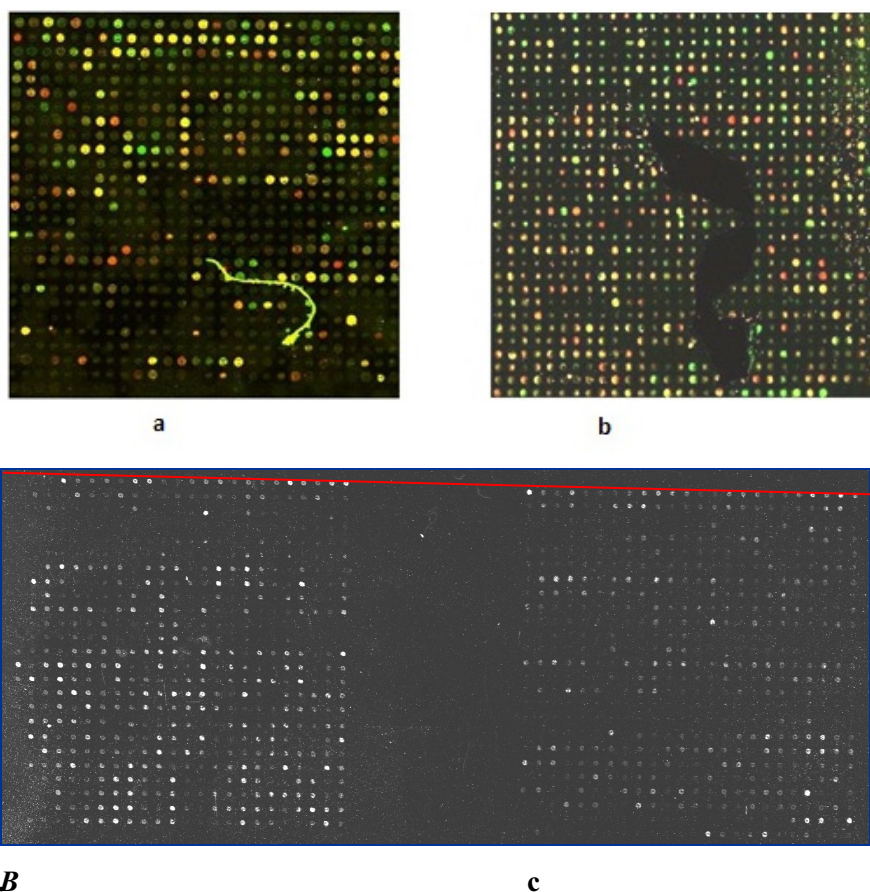
The microarray technology is a complex electrical-optical-chemical process and there are several sources of variability, that lead image processing a difficult task. Figure 3.8 shows various defective microarray images. Figure 3.7 (a) is a common problem called comet tail that arise due to either excess DNA on the slides or use of defective slides. Figure 3.7(b) shows an image with irregular spot morphology mainly due to use of damaged pins. Figure 3.7 (c) shows a high background image in combination with weak signals, due to either insufficient blocking, or precipitation of the labeled probes. Figure 3.7(d) indicates the Spot overlap, due to big amount of dehydration during post processing.



A.

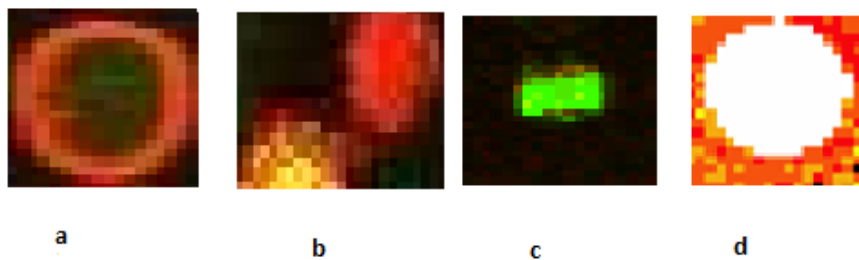
**Figure.3.7** Examples of Defective Microarrays. (A). (a) Comet tails (b) Irregular spot morphology(c) High background intensity (d) Spot overlap

Figure 3.8 shows more problems with the microarray images. Figure 3.8(a) shows particle contamination. In (b) there is large bubble in the image mainly due to poor set up technique or particulate material in hybridization solution. Figure 3.8 (c) shows misalignment of red and green channels.

**B****c**

**Figure 3.8** Examples of Defective Microarrays. (B). (a) Microarray image with particle contamination. (b) Bubbles (c) Misalignment between channels.

Spot morphology influences the measurement of gene expression level. It depends on many factors, such as amount of material carried in pins, period of time the pin in contact with the slide, composition of arrayed material. For ideal spotted microarray images circular shape is common and all the spots are of same shape and size. Variations from the ideal characteristics are a major challenge while implementing gridding and segmentation techniques. Different classes of morphological deviations are shown in Figure 3.9. A doughnut-shaped spot is shown in Figure 3.9 (a), which consists of pixels of high intensity at the perimeter and those of low intensity in the central area. Such patterns of spot images are primarily due to the non-uniform distribution of cDNA molecules while the cDNA solution dries out during the microarray printing process. 3.9(b) is a dilated spot which happens due to excessive delivery of DNA during printing or damaged pins. Fig 3.9 (c) is an irregular shaped spot and (d) is a saturated spot with spot pixel intensity value exceed the detection range of the photo multiplier tube. An image analysis method should be capable of handling all these problems.



**Figure 3.9.** *Different morphological deviations in microarray spots (a) Doughnut (b) Dilated spot (c) Irregular shape spot (d) Saturated spot*

### **3.4 Microarray Databases**

Microarray experiments produce a great deal of data, not just in terms of the data output from the scanned slide, but also in terms of recording such as how the experiment was performed, Information about the identity of the probes, methods in slide fabrication, hybridization conditions of the experiment etc. Data base help the researchers to efficiently retrieve these information for their studies. The present work mainly uses two data bases.

#### **Stanford Microarray Database (SMD)**

The Stanford Microarray Database (SMD) stores raw and normalized data from microarray experiments, and provides web interfaces for researchers to retrieve, analyze and visualize their data. The two immediate goals for SMD are to serve as a storage site for microarray data from on going research at Stanford University, and to facilitate the public dissemination of that data once published, or released by the researcher. Many useful tools are available with the database, and the software allows users to view images of microarray scans to evaluate the results visually.

#### **Gene Expression Omnibus (GEO)**

GEO is a gene expression and hybridization array data repository, as well as an online resource for the retrieval of gene expression data from any organism or artificial source. The database has a flexible design that can handle diverse styles of both unprocessed and processed data in a MIAME- (Minimum Information About a Microarray Experiment) supportive infrastructure that promotes fully annotated submissions. Several user-friendly Web-based interfaces and applications have been implemented that enable effective exploration, query, and visualization of these data, at the level of individual genes or entire studies.

## CHAPTER 4

### Development of a Fully Automatic Gridding Technique for High Density Microarray Images

---

*Ideally the spots in most microarrays are arranged in regular pattern, with multiple distinct 2D sub-arrays of spots. Gridding is the first step in the analysis of microarray images for locating these subarrays and individual spots within each subarray. This chapter describes state of the art gridding methodologies and explains a novel method for automatic grid alignment for high density microarray images using intensity projection profile of best subimage. Experimental results show that the method is capable of gridding microarray images with irregular spots, varying surface intensity distribution and contamination. Performance of the system has been evaluated in terms of gridding accuracy, robustness against noise and computation time.*

---

## 4.1 Introduction

Denser layout of high density microarray images encounters big challenges in gridding and segmentation using traditional techniques. The four common difficulties while implementing the gridding methods are:

***Uneven subarray position:*** The subarrays are not aligned with one another. This can happen if the pins are not perfectly aligned.

***Curves within the subarray:*** Glass slides are not completely horizontal or pin has moved slightly in the array and so the features (spots) are printed in a curved pattern.

***Uneven spot spacing:*** Slight variations in pin tip positions or the surface of the array is not flat.

***Irregular shape and size spots:*** More or less fluid has been deposited on the slide during manufacture.

Other parameters like misregistration of the red and green channels, rotation of the array in the image, deviation from symmetry are due to printer or scanning artifacts (Bowtell D. et al., 2003). Different background noises like Gaussian noise and sharp noise peaks that appear during the preparation of the microarrays may also introduce difficulties while gridding.

## 4.2 Literature Survey

At present there are three different types of gridding methodologies, corresponding to the degree of human intervention in the process. They are manual, semiautomatic and automatic technique for gridding.

### 4.2.1 Manual Gridding

This is essentially a computer aided image analysis method. This was the first method used in early days of microarray technology. In this method, all the parameters required for gridding have to be provided manually. Existing software

Scanalyze (M.B.Eisen, 1999) uses manual gridding method. The method is very time consuming and not applicable for high density microarray images. Moreover, considerable inaccuracy may be introduced due to human errors, particularly with arrays having irregular spacing between the spots and large variation in spot sizes.

#### **4.2.2 Semiautomatic Grid Alignment Technique**

The semiautomatic method requires some level of user interaction. This approach typically uses algorithms for automatically adjusting the location of the grid lines, or individual grid points after the user has specified the approximate location of the grid. ImaGene (Medigue.C et al., 1999), Dapple (Jeremy Buhler et al., 2000), UCSF Spot (Jain, A. N. et al. 2002), MAGIC (L. J. Heyer, 2005), Spot Finder (Saeed A. I et al., 2006) are some of the commercially available software in this category. However, these methods might not sufficient to meet the requirement of high throughput microarray image processing.

#### **4.2.3 Fully Automatic Grid Alignment Techniques**

These methods should reliably identify all spots without any human intervention. Automatic gridding algorithm utilizes the image processing techniques for calculating the parameters like spot diameter, spacing between spots and between sub arrays automatically. This will greatly reduce the human effort, minimize the potential for human error, and offer great deal of consistency in the quality of the data (Draghici, S, 2003). Automation of the gridding process is important to guarantee the repeatability of microarray image processing. i.e when the algorithm executed with the same data, result obtained should be same at each time. However, even for these algorithms, there are always limitations due to unpredictable deviations from the assumed array design, high contamination level, and large number of missing spots (Novikov E et al., 2006).

In the past several years, many gridding approaches have been proposed. Only a few state of the art methods have been proposed as providing fully

automatic gridding, but most of them do not address all requirements of fully automatic gridding, i.e. handling of irregular spots and robustness against noise, artifacts and image rotation.

Jain et al. (2002) used a gridding algorithm based on axis projection of image intensity. Integrated image intensities in both image dimensions, is used for automatic location of subarray grids. The algorithm requires a large number of bright spots and is not robust to misalignment of different grids.

Mathematical morphology is a technique used for analysis and processing geometric structures, based on set theory, topology, and random functions. It helps remove peaks and ridges from the topological surface of the images. One approach of gridding based on mathematical morphology method has proposed by Angulo, J et al. (2003) requires that, the blocks are well separated, and representative number of bright spots must be available to evaluate correctly the spot size using “spot size distribution law”.

X.H. Wang et al. (2003) reported a wavelet modulus maxima based for spot identification method. The approach is based on the detection of wavelet modulus maxima in the microarray images. The detected maxima are actually the contour of the spot.

Wang.Y. et al. (2005) has proposed a fully automatic gridding methodology using intensity projection profile of whole microarray image for estimating parameters necessary for gridding. It is found that, the method is sensitive to contaminations and large number of missing spots

Novikov E. et al. (2006) implemented a noise resistant grid finding algorithm which also uses intensity projection profile by transforming the fluctuations of the intensity of each row or column to special parameter and systematically penalizing the irregular region. Algorithm requires some basic parameters such as number of subarrays and number of spots in horizontal and vertical directions as input.

A method for detecting spot locations based on a Bayesian model has been recently proposed, and uses a deformable template to fit the grid of spots using a



posterior probability model for which the parameters are learned by means of a simulated-annealing-based algorithm (Ceccarelli B. et al., 2006).

Another method for finding spot locations uses a hill-climbing approach to maximize the energy, seen as the intensities of the spots, which are fit to different probabilistic models (Rueda L, et al., 2006).

An algorithm for recognizing distorted grids with perspective transformations is developed by Qi F, et al. (2006). The proposed approach contains three parts: (a) recognizing parameters of affinely distorted grids by fitting Gaussian mixture models (GMMs) to grid spectrums, (b) rebuilding the grid structures via a generating iteration based on the acquired parameters, and (c) eliminating nonlinear effects caused by perspective transformations with the median of infinite lines from local structures (MILLS) method.

A Radon-transform-based method that separates the sub-grids in a cDNA microarray image has been proposed by Rueda et al. (2007). The method applies Radon transform to find possible rotations of the image and then finds the sub-grids by smoothing the row or column sums of pixel intensities; however, that method does not automatically find the correct number of sub-grids, and the process is subject to data-dependent parameters.

Another approach for cDNA microarray gridding is a gridding method that performs a series of steps including rotation detection and compares the row or column sums of the top-most and bottom-most parts of the image (Wang Y et al., 2007). This method, which detects rotation angles with respect to one of the axes, either x or y, has not been tested on images having regions with high noise.

An automatic gridding method based on Evolutionary algorithm was suggested by Zacharia E et al. (2008). It is based on a genetic algorithm that discovers parallel and equidistant line segments, which constitute the grid structure. Thereafter, a refinement procedure follows which further improves the existing grid structure, by slightly modifying the line-segments.

Using maximum margin is a method for automatic gridding of cDNA microarray images based on maximizing the margin between rows and columns of

spots (Bariamis D et al., 2010). Initially, a set of grid lines is placed on the image in order to separate each pair of consecutive rows and columns of the selected spots. Then, the optimal positions of the lines are obtained by maximizing the margin between these rows and columns using a maximum margin linear classifier.

The following section describes the novel method developed for automatic gridding of high density microarray images.

### **4.3 Automatic Gridding of microarray images using intensity projection profile of best subimage.**

One of the major difficulties while implementing the automatic gridding algorithm using the intensity projection profile of the whole image is that, the parameters estimated will not be accurate, if the image consists of spots with high variability in luminance, size and shape such as blooming spots, doughnuts, comets. Accuracy of parameter estimation will also be affected, if the image has hybridization inconsistencies and other contaminations.

The developed method first locates each subarray by a global gridding technique and then identifies an optimum subimage within each subarray to accurately estimate parameters for locating each spot. The different steps involved in the gridding process are:

- (1) Pre-processing
- (2) Global gridding
- (3) Local gridding

### 4.3.1. Pre-Processing

The pre-processing steps are used to construct a binary reference image from the input image and keep the input image intact for further analysis. Different image processing techniques are used to create the binary reference image. The different pre processing steps are mentioned below:

**Step1:** Convert composite image (RGB) to gray level Image: The composite microarray image is converted to grayscale intensity image by eliminating the hue and saturation information while retaining the luminance. In a gray scale image, each pixel has only one value representing the intensity (I). RGB image can be converted to gray scale image by applying the following equation to each pixel.

$$I=0.2989 * R + 0.5870 * G + 0.1140 * B \quad (4.1)$$

Here R, G and B are the red, green and blue components of the each pixel in the composite image. For a two colour (red, green) composite microarray image, the blue component is set as zero.

**Step2:** Perform contrast enhancement using contrast-limited adaptive histogram equalization followed by intensity rescale so that it fills the data type's entire dynamic range.

#### *Histogram equalization*

Histogram of the digital image ' $f$ ' with intensity levels in the range  $[0, L-1]$  is the discrete function  $h(r_k) = n_k$ , where  $r_k$  is the  $k^{\text{th}}$  intensity value and  $n_k$  is the number of pixels in the image with intensity  $r_k$ . Normalized histogram,  $p(r_k)$  is obtained as:

$$p(r_k) = \frac{n_k}{n} \quad 0 \leq r_k \leq L \quad (4.2)$$

where 'n' is the total number of pixels in the image and *cumulative distribution function* corresponding to  $p(r_k)$  as

$$cdf(r_k) = \sum_{j=0}^k p_r(r_j) \quad (4.3)$$

which is the image's accumulated normalized histogram.

Histogram based image enhancement method uses a transformation (T) on every intensity level  $r_k$  of the input image, of the form  $s_k = T(r_k)$  to produce a new image  $f_1$ , such that its *cdf* will be liberalized across the value range, i.e. the transformed intensities will have a uniform probability density function (PDF). The new intensity level  $s_k$  is:

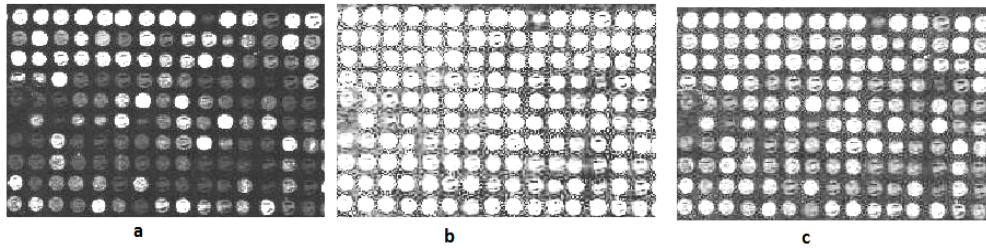
$$s_k = (L-1) \sum_{j=0}^k p_r(r_j) \quad (4.4)$$

This process is called histogram equalization and it is commonly used for image enhancement.

Adaptive histogram equalization (AHE) differs from normal histogram equalization in the respect that it operates on small regions in the image, called *tiles*, rather than the entire image. Each tile's contrast is enhanced, so that the histogram of the output region follows uniform distribution. It is therefore suitable for improving the local contrast of an image and bringing out more detail. The neighbouring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. In contrast limited adaptive histogram equalization, the contrast especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image.

Figure 4.1 (a) shows the gray level converted microarray image and (b) is the corresponding histogram equalized image. Better contrast enhancement can be achieved by adaptive histogram equalization as shown in (c). Intensity rescaling

maps the pixel's intensity into entire range. This helps to obtain maximum information from any spot.



**Figure.4.1** Preprocessing steps 1 and 2 (a) Gray scaled microarray image (b) Histogram equalized image(c) Adaptive histogram equalized image.

### Step3: Edge detection using canny method

Edge detection is the process of finding meaningful transitions in an image. The points where sharp changes in the brightness occur typically form the border between objects. Edge pixels are pixels at which the intensity of an image function changes abruptly and edges are set of connected edge pixels. Most of the edge detectors work on measuring the intensity gradient at a point in the image. Canny edge detection method (J.F. Canny, 1986) was developed to obtain an edge detector with the following properties:

1. *Good detection:* Mark as many real edges in the image as possible.
2. *Good localization:* Edges marked should be close to the edge in the real image.
3. *Minimal response:* A given edge in the image should only be marked once, and where possible, image noise should not create false edges

Canny edge detector defines edge as zero-crossings of second derivatives in the direction of the greatest first derivative. The Canny operator works in multistage process. First, the image is smoothed by a Gaussian convolution. Next, a 2D first derivative operator is applied to the smoothed image to highlight

regions of image with high spatial derivatives. Edges give rise to ridges in the gradient magnitude image. The algorithm then tracks along the top of these ridges and sets to zero all pixels that are not actually on the rigid top so as to give a thin line in the output. The tracking process exhibits hysteresis controlled by two thresholds  $T_1$  and  $T_2$  with  $T_1 > T_2$  thresholds. Tracking can only begin at a ridge higher than  $T_1$ . Tracking then continues in both directions out from that point until the height of the ridge fall below  $T_2$ . Canny edge detection is applied to microarray contrast enhanced images to detect the boundaries of the spots. After the edge detection a binary image with only boundary of the spots are generated as shown in Figure 4.2(b).

#### **Step 4. Morphological Hole Filling**

In image processing a hole is defined as a background region surrounded by a connected border of foreground pixels. The boundary of the spot created by edge detector creates a hole. Morphological filling can be applied to fill these holes.

#### **Morphological operations**

Mathematical morphology is a collection of non-linear processes that can be applied to an image. *Dilation and erosion* are fundamental operations of morphological processing. In fact, many of the morphological algorithms are based on these two primitive operations. Dilation, in general, causes objects to dilate or grow in size; erosion causes objects to shrink. The amount and the way that they grow or shrink depend upon the choice of a structuring element. Size and shape of the structuring element is determined by number of 0's and 1's in it. Morphological operations are defined by moving the structuring element over the binary image to be modified in such a way that it is centered over every image pixel at some point. When the structuring element is centered over a region of the

image, a logical operation is performed on the pixel covered by structuring element, yielding a binary output. With  $A$  as the image to be processed and  $B$  the structuring element, the dilation of  $A$  by  $B$ , denoted  $A \oplus B$ , is defined as:

$$A \oplus B = \{z \mid (\hat{B})_z \cap A \neq \emptyset\} \quad (4.5)$$

i.e. reflecting  $B$  about origin and, shifting this reflection by  $z$ . Here sets  $A$  and  $B$  are in  $Z^2$  (2D integer space). Erosion of  $A$  by  $B$  is denoted by  $A \ominus B$  is defined as:

$$A \ominus B = \{z \mid (B)_z \subseteq A\} \quad (4.6)$$

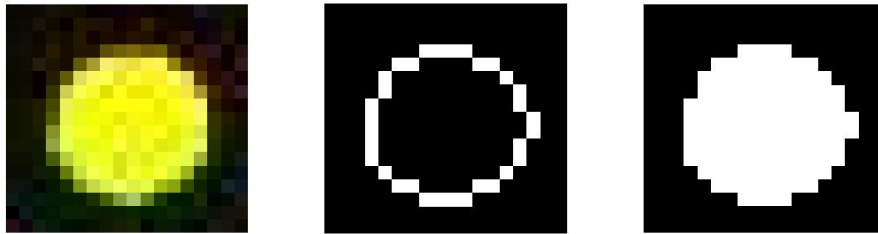
i.e., erosion of  $A$  by  $B$  is the set of all points  $z$  such that  $B$ , translated by  $z$ , is contained in  $A$ . The simplest kind of erosion is to remove any pixel touching another pixel that is part of the background. This removes a layer of pixels from around the periphery of all features and regions, which will cause some shrinking of dimensions and may create other problems if it causes a feature to break up into parts. Dilation can be used to add pixels. A combination of dilation and erosion can be used for different morphological operations.

### ***Hole filling***

Let  $A$  denote the set whose elements are 8-connected boundaries, each boundary enclosing a background region (hole). Given a point in each hole, the objective is to fill the hole with 1s. Let  $X_0$  represent an array of 0s (with size as the array containing  $A$ ), except at locations in  $X_0$  corresponding to the given edge point in each hole, which is set to 1. Then, the following procedure fills the hole with 1's which can be mathematically expressed as:

$$X_k = (X_{k-1} \oplus B) \cap A^c \quad k=1, 2, 3 \dots \quad (4.7)$$

Here also  $B$  is the structuring element. The algorithm terminates at iteration step  $k$  if  $X_k = X_{k-1}$ . The set  $X_k$  then contains all the filled holes. The set union of  $X_k$  and  $A$  contains all the filled holes and their boundaries. In Figure 4.2, (a) shows an example of a spot (b) is the resultant binary image after edge detection and (c) is the filled hole using a disc structuring element.



**Figure 4.2** Preprocessing-3 and 4 (a) spot (b) Edge detected spot (c) After filling holes.

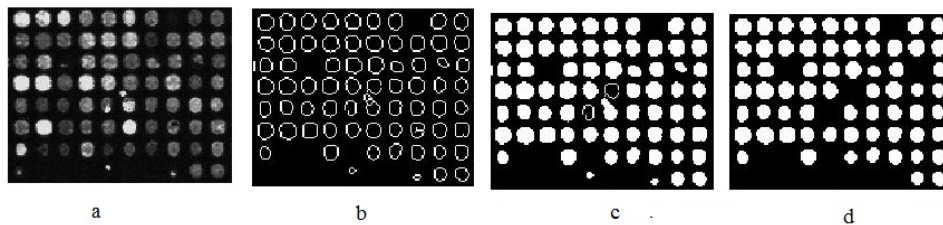
#### Step5: Morphological filtering

Morphological operations can be used to construct filters that remove noise. The filters are implemented by applying the operation called *opening*. Morphological opening are used to smoothen the contours of the object; it break narrow isthmuses and eliminate thin protrusions. Opening of an input image  $A$  by structuring element  $B$  is the erosion of  $A$  by  $B$  followed by dilation. Opening operation of input image  $A$  and structuring element  $B$  is represented as

$$A \circ B = (A \ominus B) \oplus B \quad (4.8)$$

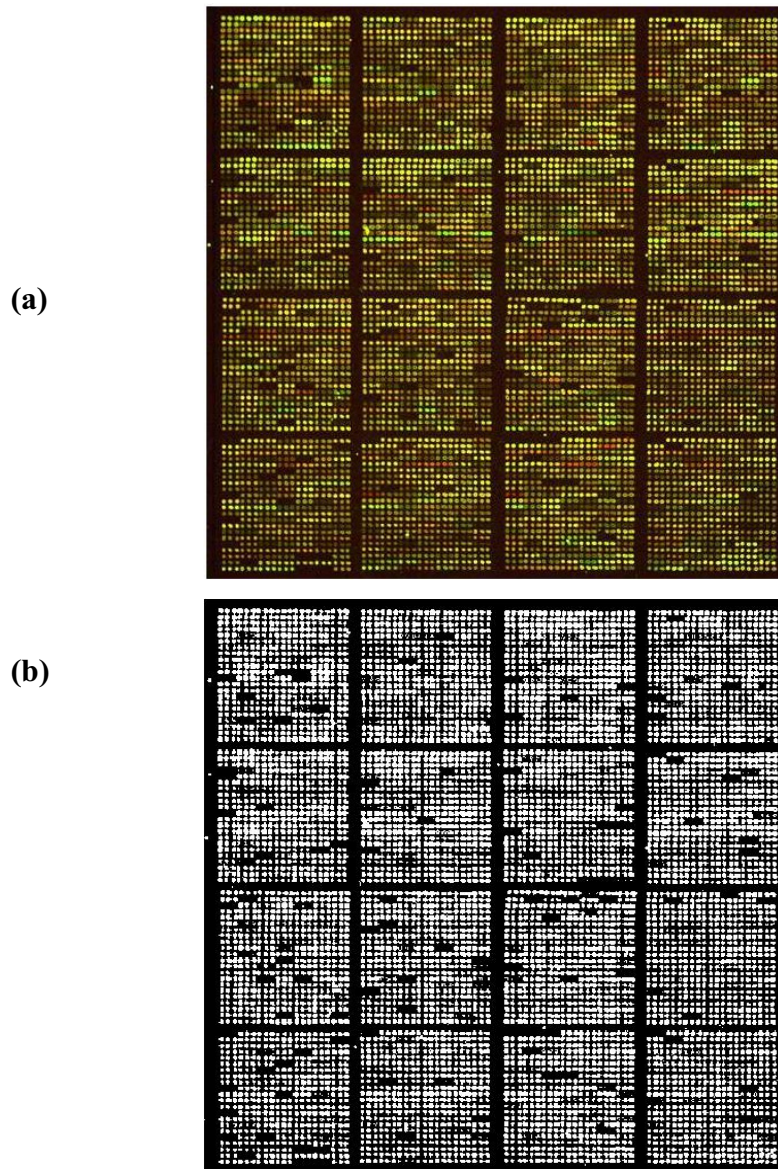


Morphological opening is used to remove islands as well as the noise that has been emphasized by the edge detection process in step3. Figure (4.3) shows the different preprocessing steps, here (a) is gray level image(b)is the image after edge detection(c)shows the filled holes and (d) is the noise eliminated image after applying the opening operation using a disk structuring element.



**Figure 4.3** Different morphological operations (a) gray level Image (b) image after Edge detection (c) Hole filling (d) Morphological opening.

Figure4.4 (a) shows an array CGH image consisting of approximately 7500 spots. Here each subarray consists of 462 spots. The preprocessing step is applied in the whole array, consists of approximately 7500 spots. The preprocessed binary image is shown in Figure 4.4 (b).This binary image generated after the five preprocessing steps is now suitable for estimation of parameters necessary for gridding.



**Figure 4.4.** (a) *Microarray image* (b) *Reference binary image*

### 4.3.2 Global Parameter Estimation (Global gridding)

Global gridding refers to the process of locating each subarray within a microarray image. The global parameters required for locating subarrays are width and height of each subarray as well as spacing between them. These parameters are estimated using the intensity projection profiles of the binary reference image generated after the preprocessing step. Horizontal and vertical intensity projection profiles of binary reference image are the sum of pixel intensities along each row and column respectively.

Let  $I_b$  indicates the binary reference image of size  $M \times N$ . Then, the intensity projection profile along  $r^{\text{th}}$  row ( $I_{pr}$ ) and  $c^{\text{th}}$  column ( $I_{pc}$ ) are computed using equation (4.9) and 4.10)

$$I_{pr} = \sum_{j=1}^N I_b(r, j) \quad (4.9)$$

$$I_{pc} = \sum_{i=1}^M I_b(i, c) \quad (4.10)$$

Figure4.5 (a) and (b) shows intensity projection profile of the reference image in Figure4.4 (b). These intensity projection profiles have to be thresholded for the estimation of the global parameters. Let  $T_r$  and  $T_c$  be the threshold values for row and column profiles respectively. Then thresholded values  $I_{pr(T)}$  and  $I_{pc(T)}$  are calculated using equations 4.11 and 4.12.

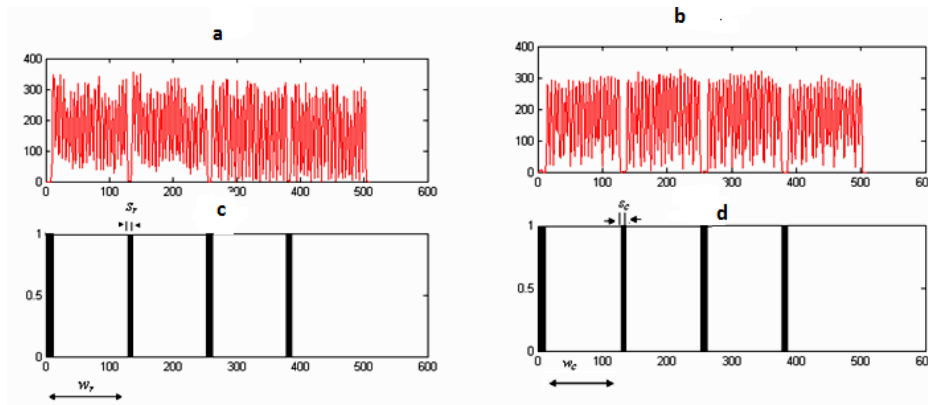
$$I_{pr(T)} = 1, \text{ if } I_{pr} \geq T_r \quad (4.11)$$

$$= 0 \text{ otherwise}$$

$$I_{pc(T)} = 1 \text{ if, } I_{pc} \geq T_c \quad (4.12)$$

$$= 0 \text{ otherwise}$$

Figure 4.5(c, d) shows the thresholded projection profile. Here ‘ $w_r$ ’ and ‘ $w_c$ ’ denotes the row and column width of the subarrays respectively. ‘ $s_r$ ’ and ‘ $s_c$ ’ are the row and column spacing between the subarrays. Regions with contaminations or other artifacts show large variation in these parameter values from their mean value.



**Figure 4.5** Intensity Projection profiles (a) Row profile (b) Column profile (c) Thresholded row profiles (d) Thresholded column parameters

Accuracy of the gridding parameters can be increased by eliminating these irregular elements. Let,  $W_R$  is the set of all  $w_r$  and  $W_C$  is the set of all  $w_c$  as given in equation 4.13 and 4.14.

$$W_R = \{w_{r1}, w_{r2}, \dots, w_{rk}\} \quad (4.13)$$

$$W_C = \{w_{c1}, w_{c2}, \dots, w_{cl}\} \quad (4.14)$$

The median values of  $W_R$  and  $W_C$  are evaluated as  $\tilde{w}_r$  and  $\tilde{w}_c$  respectively. Any row width  $w_{ri}$  (for  $i=1, 2, k$ ) and column width  $w_{cj}$  (for  $j=1, 2, \dots, l$ ) will be considered for the evaluation of final parameters of the subarray, if the following condition is satisfied.

$$0.5\tilde{W}_r \leq W_{ri} \leq 1.5\tilde{W}_r \quad (4.15)$$

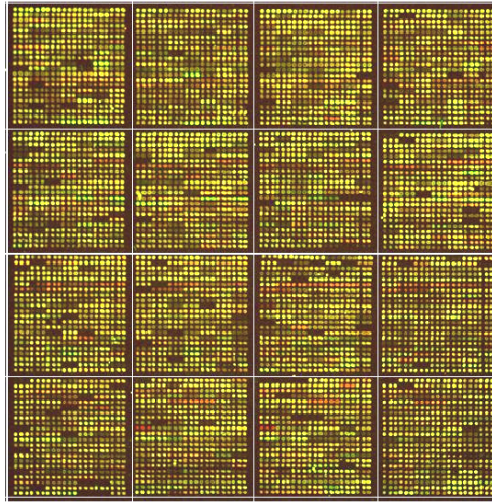
$$0.5\tilde{W}_c \leq W_{cj} \leq 1.5\tilde{W}_c \quad (4.16)$$

Thus the irregular elements in  $W_R$  and  $W_C$  are rejected. Using the selected  $w_{ri}$  and  $w_{cj}$  new median values  $\tilde{w}_{rn}$  and  $\tilde{w}_{cn}$  are estimated. The same procedure is applied for the spacing parameters  $S_{ri}$  and  $S_{cj}$  to reject irregular spacing variables. New median spacing values are denoted by  $\tilde{s}_{rn}$  and  $\tilde{s}_{cn}$ . The global gridding parameters, subarray rowwidth ( $G_R$ ) and subarray column width ( $G_C$ ) are estimated as:

$$G_R = \tilde{w}_{rn} + \tilde{s}_{rn} \quad (4.17)$$

$$G_C = \tilde{w}_{cn} + \tilde{s}_{cn} \quad (4.18)$$

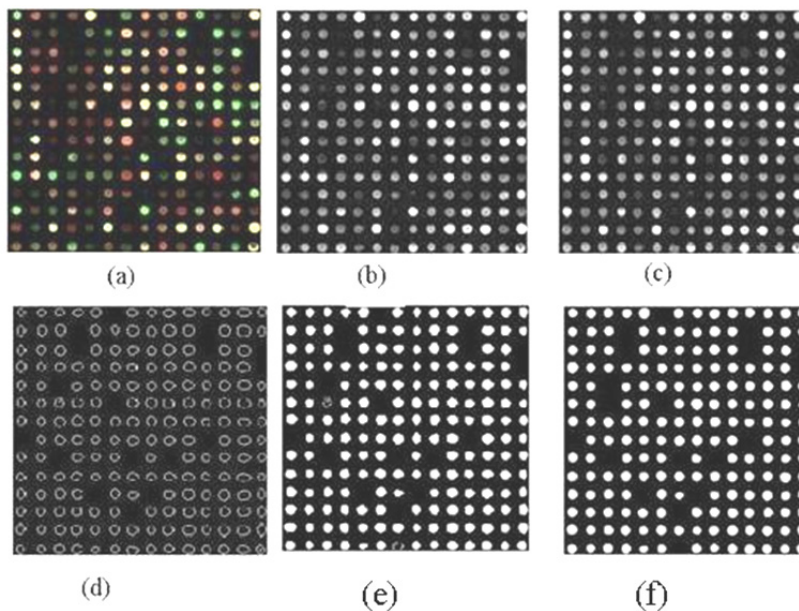
Thus, the subarray grid size is defined using rectangular window of size  $G_R \times G_C$ . The resultant image after applying the gridding algorithm is shown in Figure.4.6.



**Figure 4.6.** *Microarray image after global gridding*

### 4.3.3 Local Gridding

A typical microarray slide consists of rectangular subarrays of spots. There are variations among the individual subarray due to non-uniformity in the hybridization, artifacts on the surface of the array and gaps or dark areas where little or no hybridization has occurred. Once each subarray has been located correctly, the next step is locating each spot within the subarray. This process is called local gridding. The preprocessing steps are applied to each subarray for generating binary reference subarray. Fig.4.7 illustrates different preprocessing steps applied to a subarray consisting of 196 spots.



**Figure 4.7** Different Preprocrssing steps (a) A subarrayof microarray image (b) Grayscale image (c)Adaptive histogram equalization (d)Canny edge detection(e) Morphological filling (f)Binary reference image after applying Morphological filter.

Parameters for locating individual spots are estimated from the most suitable subimage within the reference image. Local gridding process consists of. two steps (i) Identification of an optimum subimage (ii) Parameter estimation.

#### 4.3.3.1 Identification of the best subimage

Block processing method is used to identify the best subimage within the reference image. The different steps involved in this process are explained below:

**Step 1: *Identify the optimal block size for block processing.***

Consider a subarray of size  $[m \ n]$ . To determine the row and column dimension of the optimum block (subimage) for block processing, First define the maximum size of the subimage using a scalar  $K$ . Let  $K = \text{maximum}(m/2, n/2)$ . The algorithm for determining block size ( $p1 \times p2$ ) is as follows:

- If  $m$  is less than or equal to  $K$ , return  $m$ .
- If  $m$  is greater than  $k$ , consider all values between  $\min(m/10, k/2)$  and  $k$ .
- Return the value that minimizes the zero padding required. The same algorithm is repeated for  $n$  also.

**Step2:** Once the block size has been identified as  $p1 \times p2$ , using the sliding window approach, calculate the mean intensity value of each block, as window slides pixel by pixel from the top left to bottom right of the binary reference image.

**Step4.** Select the block with maximum mean intensity  $I_{\max}$

**Step5.** Find thresholded intensity projection profile of this image block using the same procedure given by equations (4.11) to (4.14)

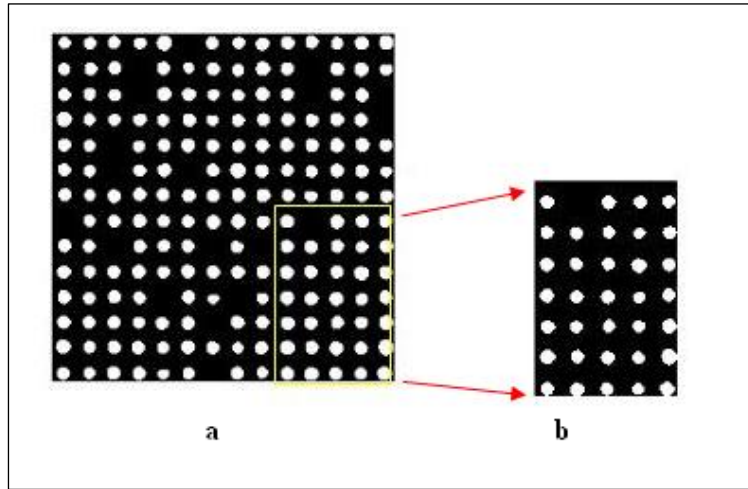
**Step 6:** Let,  $w_{rb}$  and  $w_{cb}$  be the row and column widths of the thresholded projection profiles ( shown in Fig.4.9) and let  $W_{RB}$  be s the set of all  $w_{rb}$  's and

$W_{CB}$  is the set of all  $w_{cb}$ 's with their median values  $\tilde{w}_{rb}$  and  $\tilde{w}_{cb}$  and standard deviation  $\tilde{\sigma}_{rb}$  and  $\tilde{\sigma}_{cb}$ .

$$W_{RB} = \{w_{rb1}, w_{rb2}, \dots, w_{rbp}\} \quad (4.19)$$

$$W_{CB} = \{w_{cb1}, w_{cb2}, \dots, w_{cbq}\} \quad (4.20)$$

Then, the subimage is selected as the best sub image if both  $\tilde{\sigma}_{rb}$  and  $\tilde{\sigma}_{cb}$  are less than 50% of  $\tilde{w}_{rb}$  and  $\tilde{w}_{cb}$  and respectively. Otherwise, the selected subimage is rejected. Then search next subimage with next lower mean intensity and repeat steps 5 and 6 until the optimum sub image has been identified. The median value of all the row spacing ( $\tilde{s}_{rb}$ ) and column spacing ( $\tilde{s}_{cb}$ ) in the selected subimage are estimated from the thresholded projection profile. Figure (4.8) shows a reference image and its optimum subimage. Fig (4.9) is the projection profiles of the subimage before and after thresholding.

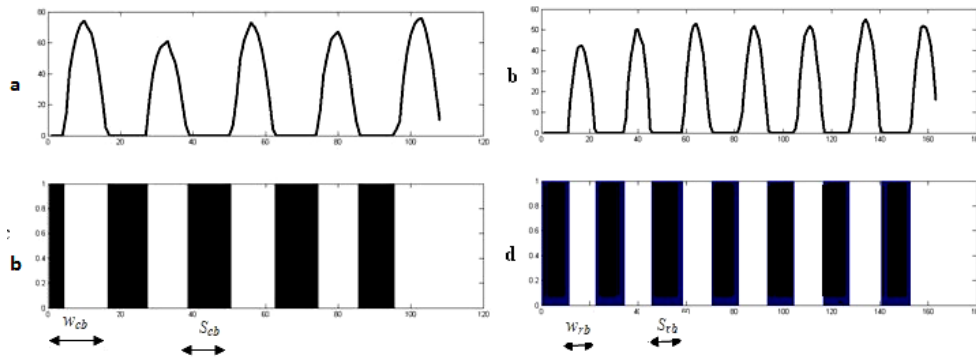


**Figure 4.8.** (a) Binary reference image (b) Identified optimum sub image



### 4.3.3.2 Parameter Estimation

Parameters required for exactly locating the spot are spot diameter and row and column spacing between spots. Using these parameters grid size is evaluated. Spot diameter ( $D$ ) is calculated using equation 4.21.



**Figure 4.9.** Intensity projection profile of optimum subimage (a) column (b) row. (c) Thresholded column profile (d) Thresholded row profile

$$D = \frac{(\tilde{W}_{rb} + \tilde{W}_{cb})}{2} \quad (4.21)$$

Row distance ( $L_R$ ) and column distance ( $L_C$ ) of each local grid are evaluated as in equation 4.22 and 4.23

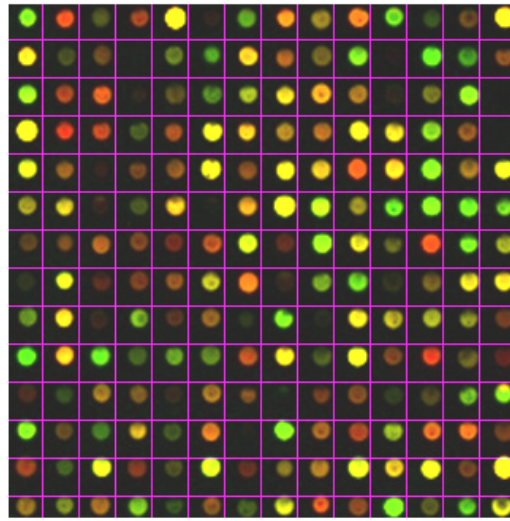
$$L_R = D + \tilde{s}_{rb} \quad (4.22)$$

$$L_C = D + \tilde{s}_{cb} \quad (4.23)$$

The local grid size for each spot is  $L_R \times L_C$ . The accuracy of the gridding algorithm is calculated as:

$$\text{Percentage Accuracy} = \frac{\text{Number of spots perfectly gridded}}{\text{Total number of spots}} \times 100 \quad (4.24)$$

Figure. (4.10) shows the resultant image after applying the gridding algorithm. Gridding accuracy obtained is 100%.



**Figure 4.10.** *Gridded Image*

#### **4.4 Implementation**

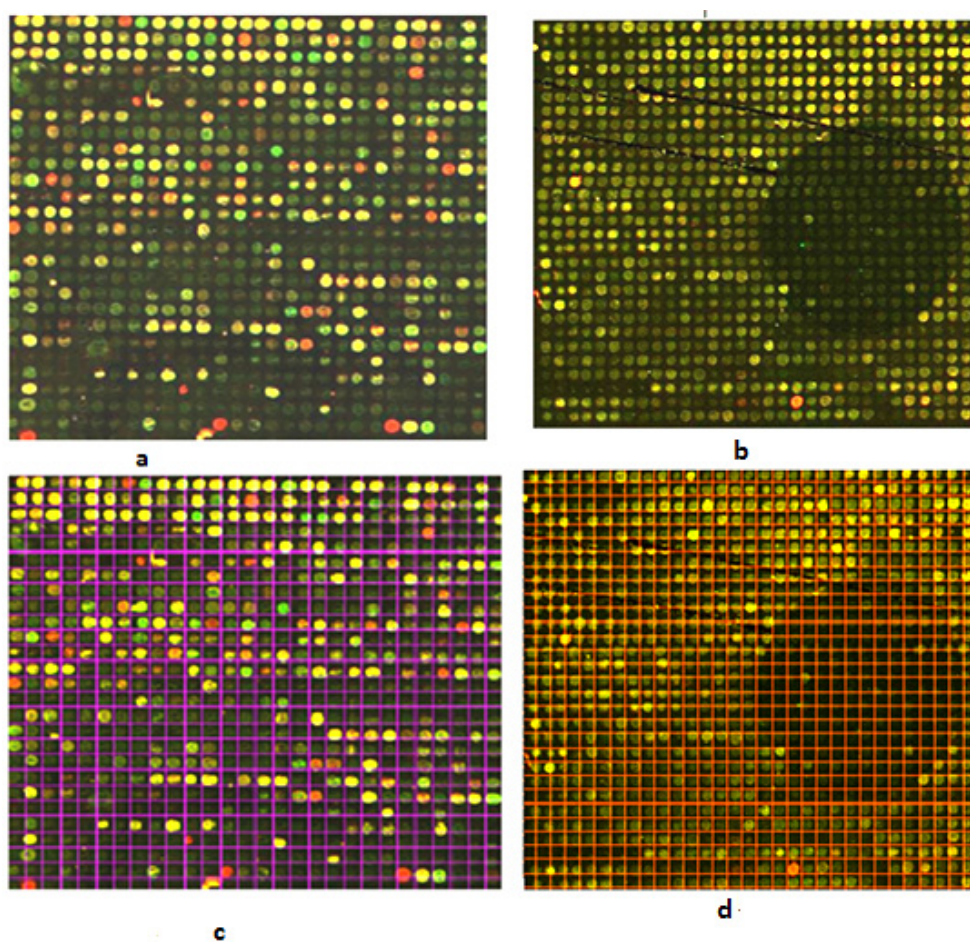
This algorithm has been implemented using MATLAB software. The microarray images available from Stanford microarray database (SMD) were used to implement the gridding algorithm. Stanford Microarray data base provide the images from each channel separately as well as in the composite RGB image format. It is a common practice that, before applying the gridding algorithm, the 16 bit tiff images from each channel are compressed to 8 bits. Then, the 24 bits composite RGB image is created using the intensities from red channel (Cy5), green channel (Cy3) and blue channel zero values Yang Y.H et al. (2001). The present work implements gridding algorithm on 24 bit composite image. The validity of the algorithm has been tested and confirmed using 10 real images with

high intensity spots(From Lung cancer studies), 20 images with different level of contaminations(images from brain tumor ,diabetics studies ),10 different noisy images (leukemia studies). The variation of gridding accuracy with the coefficient of variation (ratio of standard deviation to mean intensity) was studied on different subarrays within a microarray as well as between different arrays with varying spot size. Performance of the spot gridding algorithm was evaluated by comparing the results with the method demonstrated Wang Y.et al. (2005), which was also implemented for comparison. To study the influence of various noises that commonly occur during microarray image acquisition, artificial images were generated with known parameters. Noises that are common in microarrays like Gaussian and Salt and Pepper noise are added with these images and the gridding accuracy was evaluated and compared with the existing method.

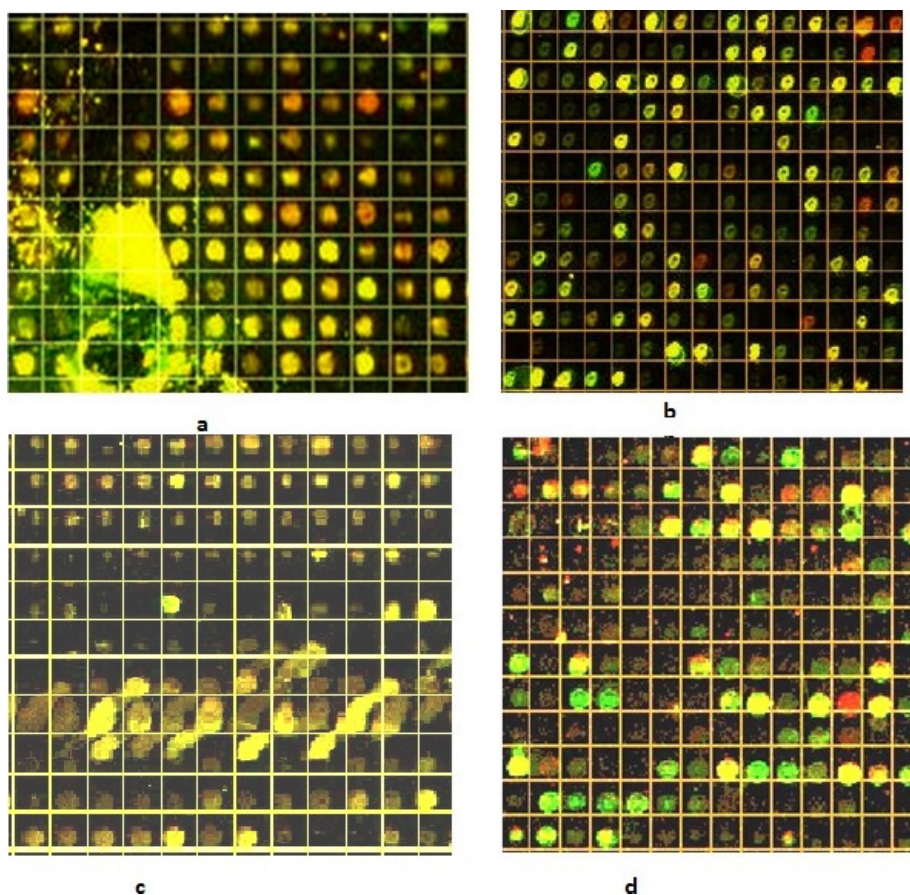
#### **4.5 Experimental Results and Performance Analysis**

The new gridding method has been tested on subarrays from 40 different microarray images with different characteristics. Figure 4.11 shows two subarrays and their gridded images. In Figure.4.11, (a) is a subarray (ID-11712) with 756 good quality spots having uniform shape and size. The gridding method was applied and spots were located with an accuracy of 98% as shown in shown in Figure 4.11 (c). Figure 4.11 (b) is a sub array (ID-18842) with a large bubble and (d) shows the corresponding gridded image with gridding accuracy of 91%.

Figure 4.12 shows part of the subarrays with different contaminations, (a) is a sub array (ID- 27746) with high background. b) is subarray from ID 27746 itself, but the spots have different size and shape. (c) is a part of the subarray consisting of 900 spots with large number of comet tail spots and (d) is a noisy subarray. Since the intensity projection profile of the best subimage was used for the evaluation of gridding parameters high gridding accuracy was obtained in all these cases.

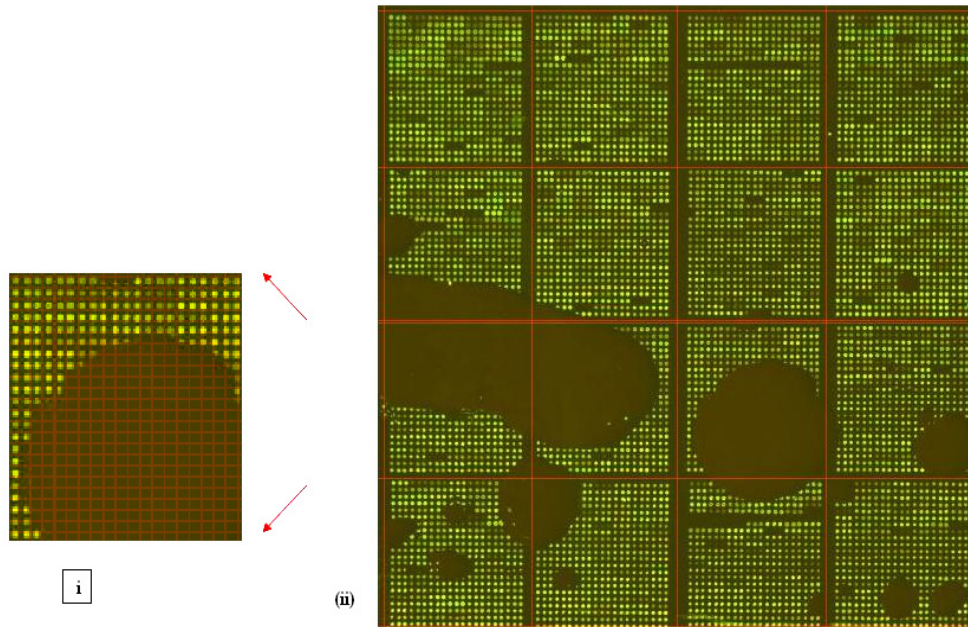


**Figure 4.11** Implementation of new gridding method on two sub arrays (a&b) and the resultant gridded images (c&d)



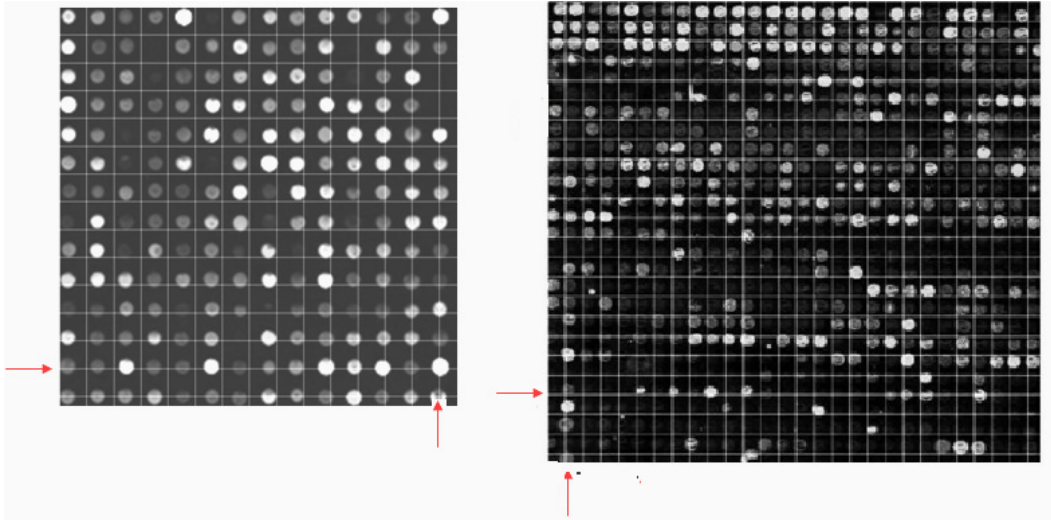
**Figure 4.12** *Different gridded subarrays with contaminations*

Fig.4.13 shows 1024×1024 pixels microarray image with 7392 spots provided by microarray image analysis software package MAIA. The contamination level of different subarrays are different as can be seen from Figure 4.13(ii).It has been shown that the algorithm was able to grid the spots with an accuracy of 97% even for a subarray with more than 50%contamination.



**Figure 4.13** (i) A gridded subarray with more than 50% contamination  
(ii) Microarray image

Performance of the spot gridding algorithm was evaluated by comparing the results with the method proposed by Wang. Y. *et al.* (2005), which was also implemented for comparison. In this method intensity projection profile of the whole image was used to estimate the parameters necessary for gridding. Figure 4.14 shows the results of applying this algorithm on subarray images shown in figure 4.7(a) and 4.11(a). Arrow shows some of the misaligned grids. Gridding accuracy of 81.72% and 64.12% were obtained for the images in 4.7(a) and 4.11(a) respectively using this previous method. The accuracy obtained with the present work while implementing on the same sub arrays were 100% and 98% respectively.



**Figure 4.14** Results of applying Gridding by using the intensity projection profile of whole subarray (Wang Y. et al.) on images in Fig 4.7(a) & 4.11(a)

Using Wang's method gridding accuracy was found very low for microarray images with contamination. Table 4.1 shows the comparison of gridding accuracy between two methods when applied to subarrays with varying number of spots and intensity levels/contamination. Since the new method selects an optimum subimage for parameter estimation high gridding accuracy was obtained. Coefficient of variation of an image is defined as ratio between standard deviation and mean intensity. Table 4.2 shows the gridding accuracy when the two methods were applied to 12 different sub arrays with same number of spots but varying coefficient of variation. As the coefficient of variation changes from 0.5 to 2.15, gridding accuracy of the earlier method decreases from 94.55% to 68.66%. But using the new method an accuracy of 99.14% to 89.43% was obtained. Figure 4.15 shows the plot between the coefficient of variation and the gridding accuracy. Result indicates the superior performance of the present work when compared with the method that uses intensity projection profile of the whole image.

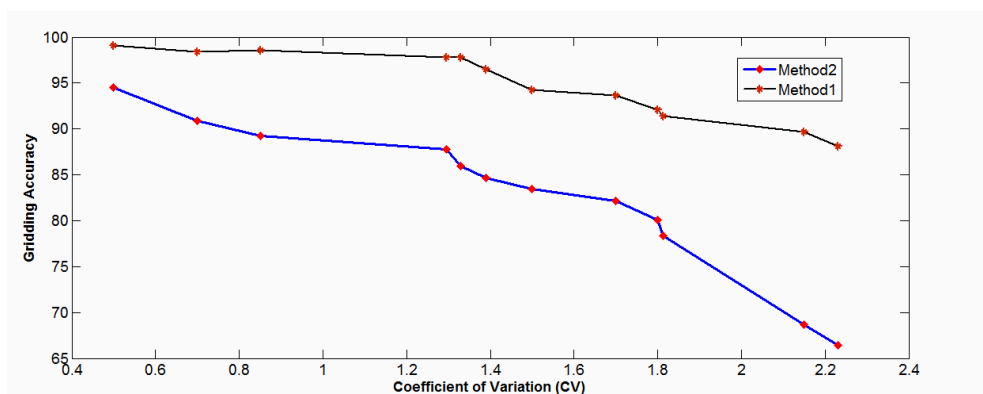
**Table 4.1** Comparison of Gridding accuracy between two methods-  
Using Different subarrays

Number of spots in Subarray	Mean intensity	Standard deviation	Coefficient of Variation (CV)	Gridding Accuracy (%)	
				New method	Wang 's method
1024	29.5	25.66	0.87	97.15	83.25
756	37.9	49.09	1.29	96.97	81.74
756	22.93	36.6	1.59	93.53	81.51
900	29.94	45.2	1.51	93.90	82.92
756	26.12	47.08	1.80	91.84	77.65
196	21.33	45.63	2.14	90.51	64.12

**Table 4.2** Comparison of Gridding accuracy between two methods for subarrays with same number of spots but different CV values

Coefficient of Variation (CV)	Gridding Accuracy (%)	
	New method	Wang 's method
0.5	99.14	94.55
0.7	98.37	90.89
0.85	98.56	89.21
1.295	97.81	87.74
1.3300	96.53	85.98
1.3900	94.21	84.63
1.5000	93.68	83.47
1.7000	92.13	82.16
1.8000	91.44	80.07
1.8130	89.69	78.39
2.15	89.43	68.66





**Figure 4.15** Gridding accuracy vs. Coefficient of Variation (CV) of subarray images -Comparison between Method 1 (New method), Method 2 (Wang.Y. et al).

### Effect on Noisy Microarrays

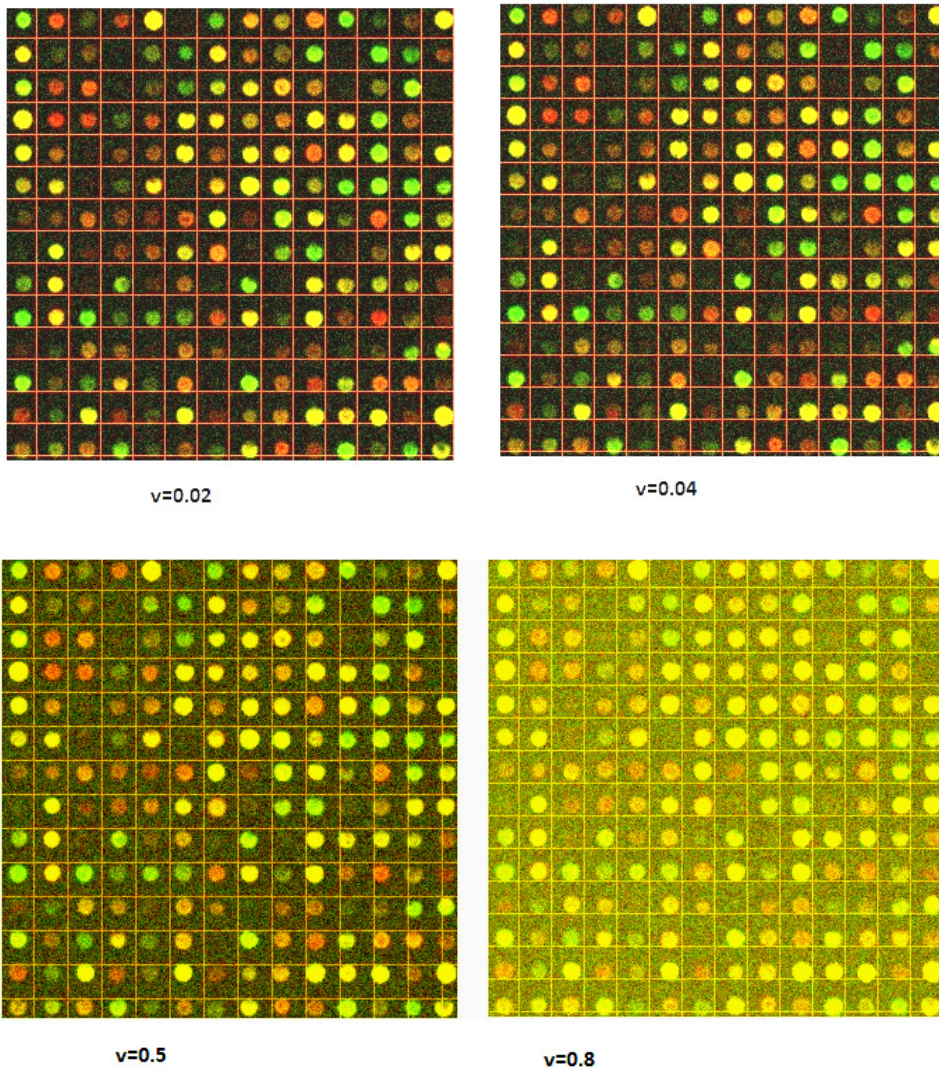
To study the influence of various noises that commonly occur during microarray image acquisition, artificial images were generated with known parameters. Background in microarray images contains many small, sharp noise peaks as well as weaker Gaussian noise (Katzner.M et al.,2003). Gaussian and Salt and Pepper noises were artificially added to the microarray images and the gridding accuracy was evaluated for varying variance values. Figure 4.16 shows the gridded noisy images when Gaussian noise at different variance levels were added. Table 4.3 shows the performance of the gridding algorithm against Gaussian and Salt and Pepper noise. Results shows that using the new method at a variance of 0.8, gridding accuracy of 91.84% was obtained for microarray images with good quality spots, while the accuracy obtained by considering the intensity projection profile of the whole image was only 18.88%.To study the effect of sharp noise peaks, Salt and Pepper noise with various density levels were added Figure 4.17(a) and (b) shows the results of applying gridding algorithm on microarray images contaminated with Salt and Pepper noise. It is clear from (b) that, with a noise

density of 0.3 the image is highly contaminated. Figure 4.18(a) shows a preprocessed image from a noisy microarray and (c) is the optimum subimage identified. Large amount of noise peaks were eliminated using morphological filters during the preprocessing step. Since parameters are estimated using the binary reference image after the preprocessing step, the effect of this noise can be greatly reduced. Figure 4.17(b) shows the gridded image using the intensity projection profile of the optimum subimage in 4.18(b). Figure 4.19 shows the gridded noisy image using previous method.

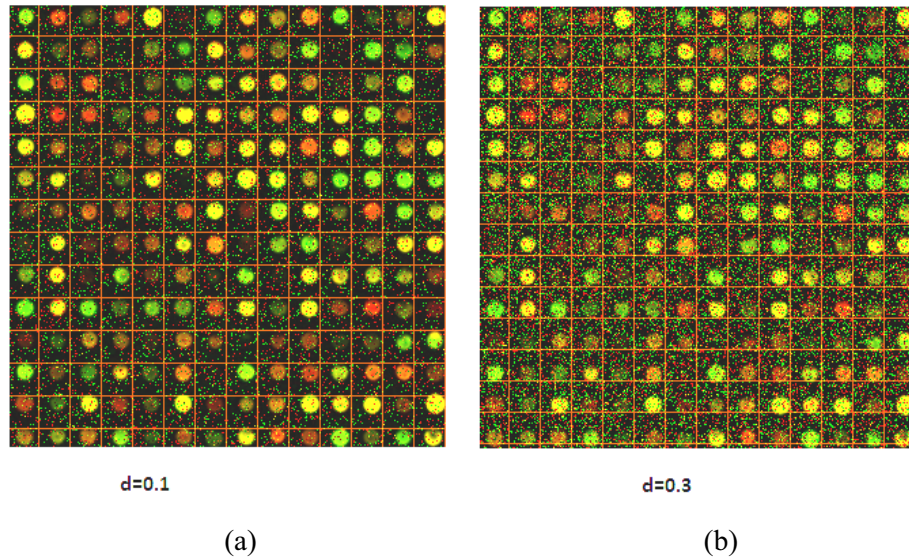
Table 4.3 shows that gridding accuracy obtained for the two methods using Gaussian noise (with variance upto 0.8) and Salt and Pepper noise (with noise density upto 0.3). It is seen that an accuracy of 100 % to 97.6% and 56.12% to 5.1% respectively were obtained, for Salt and Pepper noise with noise density varying from 0.05 to 0.3, using the new method and the Wang's method. Similarly the values are 100% to 91.84 % and 61.22 % to 18.88 % respectively with the Gaussian noise as noise variance varies from 0.05 to 0.8, for the two methods.

**Table 4. 3** Gridding accuracy for noisy Microarray images

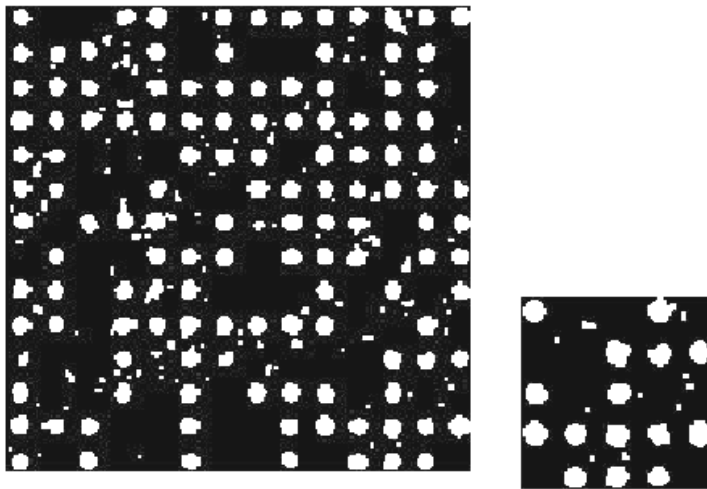
Gaussian Noise			Salt and Pepper Noise		
Variance (v)	Gridding Accuracy (%)		Noise Density (d)	Gridding Accuracy (%)	
	New method	Wang 's method		New method	Wang 's method
0.05	100	61.22	0.05	100	56.12
0.1	100	57.14	0.1	100	35.71
0.2	100	50.51	0.15	100	30.61
0.3	100	45.92	0.2	100	15.3
0.4	98	35.70	0.25	99.4	10.20
0.8	91.84	18.88	0.3	97.6	5.1



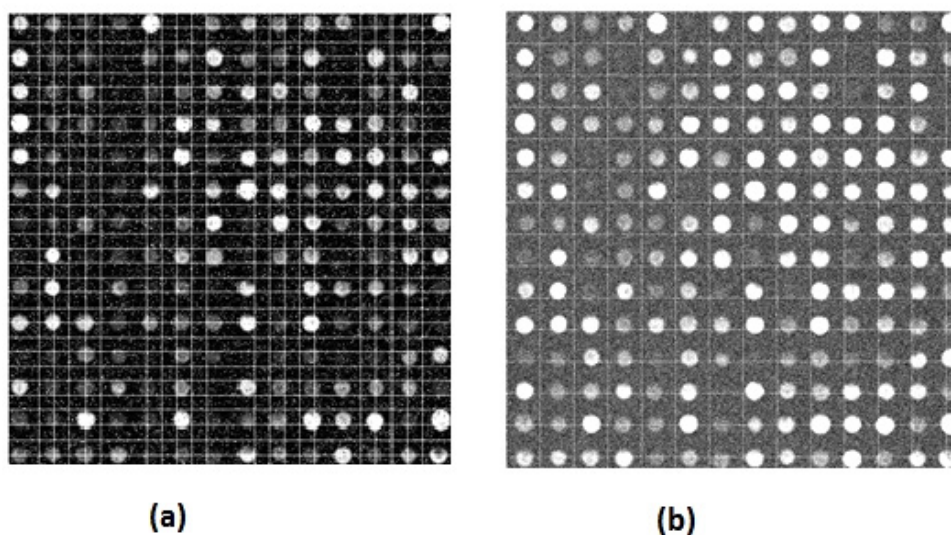
**Figure 4.16** Results of applying gridding algorithm on Gaussian Noise added Microarray image for different variance ( $v$ ) level.



**Figure 4.17** Results of implementing gridding method on Salt and Pepper noise added microarray image for different noise density ( $d$ )



**Figure 4.18** (a) Pre processed image from Salt and Pepper noise added image  
(b) Optimum subimage



**Figure 4.19** A gridded noisy image (a) Salt and Pepper noise with noise density 0.03 (b) Gaussian noises with variance 0.3 -using the projection profile of whole image

Comparison with the existing software MAGIC2.2 has shown similar performance for images with good quality spots. While locating the spots in arrays with low mean intensity as well as large contaminations the new method shows superior performance. Eventhough searching the best subimage is a time consuming task, especially if the contamination is high, using the block processing capability of the MATLAB software the computation efficiency has been improved.

Table 4.4 shows the execution time required for gridding subarrays with different number of spots using Pentium (R), Dual Core Processor 3GHz with 2GB of RAM system. The method takes less than 3 seconds for gridding a microarray image consisting of 900 spots with 1024×1024 pixels, making it suitable for gridding high-density arrays.

**Table 4.4.** Execution time for gridding sub arrays with different number spots

Number of spots in the subarray	Time required (sec)
100	0.832361
196	1.432831
552	2.294093
756	2.586965
900	2.7523924

#### 4.6. Conclusions

A new method of automatic gridding of microarray images based on intensity projection profile of best subimage has been introduced in this chapter. The method involves various tasks like preprocessing, identification of a subimage and parameter estimation. The most suitable subimage with maximum mean intensity and regular profile has been used to determine the parameters. It has been proved that accuracy is very high when compared with the existing methods that use projection profile of the entire image. It can automatically locate both subarrays and individual spots without any input parameters and human intervention. According to results obtained, the accuracy of our algorithm is between 90-100% for microarray images with coefficient of variation less than two. Experimental results show that the method is capable of gridding microarray images with irregular spots, comet tails, bubbles, dilated spots, varying surface intensity distribution and with more than 50% contamination. The method is robust with respect to different types of contamination and can tolerate a high percentage of missing spots to make it a suitable for gridding high density microarray images.

## CHAPTER 5

### Development of Automatic Adaptive Seed Region Growing Technique for Microarray Spot Segmentation

---

*One main challenge in spotted microarray image processing is the variation of results obtained by different researchers on the same microarray images even with a perfect placement of grids. Such a lack of robustness is mainly related to segmentation methods. This chapter presents state of the art segmentation techniques for microarrays and explains a novel segmentation method called automatic adaptive seed region growing. The method is fully automatic and is independent of size and shape of the spots. Experimental results show that the method is capable of accurately segmenting spots with low intensity and irregular morphology. Monte Carlo simulations on artificial spots shows that the proposed algorithm provide good segmentation accuracy and mean squared error of ratio, specifically at low signal to noise ratio (SNR) levels.*

---

## 5.1 Introduction

Accuracy of the segmentation algorithm plays a vital role in the analysis of microarray images. Irregularity in spot size and shape is a major hurdle while developing segmentation methods for microarray images. Although many software packages exist for segmenting microarray data, continual efforts are still been put for the segmentation of spots from high density microarray images. Moreover as the density of microarray increases number of pixels in the local background region decreases dramatically makes the background estimation a difficult task. In this chapter a novel spot segmentation method known as automatic adaptive seed region growing is described. Theory, implementation and performance analysis are discussed. A new local background estimation technique is developed which uses the information about the global background pixels to set a threshold for selecting local background.

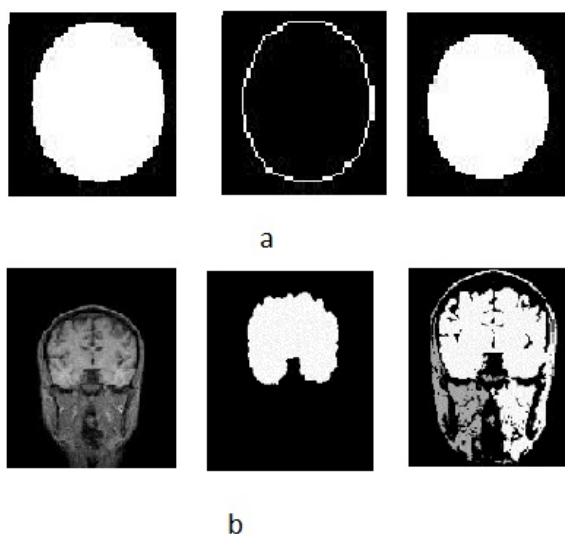
## 5.2 Image Segmentation Techniques

Image segmentation is the process of partitioning the image into a set of non-overlapping regions. The level of detail to which this partitioning is carried out depends on the problem being solved. The segmentation should stop when the object or region of interest in an application has been detected. Generally segmentation algorithms are based on one of the two basic properties of the intensity values such as discontinuity and similarity. In the first category, the approach is to partition the image based on abrupt change in the intensity, with the assumption that the boundaries of the regions are sufficiently different from each other and from the background. Edge based segmentation is the principal approach used in this category. The principle of second category is the partitioning the image into regions that are similar according to a set of predefined criteria. Region based segmentation is based on second category.

Figure 5.1 (a) shows an image which consists of an object with constant intensity superimposed on a darker background which is also of constant intensity.



The object can be segmented using edge based method. Image in Figure 5.1 (b) has lots of spurious change in intensity and edge based segmentation is not suitable for feature extraction. Third image in Figure 5.1 (b) shows the results of region based segmentation for extracting the region of interest.



**Figure 5.1** Different segmentation schemes (a) Edge based (b) Region based

The existing automatic image segmentation techniques can be classified into five approaches, namely, thresholding techniques, boundary-based methods, region-based methods, hybrid techniques and clustering-based techniques.

Seeded region growing (SRG) method of segmentation was introduced by Adams et al.(1994).This method is robust, rapid and free of tuning parameters. SRG is also very attractive for semantic image segmentation by involving the high-level knowledge of image components in the seed selection procedure. In microarray image segmentation seed region growing is the most suitable method for extracting features from irregular shaped spots.

### 5.3 Microarray Image Segmentation: A Review

Fixed circle algorithm is one of the first segmentation methods used in microarray studies. The algorithm is based on the assumption that all microarray spots are considered circular with a constant radius. After gridding, a circular mask of a fixed radius is placed on each spot location; pixels inside the mask are considered as spot foreground, everything else as background. The software tool ScanAnalyze (Eisen, 1999) uses this method for feature extraction.

Adaptive circle algorithm provides flexibility for the traditional fixed circle. Similarly, the algorithm assumes all spots as circular. However, the radius for each spot is estimated separately. This permits user interaction to adjust the radius for each spot. For high density microarrays such approach is extremely laborious and time consuming. An automated version of the adaptive circle is available in the Dapple software (Buhler et al., 2000), where the radius of each spot is estimated using edge detection. First, the outline of each spot is enhanced using the second-difference approximation of Laplacian. Thereafter, the radius of a circle, matching the given enhanced edges is identified with matched filtering.

In the software ‘Spot’ (Yang *et al.*, 2002), the seed region growing algorithm was used for microarray segmentation for the first time. The algorithm segments each spot by iteratively growing separate regions with respect to a set of predefined seed points providing a starting point for the segmentation. In each iteration, the algorithm includes the most homogenous pixels from the neighborhood to the segmented regions. Finally, the region originating from the foreground seeds is considered as the spot foreground, and the region originating from the background seeds as the background.

A segmentation algorithm based on the statistical Mann–Whitney test was first suggested by Chen et al. (1997). The algorithm iteratively computes the threshold between foreground and background using the Mann–Whitney test,

which is a non-parametric statistical test for assessing the statistical significance of the difference between two distributions. First, a circular target mask enclosing all possible foreground pixels separating them from the known background is selected. Second, a set of random pixels from the background are compared against a selected amount of pixels with the lowest intensity within the target mask using Mann–Whitney test. If the difference between the two sets is not significant, the algorithm discards some predetermined number of pixels from the target area and selects new pixels from the target area. The iteration is terminated when the two sets significantly differ from each other. Finally, the spot foreground is considered as the pixels remaining inside the target mask after iterations.

Another method is the  $k$ -means segmentation, based on the traditional  $k$ -means clustering, and was first used in connection with microarray images (Bozinov et al., 2002). The segmentation result is derived using simultaneous information from two channels.

The hybrid  $k$ -means algorithm (Rahnenführer et al., 2004) is an extended version of the original  $k$ -means segmentation approach. The algorithm uses repeated clustering to increase the number of foreground pixels. As long as the minimum amount of foreground pixels is not reached, the remaining background pixels are clustered into two groups and the group with higher intensity pixels is assigned into the foreground. In addition, the number of outlier pixels in the segmentation result is reduced with mask matching.

The Markov random field (MRF) modeling for the microarray spot segmentation was introduced by Demirkaya et al. (2005). The method models spot foreground and background intensities as exponential distributions. In addition to the intensity information, the method takes the spatial information into account by modeling the neighborhood pixel labelings with MRF. Initial classification into spot foreground and background is used as a basis for the segmentation, and the initial segmentation affects the final result given by MRF.

Model-based segmentation algorithm (Li et al., 2005) is a two-step method for spot segmentation. It consists of model-based clustering of pixel values and spatial extraction of connected components. Initially segmentation into at most three different clusters sharing similar intensity values, which are the background, the spots with background or artifact, and the foreground. Model-based clustering relies on Gaussian mixture models, and the number of clusters is defined based on data by using Bayesian Information Criterion (BIC). Spatial connected component removal is used for excluding small disconnected clusters that are assumed to be artifacts from the spot foreground pixels. Globally Optimal Geodesic Active Contours (GOGAC) is another segmentation method which was implemented in Spot (2007). Shenghua NI et al. (2009) presented an Active Contour without Edges (ACWE) method to detect objects' boundary by solving numerical finite difference equations. The major drawback of this method was the computation time.

#### 5.4 Region Growing Based Segmentation

Region in an image is a group of pixels with similar properties. Let  $R_g$  represent the entire spatial region occupied by an image. Segmentation is referred as a process that partitions  $R_g$  into 'n' subregions  $R_1, R_2, R_3, \dots, R_n$ . The basic formulations for region-based segmentation are:

- (a)  $\bigcup_{i=1}^n R_i = R_g$
- (b)  $R_i$  is a connected region,  $i=1, 2$
- (c)  $R_i \cap R_j = \emptyset$  for all  $i=1, 2, 3 \dots n$
- (d)  $P(R_i) = \text{True}$  for  $i=1, 2, 3, \dots, n$
- (e)  $(R_i \cup R_j) = \text{False}$  for  $i \neq j$

$P(R_i)$  is a logical predicate defined over the points in set  $R_i$  and  $\emptyset$  is the null set. Criterion (a) means that the segmentation must be complete; that is, every pixel

must be in a region.(b) requires that points in a region must be connected in some predefined sense.(c) indicates that the regions must be disjoint.(d) deals with the properties that must be satisfied by the pixels in a segmented region. For example  $P(R_i)=TRUE$  ; if all pixels  $R_i$  have the same gray level.(e) indicates that region  $R_i$  and  $R_j$  are different in the sense of predicate P.

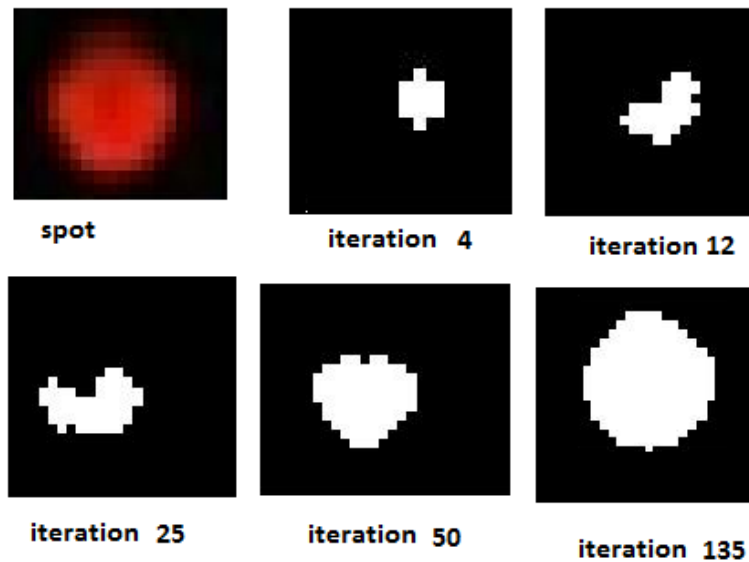
Region growing method of segmentation is a procedure that groups pixels based on predefined criteria. The basic approach starts with a set of seed points and from these grow regions by appending to each seed those neighboring pixels that have predefined properties similar to seed. Selection of seed points and the criteria (such as specific range of gray level) are based on the nature of the problem. Some researchers use edge-based method to select seeds. To get more accurate result it is better to take the center of the region or maximum intensity pixel as seed point. The connectivity or pixel adjacent information is helpful to determine the region growing criteria and seed points. The basic approach for segmenting any digital image using region growing is to start with a set of seeds  $S_1, S_2, \dots, S_n$  (R.Adams et al.1994). Sometimes individual sets will consist of single seed. The process evolves inductively from seeds. Each step of algorithm involves the addition of one pixel to one of the above set of seeds. For example, consider the region grown from seed  $S_i$  after 'm' steps. Let D be the set of all yet unallocated pixels which border at least one of the regions.

$$D = \left\{ x \notin \bigcup_{i=1}^n S_i \mid N(x) \cap \bigcup_{i=1}^n S_i \neq \emptyset \right\} \quad (5.1)$$

where  $N(x)$  is a set of immediate neighbours (8 neighbours) of pixel  $x$ . If, for  $x \in D$  and  $N(x)$  meet just one of  $S_i$ , then, an index  $i(x)$  is defined as  $i(x) \in \{1,2,\dots,n\}$  such that  $N(x) \cap S_{i(x)} \neq \emptyset$  and  $\delta(x)$  is a criteria is to measure how different  $x$  is from the region it adjoins. The simplest definition of measure of similarity is that the difference ( $\delta$ ) between a pixel's intensity  $g(x)$  and the region's mean  $g(y)$  as given in equation 5.2.

$$\delta = |g(x) - \text{mean}[g(y)]| \quad (5.2)$$

Then a given pixel  $x$  is appended to the seed if  $\delta \leq T$ , Where  $T$  is a user defined threshold. If  $N(x)$  meets two more  $S_i$ , then  $i(x)$  is chosen to be value of  $i$  for which  $\delta$  is minimized. The process is repeated until all pixels are allocated. Figure 5.2 shows an example of basic seed region growing approach of segmentation applied to a spot with center pixel is defined as the seed. Regions labeled with '1' are extracted from this single seed pixel. After each iteration the region grows.



**Figure 5.2** Basic seed region growing approach of segmentation

Region growing methods can correctly separate the regions that have the same properties. Region growing stops when no more pixels satisfy the criteria for inclusion in that region. An ideal candidate seed point should have these

properties: i) It should be inside the region and near the centre of the region ii) Assuming that most of the pixels in the region of interest (ROI) belong to the region, the feature of this seed point should be close to the region average (iii) The distances from the seed pixel to its neighbours should be small enough to allow continuous growing. SRG has been implemented in microarray processing package ‘Spot’ (Buckly M.J, 2000). In this package, the foreground seed is chosen as the center pixel of the horizontal and vertical grid line. To avoid the situation when the spot is small and the grid center is slipped out of the spot foreground, a small number of  $n \times n$  square pixels, whose center has the maximum intensity in a small area around the grid center, are taken as foreground seeds. The background seed is chosen as the point in which the grid lines intersect. After obtaining the seeds, the process is repeated simultaneously for both foreground and background regions until all the pixels are assigned to either foreground or background. Those pixels that are adjacent to a region are assigned first according to its intensity (Yang et al. 2002).

Jie Wu et al. (2008) proposed a new texture feature-based seed region growing algorithm for automated segmentation of organs in abdominal MR images, based on a cost-minimization approach. Frank et al. (2005) presents an automatic seeded region growing algorithm for color image segmentation, which satisfy the following three criteria. First, the seed pixel must have high similarity to its neighbors. Second, for an expected region, at least one seed must be generated in order to produce this region. Third, seeds for different regions must be disconnected.

In the present work, a new automatic segmentation method for high density microarray images has been developed. The following session describe the different steps in the development of the new algorithm.

## 5.5 Development of Automatic Adaptive Seed Region Growing (AASRG) Method

In microarray experiment, region based segmentation refers to classification of pixels as foreground (F) or background (B) so that fluorescence intensities can be calculated for each spot within the selected grid. This has to be done for the two channels separately. Let R represents the region of image within a grid. The segmentation partitions the region into two sub regions F and B, where F and B are two different connected regions. Since spots are accurately located during gridding stage, locating spot centre is not a difficult task. But for images with irregular shaped spots, selecting seed as centre region will not be an effective method, especially when spot surfaces exhibit non uniform intensity patches.

In the new method seed(S) and threshold (T) values are automatically selected, based on the characteristics of image within each grid. Let  $f(x, y)$  represents the image inside a grid. The different steps in AASRG algorithm are:

**Step1.** High frequency noise spikes within the grid are eliminated using  $3 \times 3$  median filter. Here each central pixel is replaced by the median value in the  $3 \times 3$  neighbourhood.

**Step2.** The coordinates of pixel (i, j) with maximum intensity in the median filtered image is considered as the locations of the seed. If there is more than one pixel with this maximum intensity, then pixel closer to the center pixel is considered as the seed.

**Step3:** If this seed pixel(S) is located within minimum distance D from the center of the grid and if, the number of pixels having intensity within 10% of the seed pixels intensity are greater than  $N_{\min}$  then the spot is considered as a *regular spot* and select S as the seed for the foreground calculation.

If the condition in step 3 is not satisfied then a circular mask with the estimated diameter of the spot (estimated during gridding process) is placed over the grid with center at the mid point of the grid. All the pixels inside the mask are selected and the mean intensity value is calculated. The pixel with intensity is



nearest to the mean value is selected as seed (S). This selection criterion is found to be useful for selecting seeds for spots with high background (black holes).

**Step4: Case1: Regular spots:** A threshold value T is calculated to test if a pixel is sufficiently similar to the seed to which it is 8-connected. For this all pixels within each grid are collected and sorted according to the intensity. Lower 10% of them are excluded from the calculation of foreground. This determines the lower cut off intensity value for foreground. The upper cut off intensity value is selected as the intensity of seed S. All pixels in the grid whose intensity values are between a lower and upper cutoff are considered for the calculation of parameters such as mean ( $\mu$ ) and standard deviation ( $\sigma$ ). These parameters determine the threshold value 'T' for segmenting the specific spot. Let  $I$  be the set of all pixels in the selected region having seed S and  $M$  be the number of selected pixels then the parameters are evaluated as:

$$\mu = \frac{\sum_{i=1}^M x_i}{M} \quad (5.3)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^M (x_i - \mu)^2}{M}} \quad (5.4)$$

where  $x_i \in I$

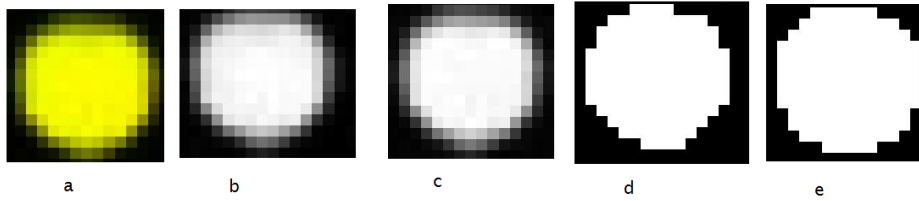
Threshold value T is calculated as:

$$\begin{aligned} T &= S - \mu \text{ if } \mu \leq S - \mu \\ &= 2 \times \sigma \text{ otherwise} \end{aligned} \quad (5.5)$$

*Region growing criteria:* A pixel in the region will pass the test if

$$|x_i - S| \leq T \quad (5.6)$$

i.e. if the absolute value of the difference between the gray level of the pixel and seed is less than or equal to the threshold value and the pixel is 8 connected to the seed and region grows. This process is repeated for all the pixels in matrix I. Resulting image is a binary image with segmented foreground pixels are labeled with value of '1' and remaining pixels labeled with '0'. The pixels labeled with 1 are used for calculating the intensity of the foreground signal. Figure 5.3(a) shows an yellow spot within the grid, (b) and (c) are the red and green channels respectively. Segmentation algorithm has been applied to each channel. Figure 5.3 (d) and (e) indicates the segmented foreground regions of red and green channels respectively.

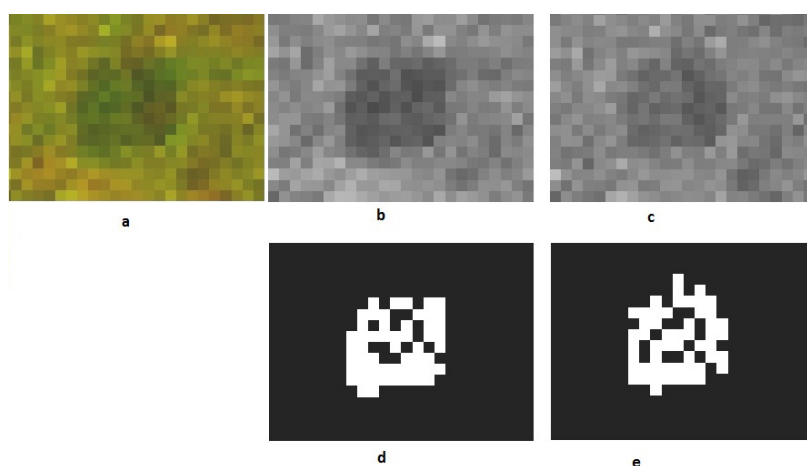


**Figure 5.3** AASRG applied to a good spot (a) Spot (b) Red channel (c) Green channel (d) Red channel foreground region (e) Green channel foreground region.

**Step 4 Case II:** Spots with high background intensity (black holes): It is a dark spot appeared after hybridization at the position of the probeDNA printed on the chip. The resulting spot will have fluorescent intensity which is less than that of the surrounding background. Figure 5.4 shows a spot with large background intensity. All the pixels within the circular mask are collected and standard deviation ( $\sigma$ ) is evaluated. The threshold in this case, T is defined as

$$T = \sigma \quad (5.7)$$

Figure 5.4 shows the result of applying the segmentation algorithm on the grid with high background intensity and the resulting red and green channel foreground pixels. The algorithm correctly identified the foreground regions.



**Figure 5.4** ASSRG applied to a black hole (a) Spot with high background intensity (b) red channel (c) Green channel (d) Red channel segmented (e) Green channel segmented

## 5.6 Background Extraction

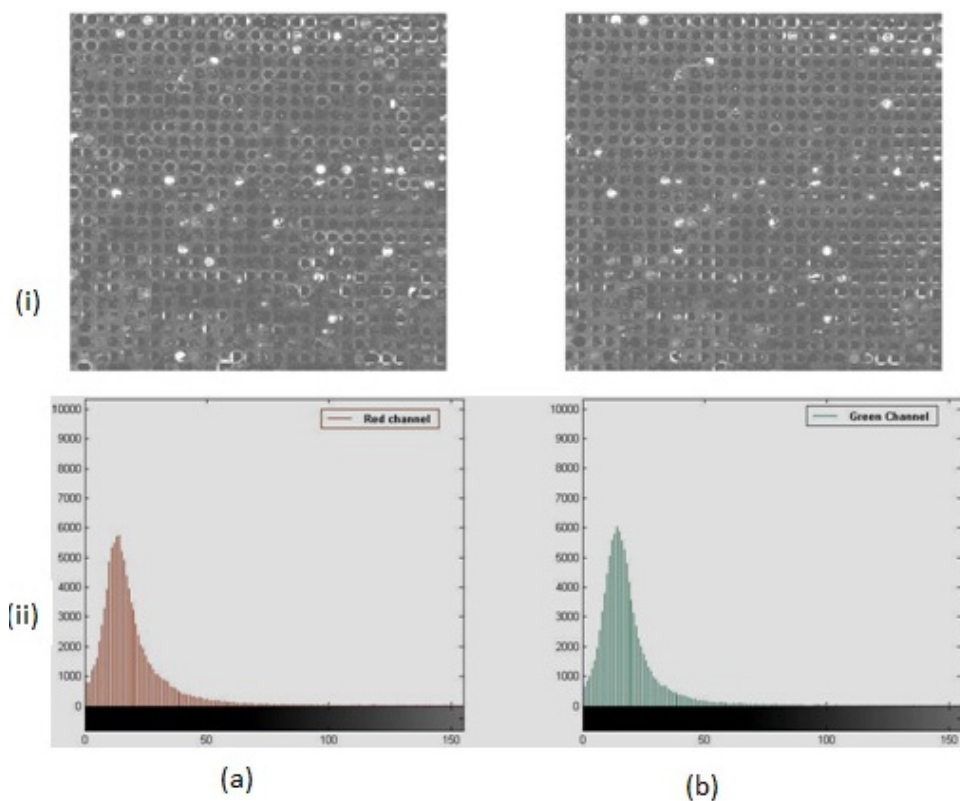
Background extraction from high density microarray is a major challenge for the researchers. Substrate noise is a factor that contributes to the background intensity. Substrate noise is the sum of all non-sample and non-instrument contributions to the background reading, including intrinsic fluorescence of the substrate and reflection off the substrate surface. Some of the existing background extraction methods are explained in section 3.2.3, which use the inter-spot pixels to extract local background. Other ways of estimating the local background are using histogram information and those using rank filters, which do not depend on the segmentation and exact position of the spot (Soille, P., 2002). Angulo, J et al.

(2003) suggested background extraction using a morphological filtering by area. But the performance of morphological opening method is found poor with spatial dependent bias (Bengtsson A. et al., 2006.). Ma M. Q. et al (2007) suggested a new method called Extended local background (ELB-Q). In ELB-Q, a large extended local background (ELB) inter spot region excluding those noise of the background pixels are used for estimating the local background by automatically determining an optimal threshold for background segmentation.

In the present work after applying segmentation algorithm using AASRG, pixels labeled with logic '0' are considered for background calculation. Corresponding pixels in the original image for each grid are collected and sorted according to the intensity. The top 10% intensity level pixels are excluded for the purpose of estimating the global background, which helps disregard the noises such as chemical contaminations in the microarray slide. The parameters such as the mean  $\mu_{\text{global}}$  and standard deviation  $\sigma_{\text{global}}$  are calculated using the remaining pixels whose intensities follow roughly a shifted Gaussian distribution. Then a local background cut-off intensity level ( $T_{\text{local}}$ ) is defined as

$$T_{\text{b\_local}} = \mu_{\text{global}} + \sigma_{\text{global}} \quad (5.8)$$

This determines is the cutoff value for selecting the background pixels in each grid. All the pixels in the local background region of each spot whose intensities are lower than  $T_{\text{local}}$  are used for calculating the local background intensity for that spot. This repeats for each spot. Fig 5.5 (i) shows the global background pixels and (ii) is the intensity distribution of global background region of red and green channels of the microarray image shown in Fig 4.11(a).



**Figure 5.5** Global back ground regions and their intensity Distribution (histogram) (a) Red channel (b) Green channel of microarray image in Fig.4.11(a).

### 5.7 Intensity Calculations

Once the pixels representing foreground and background regions are extracted, next step is the calculation of foreground and background intensity levels of each spot. Pixel intensity values from both channels are combined into a unique value representing the expression levels of a gene deposited in a given spot. Mean and median intensities in the segmented foreground region are estimated for each channel separately as  $\mu_{fr}$  (mean red channel),  $\mu_{fg}$  (mean green channel),  $m_{fr}$

(median- red channel) and  $m_{fg}$  (median- green channel). The background intensity is subtracted from the foreground intensity to provide a more reliable estimate of hybridization intensity to each spot. However if the background intensity is higher than the feature intensity, the result would be negative, which would not be meaningful. Mean and the median intensities of background are calculated using the selected pixels of red and green layer separately as  $\mu_{br}$  (mean-background- red channel),  $\mu_{bg}$  (mean-background- green channel),  $m_{br}$  (median – background-red channel) and  $m_{bg}$  (median –background-green- channel). Now the True mean signal intensity is calculated for each spot for red(R) and green layers (G) as

$$R = \mu_{fr} - \mu_{br} \quad (5.9)$$

$$G = \mu_{fg} - \mu_{bg}$$

For noisy microarray images median intensity over the mean is preferred because median is more robust to outliers pixels than mean. The median intensity for two channels ( $\tilde{R}$  and  $\tilde{G}$ ) are evaluated as

$$\tilde{R} = m_{fr} - m_{br} \quad (5.10)$$

$$\tilde{G} = m_{fg} - m_{bg}$$

### 5.7.1 Intensity Ratio

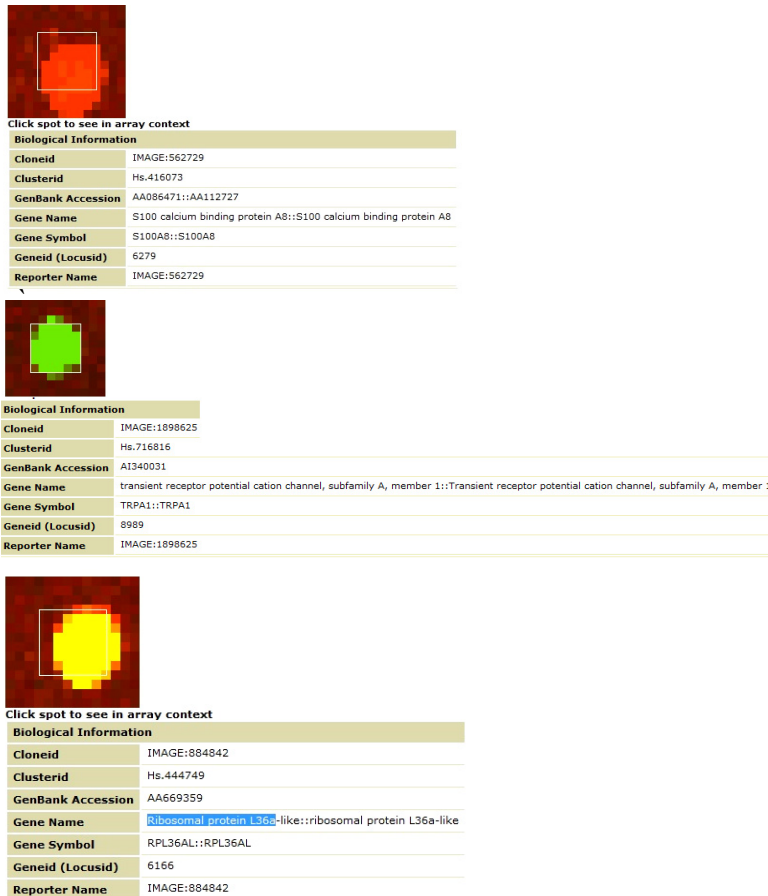
One basic purpose of a cDNA microarray experiment is to identify those genes which demonstrate a significant change in the expression level under the impact of certain experimental conditions, such as presence of cancerous tumors. The foreground and background intensity levels do not give meaningful information about the gene expression. In order to obtain estimate of gene expression, the data need to be further processed. In a two channel cDNA microarray experiment, in the composite RG image, red (R-Cy5 channel) is used to

indicate particular genes expressed in the experimental (abnormal) condition, while the green channel (G-Cy3) indicates genes expressed in control (normal) condition. For a given spot, true ratio ( $r$ ) of background subtracted mean intensity between two channels is calculated as:

$$r = \frac{R}{G} \quad (5.11)$$

For a red spot the ratio is greater than 1 indicating gene in the corresponding location is expressed more in abnormal tissue than normal tissue (up regulated), for green spot it is less than 1 (down regulated) and an ideal yellow spot the ratio is 1 (equally expressed) (Draghici, S. et al. (2003), Schena, M. (2003), Stekel, D. (2003)). The microarray data set used for the present study is available at the Stanford Microarray Database (SMD) <http://microarray-pubs.stanford.edu/eczema> and at the NCBI Gene Expression Omnibus (GEO) database <http://www.ncbi.nlm.nih.gov/geo/info/linking.html>.

For example, Fig 5.6a shows a spot from the array available at SMD public database with Exp-ID 68841 related to a study on gene expression profile associated with skin cancer. This red spot corresponds to the S100A8 protein coding gene, which exhibited a strong up-regulation in advanced stages of skin cancer in mouse and human. The ratio ( $r$ ) evaluated using as equation 5.11 as 6.1 indicating an up regulated gene. Similarly Fig 5.6 (b) shows a green spot representing down regulated gene in the same array, representing TRPA1 gene. TRPA1 is an ion channel gene located on the plasma membrane of many human and animal cells. This ion channel is best known as a sensor for environmental irritants, pain, cold and stretch. The ratio of intensity of this spot is 0.2462, indicating that the gene is expressed only in normal skin cell. Spot 3 is a yellow spot with ratio( $r$ ) 1.0036 representing Ribosomal protein L36a. Yellow spots indicate the genes that are not differentially expressed.



**Figure 5.6** Three Different spots from SMD (EXP -ID 8841) with their biological information



## 5.8 Performance Analysis of segmentation method using Monte-Carlo simulations

To analyze the robustness of the segmentation algorithm to noise, Monte-Carlo simulations have been conducted. Spots were artificially created with known ground truth. These spots are having known mean intensities for red and green channels as well as for background. The segmentation algorithm was applied to different spots. For a given spot true ratio ( $r$ ) of intensity between two channels is calculated. Next, the images were corrupted with additive Gaussian noise at ten different levels. For statistical significance, each noise level was repeated ten times. The ratio of the intensity of the spot in the two channels  $\hat{r}$  is estimated over ten repetitions at same noise level. Ten different noise levels are used in this study. Two parameters evaluated to measure the performance were (i) Mean square error (MSE) of ratio and classification error (CE). MSE between true ratio  $r$  and the ratio estimated at each noise level  $\hat{r}$  is calculated as:

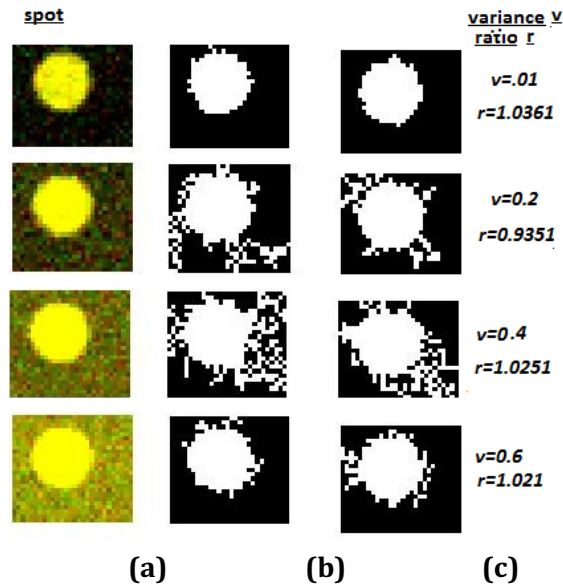
$$MSE = \frac{1}{10} \sum_{i=1}^{10} \left( \hat{r} - r_{true} \right)^2 \quad (5.12)$$

Classification error (CE) is defined as percentage of misclassified pixels within the spot after segmentation. Let  $\sigma_f^2$  represent the variance of signal intensity distribution of foreground and  $\sigma_b^2$  represents variance of pixels intensities of background signals then signal to noise ratio (SNR) is calculated as:

$$SNR = 10 \log_{10} \frac{\sigma_f^2}{\sigma_b^2} \text{ dB} \quad (5.13)$$

For a given spot the two parameters MSE and CE are calculated at different Signal to Noise Ratio by varying the noise levels. Figure (5.7) shows the segmented

foreground region when Gaussian noise at four different noise levels are added to an artificial spot.



**Figure 5.7** (a) Microarray spot after adding noise at different variance level (b) Red channel- foreground region (c) Green channel- foreground region after segmentation.

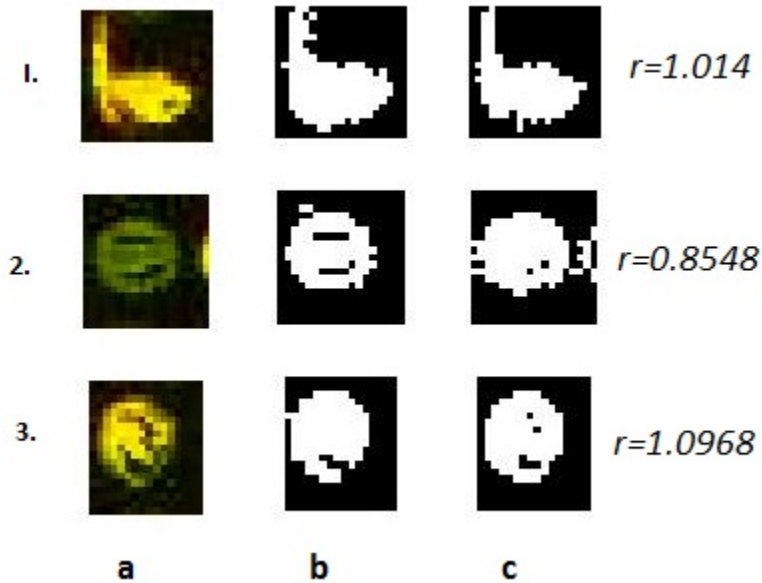
## 5.9 Implementation

The new segmentation algorithm has been implemented on different real microarray from SMD. Microarrays with different size, shapes and contaminations were used for this purpose. The segmentation method was applied each spot within the grid. The ratio estimated using AASRG method was compared with the conventional SRG used in MAGIC tool. The 16 bit uncompressed Tiff images were used for ratio estimation. The segmentation result was also compared with the existing segmentation methods such as fixed circle, adaptive circle and Seed region

growing. Gaussian noise at different variance levels were added to artificial spots and Monte-Carlo simulation were conducted to study the segmentation accuracy and classification error of the new segmentation method.

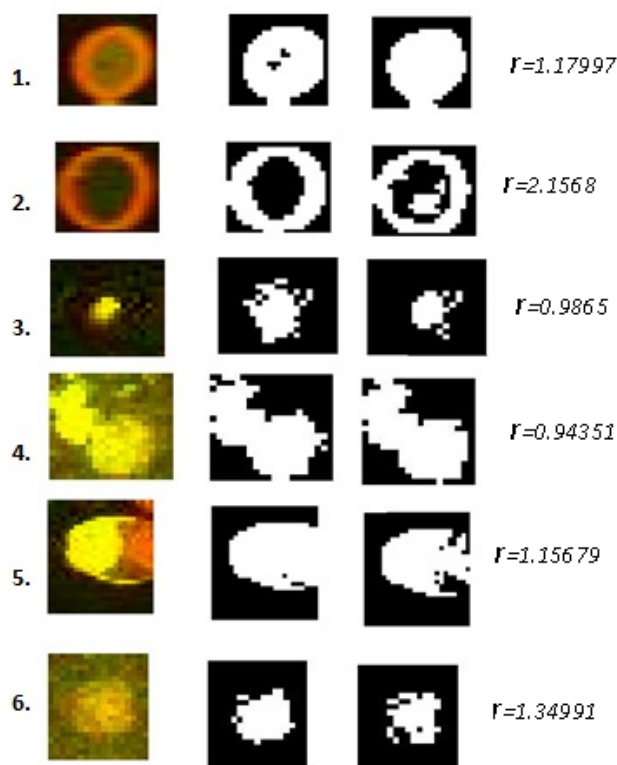
### 5.10 Result and Discussions

Real microarray spots with different intensity characteristics were used to study the segmentation accuracy of AASRG method. Figure.5.8 (a) shows spots with different defects. In Figure 5.8, (b) and (c), the region labeled with '1' is the segmented foreground region of red channel and green channel respectively. The average intensity of pixels in the corresponding red and green channels was calculated. Average back ground intensity for each channel was also calculated using the method explained in section 5.6. Using these intensity values the intensity ratio ( $r$ ) for each spot was evaluated. The results indicate that the method correctly identified the foreground regions of irregular shaped spots.



**Figure 5.8** Spots with irregular shape and the segmented foreground regions and ratio estimates (using AASRG method). (a) Spot (b) red channel segmented foreground region (c) Green channel foreground region

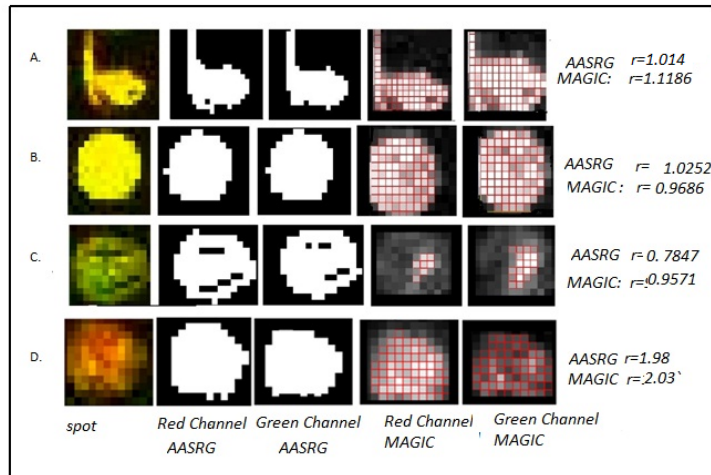
Here spot1 is a comet shaped yellow spot. The ratio estimated is 1.01 indicates that both channels expressed equally. Similarly the spot2 is a green spot having ratio estimate less than 1. Spot 3 is an yellow spot with contamination. The foreground regions are identified correctly and the intensity ratio is 1.09. Figure 5.9 shows six different spots with irregular shapes and size and the segmented foreground regions.



**Figure 5.9** Different spots with irregular shape and sizes-Segmented foreground regions and ratio estimates (using AASRG method) (a) spot (b) Red channel(c) Green channels.

Here spot 2 is a doughnut shaped red spot, the red and green channels are segmented and the intensity ratio evaluated is 2.156. Spot3 is an yellow spot with irregular shape. Foreground region is extracted and ratio is evaluated is 0.9865. The adaptive method for selecting seed and threshold values of AASRG increases the segmentation accuracy of irregular spots. Spot4 is a dilated yellow spot having area greater than the normal area. The quality control algorithm can be used to reject such spots from further analysis. Spot 6 is having large background intensity than the normal spot. AASRG method is capable of segmenting the foreground region of such spot and the intensity ratio estimated is 1.3499. Figure 5.10 shows a comparison of segmentation results with Microarray Genome Imaging and Clustering Tool (MAGIC Tool). MAGIC tool is widely used for academic purpose to analyze all types of gene expression data on all major operating systems (Windows, Mac OS X, Linux and Solaris). MAGIC provides semiautomatic gridding and interactive segmentation method. Segmentation can be performed with three algorithms: fixed circle, adaptive circle or seeded region growing. The Seeded Region Growing algorithm (Adams and Bischof, 1994) connects each pixel to a background or foreground region, continuing until all pixels are assigned. A user-specified threshold and geometric considerations (i.e. foreground near the center, background near the corners) determine which pixels are used to 'seed' the regions. Maximum intensity pixel within the grid is generally assigned as seed pixel. Ratio computation method can be selected as pixel average or total, with or without background subtraction. Figure 5.10 shows a comparison between conventional SRG based segmentation algorithm used in MAGIC and the AASRG method. Here the red squares indicate which pixels are used for foreground calculation. Pixels outside the red squares are considered for background calculation. Estimated ratio values using the two methods are also shown in Figure 5.10. The intensity ratio calculated using AASRG for spot A is 1.014, while using MAGIC is 1.118. For a low intensity spot (c), the MAGIC selects only bright pixels for foreground intensity calculation while AASRG selects more low intensity foreground pixels and hence gives a better intensity ratio value of 0.7847

for the green spot with respect to the value 0.9571 (value nearer to '1' i.e. for a yellow spot) obtained using MAGIC. Figure 5.11 (a) shows a spot with high background intensity. (b) and (c) are the foreground red and green channel regions separated using AASRG while (d) and (e) are the corresponding foreground regions using MAGIC (SRG method). In AASRG the seed values are selected using the circular mask and the low intensity foreground region is separated correctly, while using MAGIC the all the pixels within the grid is used for the calculation of intensity ratio.



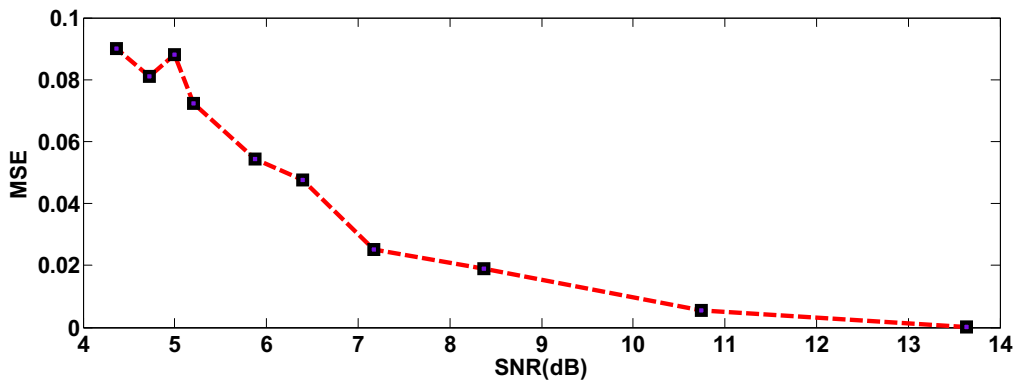
**Figure 5.10** Comparison of segmentation methods applied on different spots – AASRG method and conventional SRG method used in MAGIC tool.



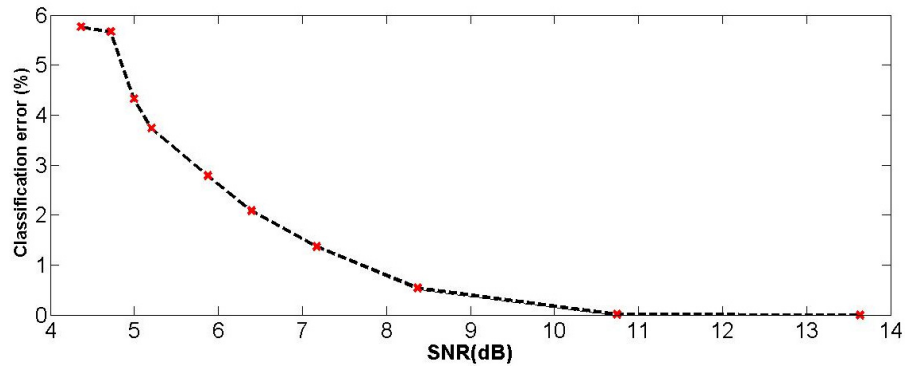
**Figure 5.11.** Segmentation results of a spot with high background intensity .(a)spot (b)Red channel Using AASRG method (c) Green channel using AASRG(d)Red channel Using MAGIC(e)Green channel using MAGIC.

### Monte Carlo Simulation Results

MSE and CE curves as functions of SNR are used to illustrate the performance of the segmentation algorithm. In both curves, the axis corresponds to the signal-to-noise ratio (SNR) calculated in decibel units. Figure 5.12 shows the results of MSE. It is clear that the MSE ratio is only 0.09 at 5dB SNR and very low at high SNR levels. Segmentation is used to group the pixels within each grid into foreground or background. Classification error is an indication of number of pixels that are correctly classified. Table 5.1 shows the average classification error (in percentage) when Gaussian noise at different levels were added to 10 different artificial spots with radius 6 pixels (with known foreground and background intensity levels). It is clear that the classification error is only 5.76% at variance level of 0.045. Figure 5.13 shows the classification error vs SNR plot.



**Figure 5.12** Mean Square error vs. SNR (dB)



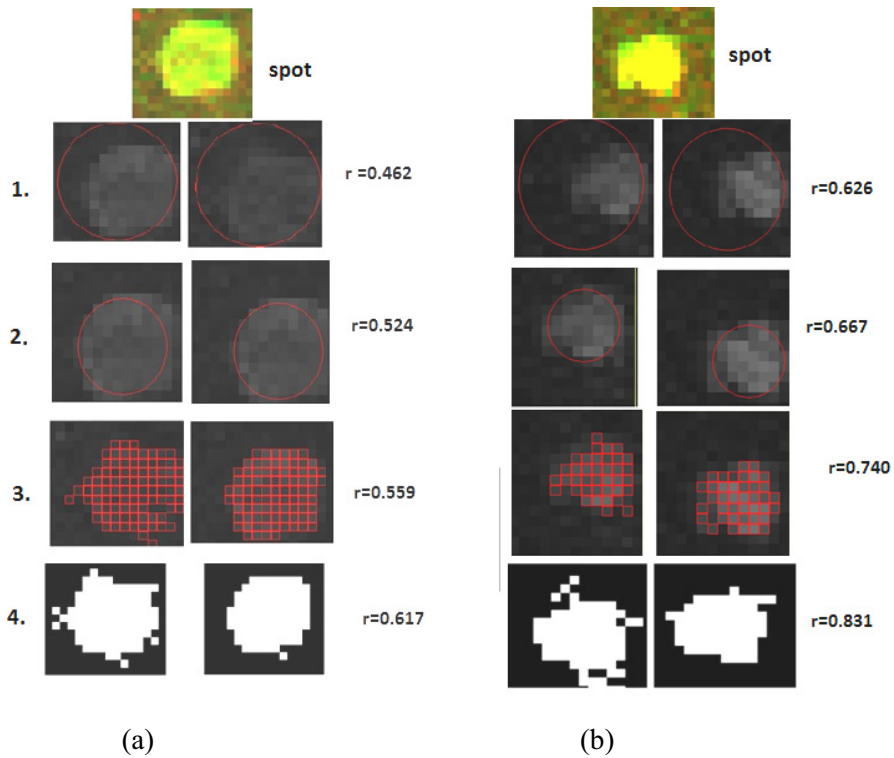
**Figure 5.13** Classification error vs. SNR

**Table 5.1** classification error at different noise levels

Gaussian noise Variance (V)	SNR(dB)	Average Classification error (%) (for spots with radius 6 pixels)
0.002	13.63991	0
0.005	10.74597	0.0143
0.01	8.374311	0.5391
0.015	7.168986	1.3716
0.02	6.392506	2.0874
0.025	5.875719	2.7919
0.03	5.200543	3.7422
0.035	4.995499	4.3307
0.04	4.720615	5.6716
.045	4.367898	5.7659

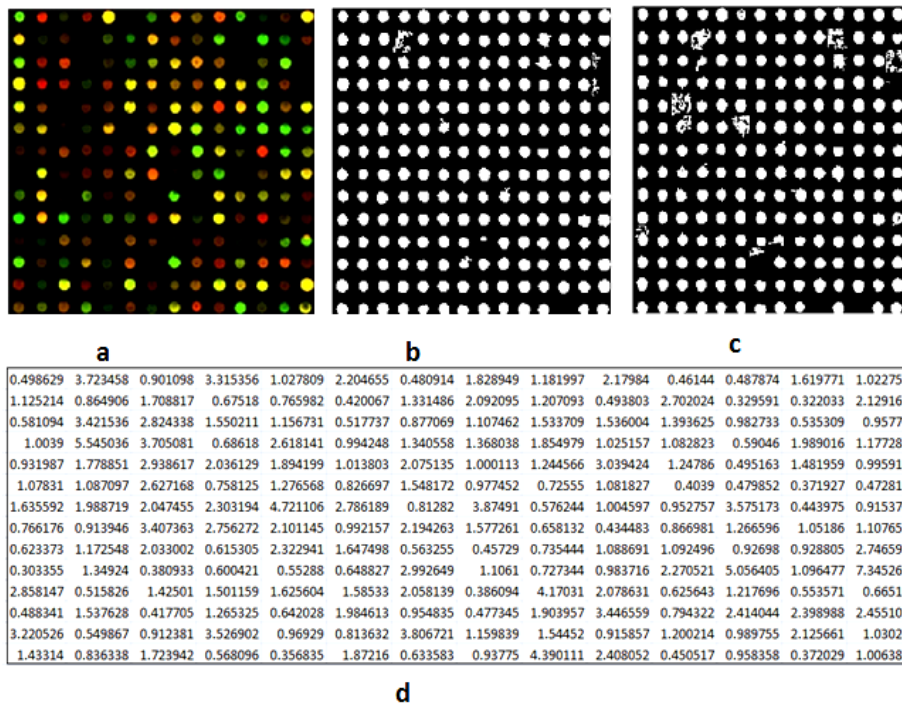


Figure 5.14 shows a comparison of existing segmentation techniques such as fixed circle, adaptive circle, SRG and AASRG. Fixed circle segmentation was implemented using a circle with radius of 6 pixels. Adaptive circle method uses two circles with minimum radius 3 pixels and maximum radius 8 pixels. SRG method is the conventional SRG method. Figure 5.14(a) shows a regular spots with radius 6 pixels. while (b) is an irregular shaped spot with a radius of 3 pixels.



**Figure 5.14** Different existing microarray segmentation methods applied to two spots and the corresponding true ratio( $r$ ). 1) Fixed circle 2) Adaptive circle 3) SRG 4) AASRG.

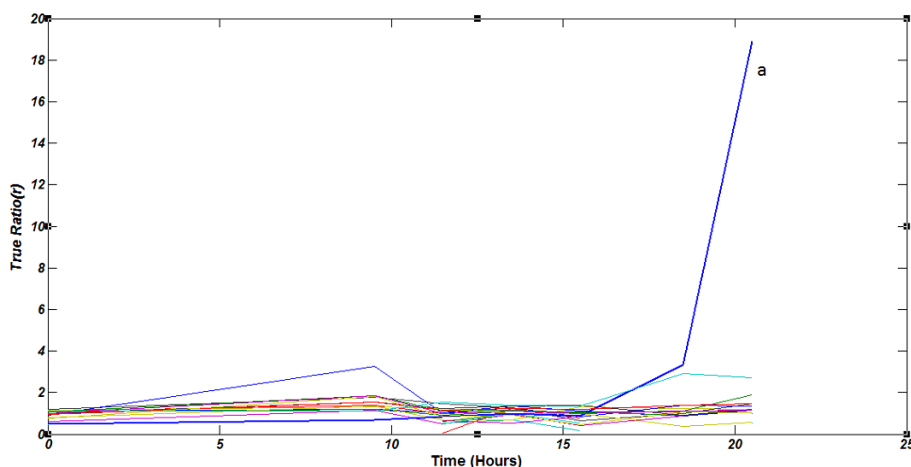
Results show that AASRG method give better segmentation accuracy for both regular and irregular shaped spots. Figure 5.15(a) shows a microarray subarray and the segmented foreground regions. (b) and (c) represents the segmented red and green channels. Fig 5.15(d) is the intensity ratio evaluated for all the 196 spots in the array using AASRG method.



**Figure 5.15.** Intensity ratio evaluated using AASRG method on a microarray subarray with 196 spots (a) subarray (b) Red channel segmented (c) Green channel segmented (d) Intensity Ratio ( $r$ ) between two channels.

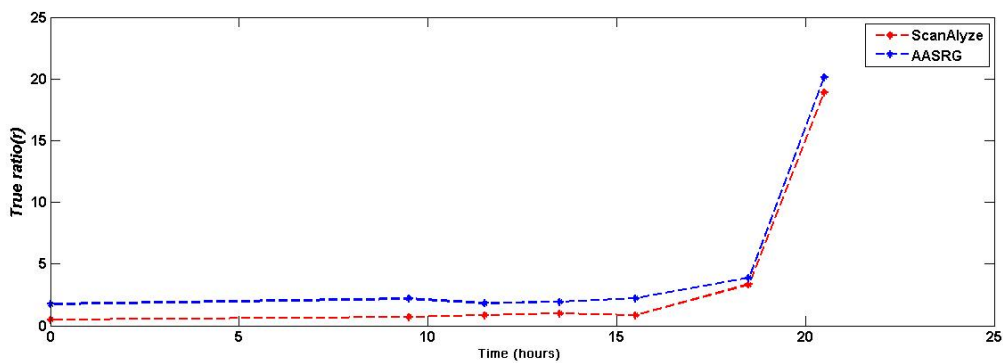
The developed segmentation algorithm has been applied to the microarray experiment on budding yeast -*Saccharomyces cerevisia* (DeRisi et al., 1997). The experiment and data are publically available at SMD database. The authors used DNA microarrays to study temporal gene expression of almost all genes (6400) in

*Saccharomyces cerevisiae* during the metabolic shift from fermentation to respiration. Expression levels were measured at seven time points during the diauxic shift. The full data set can be downloaded from the Gene Expression Omnibus website. This data can be loaded into MATLAB as yeastdata.mat which contains gene expression data from the ‘seven time steps’ in the experiment, the names of the genes, and an array of the times at which the expression levels were measured. ScanAlyze version 2.41 scanning software was used for the evaluation of these parameters. To study the performance of AASRG, true ratio was evaluated using the microarray images of the experiment from SMD. The true ratio obtained using AASRG was compared with data publically available. For example, Plot A shows the true intensity ratio ( $r$ ) of the two channels vs the time shift corresponds to first twenty spots in the array.



**Plot A**-Gene expression (true ratio) of first 15 genes from the *Saccharomyces cerevisiae* microarrays during the metabolic shift from fermentation to respiration.

Here only one gene (spot 15) representing the YAL054C (ACS1 Acetyl-coA synthetase isoform) appears to be strongly up regulated during the diauxic shift (plot -a). Plot B shows the comparison between true intensity ratio evaluated using AASRG and ScanAlyze for this gene alone. It is clear that the AASRG shows better ratio for low intensity spots corresponding to lower time steps.



**Plot B-** Comparison between true ratio evaluated using AASRG and ScanAlyze (SMD) for gene YAL054C (Spot-15) at seven time steps.

One of the disadvantages of SRG method is the computation time. Computation time required for implementing the segmentation algorithm was evaluated for various microarray images. The AASRG method was applied automatically to each grid using block processing method and the true intensity ratios were evaluated. Table 5.2 shows the execution time calculated for segmentation using AASRG method for four different subarrays with the system specification as given in section 4.5. The computation time was compared with that of existing software MAGIC 2.2 which uses SRG method. Microarray Images with different number of spots and contaminations were applied as the input and the computation time for segmentation alone was calculated and compared with the AASRG method. The comparison is done under identical conditions. It is found that the AASRG is approximately 10 times faster than the convention SRG method

applied in MAGIC software. For a microarray consist of 7300 spots the AASRG method requires 19.6seconds for the calculation of intensity ratio.

**Table 5.2** Comparison of Computation time using AASRG and SRG (MAGIC)

<b>Number of spots</b>	<b>Execution time (sec) using AASRG</b>	<b>Execution time (sec)using MAGIC</b>
100	1.4	16
552	2.1	19
729	3.2	34
1024	4.18	44

## 5.11 CONCLUSIONS

A novel method for automatic segmentation of microarray images using automatic adaptive region growing approach (AASRG) has been developed. When compared with conventional SRG method, the AASRG results in better segmentation accuracy especially for low intensity irregular shaped spots. Seed and the threshold values are selected automatically and adaptively for each spot, the method is found suitable for accurately segmenting doughnut shaped spots as well as spots with high background intensity. Block processing approach has been implemented which results in faster computation time. Monte-Carlo Simulation results show that the developed method is robust against additive noise, which is

common in microarray images. The local background region is extracted considering the global background characteristics of the image. This increases the accuracy of background calculation for high density microarrays. Since the seed points and threshold values were selected for each spot adaptively the method is independent of the shape and size of individual spots.

## CHAPTER 6

### Microarray Spot Intensity Quantification and Normalization

---

*The complexity of microarray experimentation process introduces systematic biases into the intensity measurements. Intensity quantification and normalization are major preprocessing steps to be done before the analysis of microarray data. Spot quantification combine the pixel intensity values into a unique quantitative measure that can be used to represent the expression level of a gene deposited in a given spot. The purpose of normalization is to remove the effect of any systematic source of variation introduced in the microarray experiment. This chapter presents spot intensity quantification techniques and common normalization techniques used in microarray based studies. Experimental results of Linear and Lowess regression based normalization techniques are explained. Various graphical tools used for the analysis of microarray data are also described in this chapter.*

---

## 6.1 Introduction

Microarray technology has resulted in the generation of large amount complex data sets. Quantifying the intensity information and data analysis has become one of the major bottlenecks in the utilization of array technology. The amounts of data obtained from each experiment require powerful computing techniques for identifying clusters of genes and interpreting overall patterns of expression. At present it seems that the greatest challenges in microarray research are not the arrays itself but the way by in which the resulting data matrices are handled and analyzed (Fadiel, A. et al., 2003). Before analyzing the data a number of preprocessing steps and normalization are commonly taken to ensure that the data is of high quality and is suitable for analysis. Typical example of preprocessing is log transformation of intensity values. Normalization techniques are collection of methods that are used to resolve systematic error and bias introduced in microarray experiments. This chapter explains the intensity quantification and normalization procedure currently used. Several selected techniques used for explorative and confirmative visualization of microarray data are also explained in the following sections.

## 6.2 Log transform

Most microarray experiments investigate relationships between related biological samples based on patterns of expression, and the simplest approach looks for genes that are differentially expressed. The final goal of the image processing is to compute a unique value that is directly proportional with quantity of mRNA present in the solution that hybridized the chip. For each gene deposited on the slide, this unique value is required. Such unique value can be obtained by quantifying the intensity information from each spot within the array. Typically, spots are quantified by taking the mean and median of the intensity of both



foreground and background regions as explained in section 5.7. The microarray data generated by the segmentation algorithm is typically in the form of text files consist of intensity ratio of every spot in the slide. It is usual to transform the microarray raw intensity data to logarithmic scale.

The logarithmic transformations have been used for preprocessing the raw intensities to provide values that are more easily interpretable and more meaningful from a biological point of view. The intensities in a microarray experiment span the full 16 bit range, from 0 to 65,535 units, with majority data in the lower range of values. If the data is not transformed, they must by necessity be presented in a very compressed form in the low range. Taking the log spread the values more evenly across the range and provides a more visually appealing data. Another reason for log transformation is that log transformation makes the distribution symmetrical and almost normal (T.P.Speed, 2000). Third, the random variation (as measured by standard deviation of intensities) typically increases approximately linearly with the average signal strength. Taking the log tends to make variability more constant. Fourth, the ratio of raw Cy5 and Cy3 intensities which is the final parameter evaluated is transformed into the difference between the log of intensities of Cy5 and Cy3 channels. i.e.,  $\log_2(R/G)$  is the difference  $\log_2(R) - \log_2(G)$ . It is usual in microarray data analysis to use logarithm to base 2. Log (base 2) ratio of 1, 2, 3 corresponds of 2, 4, 8 fold changes respectively. For example in a cDNA microarray experiment, if a specific gene at location A is expressed two times more in tumor sample than in normal sample (i.e,2 fold up regulated) then a log ratio of +1 is obtained. Similarly a gene expressed two times more in normal sample compared to tumor sample (2 fold down regulated) corresponds to a log ratio of -1. Genes that are not differentially expressed (i.e,expressed equally in both samples) corresponds to a log ratio of 0.

Log differential ratio (expression ratio) is usually denoted by M and is defined as

$$M = \log_2 \frac{R}{G} \quad (6.1)$$

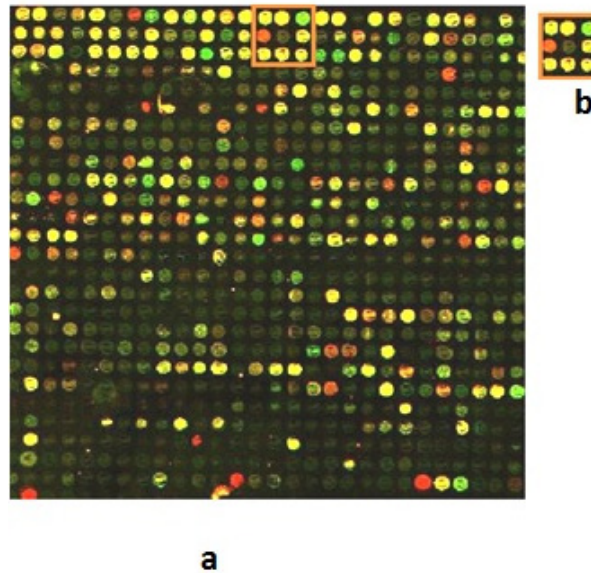
Letter M is a mnemonic of *minus* as,

$$M = \log_2 R - \log_2 G \quad (6.2)$$

Where, R and G are the background subtracted mean intensity for each spot in the red layer and green layer respectively as mentioned in section 5.7. Table 6.1 shows the log ratios for a range of fold differences. Gene at location (2, 3) has an expression value of 0.1241 indicating a low intensity signal and the gene is not expressed.

**Table 6.1** Conversion from fold ratios to log (Base 2) ratios.

Fold ratio	Log ratio (base2)
4- fold down regulated	-2
3- fold down regulated	-1.58
2- fold down regulated	-1
1.5 fold down regulated	-0.58
No change	0
1.5 fold up regulated	0.58
2- fold up regulated	1
3- fold up regulated	1.58
4- fold up regulated	2



**Figure 6.1** *cDNA Microarray image with 756 spots (b) A Sub image with 6 spots.*

**Table 6.2** Expression ratio ( $\log_2$  ratio) of spots in (b)

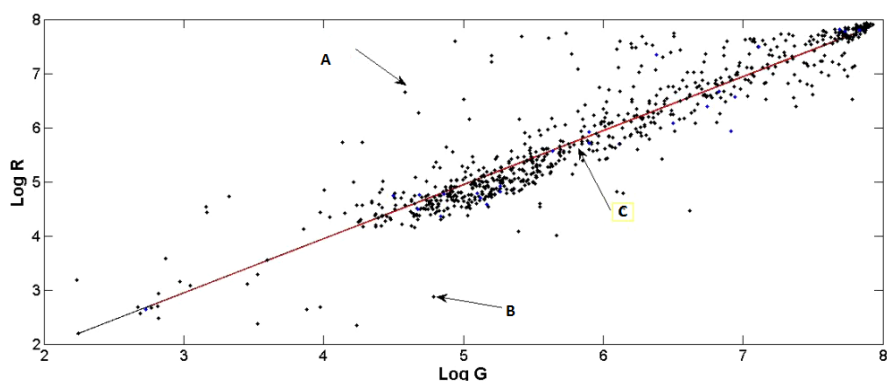
Expression ratio( $\log_2(R/G)$ )			
	column1	Column2	Column 3
Row 1	<b>-0.0663</b>	<b>0.0464</b>	<b>-0.659</b>
Row 2	<b>1.2408</b>	<b>0.1241</b>	<b>0.0267</b>
Row 3	<b>0.0974</b>	<b>0.0113</b>	<b>0.0478</b>

## 6.3 Visualization of Microarray data using various Representations

Visualization can be an effective tool for summarizing and interpreting data sets, describing their contents and revealing significant features. Different graphical tools are used to analyze the results of microarray experiments. Such tools can assist in deciding whether the experiment was successful. Scatter plot are the most common graphical tool used to analyze expression data of cDNA microarray experiments.

### 6.3.1 Scatter Plots: Diagonal and MA Plots

A simplest scatter plot is the diagonal plot in which  $\log_2$  of background corrected intensity value of one channel (Cy3) is plotted against the other channel (Cy5) i.e.  $\log_2 R$  versus  $\log_2 G$ . Suppose a gene 'g<sub>i</sub>' has an expression level (intensity level)  $x_i$  in Cy3 channel and  $y_i$  in the Cy5 channel, the point representing  $g_i$  will be plotted at coordinates  $(\log x_i, \log y_i)$  in the scatter plot. In this way all the genes in the data set can be plotted on a graph. In majority of the experiments involving living organisms, most genes are expected to be expressed equally in normal and abnormal tissues. So these genes that are expressed equally (yellow spot) will lie on the diagonal line  $y = x$  in the scatter plot. Genes that are expressed differentially will lie away from the diagonal line. For example, a gene expressed more in abnormal tissue (corresponding to red spot) will lie above the diagonal line and a gene expressed more in normal tissue (green spot) lie below the diagonal line. Points located at a higher distance from the diagonal represent genes that are more differentially expressed. Fig.6.2 shows the scatter plot with 756 points corresponds to 756 spots in the microarray image shown in Figure 6.1(a).



**Figure 6.2.** Scatter plot of log intensities

In this plot gene A corresponds to an up regulated gene. Gene B is down regulated and gene C is not differentially expressed. Genes with two fold change will be at a distance of at least 1 from the diagonal line. In general for given threshold  $T$ , the fold change method reduces to drawing lines parallel to diagonal at a distance  $\pm \log_2(T)$  and selects the genes outside the lines. Majority of the genes are expected to fall in the vicinity of the line  $y=x$ . If the plot does not exhibit these characteristics, then the data from one channel are consistently lower or higher than the data from the other channel, suggesting the need for preprocessing and normalization (Zhang A.2006). Although such a plot is straightforward, the very high correlation between the two channel intensities always dominates the plot making the more interesting features of the plot difficult to discern. Scatter plot visualizes data on a linear scale. Human eye and brain are better in preserving deviations from horizontal and vertical lines than diagonal lines (D.Stekel, 2003). Since the interest lies in deviations of the points from the diagonal line, it is beneficial to rotate the plot by 45 degrees and to rescale the axes as in the MA-plots (Dugout et al., 2002).

### MA Plots

These are an alternative type of scatter plots in which y axis is the log ratio  $M$  (defined in eqn. 6.1) and x axis is the average log intensity ( $A$ ) defined as :

$$A = \frac{\log_2(R) + \log_2(G)}{2} \quad (6.3)$$

$A$  is the mnemonic for *add*. The MA-plot serves to increase the room available to represent the range of differential expression and makes it easier to see non-linear relationships between the log intensities it also displays the important relationship between differential expression and intensity, which is used in later analysis steps. It is also convenient to use base 2 logarithms for both  $M$  and  $A$  so that  $M$  is the units of two fold change and  $A$  is the unit of two fold increase in brightness. In MA plots genes with a particular fold change can be selected using horizontal line. Fig.6.3 shows the MA plot for the microarray image in Figure 6.1(a). In microarray experiments most of the assumption is that majority of the genes are not differentially expressed, which means majority of the genes are expected to fall in the vicinity of the line  $y=0$  as seen in Figure 6.3.

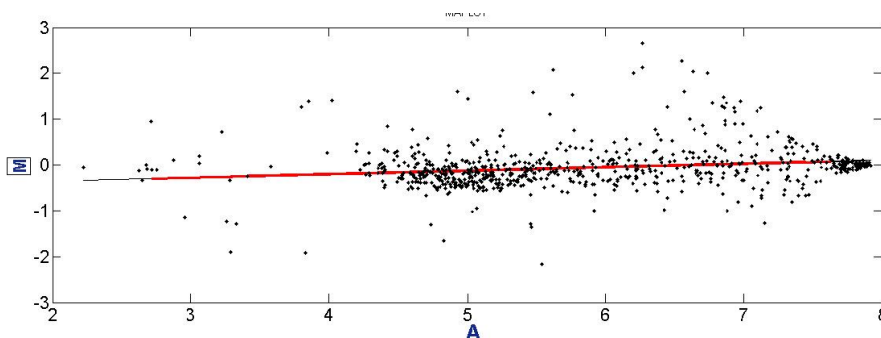
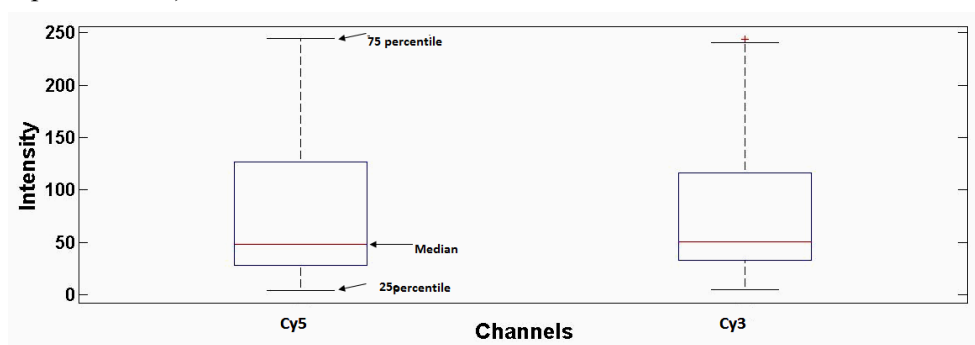


Figure 6.3 MA plot

### 6.3.2 Box plots

Box plots are simple graphical representations of several descriptive statistics, such as mean (or median) and variance for a given data set. In general, vertical axis of the box plot is formed by a response variable, while the horizontal axis corresponds to the factor of interest. Figure. 6.4 show an example of a box plot used for analysis of microarray data. Vertical axis represents intensity levels (8 bit tiff image) while the horizontal axis contain the of two channels (8 bit representation).



**Figure 6.4** Box plot for intensity levels of two channels

Box plot is characterized by a central box and two tails. The center line (red) in the box indicates the position of the median value of the data set. The upper and lower boundaries of the box represent the locations of the upper quartile and lower quartile which are 75<sup>th</sup> and 25<sup>th</sup> percentiles of the intensity values respectively. Thus, the box will represent the interval that contains the central 50 percent of the data. The box plot is an important tool for determining whether a factor has a significant effect on the response with respect to the location or variation. For example, juxtaposing the distributions of measurements from Cy3 and Cy5 channels in the box plot may reveal whether the measurement of Cy5 and Cy3 channels in the box plot may reveal whether the measurement of Cy3 is

systematically higher or lower than that of Cy5 so that a normalization method is required to correct the bias.

## **6.4 Normalization Techniques**

The purpose of data normalization is to minimize the effects caused by technical variations and, as a result, allow the data to be comparable in order to find actual biological changes. Replicate slides, hybridized with RNA from the same extractions are used to reduce the variability. Similarly the spots are replicated within the array by duplicating the DNA sequence by printing them in adjacent location within the array. The precision of the particular measurement will increase if the spot intensities of replicate spots are averaged. The normalization methods are applied within the array as well as between arrays. Several normalization approaches have been proposed, most of which derive from studies using two-color spotted microarrays. Some authors proposed normalization of the hybridization intensity ratios; others use global, linear methods, while others use local, non-linear methods.

### **6.4.1 Within Array Normalization**

In two colour microarray experiments, the two samples (test and reference) have been labeled with two different fluorescent dyes in two separate chemical reactions and the intensity is measured with two different lasers operating at two wavelengths. In addition the features on the array are distributed on different part of the surface of the array. One of the common problems is response of the Cy3 and Cy5 channels may vary at different intensities, and red intensities usually tend to be lower than the green intensities. This can be corrected by applying the following methods:

1. Linear regression of Cy5 against Cy3.
2. Linear regression of log ratio against average intensity.
3. Non-linear (Lowes) regression of log ratio against average intensity.



All these methods use the assumption that the majority of the genes on the microarray are not differentially expressed.

#### **6.4.1.1. Linear Regression of Cy5 against Cy3**

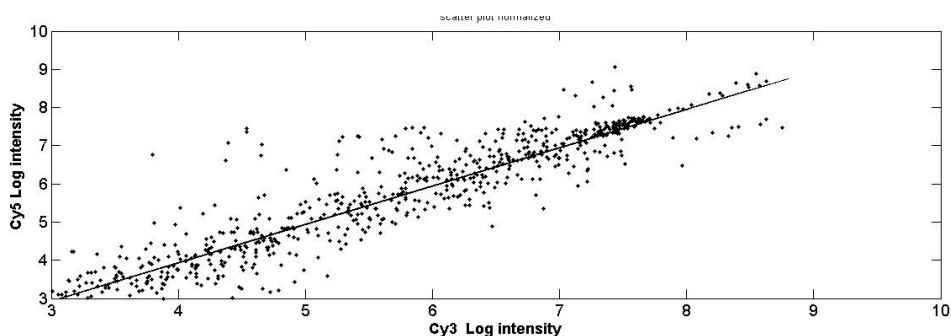
This is the simplest method used to check whether the Cy3 and Cy5 channels are behaving in a comparable manner via scatter plot of the two channels. If the Cy3 and Cy5 channels are behaving similarly, then the clouds of points on the scatter plot is approximately a straight line and the linear regression line through the data should have gradient of 1 and intercept of 0. Variations from these values represent different response of Cy5 and Cy3 channels. They are:

1. A nonzero intercept represent one of the channels being consistently brighter than other.
2. A slope not equal to 1 represents one channel responding more strongly at higher intensity than other.
3. Deviations from a straight line represent nonlinearities in the intensity response of the two channels.

The steps involved in evaluating the linear regression plots are:

- (i). Plot Cy3 vs. Cy5 scatter plot.
- (ii). Fit a regression line through the scatter plot and identify the gradient and intercept.
- (iii) Replace Cy3 values with the fitted value on the regression line.

This method of normalization works well for the data where the linear fit is good and reasonable. One of the main disadvantages of this type of plot is the difficulty to see the nonlinearities in the data. Figure.6.5 shows the scatter plot of the normalized log intensities of the microarray image in Figure.6.1.



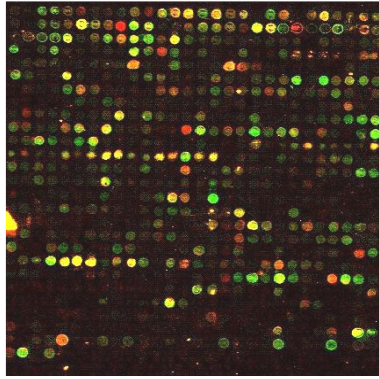
**Figure.6.5** Scatter plot of the normalized log intensities of the microarray image in Figure.6.1.

#### 6.4.1.2 Linear Regression of Log Ratio Against Average Intensity

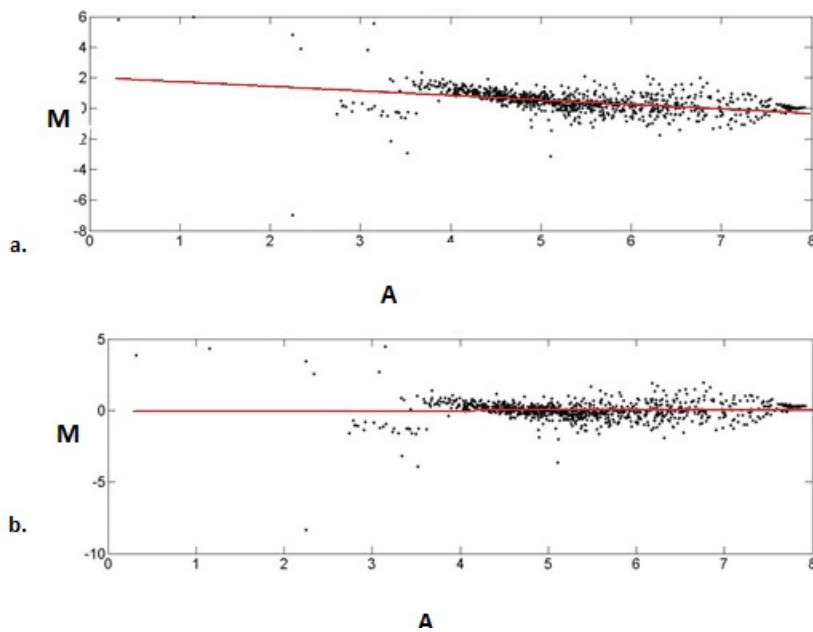
A common method for evaluating how well normalized an array is, to plot an MA plot of the data. If the data is not symmetrical about the horizontal line, it implies that one channel is responding more strongly than the other and normalization is required. Intensity dependent variations can be corrected by generating a best fit curve through the middle of the MA plot, and this becomes the new zero line for the vertical axes. In this analysis, the M value of the pair (A, M) is shifted by a quantity  $c(A)$  depending on the value of A as:

$$M = \log_2 \frac{R}{G} - c(A) \quad (6.4)$$

Figure 6.6 shows a microarray image and its MA plot is shown in Figure 6.7 (a). The plot shows that the average trend of the log ratio as a function of intensity. It is clear from the plot that at low intensities Cy5 channel is responding more strongly and at high intensity Cy3 channel is responding strongly. Plot (b) is the MA plot after applying the normalization.



**Figure 6.6** A Microarray image

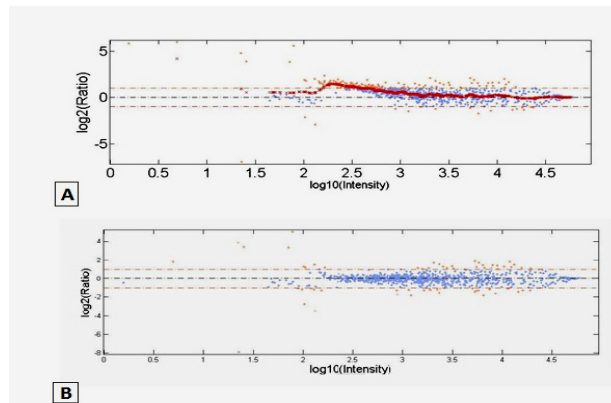


**Figure 6.7** (a) MA plot -linear regression fit through the data (b) MA plot after linear normalization

### 6.4.1.3 Lowess Regression of Log Ratio Against Average Intensity

Some microarray expression levels may have large dynamic range that will cause scanner systematic deviations such as nonlinear response at lower intensity range and saturation at higher intensity. Although data falling into these ranges are commonly discarded from further analysis, the transition range, without proper handling, may still cause some significant error in differential expression gene detection. To account for this deviation, locally weighted linear regression (Lowess) is regularly employed as a normalization method for such intensity dependent effects (Yang.Y.H. et al., 2002 G.C. Tseng et al., 2001). Lowess detects systematic deviations in the M-A plot and corrects them by carrying out a local weighted linear regression as a function of the  $\log_{10}$  (intensity) and subtracting the calculated best-fit average  $\log_2$  (ratio) from the experimentally observed ratio for each data point.

Figure 6.8(A) shows the MA plot and the nonlinear Lowess fit through the microarray expression data of Figure 6.6. Figure 6.8. (B) is the data after Lowess normalization. Here x axis is plotted in logartimic scale with base 10.



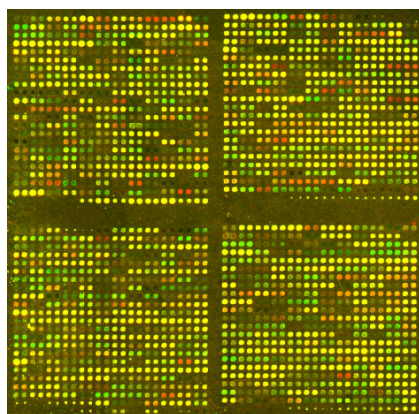
**Figure 6.8** (A) MA plot and the Lowess fit through the microarray data. (B) Plot after Lowess normalization

## 6.5 Implementation

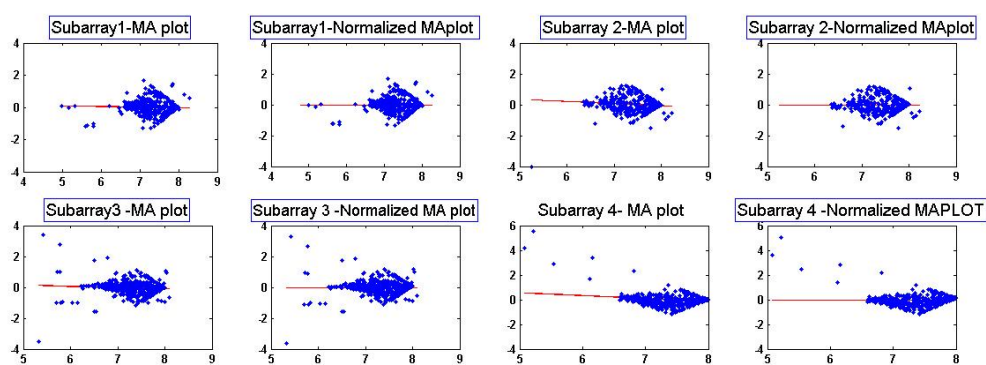
Microarray image quantifications and normalization have been tested on various microarray images. The newly developed fully automatic gridding and segmentation techniques were initially implemented and spot intensity ratios were estimated. These ratios are log transformed and Scatter plots and MA plots were used to visually analyze the expression data. Yeast genome microarray slide provided by MAGIC was also analyzed. Different slides from SMD with spatial bias were also used for normalization. The normalization techniques were applied to eliminate spatial bias within the array.

## 6.6 Experimental Results

cDNA microarray image corresponding to yeast genome provided by MAGIC is shown in Figure 6.9. The microarray consists of 4 subarrays. Each subarray consists of 552 spots in  $23 \times 24$  format. Gene expression values obtained after applying AASRG based segmentation are quantified and normalization techniques were applied. Figure 6.10 shows the MA plots before and after applying linear regression based normalization.

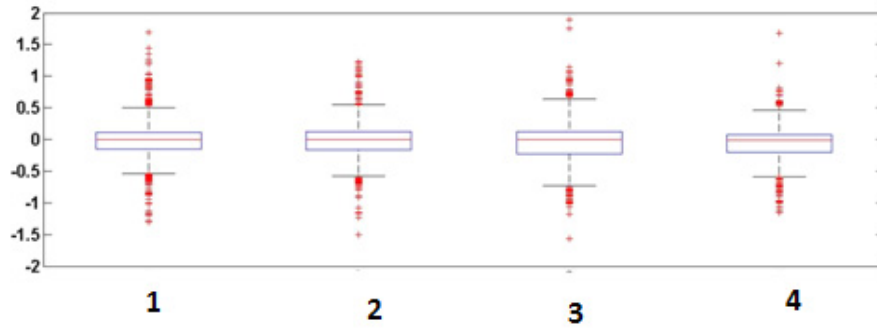


**Figure 6. 9** A Microarray Image (Yeast genome)

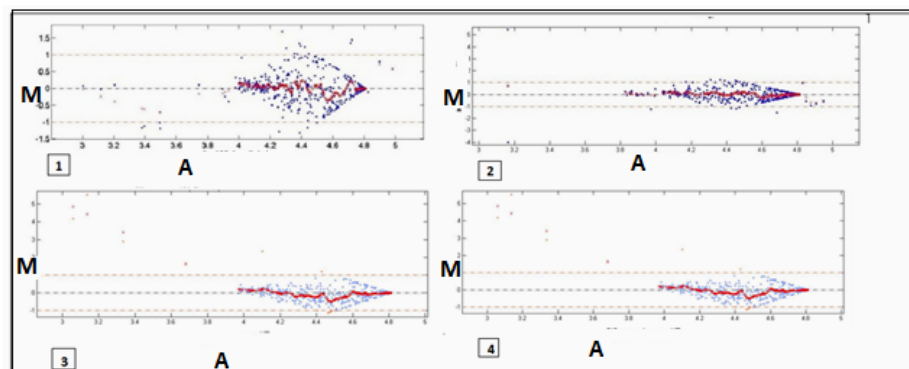


**Figure 6.10** MA plot for four subarrays before and after linear regression based normalization.

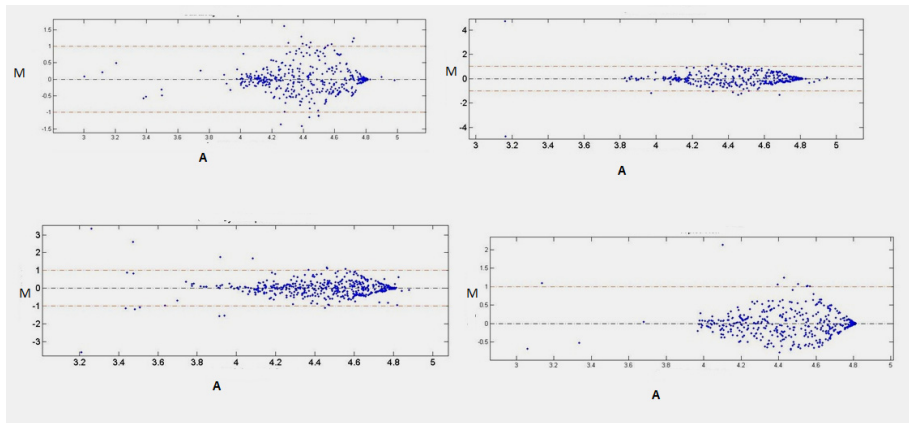
The straight line that has been fitted through the data points demonstrates the trend in Cy5 and Cy3 responses. From the plots of subarray 2 and 4 it is clear that at low intensities Cy5 channel respond more strongly and at high intensity Cy3 channel is responding strongly. This effect is due to the experimental artifacts and should be removed. The data has been normalized by applying linear regression which transforms the horizontal line through zero. Points with the highest intensities lie above the line .Figure 6.11 shows the box plot displaying the log ratio distribution in the four subarrays. To normalize the spatial biasing effect, lowess regression method was applied to each subarray. Figure 6.12 shows the lowess regression fit applied through the data. Figure 6.13 shows the MA plot after applying the normalization method. Figure 6.14 shows the MA plot for the whole data in the array before and after applying linear normalization. It is clear from Figure 6.14 (a) that the at low intensities Cy5 channel is responding more strongly while at high intensities Cy3 channel is more responding and the fitted line is not horizontal . The log ratio is normalized by subtracting the fitted value on the straight line from each log ratio. The MA plot obtained for the whole array using MAGIC tool is shown in Figure 6.14(c).



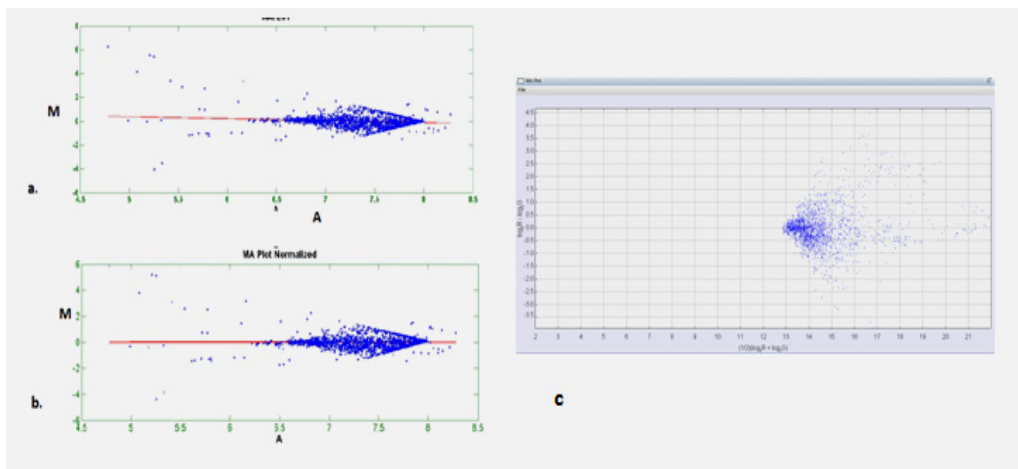
**Figure 6.11** Box plots of  $\log_2$  ratios



**Figure 6.12** Lowess regression (curve in red colour) on  $\log_2$  ratio



**Figure 6.13** MA plot for Normalized data of four subarrays

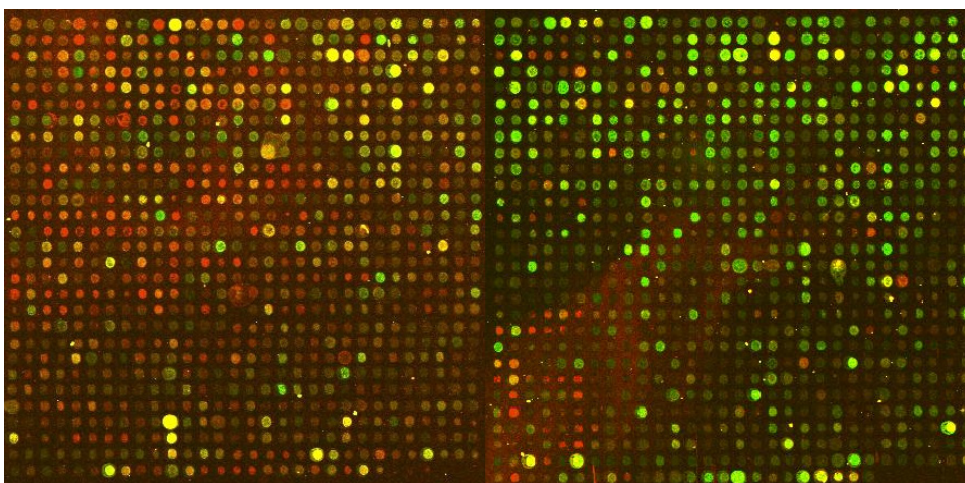


**Figure 6.14** (a) MA plot for the whole array (Fig.6.9) (b) MA plot after linear normalization (c) MA plot for the whole array using MAGIC tool.

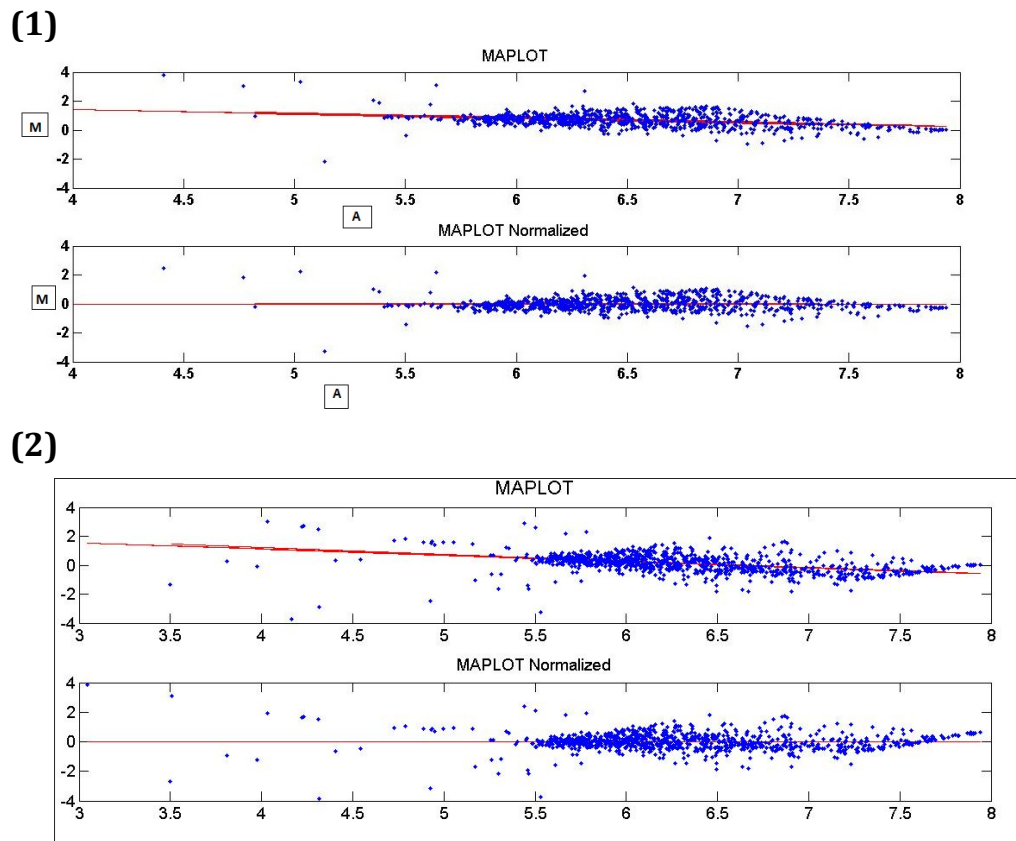


### Correcting for Spatial Bias

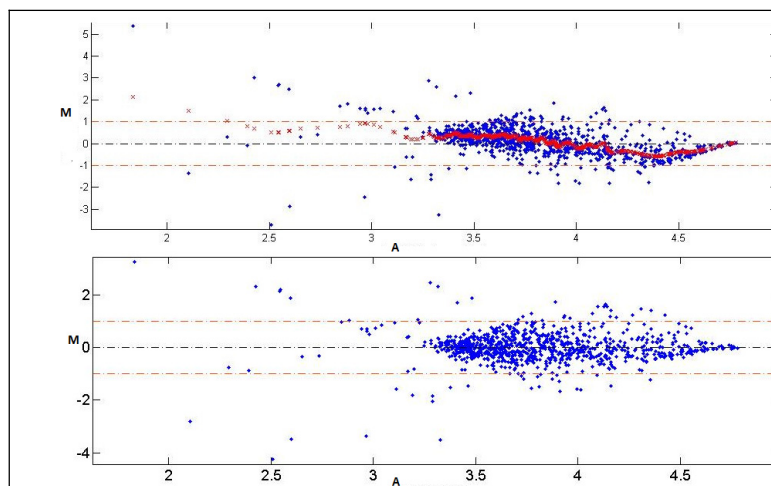
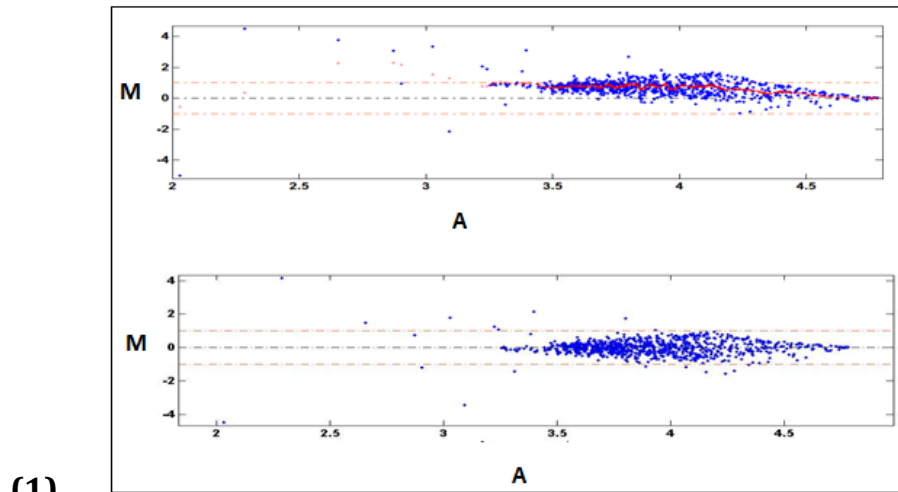
In some microarray experiments there is a spatial bias of the two channels i.e, in some region of the array the Cy3 channel is brighter and the other region Cy5 channel is brighter. This can result from the array not being completely flat or horizontal in the scanner. This can affect the log ratios with some regions of the array having positive log ratio and other regions have negative log ratio. The block by block normalization techniques can be used to eliminate the effects of spatial bias. Figure 6.15 shows two subarrays of a microarray image from SMD data base. The spatial bias is clear in the images. Figure 6.16 and Figure 6.17 shows the MA plots of each subarray. The shape of the plot indicates the spatial biasing effect. Both linear and lowess regression methods were applied to eliminate the spatial biasing effect are shown in Figure 6.16 and Figure 6.17.



**Figure 6.15** *Two subarrays with different spatial bias*



**Figure 6.16** MA plot before and after applying linear regression (1) first subarray (2) second subarray of Fig.6.15



**Figure 6.17** (1) MA plot data before and after applying *lowess* regression (1) first subarray (2) second subarray

Lowess transformation divides the data into a number of non overlapping intervals and fit the function. This method works well for data where there is a non linear relation between the response of Cy3 and Cy5 channels.

## **6.7 Conclusions**

In this chapter microarray quantification methods, normalization techniques and the various graphical tools used for visual inspection of data are presented. Box plots, Scatter plots and MA plots are some of the common visualization tools. These techniques provide an indication of the quality of the slides and the experiments and help to identify systematic variations. Different normalization techniques are used to remove such systematic effects so that the real biological difference can be easily distinguished. Within the array normalization is used to remove the dye bias and spatial bias. Normalization techniques using linear regression method and Lowess regression method were implemented to remove this bias within arrays.

## CHAPTER 7

### Spot Quality Evaluation

---

*Success of microarray technology is characterized by accuracy of data measurements at various steps. Identifying and removing unreliable data is crucial to prevent the risk of receiving deceptive analysis results. Various quality measures are used during image processing steps to evaluate the quality of individual spots so that bad spots can be excluded from further analysis. In this chapter a set of quality measures are defined for testing the quality of the spot. Quality values assigned to each spot are based on its intensity characteristics and spatial information. A composite quality score is then assigned to each spot based on individual quality values to give an overall assessment of spot quality. The quality control method has been tested on various microarray spots and Receiver operating characteristics was used to study its performance.*

---

## 7.1 Introduction

Spot quality control is one of the essential steps in automated microarray image analysis systems. To determine spot quality, a clear definition of a good spot, or a list of all possible distortions that may affect the quality of the spot is required. The diversity of instrumental platforms and biological factors that may influence the quality of the spot makes the formalization a difficult task. Several attempts have been made to approach this problem. Generally a number of quality measures characterizing the spot, such as signal-to-noise ratio, size and circularity are used. In this chapter six quality measures are used to define a composite quality score for each spot. Composite quality scores give an overall assessment of the spot quality. Spots with undesired quality are discarded from subsequent data analysis.

## 7.2 Literature Review

Many of the early quality analyses were based on manual editing combined with image analysis diagnosis using semi-automatic software. Since manual editing is a tedious task, now most of the commercial systems for microarray image analysis, such as GenePix, ImaGene and MAGIC have developed features for quality measures as flags. As these flags usually focus on the correctness of the analysis system rather than the quality of images, they cannot represent the quality of microarray images. Gabriel et al. (1999) defined a simple measure for noises, to find the regularity of a microarray image in the auto gridding process, which was the first attempt to evaluate noise levels for microarray images. Kuklin et al. (2001) developed separate quality measures for each parameter such as signal to noise ratio, the roundness of spots etc. instead of a single quality value for an image. Brown *et al.* (2001) showed that features of images vary as intensities of spots increase and suggest the spot ratio variability as a simple measure of irregularity of the spot. The composite quality score introduced by Wang et al., (2001) provides a very comprehensive quality assessment of microarray data. Wang et al., (2003) & Tran et al., (2002) suggested a correlation between mean and

median of pixel intensities as a measure for spot quality. Hautaniemi et al.(2003) developed Bayesian network based spot quality control shows what factors have an effect on the qualities of microarray images. In this method a training data set is required, which may times be difficult to obtain.

A quality measure model was suggested by Kim et al. (2005) where, five functions are defined for quality measures such as signal noise, background noise, scale invariant, spot regularity, and spot alignment. Novikov. E et al. (2005) developed an algorithm for automatic evaluation of spot quality of microarray images and assigns a quality score to each ratio estimate. This score is calculated from 10 main quality characteristics reflecting different spot properties within the microarray. The quality values assigned to each spots used to eliminate spot or to weight contribution of each ratio estimate. The microarray quality control project-phase I (MAQC, 2006) provide a quality control (QC) tool to the microarray community to avoid procedural failures, variation between different platforms, as well as variation between subarrays. An implementation of semi automated multivariate quality control assessment for cDNA microarray was suggested by Bylesjo et al. (2005).

Two quality estimates considered by Tang et.al (2007) were the local similarity between the 'red' and the 'green' images, and the homogeneity of the spot and suggested that images should be registered before the analysis so as to reduce the variability. To avoid problems in quality analysis pre-censoring has been conducted to remove all poorly expressed replicates using  $t$  values. Bergemann, T. L. (2010) described two signal quality estimates that capture the reliability of each spot printed on a microarray. A parametric estimate of within-spot variance that assumes pixels follow a normal distribution and a non-parametric estimate of error, called the mean square prediction error (MSPE), assumes that spots of high quality possess pixels that are similar to their neighbours. Wu J. (2012) proposed a multi variant method of quality control that was needed to assess classification errors associated with simple model of spot

quality. The following section describes various factors that affect the quality of the spot.

### 7.3. Factors Affecting the Spot Quality

The Number technical issues during the microarray experiment affects of the quality of the spots in microarray images. They are broadly classified as:

- **Low intensity spots:** This is the common problem that the signal is only few fold above the background. The main causes for this situation are the low expression levels, surface properties of the slide, poor labeling, incomplete or irregular hybridization conditions, scanner defect etc.
- **Non uniform Intensity distribution:** The variation in intensity distribution is the consequence of particle contamination, non specific binding, irregular distribution of printed materials on the slide, printing defects etc. results in regions of pixels containing signals deviates from average signal.
- **Morphological issues:** This refers to unexpected shape variations of the spot foreground region. This includes large variation in spot size from the average spot size may be due to poor pin design, impurities in the printing solution, washing problems, impurities etc.
- **Background Fluorescence:** The intensity fluctuations in the local background region of a spot compared to the global background of the slide typically, results from dye contaminants due to non-specific binding or incomplete washing, drying during the washing etc.

### 7.4 Quality measures

Many image analysis programs collect a number of quantitative measurements associated with each spot. These includes morphological measures such as area, perimeter uniformity measures such as standard deviation of



foreground and background intensities in each channel and spot brightness measures. Sometimes some quality measures are derived from the basic measures. Shape (area/perimeter) and signal to noise ratio measures are examples. The following sections describe a set of quality parameters, characterizing different features of the spot. These parameters are scaled between 0 (bad spot) and 1 (good spot) to facilitate further quality analysis.

#### 7.4.1 Signal Level (S)

As mentioned in section 5.7 intensity level for each channel is computed as  $R$  (Cy5) and  $G$  (Cy3) respectively as per equations given below

$$R = \mu_{fr} - \mu_{br} \quad (7.1)$$

$$G = \mu_{fg} - \mu_{bg} \quad (7.2)$$

Where  $\mu_{fr}$  and  $\mu_{fg}$  are the is mean intensity of the pixels in the foreground region of each channels and  $\mu_{br}$ ,  $\mu_{bg}$  the **corresponding** mean intensities of the local background regions. The following parameters are used to define the quality of the foreground signal.

(i) *Signal level*  $S_m = \max(R, G)$  (7.3)

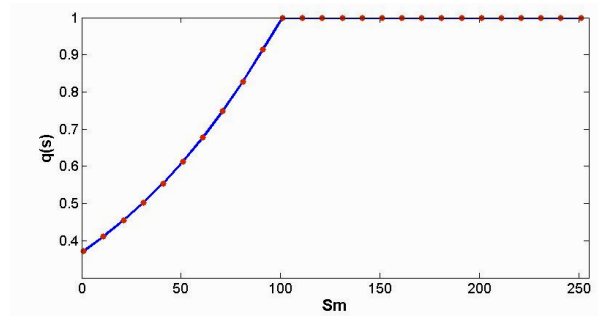
(ii) *Mean background level*  $B_m = \frac{\mu_{br} + \mu_{bg}}{2}$  (7.4)

(iii) *Threshold level of signal*  $T_s = c \times B_m$ , where  $c$  is a constant.  
Then the quality measure  $q(s)$  is defined as

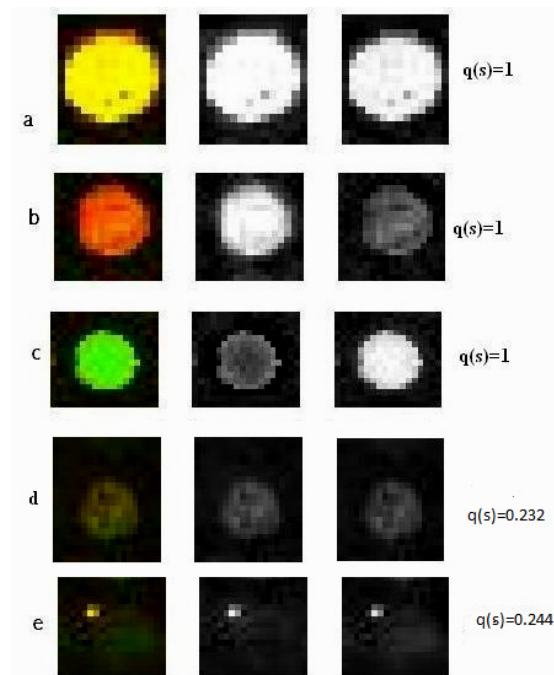
$$q(s) = 1, \quad \text{if } S_m \geq T_s \quad (7.5)$$

$$q(s) = e^{((S_m/T_s)-1)} \quad \text{if } S_m < T_s \quad (7.6)$$

Figure 7.1 shows the relationship between signal level ( $S_m$ ) and quality measure  $q(s)$  for a mean back ground level of 10. The plot indicates that for  $S_m/T_s=0.5$ ,  $q(s)$  become 0.6065. Setting this as the cut off value for signal quality measure, most of the low quality spots can be eliminated from further analysis. Figure 7.2 shows five different spots and their quality measure value  $q(s)$ .



**Figure 7.1** Variation of  $q(s)$  with signal intensity ( $S_m$ )



**Figure 7.2.** Five different spots and their  $q(s)$  values. (a,b,c) are spots with high signal levels (d) low signal spot (e) low intensity noisy signal.

Spots (a), (b) and (c) have high signal levels. So the corresponding quality value is '1'. Spot (d) and (e) have low signal levels and the hence the quality value are 0.232 and 0.244 respectively.

Quality matrices for the intensity variations within the spot is defined using two parameters coefficient of variation ( $CV$ ) and coefficient of determination( $CD$ ).

#### 7.4.2 Coefficient of variation

A measure of homogeneity of intensity of the pixels of an image is the ratio between the standard deviation ( $\sigma$ ) and mean ( $\mu$ ) of pixel intensities, called Coefficient of variation ( $CV$ ). For a microarray spot  $CV$  is evaluated for the red and green channels separately as  $CV_r$  and  $CV_g$ .

$$CV_r = \frac{\sigma_r}{\mu_r} \quad (7.7)$$

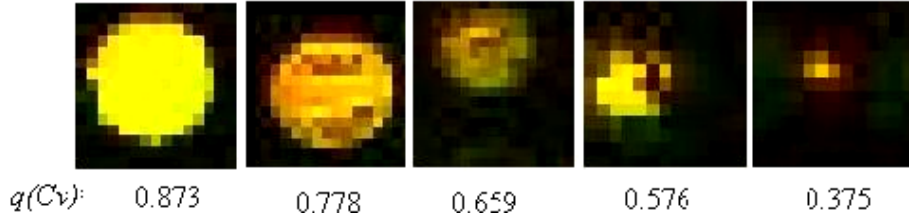
$$CV_g = \frac{\sigma_g}{\mu_g} \quad (7.8)$$

where  $\sigma_r$  and  $\sigma_g$  are the standard deviations of signal intensities of red and green channels respectively. The mean value  $\overline{CV}$  is used for calculating the quality measure  $q(CV)$ .

$$\overline{CV} = \frac{CV_r + CV_g}{2} \quad (7.9)$$

$$q(CV) = e^{-\overline{CV}} \quad (7.10)$$

Figure 7.3 shows the spots with different intensity profile and their quality values estimated using equation (7.10).



**Figure 7.3** Spots with different quality measures  $q$  (CV) indicating the homogeneity of intensities

Coefficient of variation value close to '0' corresponds to a regular (homogenous) spot. It is found that for spots with good morphology coefficient of variation is less than 0.3, and spots with irregular shapes and doughnut shapes have coefficient of variation greater than 0.5 (Sauer et al.,2005).Setting a threshold as 0.35, the cutoff value for the  $q(cv)$  is 0.7047.

#### 7.4.3 Coefficient of Determination (CD)

Coefficient of determination of linear regression indicates degree of the linear relationship between intensities in Cy5 and Cy3 channels (E. Novikov et al., 2005). It is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The farther the line is away from the points, the less it is able to explain. It is defined as the square of the correlation coefficient ( $\gamma$ ).

$$\gamma = \frac{Cov(R, G)}{\sigma_R \cdot \sigma_G} \quad (7.11)$$

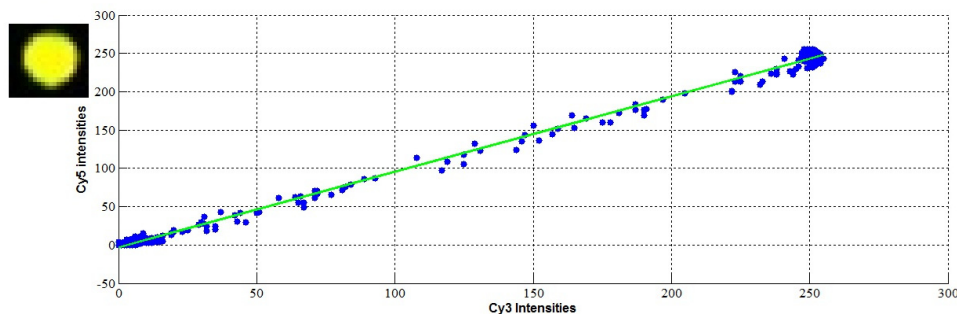
where  $Cov$  is the covariance function.

$$\gamma = \frac{\sum R.G - n\mu_{fr}\mu_{fg}}{\sqrt{\sum R^2 - n\mu_{fr}^2} \sqrt{\sum G^2 - n\mu_{fg}^2}} \quad (7.12)$$

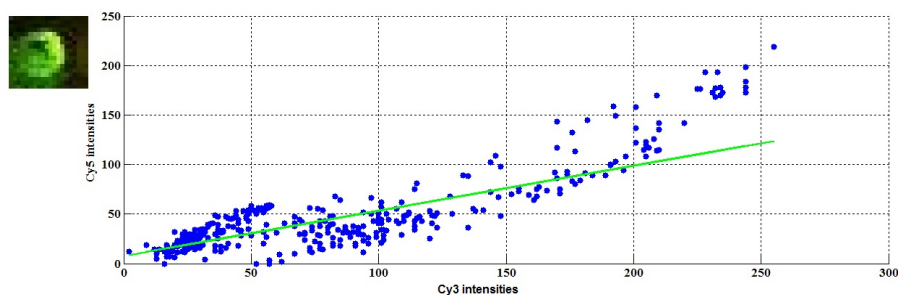
where ‘n’ is the number of pixels, R and G are the background subtracted mean intensity levels of each channel. High value of  $CD$  are expected for good spots and low values suggest either bright and uncorrelated contamination or strong noise content. A closer  $CD$  value to 1 indicates closer the correlation between pixels in the two channels. The spots that have low  $CD$  values must be flagged out using the quality measures. The quality measure is defined as

$$q(CD) = CD \quad (7.13)$$

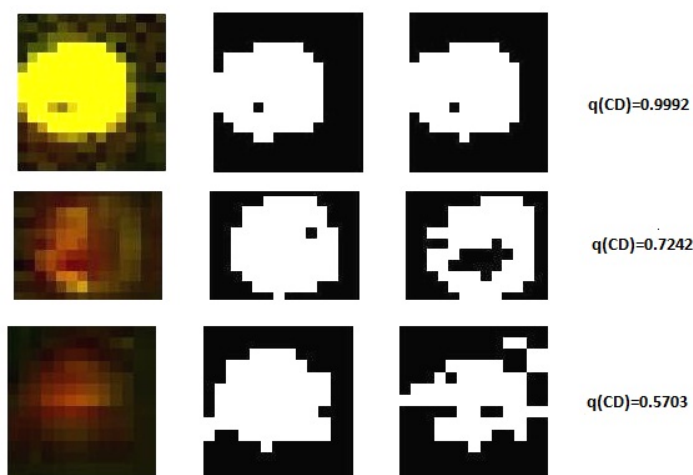
The true intensity ratio( $r$ ) can be represented as the slope of the linear regression fit of the pixel intensities in the two channels. The signals from the two channels are separated and the outliers are removed using median filters for increasing the accuracy of the ratio estimation and the values of coefficient of determination. Figure 7.4 shows the linear regression plot for a good quality spot with  $CD$  value 0.9970. While Figure 7.5 is that of a spot with contamination. The  $CD$  for this spot is 0.7407. Figure 7.6 shows spots with different quality measure  $q(CD)$ . Setting a cutoff value  $q(CD)$  as 0.75 is sufficient to eliminate most of the spots with low  $CD$  value.



**Figure 7.4** Scatter plot of pixel intensities of two channels and the linear regression fit (green line) for a good quality spot (left).



**Figure 7.5** Scatter plot of pixel intensities of two channels and the linear regression fit (green line) for spot with contamination (left)

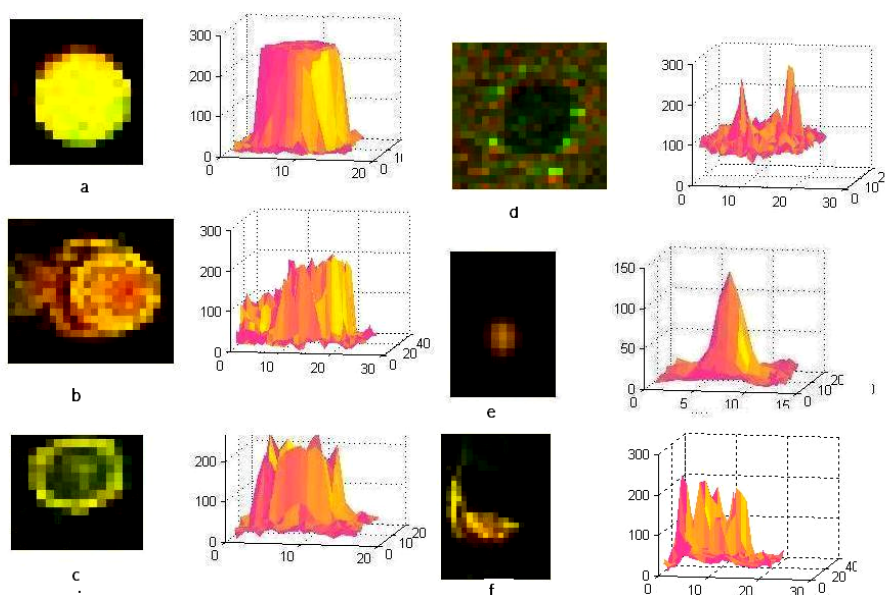


**Figure 7.6** Spots with different  $q$  (CD) values

#### 7.4.4 Spot Area

The spots within a microarray are expected to be circular in shape. Spots with large variation from normal size should be rejected from further analysis, because they are more likely due to noise than fluorescence form the dye due to hybridization. Similarly the spots with very large diameter usually indicate dilation, particle

contamination etc. These spots also should be penalized. Figure 7.7 shows a gallery of microarray spots with different morphology and three dimensional views of their pixel intensities.



**Figure 7.7** Gallery of spots from cDNA based Microarray and their red green intensities graphed three dimensionally.

Here (a) is a high quality spot with uniform intensity (b) is a dilated spot (c) a doughnut spot (d) black hole (e) spot peak with less number of pixels (f) spot with comet tail. The morphology of the spot is clear from the three dimensional surface plot. Spots with smaller than usual size should be penalized since they are more likely due to isolated noise than dye incorporation due to hybridization; spots with excessively large diameter, may indicate contaminants and/or spots likely to be too close to its neighbor. These are also eliminated from further analysis. Quality measures are defined for detecting the spot irregularity. Quality measures used for the morphological analysis in the present work is the area of the spot.

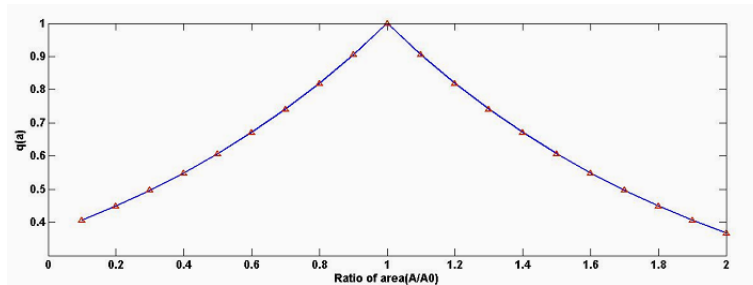
The area ( $A$ ) of a spot under study is calculated as the number of pixels within the foreground region. Average diameter ( $D$ ) of the spots in microarray is estimated during the gridding process using equation 4.21. Let average area, of the spots in the microarray image be  $A_0$  is evaluated as

$$A_0 = \frac{\pi * D^2}{4} \quad (7.14)$$

Then quality measure for area,  $q(A)$ , is defined by Wang X *et al.* (2001) as

$$q(a) = e^{-\frac{|A-A_0|}{A_0}} \quad (7.15)$$

Figure 7.8 shows the relationship between ratio of area ( $A/A_0$ ) and  $q(a)$ . For an ideal spot  $\frac{A}{A_0} = 1$ , So selecting  $\frac{A}{A_0}$  between 0.5 and 1.5, the cut off value for quality measure  $q(a)$  is obtained as 0.6065. i.e, by selecting the cut off value for  $q(a)$  is 0.6065 all spots having area ratio between .5 and 1.5 are selected as good quality spots.



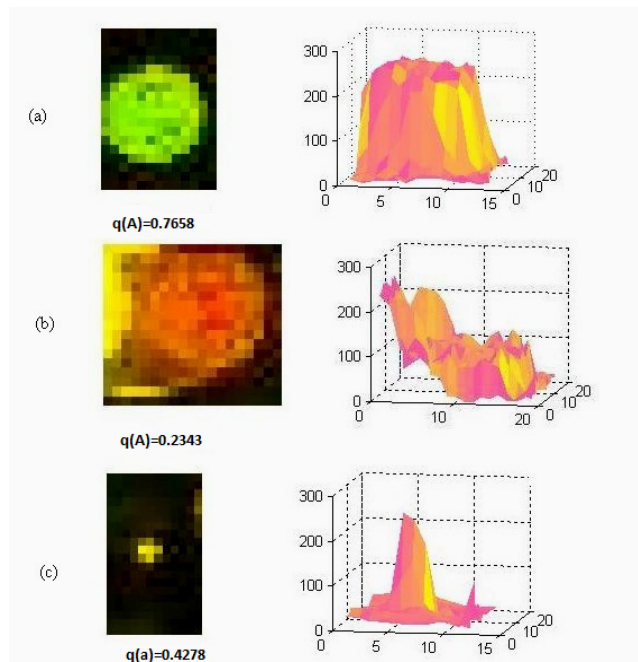
**Figure 7.8** Relation between ratio of area ( $A/A_0$ ) and quality  $q(a)$

Quality measure  $q(a)$  is evaluated for the two channels separately as  $q_r(a)$  (red channel) and  $q_g(a)$  (green channel). The quality of the spot is defined as average of  $q_r(a)$  and  $q_g(a)$ .



$$q(a) = \frac{(q_r(a) + q_g(a))}{2} \quad (7.16)$$

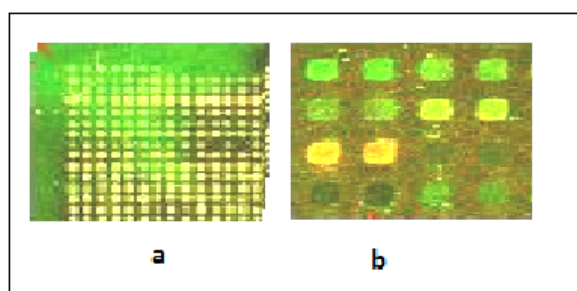
Figure 7.9 shows three different spots with different morphology and their quality measure value  $q(a)$ . Figure 7.9 (a) is a good quality spot with uniform intensity distribution.  $q(a)$  value for this spot is 0.7658, while (b) is a dilated spot. The spread of intensity is clear from the surface plot. The quality value for the spot is 0.2343. (c) is a spot with less number of pixels. The quality values for this spot is 0.4278.



**Figure 7.9** Spots with different area and the corresponding quality measures  $q(a)$  and 3D intensity plot.

### 7.4.5 Coefficient of variation in the Local background

Variations in the background reflect as either high background (fluorescence) or black holes. High background occurs mainly due to the fluorescence of the slide substrate, failure to adequately remove unhybridized molecules, drying problem etc. Sometimes the spots appear as black holes with fluorescent intensity of the spot beneath that of the surrounding background. Figure (7.10) shows the two conditions. Figure 7.10 (a) indicate high background condition while (b) indicate the black holes.



**Figure 7.10** Background variations (a) High intensity background (b) Black holes

Quality measures are used to measure the variations in the local and global background signal and reject spots with large back ground variations. The local background region of each spot is extracted after applying the segmentation algorithm. The quality measure of the variation of the intensity of background region is defined in terms of coefficient of variation as:

$$CV_b = \frac{\sigma_b}{\mu_b} \quad (7.17)$$

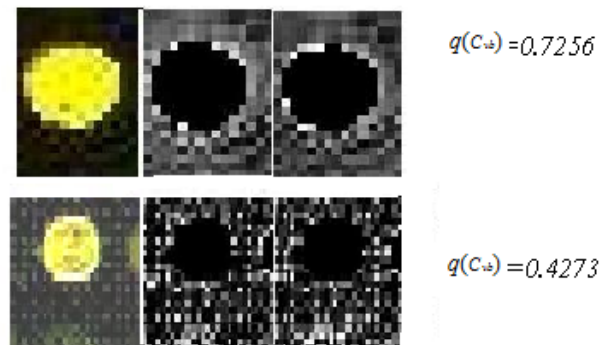
where  $\sigma_b$  is the standard deviation and  $\mu_b$  is the mean of pixel intensities of the local background region. The quality measure is evaluated for the two channels separately as  $q(CV_{br})$ ,  $q(CV_{bg})$ . The geometric mean value of these two quality measures is used for defining the quality measure of local background.

$$q(CV_{br}) = e^{-\overline{CV_{br}}} \quad (7.18)$$

$$q(CV_{bg}) = e^{-\overline{CV_{bg}}} \quad (7.19)$$

$$q(CV_b) = [q(CV_{br}) \times q(CV_{bg})]^{1/2} \quad (7.20)$$

Figure (7.11) shows two spots with their local background from two channels and corresponding quality measures. Similar to foreground region, coefficient of variation value close to '0' corresponds to a regular (homogenous) spot. By setting a threshold as 0.5 for  $CV_b$ , the cutoff value for the  $q(CV_b)$  is 0.6065



**Figure 7.11** Two spots and their local background from two channels and corresponding quality measures  $q(CV_b)$

### 7.4.6 Deviation from Global background

Problematical regions on the microarray have higher local background compared to the global background. Quality measures are used to identify spots with such regions. The ratio of absolute difference between the median intensity values of local and global regions to the median intensity of global region is used as a quality measure defined as

$$r_{gb} = \frac{|\tilde{l}_b - \tilde{g}_b|}{\tilde{g}_b} \quad (7.21)$$

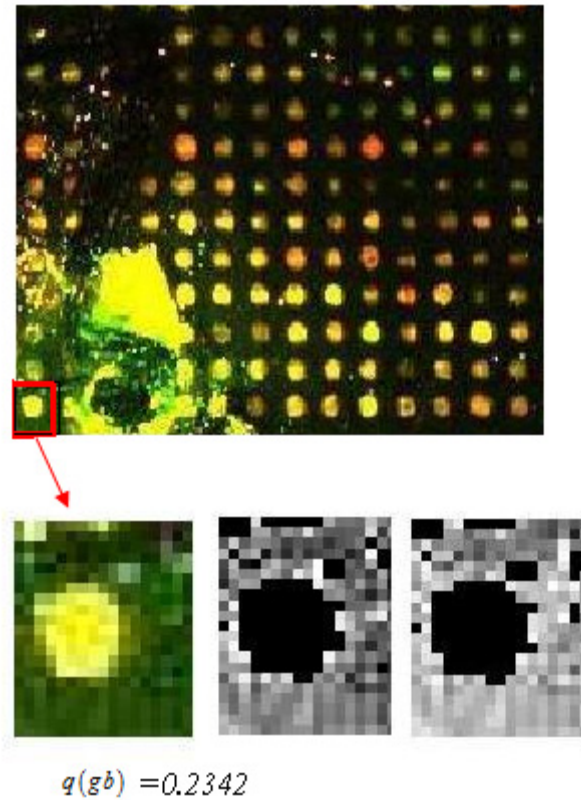
Where  $\tilde{l}_b$  is the median intensity level of local background region and  $\tilde{g}_b$  is the median intensity level of the global region. The parameter is evaluated for two channels separately and the average value ( $\overline{r_{gb}}$ ) used to define the quality measure as

$$q(gb) = e^{-\overline{r_{gb}}} \quad (7.22)$$

Figure (7.12) shows such a spot with large coefficient of variation in the local background region with respect to the global background in a microarray.

The quality measure for global background is evaluated for the two channels separately as  $q_r(gb)$  (red) and  $q_g(gb)$  (green) respectively. Then the quality measure of global background deviation for the spot is defined as

$$q(gb) = \min[q_r(gb), q_g(gb)] \quad (7.23)$$



**Figure 7.12** A spot with high background compared to global background

### 7.5 Saturated Spots

Saturation occurs when spot pixel intensity values exceed the detection range of the photomultiplier tube or the electron detector. This happens in spots of highly expressed genes or spots that contain contaminations. Saturation issue poses a different problem when compared with the previous issues. When saturation happens, there is no prior reason for the variability in measurements to be high. Instead, the measurement distribution is shifted from that of no saturation,

especially when the instrument setting have not been adjusted to give a ratio of 1 for a differential expression of 0. Instead of constructing a continuous function  $q_{sat}$  a threshold of tolerance is defined. A typical spot of microarray image consist of 150 -250 pixels. With a threshold of 10% of pixels as the cut off value the quality measure for saturation is defined based on the number of saturated pixels with respect to total number of foreground pixels. It is defined as:

$$q (sat) = 1, \text{ if saturated pixels is } <10\% \quad (7.24)$$

$$=0, \text{ if the saturated pixels } \geq 10\%$$

## 7.6 Composite Quality Factor

All the above six quality measure and quality of saturation are combined to generate a composite quality score value for each spot. The composite quality score is defined as:

$$q_{com} = [q(s) \times q(CV) \times q(CD) \times q(a) \times q(CVb) \times q(gb)]^{1/6} \times q(sat) \quad (7.25)$$

This value can be used to flag out directly spots, with low a quality lower than a defined threshold.

## 7.7 Implementation of Spot Quality Evaluation Method


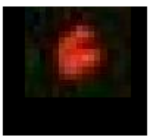

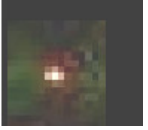

The quality control algorithm is developed with the primary task is to classify spots into two classes as faulty and good. The fundamental interest is the classification of good spot as good and faulty spot as faulty, but the situation becomes complicated by considering good spot as faulty and faulty spot as good. Classifying faulty spot as good is considered harmful, since these faulty spot will be also include in the subsequent analysis of the microarray. The new quality

control method has been tested on several of spots different morphological characteristics, intensity variations and background intensity variations. A microarray subimage consist of 195 spots were tested and the results are explained in the following section. A composite quality score  $q_{com}$  was evaluated for each spot. A threshold value was defined to filter out spots with low  $q_{com}$ .

## 7.8 Experimental Results

The quality analysis method has been implemented on different real subarrays. Table 7.1 shows different spots; their quality values evaluated using equations 7.1 to 7.24. Composite quality score  $q_{com}$  was calculated for each spot using equation 7.25. Spot 1 is a regular spot with  $q_{com}$  0.7979, spot 2 has less number of pixels as well as non uniform intensity;  $q_{com}$  value for the spot is 0.6349. Spot 3 is a spot has large coefficient of variation and  $q_{com}$  value is 0.5835, spot 4 is spot with high back ground intensity, a lower CD value indicates that low correlation between the two channels also less,; the  $q_{com}$  value for the spot is 0.5160. Spot 6 is a dilated spot and has low CD value the  $q_{com}$  for this spot is 0.6741. A threshold value was calculated using the cut off values for each quality measures. Table 7.2 shows the cut off values for each quality measures. Geometric mean of all these cut off values is taken as the threshold value to reject low quality spots.

**Table 7.1 Spots with Quality measures**

Quality measures					
	Spot 1	Spot2	Spot3	Spot4	Spot5
q(s)	1	0.836688	0.613811	0.425064	0.971711
q(cv)	0.61122	0.545971	0.235812	0.583387	0.696236
q(CD)	0.994071	0.502469	0.945443	0.369605	0.47985
q(a)	0.861635	0.600086	0.581504	0.556706	0.573794
q(cvb)	0.698049	0.703434	0.62941	0.679986	0.703371
q(gb)	0.706222	0.676174	0.788128	0.54406	0.716531
q(sat)	1	1	1	1	1
<b>q<sub>com</sub></b>	<b>0.7979</b>	<b>0.6349</b>	<b>0.5835</b>	<b>0.5160</b>	<b>0.6741</b>

**Table 7.2 Cut off values of quality measures**

Quality measures	q(s)	q(cv)	q(CD)	q(a)	q(cvb)	q(gb)
Cutoff values	0.6065	0.7047	0.75	0.6065	0.6065	0.6065

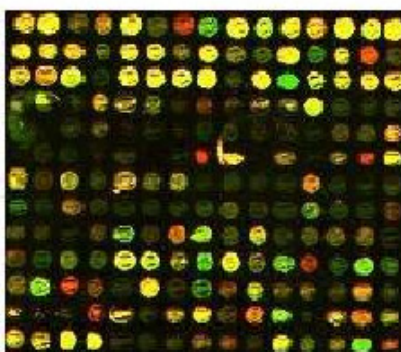
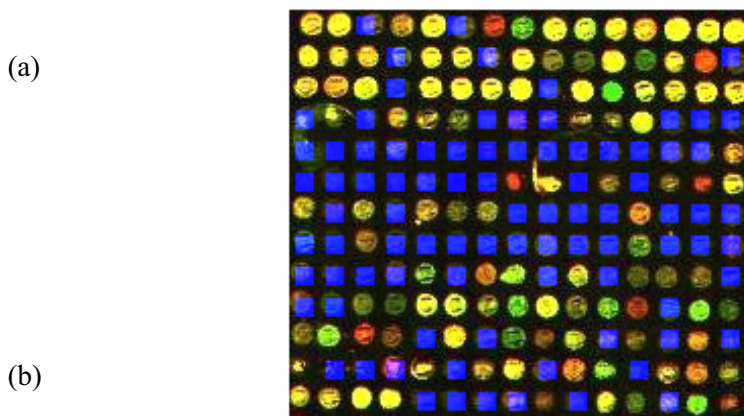
**Figure 7.13** A subarray consists of 195 spots.



Figure 7.13 shows a subarray consist of 195 spots. Figure 7.14 to 7.20. shows the result of applying different quality measures in the spots. Figure 7.14(a) shows quality values for each spot obtained when quality measures for signal applied. A cut off value of 0.6065 rejects spots with low signal q(s) as shown in (b). The spots having low quality are labeled in blue .

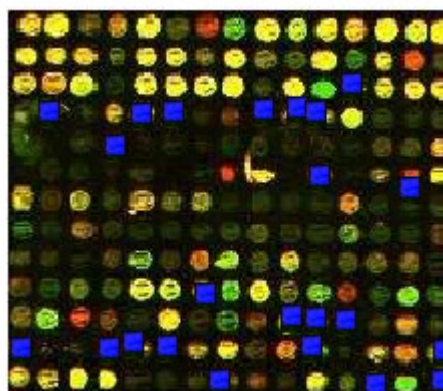
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	1	0.747183	0.56485	0.766893	1	0.436279	0.886618	0.93058	1	0.901374	1	0.985881	1	0.90445	1
2	1	1	1	0.591011	1	1	0.59561	1	0.660349	0.631964	1	0.637051	1	1	0.460925
3	1	1	1	0.498052	1	1	1	1	0.550376	1	1	1	1	1	0.957384
4	0.47767	0.756353	0.451676	1	0.734024	0.668318	0.441003	0.590791	0.516823	0.648368	1	0.522776	0.457201	0.473232	1
5	0.479963	0.409791	0.438548	0.552995	0.401666	0.434982	0.420139	0.438297	0.422835	0.410331	0.426338	0.431925	0.521349	0.539887	1
6	0.378282	0.441302	0.43738	0.406197	0.453772	0.409443	0.413876	0.933244	0.962707	0.388766	0.640579	0.406644	0.637143	0.964	1
7	1	0.473532	0.89192	0.48797	0.876953	0.667185	0.86655	0.427937	0.472268	0.420158	0.428325	1	0.474463	0.423865	0.478424
8	0.566517	0.431961	0.708744	0.480121	0.46531	0.501173	0.510818	0.51329	0.584721	0.460523	0.453987	0.783642	0.463593	0.455856	0.518262
9	0.518386	0.45215	0.435959	0.575369	0.995826	0.538782	0.900951	1	0.573032	1	0.467586	0.636764	0.666936	0.631432	0.416587
10	0.436477	0.516531	0.737094	0.634532	1	1	0.874214	1	1	0.711945	1	0.698613	0.595855	1	0.699154
11	0.61375	1	0.858364	0.701463	0.501684	1	0.518647	0.692587	0.967424	1	0.568318	0.622917	0.534833	1	0.436358
12	0.803158	0.42137	0.425781	0.493072	1	0.552112	1	1	0.602021	1	1	0.460554	0.792846	1	0.908687
13	1	1	1	1	0.468718	0.475779	0.534821	0.524114	0.639792	0.477462	1	0.64279	0.5831	1	0.776855



**Figure 7.14** Results of applying quality measure for signal (a) Quality value (b) rejected spots (blue colour)

Figure 7.15 (a) shows quality values obtained when quality measures for Coefficient of variation q (CV) applied. A cut off value of 0.7047 rejects spots with large variation of intensity within the spot as shown in (b).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0.741252	0.787304	0.747684	0.805271	0.815556	0.880891	0.744459	0.75854	0.781387	0.803211	0.780817	0.797464	0.830836	0.813514	0.794145
2	0.713742	0.726486	0.749505	0.740281	0.753209	0.744966	0.768282	0.780042	0.767842	0.75242	0.79918	0.823433	0.723129	0.787375	0.791202
3	0.758157	0.76476	0.811604	0.759123	0.750508	0.787549	0.748716	0.782908	0.749126	0.733541	0.762409	0.740148	0.754452	0.733533	0.772365
4	0.747318	0.610964	0.718596	0.710733	0	0.696458	0.775037	0.722061	0.621886	0.63332	0	0.772986	0.782292	0.793459	0.766892
5	0.782248	0.799745	0.819985	0.695711	0.855349	0.807766	0.81127	0.868494	0.920495	0.817774	0.803237	0.807995	0.78903	0.75248	0.70351
6	0.930261	0.818295	0.814075	0.84106	0.746693	0.841713	0.838568	0.678237	0.731018	0.875012	0.686427	0.863005	0.641534	0.641211	0.709887
7	0.75505	0.797352	0.745169	0.766534	0.753627	0.724263	0.755765	0.827102	0.795952	0.833216	0.816816	0.75712	0.803047	0.815969	0.779181
8	0.836729	0.82099	0.743469	0.794585	0.791929	0.777402	0.775039	0.794542	0.784643	0.799247	0.809277	0.717868	0.899042	0.7987	0.770817
9	0.726366	0.743735	0.805917	0.77948	0.690241	0.788014	0.775539	0.756922	0.797521	0.756794	0.798968	0.75432	0.774009	0.752132	0.879258
10	0.886295	0.808009	0.775712	0.765168	0.750135	0.756659	0.692567	0.73987	0.773532	0.741379	0.741563	0.833523	0.75167	0.780865	0.742232
11	0.755011	0.76991	0.788194	0.73217	0.770583	0.795877	0.789912	0.861084	0.74466	0.73839	0.675768	0.685765	0.754261	0.726818	0.868428
12	0.500332	0.818337	0.827203	0	0.608595	0.733961	0.667172	0.732539	0.773381	0.752427	0.686647	0.708604	0.723958	0.736664	0.645017
13	0.721521	0.724997	0.777757	0.733083	0.744105	0.801877	0.732675	0.88047	0	0.779022	0.71283	0.768435	0	0.754482	0

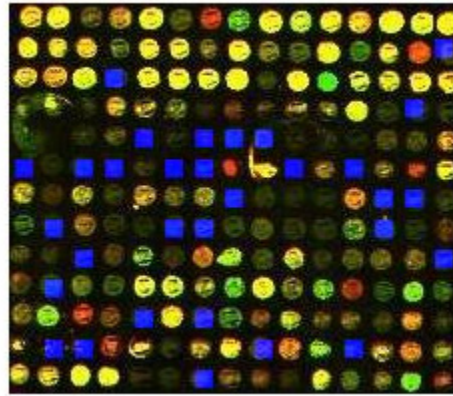


**Figure 7.15** Results of applying quality measure for coefficient of variation  $q(cv)$   
 (a) Quality value (b) Rejected spots (blue color)

Figure 7.16 (a) shows quality values obtained when quality measures for Coefficient of determination  $q$  (CD) applied. A cut off value of 0.75 rejects spots with large variation of intensity within the spot as shown in (b).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0.969008	0.981949	0.943827	0.94593	0.982264	0.99186	0.809987	0.771908	0.979086	0.982472	0.964104	0.9677	0.967884	0.976926	0.987156
2	0.990638	0.9829	0.978263	0.961631	0.989208	0.984204	0.965101	0.976132	0.955281	0.933263	0.982205	0.849306	0.977432	0.906203	0.734623
3	0.981648	0.969088	0.976633	0.616767	0.993324	0.976732	0.98113	0.990472	0.92792	0.992013	0.873666	0.977259	0.990335	0.985269	0.975086
4	0.955045	0.992139	0.894888	0.954169	0.951892	0.965246	0.820265	0.933713	0.870416	0.952436	0.9662	0.988604	0.920253	0.59259	0.894102
5	0.855012	0.948313	0.894723	0.923188	0.454666	0.887707	0.354082	0.439182	0.736986	0.827988	0.819566	0.846501	0.980218	0.958149	0.979104
6	0.239523	0.873033	0.660886	0.692899	0.928382	0.315415	0.125354	0.779875	0.977377	0.272255	0.960544	0.429997	0.962128	0.881277	0.972763
7	0.969723	0.911162	0.971705	0.880287	0.982405	0.961356	0.982361	0.367461	0.838247	0.827666	0.795142	0.945597	0.526324	0.685189	0.82604
8	0.87342	0.687193	0.960016	0.867706	0.823407	0.690408	0.545301	0.844184	0.910909	0.804055	0.872282	0.845415	0.748524	0.918049	0.949352
9	0.904022	0.896014	0.679008	0.931933	0.960773	0.791999	0.923309	0.981594	0.926733	0.9497	0.802636	0.93277	0.977643	0.947964	0.674195
10	0.800756	0.681487	0.929478	0.927241	0.985057	0.978593	0.921968	0.918495	0.977031	0.939497	0.95509	0.862207	0.861205	0.946889	0.936009
11	0.91801	0.958394	0.882741	0.90044	0.605055	0.96827	0.647795	0.917969	0.852704	0.973469	0.91747	0.94324	0.895024	0.955071	0.89197
12	0.950498	0.392092	0.464324	0.898801	0.962638	0.833122	0.957434	0.953656	0.590073	0.926746	0.962995	0.584203	0.935402	0.95636	0.953429
13	0.947367	0.980153	0.98138	0.992949	0.921635	0.86426	0.729113	0.899981	0.937702	0.864065	0.983031	0.952545	0.923122	0.914014	0.861648

(a)



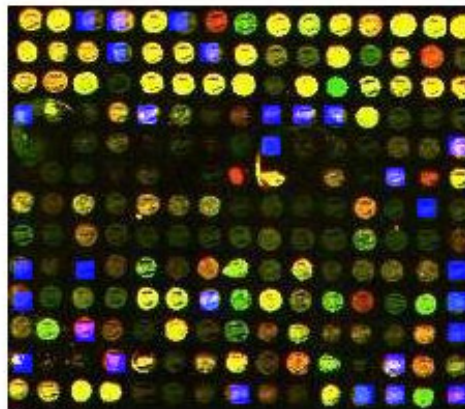
(b)

**Figure 7.16** Results of applying quality measure for coefficient of determination (a) Quality value (b) Rejected spots

Figure 7.17 (a) shows quality values obtained when quality measures for area q (a) applied. A cut off value of 0.6065 rejects spots with large variation of intensity within the spot as shown in (b)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0.890787	0.974699	0.420317	0.606254	0.886728	0.533485	0.805174	0.834161	0.931052	0.829824	0.848446	0.894673	0.930943	0.815311	0.902119
2	0.90276	0.790443	0.85333	0.395832	0.829954	0.84838	0.605455	0.898874	0.818899	0.863583	0.955968	0.644356	0.79765	0.891205	0.808386
3	0.878922	0.947515	0.994742	0.761178	0.914732	0.906574	0.890648	0.986831	0.780051	0.964571	0.791061	0.759602	0.837194	0.875173	0.902619
4	0.531601	0.848778	0.827141	0.776587	0.432388	0.833574	0.924011	0.763154	0.524847	0.367879	0.367879	0.914625	0.880644	0.845911	0.81169
5	0.869698	0.657001	0.623595	0.819091	0.723953	0.708913	0.864529	0.733553	0.510193	0.955968	0.856115	0.919439	0.878991	0.883118	0.577414
6	0.782119	0.957577	0.95608	0.667764	0.794061	0.691711	0.762191	0.667041	0.941349	0.675354	0.756259	0.67158	0.547316	0.705769	0.829824
7	0.833574	0.923037	0.797525	0.871253	0.878991	0.852264	0.815406	0.808767	0.956889	0.859698	0.80965	0.859698	0.729528	0.512639	0.879128
8	0.632342	0.769243	0.90668	0.926809	0.835556	0.894178	0.857692	0.891519	0.951946	0.95608	0.917455	0.898558	0.743885	0.822551	0.899823
9	0.587979	0.805866	0.56516	0.863583	0.731178	0.96025	0.826818	0.819091	0.875173	0.804545	0.808892	0.918723	0.882842	0.859833	0.475872
10	0.47525	0.722104	0.815566	0.894603	0.878922	0.890787	0.584434	0.815566	0.960175	0.617606	0.790813	0.756764	0.853763	0.812071	0.53858
11	0.894813	0.855914	0.590955	0.834161	0.872754	0.90668	0.891205	0.687175	0.723723	0.759632	0.619484	0.770654	0.84838	0.780295	0.569889
12	0.598867	0.871594	0.614292	0.419551	0.797837	0.636382	0.787225	0.852264	0.931868	0.895023	0.730835	0.867639	0.415421	0.863516	0.776769
13	0.773184	0.808102	0.837227	0.848446	0.804608	0.944625	0.647623	0.403461	0.626062	0.800988	0.783484	0.563741	0.540946	0.790443	0.393967

(a)



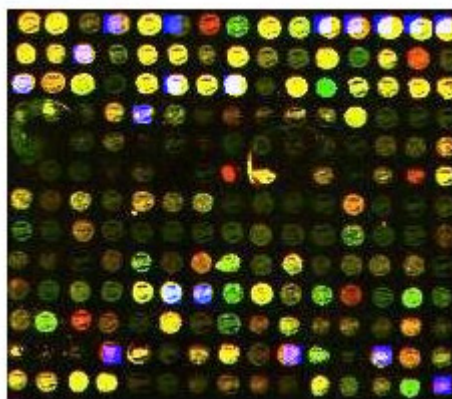
(b)

**Figure 7.17** Results of applying quality measure for area (a) Quality value  
(b) Rejected spots

Figure 7.18 (a) shows quality values obtained when quality measures for local background variations  $q_{(CV_b)}$  applied. A cut off value of 0.6065 rejects spots with large variation of local background (b)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0.66788	0.648456	0.658156	0.576303	0.62513	0.577559	0.724051	0.630007	0.654618	0.628816	0.565106	0.600872	0.599429	0.569851	0.579598
2	0.640143	0.629437	0.574359	0.625206	0.680481	0.652669	0.646335	0.638536	0.645416	0.659908	0.657957	0.63471	0.627631	0.607866	0.646208
3	0.575806	0.61163	0.635305	0.654018	0.655538	0.602822	0.631719	0.595491	0.732723	0.668335	0.632863	0.624465	0.661933	0.627228	0.646942
4	0.702001	0.723186	0.762858	0.721077	0.604074	0.655193	0.780067	0.727208	0.733792	0.659532	0.617297	0.646461	0.642787	0.780102	0.721744
5	0.765176	0.878564	0.81308	0.777832	0.746919	0.80282	0.823829	0.764838	0.717204	0.848259	0.756939	0.743309	0.771904	0.780209	0.724431
6	0.754059	0.822039	0.821042	0.864102	0.80513	0.790396	0.757833	0.670034	0.726227	0.886748	0.828687	0.781086	0.785883	0.723187	0.673722
7	0.636013	0.804103	0.687203	0.728737	0.704309	0.686716	0.673907	0.815265	0.768811	0.846595	0.813888	0.691027	0.750962	0.829169	0.752638
8	0.678675	0.831963	0.761117	0.706232	0.801607	0.749911	0.737398	0.750022	0.679787	0.787002	0.806724	0.689834	0.69743	0.809168	0.809179
9	0.765093	0.838238	0.784489	0.764665	0.687474	0.675187	0.652474	0.664343	0.69749	0.621761	0.745276	0.672556	0.71751	0.727506	0.739767
10	0.6726	0.707384	0.707373	0.691288	0.678768	0.603559	0.676051	0.661789	0.630056	0.678069	0.69159	0.688286	0.684435	0.655406	0.701751
11	0.680885	0.647831	0.669606	0.700133	0.665039	0.622938	0.680752	0.604161	0.657612	0.623888	0.715027	0.789126	0.677966	0.633041	0.737571
12	0.711519	0.661071	0.626431	0.586367	0.619339	0.70455	0.640841	0.671901	0.692529	0.59352	0.68162	0.74208	0.581189	0.628424	0.659617
13	0.623617	0.665354	0.643638	0.685417	0.802171	0.779306	0.68805	0.64407	0.736929	0.759042	0.695469	0.767291	0.648366	0.618102	0.644691

(a)



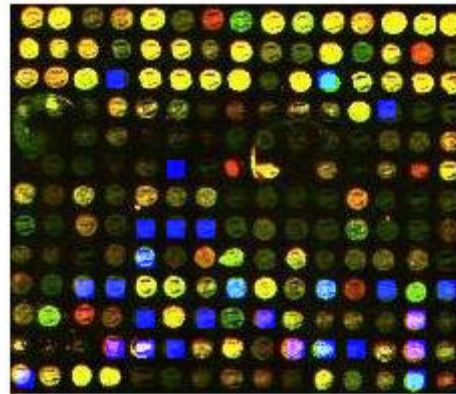
(b)

**Figure 7.18** Results of applying quality measure for local background  $q$  ( $CVb$ )  
 (a) Quality value (b) Rejected spots

Figure 7.19 (a) shows quality values obtained when quality measures for global background variations  $q$  ( $gb$ ) applied. A cut off value of 0.6065 rejects spots with large variation of intensity in the global background is shown in (b)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0.656356	0.606531	0.810158	0.788128	0.931063	0.909156	0.729213	0.768621	0.909156	0.948729	0.716531	0.85394	0.729213	0.751477	0.733796
2	0.710273	0.729213	0.651439	0.729213	0.768621	0.768621	0.826565	0.68321	0.807118	0.68321	0.810158	0.826565	0.788128	0.768621	0.948729
3	0.656356	0.673857	0.651439	0.564718	0.699673	0.656356	0.768621	0.710273	0.651439	0.768621	0.59226	0.621145	0.826565	0.729213	0.810158
4	0.768621	0.716531	0.729213	0.729213	0.699673	0.691826	0.68321	0.622704	0.673857	0.826565	0.948729	0.691826	0.59226	0.651439	0.68321
5	0.866878	0.768621	0.751477	0.716531	0.656356	0.68321	0.621145	0.716531	0.788128	0.807118	0.788128	0.651439	0.788128	0.68321	0.691826
6	0.751477	0.651439	0.68321	0.691826	0.729213	0.531752	0.651439	0.656356	0.768621	0.656356	0.85394	0.651439	0.733796	0.656356	0.716531
7	0.651439	0.729213	0.807118	0.716531	0.768621	0.651439	0.716531	0.651439	0.68321	0.716531	0.68321	0.768621	0.656356	0.751477	0.651439
8	0.751477	0.716531	0.729213	0.621145	0.564718	0.564718	0.538457	0.68321	0.716531	0.751477	0.651439	0.68321	0.68321	0.636112	0.729213
9	0.751477	0.699673	0.606531	0.691826	0.564718	0.621145	0.729213	0.729213	0.68321	0.639308	0.729213	0.622704	0.656356	0.789116	0.716531
10	0.900088	0.691826	0.59226	0.538457	0.788128	0.729213	0.729213	0.578325	0.68321	0.729213	0.59226	0.729213	0.578325	0.68321	0.564718
11	0.807118	0.751477	0.768621	0.68321	0.513417	0.636112	0.59226	0.622704	0.590778	0.651439	0.729213	0.651439	0.621145	0.590778	0.716531
12	0.68321	0.622704	0.621145	0.560488	0.590778	0.531752	0.622704	0.691826	0.622704	0.590778	0.564718	0.564718	0.751477	0.590778	0.639308
13	0.560488	0.68321	0.748657	0.622704	0.621145	0.621145	0.606531	0.789116	0.68321	0.622704	0.68321	0.68321	0.768621	0.538457	0.651439

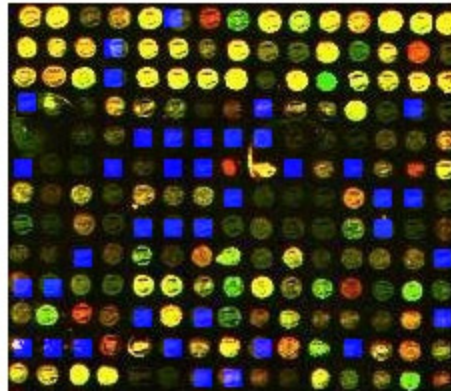
(a)



(b)

**Figure 7.19** Results of applying quality measure for global background  $q$  (gb) (a) Quality value (b) Rejected spots

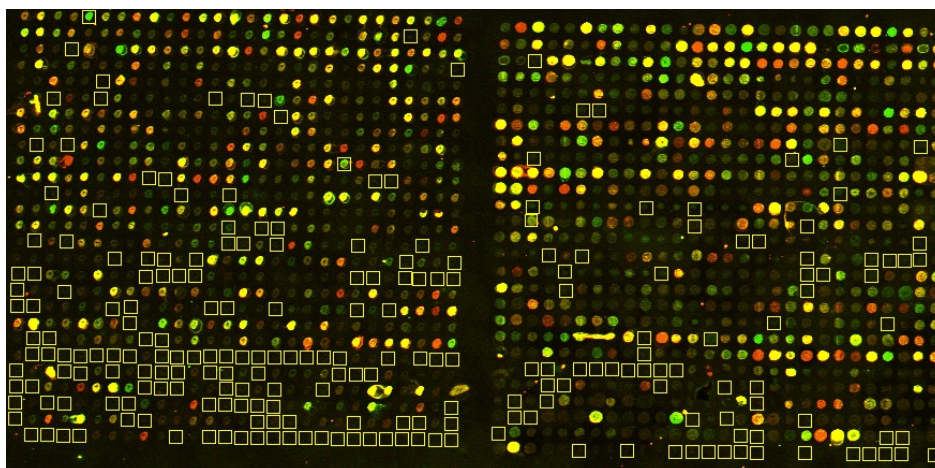
Figure 7.20 shows the rejected spots with composite quality less than the threshold value of 0.6442 calculated from table 7.2 using equation 7.25.



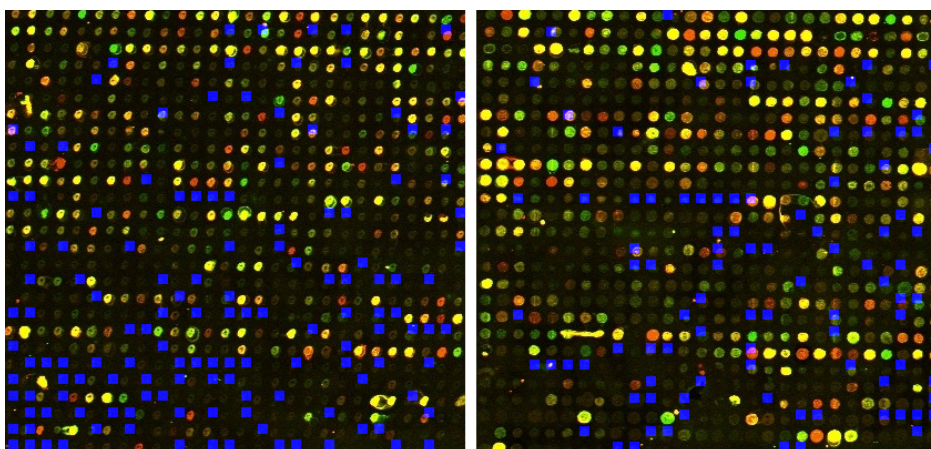
**Figure 7.20** Results of applying  $q_{com}$  and rejected spots

The quality control algorithm has been applied to spots in 10 different real microarrays with different level of contaminations and background intensity. The images were taken from Stanford Microarray Database. The SMD provides the facility to add control flags which is used to eliminate bad spots from further analysis. For example, Figure.7.21 shows two subarrays of the microarray used for lung cancer study (Experiment ID-11712) and the filtered spots (yellow square) using SMD tool.

The developed quality control method has been applied to the same subarrays and the results were analyzed for different threshold values. Figure 7.23 shows the flagged spots (blue square) while implemented the developed algorithm for a threshold of 0.5. The accuracy of the proposed scheme to detect faulty spots and how well the individual spots are classified correctly and how often the spots are misclassified in the two possible directions (good as faulty and faulty as good) was analyzed using Receiver Operating Characteristics (ROC) curves.



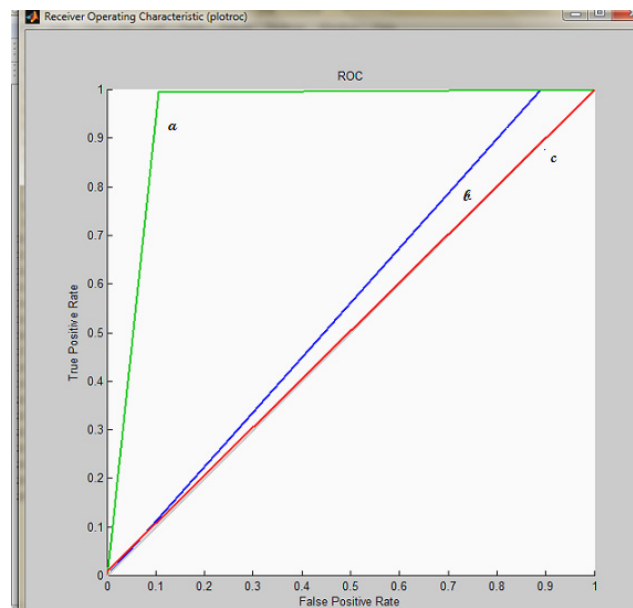
**Figure .7.21** *Flagged spots using SMD tool Miroarray (Exp. ID-11712),subarray 1, 2)*



**Figure .7.22** *Flagged spots using new quality control algorithm on a Microarray (Exp. ID-11712 -subarray (1, 2)) for threshold 0.5*



ROC visualizes the tradeoff between false alarms and detection, helping the user in choosing an optimal decision function. The spots were determined to be either good or faulty using the SMD tool enabling the derivation of the class separating discriminate functions. Data consisting of 1512 spots, of which 1285 were found as valid spots and 227 as faulty using SMD tool. Each test spot was considered to be an independent sample. The results are presented with a ROC curve for three different threshold value for the composite quality score such as 0.5, 0.6442 and 0.75. As shown in Figure. 7.23. Results show that the optimal working point of the classifier is found when threshold is 0.6442 plot (a) which agrees with the result obtained from equation 7.25.

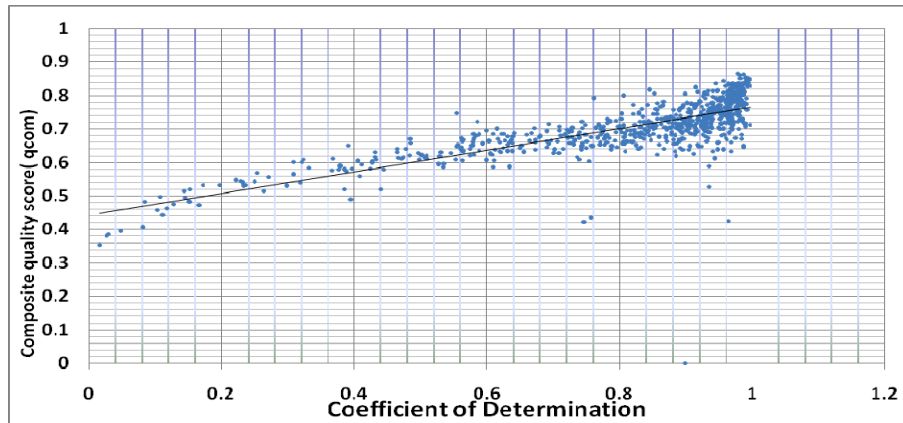


**Figure.7.23** ROC plot for different thresholds (a) 0.6442(b) 0.75(c) 0.5

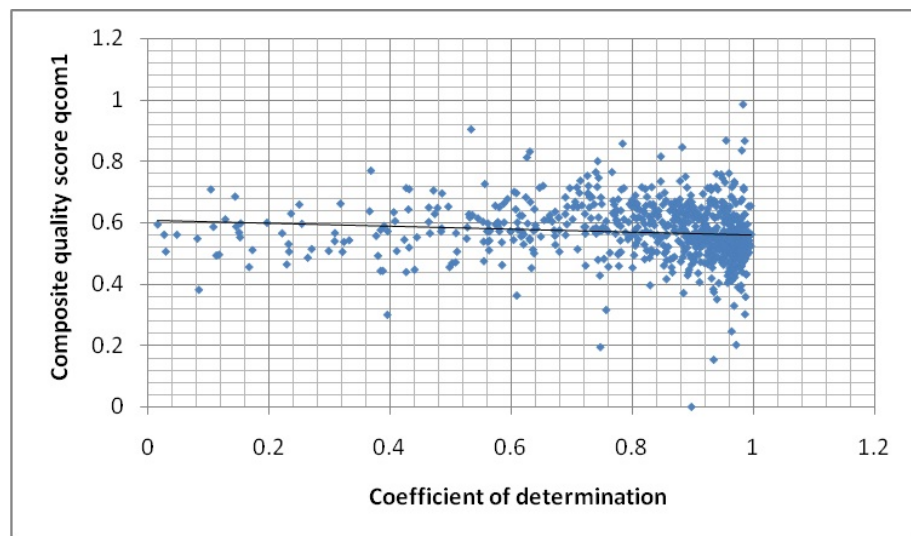
The performance of the quality control algorithm was also compared with the work done by Sauer, U et al. (2005) in which a sets of quality scores are designed considering the individual quality factors such as number of foreground pixels per spot ( $q_{size}$ ), signal to noise ratio( $q_{sig-noise}$ ),variability of local back ground ( $q_{bkg1}$ ) and variability of background with respect to the global average background ( $q_{bkg2}$ ) coefficient of variation of pixel intensities( $q_{cv}$ ) and quality factor corresponding to saturation level as given in equation 7.26.

$$q_{com1} = [q_{size} \times q_{sig-noise} \times q_{bkg1} \times q_{bkg2} \times q_{CV}]^{1/5} \times q(sat) \quad (7.26)$$

The developed quality analysis method includes a new quality parameter  $q_{CD}$  that indicates the degree of linear relation between the intensities of two channels which greatly improve the filtering of spots with noise or other contaminations. Fig 7.24 shows the plot between  $q_{com}$  and coefficient of determination (CD) when new method is applied to a subarray consists of 756 spots. Figure 7.24 shows that variation of  $q_{com}$  for spots having different Coefficient of determination. It is clear that  $q_{com}$  decreases from 0.85 to 0.35 when CD decreased from 0.99 to 0.01 as required. Fig.7.25 shows the  $q_{com1}$  vs CD plot while applying quality measures on the same subarray using method suggested by Sauer, U et.al (2005) where the quality measure for CD was not considered. Hence the developed method shows better performance for rejecting spots with low CD values.



**Figure 7.24** Variation of  $q_{com}$  for spots with different CD (New method)



**Figure 7.25** Variation of  $q_{com}$  for spots with different CD method suggested by Sauer U et al.

## 7.9 Conclusions

The definitions of spot quality measures are a difficult task in a microarray image analysis. There are factors causing variability that cannot be captured using composite quality scores. In this chapter an automatic spot quality assessment techniques have been developed, which on implementation found capable of filtering out spots with unusual morphology, non uniform intensity, low signal levels and large background variations. For each spot six quality measures are defined for evaluating a composite quality score. The saturation problem is considered as a separate case and spots with more than 10% saturated pixels are rejected from further analysis. Cut off values of each quality score is computed and using these cutoff values a threshold value is formulated. All spots with composite quality score less than threshold are rejected. The method has been implemented on real microarray images available at SMD database. Using SMD tool bad spots were flagged and compared the results with the developed quality analysis method. An automated classification of microarray image spots to classes faulty and good based quality measures were conducted. The assessment was presented for classification of individual spots using ROC analysis for different threshold values. Results show that the quality analysis method is capable of filtering the low quality spots efficiently in high density microarrays.

## **CHAPTER 8**

# **Implementation of Image Analysis Method on array CGH Images**

---

*Array based comparative genomic hybridization (array CGH) has been emerged as an efficient molecular cytogenetic technique for the detection of chromosomal imbalances. Studies based on DNA copy number variations provide insights into cancer and many genetic disorders. Array CGH log<sub>2</sub>-based intensity ratios provide useful information about genome-wide Copy number variations. This Chapter presents the implementation of new fully automated image analysis method on arrayCGH microarrays. The copy number variations at different chromosomes are identified and compared with the known results.*

---

## 8.1 Introduction

Comparative genome hybridization (CGH) identifies and maps sites of variation in DNA copy number throughout the genome (Chen, W. et al, 2005). Cytogenetic (chromosome) testing can detect if there is too much (gain) or too little (loss) of chromosomes or pieces of chromosomes. People with changes in their DNA or in the number or structure of their chromosomes may have an increased risk of birth defects, mental retardation developmental delay, behavioral problems and intellectual disability. Conventional cytogenetic analysis can detect unbalanced structural rearrangements within the limits of resolution of the technique. The resolution of the current conventional cytogenetic analyses lies in the range of 3–10 Mb (1 Mb = 1 million base pairs) and requires dividing cells. Therefore, chromosomal micro deletions or micro duplications (those smaller than 3Mb) will go undetected with conventional cytogenetic analyses (ACOG committee, 2009). Fluorescence in situ hybridization technology (FISH) can be used to detect chromosomal abnormalities smaller than 3 Mb (DiGeorge syndrome for example), but because of technical limitations, it can only screen for a limited number of chromosomal abnormalities at one time.

Microarray implementation of CGH, (Pinkel et al. 1998) have the potential to overcome many of the limitations of traditional cryptogenic CGH. arrayCGH improves resolution in detecting chromosomal abnormalities smaller than 3 Mb. In addition, array CGH has been a useful tool in discovering underlying genetic mutations in known, but genetically undefined, human genetic syndromes. arrayCGH does not require dividing cells. The disadvantages of array CGH include the inability to detect balanced inversions or translocations as well as certain forms of triploid, and array CGH costs significantly more than conventional karyotype analysis.

The two types of arrays currently available are targeted and genome-wide arrays. Targeted arrays are currently preferred in clinical genetic practice because they can detect chromosomal abnormalities for known genetic syndromes.

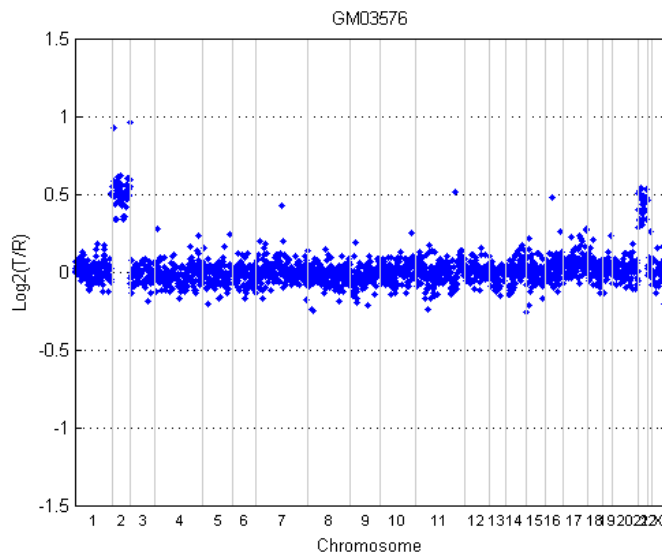
Genome-wide arrays, however, are designed to cover a greater portion of the human genome than targeted arrays. Genome-wide arrays have been particularly useful in research efforts to discover new submicroscopic syndromes. In this chapter the new automatic image analysis method has been implemented in arrayCGH microarray.  $\log_2$  ratio were validated with on known data. The copy number variations at different chromosomes are identified and compared with the known results.

## 8.2 Comparison between arrayCGH and cDNA arrays

The main difference between array CGH and cDNA is that, genomic DNA rather than mRNA transcripts are hybridized in array CGH (Lai, W. R. et al, (2005). In Array CGH microarrays DNA from the test cell is directly compared with the DNA from the normal cell using several thousands of small DNA fragments, with known identity and genomic position known as BAC (bacterial artificial chromosome). DNA fragments or clones from the test sample and reference sample are differentially labeled with dyes (typically Cy3 and Cy5) are hybridized to each probe on the microarray. Clones with normalized test intensities significantly greater than the reference intensities indicate copy number gain in test sample at those positions. Similarly, significantly lower intensities in the test sample are signs of copy number loss. BAC (clone based CGH arrays have a resolution in the order of one million base pairs (1Mb) Snijders, A.M., Pinkel, D., et al. (1996) Oligonucleotide and cDNA arrays provide a higher resolution of 50-100kb Snijders, A.M., (2001). Array CGH  $\log_2$ -based intensity ratios provide useful information about genome-wide CNVs. In humans, the normal DNA copy number is two for all the autosomes. In an ideal situation, the normal clones would correspond to a  $\log_2$  ratio of zero. The  $\log_2$  intensity ratios of a single copy loss would be -1, and a single copy gain would be 0.58. The goal is to effectively identify locations of gains or losses of DNA copy number.

Figure 8.1 shows  $\log_2$  intensity ratios for cell line GM03576 for chromosomes 1 through 23 provided by Coriell cell line BAC arrayCGH data

analyzed by Snijders et al. (2001). Coriell cell line data is widely regarded as a "gold standard" data set. The Coriell Cell Repositories provide essential research reagents to the scientific community by establishing, verifying, maintaining, and distributing cell cultures and DNA derived from cell cultures. These collections, supported by funds from the National Institutes of Health (NIH) and several foundations, are extensively utilized by research scientists around the world. A cell line is a product of immortal cells that are used for biological research. Cells used for cell lines are immortal, that happens if a cell is cancerous. The cells can perpetuate division indefinitely which is unlike regular cells which can only divide approximately 50 times. In the plot, borders between chromosomes are indicated by grey vertical bars. The plot indicates that the GM03576 cell line is trisomic (A diploid cell with an extra chromosom) for chromosomes 2 and 21 (gain of 0.5).

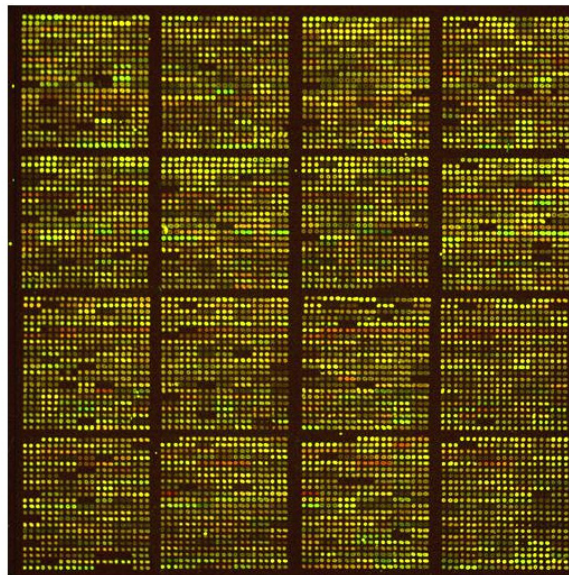


**Figure 8.1** *Log<sub>2</sub>ratios vs. chromosome plot*



### 8.3 arrayCGH Image Analysis - Case Study

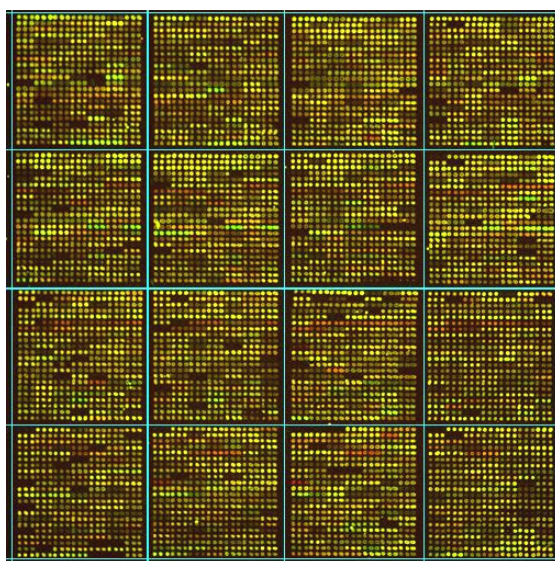
The new fully automatic image analysis method has been implemented in a array CGH microarray image used by Snijders et al. (2001). The array was assembled with 2463 BAC clones in triplicate. There are 16 subarrays (4 x 4) with 462 spots in each subarray. Spot diameter is approximately 70-100  $\mu\text{m}$ , In the array each DNA solution was printed in triplicate to create an array of  $\sim 7500$  elements in a 12 mm square area. Figure 8.2 shows the arrayCGH microarray used for the CNV analysis of colorectal cancer (colorectal adenocarcinoma). In this experiment, the tumor cell line HT-29 was derived from a primary adenocarcinoma of the recto sigmoid colon. HT-29 is hypertriploid ( $3n+$ ) and has accumulated numerous chromosomal structural aberrations. In the arrayCGH experiment cell line HT29 is the test sample labeled with Cy3 (green) and normal reference genomic DNA (red) is labeled with Cy5. The arrays provide precise measurement (S.D. of  $\log_2$ ratios of 0.05-0.10). Different image analysis steps used are explained in the following section.



**Figure 8.2** *Microarray image with 2463 BAC clones in triplicate*

**Step1: Global Gridding**

Figure 8.3 shows the image after implementing the new global gridding method using thresholded intensity projection profiles. Threshold value was selected as 10% maximum of the row and column intensity projection profiles.



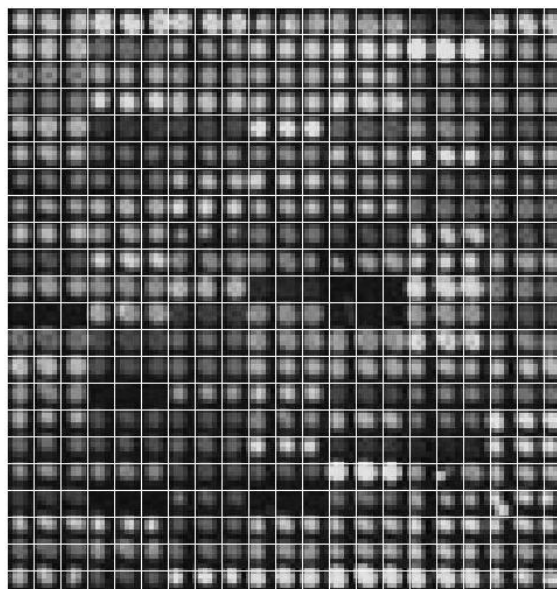
**Figure 8.3** *Image after global gridding*

**Step2: Local gridding**

Individual spots within each subarray are identified using intensity projection profile of best subimage, with maximum block size half the size of each subarray. Figure 8.4 shows locally gridded image.

**Step3: Segmentation**

AASRG based segmentation method has been implemented on each spot within the subarray and  $\log_2$  ratio was evaluated.



**Figure 8.4** *Image after local gridding*

#### **Step4: Quality Control**

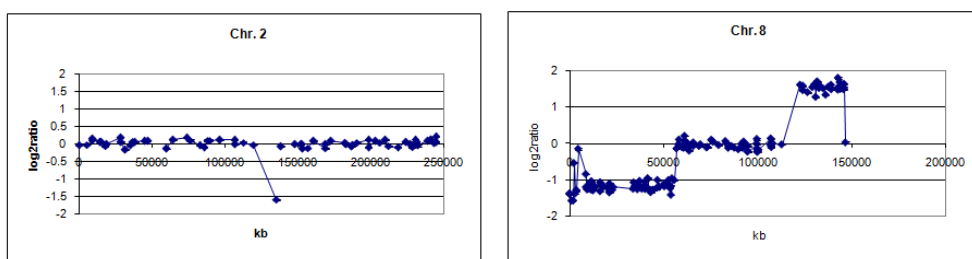
Quality control method explained in chapter 7 was applied to each spot within the grid. In the Table 8.1 and 8.2, positions marked with ‘×’ symbol indicate position of the spots with low quality.

#### **Step 5: Intensity Quantification and Normalization**

Background subtracted mean intensity was calculated for all spots in the array. This has been done for both red and green channels separately. Lowess normalization technique was used to eliminate the spatial bias.  $\log_2$  ratio obtained using the image analysis algorithm was confirmed with existing well defined results provided by (Snijders, A.M., 2001). \

## 8.4 Experimental Results

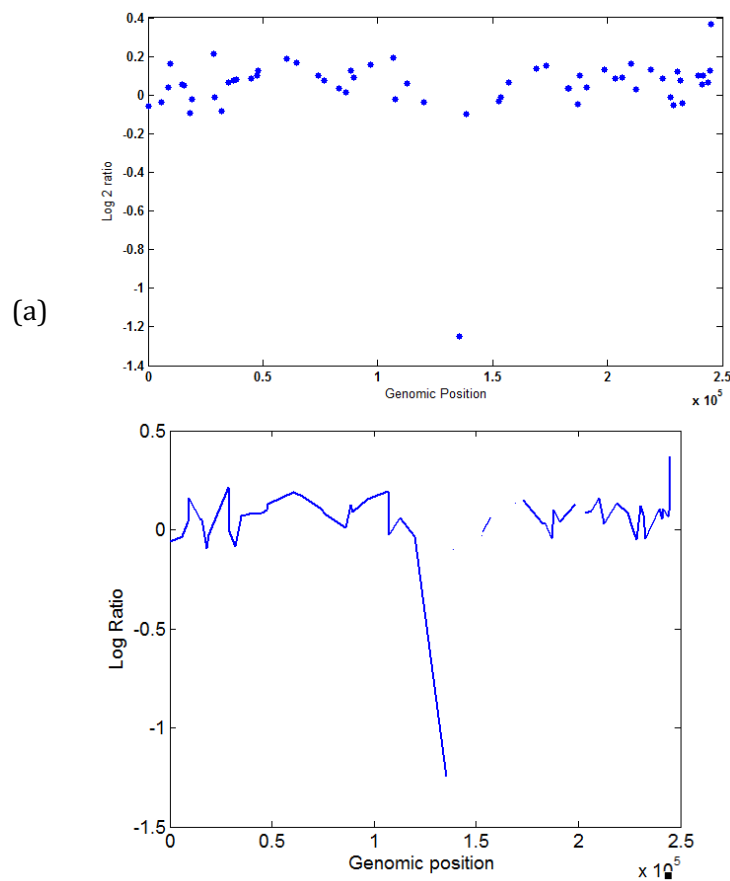
Chromosome2 and Chromosome 8 were considered in the present study. The  $\log_2$  ratio plots for these chromosomes shows breakpoints. Figure 8.5 shows the plot between the log ratio and the genomic position provided by Snijders, A.M., (2001) for chromosome 2 and chromosome 8. In Figure 8.5 a deletion in the chromosome 2 has been identified. In chromosome 8 known deletions across the 8p arm, amplification along the 8q arm in the 8q23.3-24.23 and a focal deletion in 8q23.1 have been identified.



**Figure 8.5**  $\log_2$  ratio vs. genomic position plot for chromosome2 (left) and chromosome 8.

The developed image analysis algorithm has been implemented on the arrayCGH image and  $\log_2$  ratio was evaluated. The average  $\log_2$  ratio of the triplicates was evaluated to increase the accuracy in ratio estimation. Table 8.1 and 8.2 shows the genomic position, known  $\log_2$  ratio, the location of the corresponding spots in the array,  $\log_2$  ratio calculated from the three replicate spots and the average  $\log_2$  ratio obtained using AASRG method for chromosome2 and chromosome 8 respectively. The known  $\log_2$ -based ratios and the supplemental table of known karyotypes can be downloaded from [http://www.nature.com/ng/journal/v29/n3/supinfo/ng754\\_S1.html](http://www.nature.com/ng/journal/v29/n3/supinfo/ng754_S1.html).

This information is also given in table 8.1 and 8.2. Figure 8.6(a) shows the  $\log_2$  ratio vs Genomic position plot for chromosome 2.

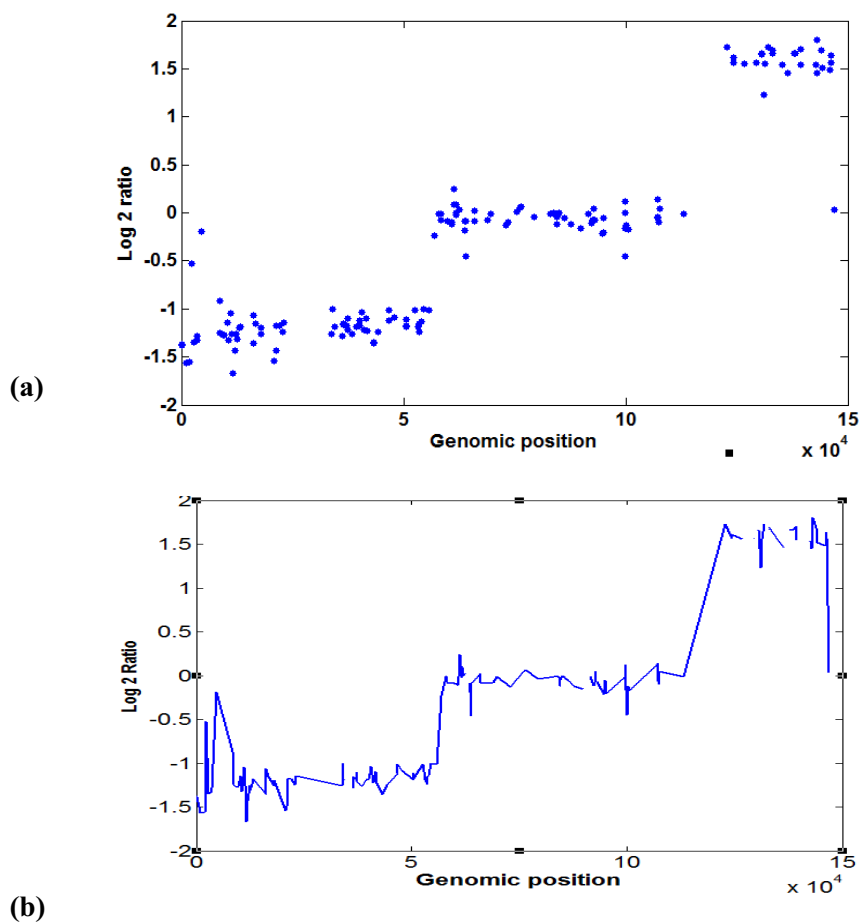


(b)

**Figure 8.6** (a)  $\log_2$  ratio vs. genomic positions plot for chromosome 2 (b)  
*Plot after smoothing*

Plots of array Comparative Genomic Hybridization (CGH) data often show special patterns: stretches of constant level (copy number) with sharp jumps between them.

There can also be much noise. Simple approach to smoothing is to use a median filter. The median filter removes outliers while preserving sustained changes in the input data. Figure 8.6 (b) shows the smoothed plot of log ratio given in Figure 8.6 (a). Figure 8.7(a) and 8.7(b) are the log ratio and smoothed plots respectively for chromosome 8.



**Figure 8.7** (a) *log<sub>2</sub> ratios vs. genomic positions plot for chromosome 8*  
(b) *plot after smoothing*

The smoothed log ratio data can be used for better visualization and later validation of the locations of copy number changes using cytoband information available at National Center for Biotechnology Information (NCBI). Results shows that known deletion across the 8p arm, amplification along the 8q arm in the 8q23.3-24.23 region and a focal deletion in 8q23.1 have been correctly identified. Similarly the deletion in chromosome 2 identified correctly. Spectral karyotyping studies by Abdel-Rahman et al.,(2001) and studies with Oligonucleotide Array CGH Analysis of a Robust Whole Genome Amplification Method (Buhler, J. et al.) as well as CGH Analytics views of Agilent Human Genome CGH 44B microarrays of chromosome 8 in the human colon carcinoma cell line HT29 vs. normal female revealed identical known deletion across the 8p arm amplification along the 8q arm in the 8q23.3-24.23 region, and a focal deletion in 8q23.1 agrees with the results obtained in the present study.

### **8.5 Conclusions**

In this chapter the developed image analysis method has been implemented on array CGH microarray. Gridding, segmentation and quantification quality control and normalization was done automatically.  $\log_2$  ratio of the results was compared with known data. Two chromosomes were selected specifically to validate the results. Experimental results shows that the developed method correctly locate the amplification and deletion regions within the chromosomes.

Table 8.1

chromosome 2 - Data												
Clone	Position Genome, kb	Log2Ratio Known data	Log2SsdDev	Map FISH	SPOT_ COL	SPOT_ ROW	SUBARRAY COL	SUBARRAY_ ROW	rep1	rep2	rep3	Log2Ratio (AASRG)
GS1-8L3	0	-0.047222	0.008809	2 p tel	7	9	4	1	0.0801	-0.0284	-0.0931	-0.05975
RP11-200t14	5906	-0.034639	0.007265	2p25.2	7	2	4	4	0.0752	0.0522	-0.123	-0.0354
RP11-185p14	9000	0.094196	0.013016	2p25.1	10	2	1	4	0.0733	0.0181	0.06	0.03905
RP11-87t07	9597	0.133141	0.004272	2p25.1	10	1	3	3	0.1267	0.2013	0.1215	0.1614
RP11-81t24	14809	0.038027	0.013118	2p24.3	10	2	2	4	0.0854	0.0253	0.0823	0.0538
RP1-254n16	15506	0.043718	0.010856	2p24.1	16	8	1	2	0.01	0.0254	0.0751	0.05025
RP11-111J06	18169	-0.06881	0.011453	2p24	13	10	1	2	-0.0486	-0.0995	-0.0896	-0.09455
RP11-17t13	19001	-0.027678	0.0089	2p24.1	10	2	3	4	-0.0372	-0.0585	0.0112	-0.02365
RP11-57f09	28689	0.169324	0.015668	2p23	10	1	4	3	0.1435	0.2358	0.1876	0.2117
RP11-72t12	29000	0.043784	0.007332	2p23	13	10	2	2	0.148	-0.1232	0.1002	-0.0115
RP11-71t07	31897	-0.168433	0.00759	2p22.3	13	1	1	3	-0.094	-0.0267	-0.1429	-0.0848
RP11-77q15	35196	-0.056964	0.01399	2p22.3	10	2	4	4	-0.1429	0.0674	0.0674	0.0674
RP11-163a23	37247	0.061546	0.005278	2p22	13	2	1	4	0.0942	0.0984	0.0951	0.07675
RP11-163n11	38637	0.037064	0.011564	2p22	13	2	2	4	0.0551	0.089	0.0729	0.08085
RP11-128c05	45000	0.071629	0.005022	2p21	13	10	3	2	0.0979	0.0673	0.1025	0.0849
RP11-119t12	47379	0.071822	0.005588	2p21	13	2	3	4	0.1027	0.1186	0.0843	0.10145
RP11-180t1	47951	0	0	2p16-2p21	13	10	4	2	*	*	0.128	0.128
RP11-81t07	60425	-0.152764	0.004526	2p16.1	4	2	3	2	0.6547	0.2548	0.1204	0.1876
GS-274p9	64772	0.106785	0.010689	2p13p12	1	9	4	4	0.2035	0.1902	0.1446	0.1674
RP11-24p23	73904	0.17287	0.013938	2p14	13	2	4	4	0.1824	0.123	0.0826	0.1028
RP11-87c16	76537	0.090912	0.026259	2p12-2p13	16	10	1	2	0.1182	0.0176	0.1307	0.07415
RP11-63t13	83078	-0.043604	0.019833	2p12	16	2	1	4	-0.0768	-0.0957	0.1597	0.032
RP11-113n17	85979	-0.107456	0.013217	2p12	13	1	3	3	-0.1719	-0.0203	0.0444	0.01205
RP11-84m16	88378	0.094877	0.023864	2p11.2	16	10	3	2	0.0033	0.109	0.1404	0.1247
RP11-82b19	89558	0.088923	0.02739	2p12-2p11.2	16	10	2	2	0.1143	0.0417	0.1138	0.089833333
RP11-67t23	97000	0.112019	0.023905	2q11.1-2q11.2	16	10	4	2	0.2493	0.1313	0.0867	0.155766667
RP11-8c18	106856	0.099962	0.01486	2q12	16	2	2	4	0.1894	0.2027	0.1833	0.1918
RP11-95h05	107408	-0.019826	0.023876	2q12-q13	19	10	1	2	0.0301	-0.0532	-0.0452	-0.022766667
RP11-218t21	112849	0.026457	0.028294	2q14.1	19	10	2	2	0.0528	0.095	0.0888	0.058686667
CTD-2024c11	120000	-0.037157	0.003441	2q13-2q14.1	7	19	1	4	-0.0351	0.0972	-0.1721	-0.036666667
RP11-9t19	135555	-1.598934	0.017059	2q21	16	2	3	4	-1.1851	-1.336	-1.2452	-1.248766667



chromosome 2 – Data (continue.....)

Clone	Position Genome, Mb	Log2Ratio Known data	Log2StdDev	Map FISH	SPOT_ COL	SPOT_ ROW	SUBARRAY COL	SUBARRAY_ ROW	rep1	rep2	rep3	Log2Ratio (AASRG)
RP11-1122	136610			2q21.2	19	10	3	2				
RP11-1122	136610		0	2q21.2	19	10	3	2	x			
RP11-29n17	138570	-0.076907	0.022219	2q22	16	2	4	4	-0.0201	-0.1276	-0.1448	-0.097433333
RP11-19N03	142450		0	2q21.3-2q22	19	10	4	2				
RP11-72h23	145137			2q22-2q24	19	2	1	4				
RP11-67f02	148050	-0.024046	0.0032	2q22-2q23	1	11	1	2	x	x	x	x
RP11-67f02	148050			2q22-2q23	1	11	1	2				
RP11-13c20	152523	-0.022752	0.016025	2q22-2q23	1	11	2	2	-0.0401	0.0414	-0.0906	-0.029766667
RP11-13c20	152523	-0.010459	0.007292	2q22-2q23	1	11	2	2	x	x	x	x
RP11-185m22	153416	-0.148155	0.001119	2q23-2q24.1	1	11	3	2	0.0133	-0.1448	0.104	-0.009166667
RP11-191f09	157033	-0.132654	0.004719	2q24.1	1	11	4	2	-0.1057	-0.0236	x	0.0645
RP11-50f20	160973	0.078614	0.0205	2q24.2	19	2	2	4	x	x	x	x
RP11-39a08	169210	-0.129323	0.010829	2q24.3	4	11	1	2	0.1198	x	0.153	0.1364
RP11-39a08	169210	-0.025305	0.017265	2q24.3	4	11	1	2	x	x	x	x
RP11-218d15	173121	0.067699	0.002182	2q24.3	4	11	2	2	0.1879	0.144	0.1226	0.1515
RP11-61n23	182639	0.014333	0.00868	2q31	19	2	3	4	-0.0233	0.1634	-0.0443	0.031933333
RP11-12n07	183417	-0.009899	0.012008	2q31.2	4	11	3	2	-0.0333	0.0136	0.1236	0.0346
RP11-250b10	187218	-0.072752	0.021796	2q31-2q32	4	11	4	2	-0.038	-0.0121	-0.0897	-0.0466
RP11-36f11	187855	-0.054866	0.010955	2q32.1	7	11	2	2	0.0908	0.0566	0.1528	0.100066667
RP11-69g04	190878	0.025398	0.00437	2q32.1	7	11	1	2	0.01142	0.0947	0.01035	0.038823333
RP11-30m01	198680	0.114396	0.013932	2q32.3	7	11	3	2	0.1592	0.141	0.0877	0.1293
RP11-30m01	198680	-0.111292	0.008616	2q32.3	7	11	3	2	x	x	x	x
RP11-34n13	203416	0.071974	0.025208	2q32.3-2q33	19	2	4	4	0.0771	0.029	0.1521	0.086066667
RP11-57C06	206432	0.030201	0.018908	2q33	7	11	4	2	0.0887	0.1034	0.084	0.092366667
RP11-47e06	210314	0.105757	0.014375	2q33	10	11	1	2	0.1778	0.1413	0.1657	0.1616
RP11-225h15	212567	-0.070201	0.023579	2q33-2q34	10	11	2	2	0.0041	0.0367	0.045	0.0286
CTD-2149f10	216783	-0.120107	0.012995	2q34	10	20	3	3	0.0899	0.2355	0.0785	0.133966667
RP11-4h06	224174	0.042284	0.02085	2q34-2q35	10	11	3	2	0.1075	0.1325	0.0199	0.086633333
RP11-53p06	227283	-0.02281	0.006892	2q35	1	3	1	4	0.0412	-0.0183	-0.0565	-0.0112
RP11-23g02	228708	-0.089841	0.023792	2q35-2q36	10	11	4	2	0.0171	-0.0356	-0.1323	-0.050266667
RP11-247e23	230615	0.097984	0.02351	2q35-2q36	13	11	1	2	0.1128	0.1331	0.1119	0.119266667

chromosome 2 – Data (continue....)

Clone	Position Genome, kb	Log2Ratio known data	Log2StdDev	Map FISH	SPOT COL	SPOT ROW	SUBARRAY COL	SUBARRAY ROW	rep1	rep2	rep3	Log2Ratio (AASRG)
RP11-183m07	231708	0.040051	0.013518	2q36	13	11	2	2	0.0363	-0.0248	0.2144	0.0753
RP11-68h19	232718	-0.058213	0.024838	2q36	13	11	3	2	-0.0105	-0.132	0.0151	-0.042466667
RP11-71h20	239473	0.07295	0.018394	2q36-2q37	13	11	4	2	0.0923	0.0634	0.1318	0.1025
RP11-176L22	241287	0.106465	0.00507	2q37.1-2q37.2	16	11	1	2	-0.0233	0.0695	0.103	0.0564
RP11-186b21	241459	0.107537	0.011765	*2q37	16	11	2	2	0.075	0.0663	0.1678	0.103033333
RP11-21k01	243800	0.021356	0.0232	2q37.2	16	11	3	2	0.1142	0.1004	-0.0143	0.066766667
RP11-116m19	244636	0.052841	0.025519	2q37.2	16	11	4	2	0.1129	0.1429	0.1429	0.1279
CTB-172113	245000	0.190113	0.007369	2 q tel.	10	9	2	1	0.3997	0.3998	0.363	0.367166667

Table 8.2

Clone	Position Genome, kb	Log2Ratio	Log2StdDev	Map FISH 8 p tel	SPOT_COL	SPOT_ROW	SUBARRAY_COL	SUBARRAY_ROW	ASRG
GS1-77L23	0	-1.387435	0.00785		10	9	1	2	-1.374
RP11-240A17	0	-1.368309	0.019822	8p23.2		13	2	4	-1.382
RP11-117P11	1000	-1.569037	0.011704		13	5	3	4	-1.5683
RP11-82K08	1949	-1.568336	0.004215	8p23.3	13	6	4	2	-1.5578
RP11-246G24	2258	-0.536518	0.015345		13	5	1	4	-0.5281
RP11-121F07	2856	-1.381451	0.002796	8p23.2	10	5	3	4	-1.3526
RP11-113B07	3527	-1.315106	0.008368	8p23.2	10	5	2	4	-1.3217
RP11-140K14	3569	-1.278282	0.022338	8p23.2	16	6	1	2	-1.2786
RP11-112008	4600	-0.145181	0.025447	8p23.2	7	5	3	4	-0.1891
RP11-218N24	8634	-0.839422	0.009186	8p22	16	6	3	2	-0.9213
RP11-235O05	8713	-1.196753	0.020838	8p23.1	16	6	2	2	-1.2467
RP11-254E10	9373	-1.260497	0.007079	8p22-8p23.1	7	5	4	4	-1.2768
RP11-241I04	9700	-1.246804	0.025073	8p23.1	16	6	4	2	-1.2673
RP11-277K10	10336	-1.147512	0.007973	8p22-8p23	10	5	1	4	-1.1457
RP11-235F10	10535	-1.139258	0.005697	8p23.1	19	6	1	2	-1.3248
RP11-262B15	11052	-1.073927	0.017055	8p22	7	5	2	4	-1.0456
RP11-112609	11254	-1.263947	0.003805	8p22	10	5	4	4	-1.2576
RP11-235O05	11600	-1.035078	0.025676	8p22-8p23	19	6	2	2	-1.6701
RP11-252K12	11959	-1.2487	0.011334	8p23.1	19	6	3	2	-1.4321
RP11-122N11	12400	-1.223438	0.016008	8p23.1	19	6	4	2	-1.2567
RP11-287P18	12500	-1.287103	0.012607	8p22-8p23	1	7	1	2	-1.3121
RP11-31B07	13167	-1.186912	0.01519	8p22-8p23.1	1	7	2	2	-1.1954
RP11-92C01	13267	-1.164542	0.003271	8p22-8p23.1	1	7	3	2	-1.1824
RP11-236O01	16300	-1.311445	0.002137	8p22	1	7	4	2	-1.3542
RP11-182G02	16314	-1.064003	0.019396	8p22	4	7	1	2	-1.0652
RP11-274K12	16816	-1.1356	0.003481	8p22	4	7	3	2	-1.1567
CTD-2105H03	17993	-1.161824	0.009953	8p21.3-8p22	4	20	4	3	-1.2589
RP11-107H5	17993	-1.189563	0.017555	8p21.3-8p22	10	9	2	2	-1.1973
RP11-51C01	20884	-1.257241	0.018521	8p21.3	4	7	2	2	-1.5421
RP11-191P09	21222	-1.344948	0.018423	8p21.3	4	7	4	2	-1.4329
RP11-233H21	21275	-1.100591	0.008047	8p21-8p22	7	7	1	2	-1.1765
RP11-93D21	21988	-1.170754	0.01556	8p21.3-8p22	7	7	2	2	-1.1798
RP11-50E20	22889	-1.273415	0.001471	8p21.3	7	7	3	2	-1.2451
RP11-110I16	23170	-1.188868	0	8p21.3	4	6	1	3	-1.1478

RP11-274M09	33872	-1.243961	0.019098	8p21.2	7	7	4	2	-1.2585
RP11-89M08	34131	-1.066517	0.005175	8p21.3	4	6	4	3	-1.0062
RP11-204M16	34535		0	8p21.3	4	6	3	3	×
RP11-204M16	34535	-1.176971	0.033086	8p21.3	4	6	3	3	-1.1871
RP11-288M10	34614			8p21	4	6	2	3	
RP11-158F09	36163	-1.258992	0.004936	8p21.2	7	6	1	3	-1.2874
RP11-788I2	36438	-1.127446	0.004995	8p21.2	7	6	2	3	-1.1543
RP11-164M24	36585	-1.116902	0.019914	8p21.2	7	6	4	3	-1.1632
RP11-70L01	37120	-1.16142	0.013503	8p21.2	7	6	3	3	-1.1786
RP11-199M14	37358	-1.025532	0.030046	8p21.2	10	6	1	3	-1.1025
RP11-232J22	37494	-1.198029	0.015217	8p12	10	6	2	3	-1.2131
RP11-138J02	38510	-1.242752	0.006993	8p21	10	6	3	3	-1.2617
RP11-116F09	38391	-1.168733	0.004824	8p12.8p21.1	10	6	4	3	-1.1816
RP11-275K07	40136	-1.168277	0.006026	8p12	13	6	1	3	-1.1765
RP11-277I21	40211	-1.250762	0.016072	8p12	13	6	3	3	-1.1239
RP11-253I17	40500	-1.027447	0.007173	8p12	13	6	4	3	-1.0367
RP11-5J20	41000	-1.224985	0.015821	8p12.8p21.1	16	6	1	3	-1.2219
RP11-287M19	41655	-0.952089	0.020485	8p12	16	6	3	3	-1.0981
CTD-2020E14	41933	-1.233304	0.022046	8p12	7	19	4	3	-1.2312
RP11-57I03	43232	-1.339579	0.017601	8p12	16	6	4	3	-1.3568
RP11-122D17	43284	-1.284143	0.015928	8p12	19	6	1	3	-1.3451
RP11-258M15	44372	-1.26953	0.012117	8p12	19	6	2	3	-1.2451
RP11-274F14	46616	-1.214997	0.043656	8p12	19	6	3	3	-1.1258
RP11-237M13	46798	-1.001145	0.026144	8p11.2.8p12	19	6	4	3	-1.0098
RP11-210F15	47866	-1.19431	0.01829	8p12	1	7	1	3	-1.0871
GS-566K20	50513	-1.162394	0.03291	8p11.2.p11.1	18	8	4	1	-1.18761
RP11-1008I16	50515	-1.102324	0.008985	8p11.2.8p12	1	7	4	3	-1.1107
RP11-285K05	50528	-1.174959	0.017197	8p12	1	7	3	3	-1.1713
RP11-282I23	52662	-1.003641	0.010829	8p11.2	4	7	1	3	-1.0086
RP11-133007	53200	-1.165279	0.019726	8p11.2	4	7	2	3	-1.1876
RP11-284J03	53600	-1.228216	0.01297	8p11.2	4	7	3	3	-1.2376
RP11-64C22	53626	-1.156308	0.011942	8p11.1.8p11.2	7	7	2	3	-1.1542
RP11-480I21	54000	-1.411001	0.01336	8p11.2	4	7	4	3	-1.1345
RP11-282J24	54540	-0.957055	0.00872	8p11.2	7	7	1	3	-1.0062

RP11-73M19	55800	-1.004723	0.0215822	$8q_{11.1}8q_{11.2}$	7	7	4	3	<b>-1.0086</b>
RP11-217N16	57000	-0.139557	0.011453	$8q_{11.1}8q_{11.2}$	4	6	1	4	<b>-0.2341</b>
RP11-12L15	58000	-0.018925	0.007417	$8q_{11.1}$	4	6	2	4	<b>-0.0087</b>
RP11-268N02	58300	0.107324	0.002361	$8q_{11.2}$	4	14	4	3	<b>-0.0124</b>
RP11-268N02	58300	-0.051361	0.012376	$8q_{11.2}$	4	14	4	3	<b>-0.0765</b>
RP11-149G12	58880	-0.096791	0.018114	$8q_{11.21}$	4	6	4	4	<b>-0.0877</b>
RP11-149G12	58880	-0.085359	0.035985	$8q_{11.21}$	4	6	4	4	<b>-0.0876</b>
RP11-188C10	60800	-0.060992	0.026376		7	6	1	4	<b>-0.0987</b>
RP11-188C10	60800	-0.11982	0.019884		7	6	1	4	<b>-0.1145</b>
RP11-221P07	61244	0.071442	0.003472	$8q_{11.2}$	7	6	4	4	<b>0.0876</b>
RP11-221P07	61244	0.208465	0.010345	$8q_{11.2}$	7	6	4	4	<b>0.2431</b>
RP11-175H20	61744	-0.013403	0.009097	$8q_{12}$	7	14	2	3	<b>-0.0211</b>
RP11-175H20	61744	0.054353	0.008651	$8q_{12}$	7	14	2	3	<b>0.0012</b>
RP11-830I4	61898	0.03954	0.007394	$8q_{12.1}$	7	14	1	3	<b>0.0841</b>
RP11-830I4	61898	-0.007786	0.010104	$8q_{12.1}$	7	14	1	3	<b>-0.0053</b>
RP11-105K05	62671	0.021373	0.010269	$8q_{11.22}$	10	6	1	4	<b>0.0321</b>
RP11-105K05	62671		$8q_{11.22}$		10	6	1	4	
RP11-99M06	63738	-0.080495	0.003306	$8q_{12.1}$	10	6	3	4	<b>-0.0865</b>
RP11-99M06	63738	-0.198209	0.012308	$8q_{12.1}$	10	6	3	4	<b>-0.1874</b>
RP11-172D02	63980	-0.049524	0.017378	$8q_{12}$	7	14	4	3	<b>-0.4567</b>
RP11-172D02	63980	-0.060365	0.011325	$8q_{12}$	7	14	4	3	<b>-0.0876</b>
RP11-221I4	65902	0.029199	0.009989	$8q_{11}$	19	8	3	4	<b>0.02311</b>
RP11-213K11	65965	-0.067014	0.00696	$8q_{11.2}8q_{12}$	13	6	2	4	<b>-0.08761</b>
RP11-228N04	69000	-0.068182	0.012791		13	6	3	4	<b>-0.07651</b>
RP11-2588I4	69732	-0.021341	0.012705	$8q_{12}$	13	6	4	4	<b>-0.0128</b>
RP11-234F08	72987	-0.107768	0.011957	$8q_{12}8q_{13}$	16	6	2	4	<b>-0.1267</b>
RP11-252M13	73429	-0.094738	0.015534	$8q_{12}$	16	6	3	4	<b>-0.09981</b>
RP11-92M10	75500	0.101084	0.016586	$8q_{12.3}$	16	6	4	4	<b>0.0112</b>
RP11-282D10	76300	0.061052	0.012718	$8q_{13}$	19	6	1	4	<b>0.0576</b>
RP11-212P10	76507	0.050097	0.021044	$8q_{13}$	19	6	2	4	<b>0.0612</b>
RP11-120M14	79436	-0.045345	0.027094	$8q_{13.2}$	4	14	3	3	<b>-0.0412</b>
RP11-148M07	83046	0.06348	0.007106	$8q_{13}8q_{21}$	19	6	4	4	<b>-0.0145</b>
RP11-107F03	83674	-0.00335	0.001333	$8q_{21.1}$	1	7	1	4	<b>-0.00147</b>
RP11-117N14	84403	-0.023212	0.004636	$8q_{21.1}$	4	7	1	4	<b>-0.0323</b>
RP11-33D07	84555	-0.120141	0.007719	$8q_{13}8q_{21.1}$	4	7	2	4	<b>-0.1207</b>

RP11-203C23	84563	-0.040621	0.029101	8q21.1	1	7	4	4	<b>-0.0456</b>
RP11-225J06	85000	-0.072013	0.042932	8q21.1	1	7	2	4	<b>-0.0012</b>
RP11-65J24	86296	-0.070312	0.028669	8q21.1	1	7	3	4	<b>-0.0561</b>
RP11-93J13	87775	-0.117163	0.011785	8q21.1	4	7	3	4	<b>-0.1189</b>
RP11-90B07	89969	-0.15925	0.007679	8q21.1	4	7	4	4	<b>-0.15925</b>
RP11-195P03	91000		8q21.2	7	7	7	1	4	
RP11-214E11	91573	-0.009907	0.007862	8q21.1	7	7	2	4	<b>-0.009965</b>
RP11-115D10	92300	-0.120794	0.005383	8q21.1	1	22	2	3	<b>-0.11207</b>
RP11-257P03	92523	-0.08225	0.003796	8q21.2,8q21.3	1	22	4	3	<b>-0.078288</b>
RP11-80C11	92829	0.05297	0.018286	8q21.1,8q21.2	7	7	3	4	<b>0.0461</b>
RP11-107C01	92900	-0.066618	0.009984	8q21.1	7	7	4	4	<b>-0.0721</b>
RP11-93E11	93160	-0.075065	0.007461	8q21.1,8q21.2	10	7	1	1	<b>-0.0712</b>
RP11-118008	94810	-0.221572	0.003482	8q21.3	10	7	3	1	<b>-0.2156</b>
RP11-3J21	95000	-0.06615	0.007185	8q21.3	10	7	2	1	<b>-0.05612</b>
RP11-2715	95100	-0.2133	0.014888	8q21.3	10	7	4	1	<b>-0.2019</b>
RP11-102K07	99764	0.000212	0.015298	8q22.2	13	7	1	1	<b>-0.001207</b>
RP11-10510	99811	0.125381	0.023385	8q22.2	13	7	2	1	<b>0.12387</b>
RP11-10610	99811	-0.149088	0.018476	8q22.2	13	7	2	1	<b>-0.1654</b>
RP11-142F22	99963	-0.228889	0.023757	8q22.2	13	7	4	1	<b>-0.4512</b>
RP11-208E21	100226	-0.223705	0.017074	8q22	13	7	3	1	<b>-0.1251</b>
RP11-136H10	100534	-0.163284	0.014036	8q22	16	7	1	1	<b>-0.1681</b>
RP11-238H10	107253	0.139584	0.014995	8q22	16	11	2	1	<b>0.1371</b>
RP11-238H10	107253	-0.03509	0.009507	8q22	16	11	2	1	<b>-0.0461</b>
RP11-131016	107286	-0.058658	0.002017	8q22	1	14	1	2	<b>-0.05912</b>
RP11-1314I2	107526	-0.100577	0.025714	8q22.2	16	11	3	1	<b>-0.10021</b>
RP11-125021	107600	0.038966	0.003594	8q22	16	13	2	2	<b>0.04311</b>
CTD-2013021	113047	-0.026195	0.008312	8q23.1	10	19	3	3	<b>-0.01267</b>
RP11-99I09	122871	1.611785	0.017269	8q24.1	13	21	3	4	<b>1.7241</b>
CTD-2178F17	124312	1.463788	0.009429	8q23	10	19	1	3	<b>1.5632</b>
RP11-24c23	124312	1.58787	0.014972	8q24.1	10	9	3	2	<b>1.6123</b>
CTD-2005M12	126603	1.408636	0.014154	8q23-8q24.1	10	19	2	3	<b>1.5573</b>
RP11-158K01	129424		0	8q24.1	16	13	4	2	<b>x</b>
RP11-229E23	129424	1.547752	0.008276	8q24.1,2,8q24.2	16	11	4	1	<b>1.5632</b>
RP11-150N13	130202		0	8q24.1	19	13	4	2	<b>x</b>
RP11-142P05	130569	1.62522	0.021348	8q24.1	19	13	2	2	<b>1.6549</b>

RP11-495D04	130569	1.608408	0.013616	8 <sub>24</sub> .1	16	22	1	1	<b>1.66321</b>
RP11-65D17	131171	1.278818	0.001595	8 <sub>24</sub> .2	16	21	4	3	<b>1.23371</b>
RP11-126618	131417	1.567078	0.010326	8 <sub>24</sub> .1	19	13	1	2	<b>1.55621</b>
RP11-110015	132000	1.704715	0.01628	8 <sub>24</sub> .1	16	21	3	3	<b>1.72318</b>
RP11-145510	132300			8 <sub>24</sub> .1	16	13	3	2	
RP11-237124	133000	1.520831	0.018831	8 <sub>24</sub> .2	16	11	1	1	<b>1.66412</b>
DMPG-HFF#1-71e5	133000			8 <sub>24</sub> .12-q <sub>24</sub> .13	16	8	4	2	
DMPG-HFF#1-71e5	133000	1.606433	0.01699	8 <sub>24</sub> .12-q <sub>24</sub> .13	16	8	4	2	<b>1.68911</b>
RP11-227F07	135133	1.513994	0.003326	8 <sub>24</sub> .2	4	14	2	2	<b>1.54461</b>
RP11-128P09	136497	1.389527	0.005487	8 <sub>24</sub> .2	4	14	3	2	<b>1.45667</b>
RP11-94M13	137852			8 <sub>24</sub> .2	10	22	1	1	
RP11-94M13	137852	1.55463	0.007058	8 <sub>24</sub> .2	10	22	1	1	<b>1.6654</b>
RP11-158601	138290	1.548874	0.007765	8 <sub>24</sub> .2	1	14	4	2	<b>1.66211</b>
RP11-184M21	138363	1.610983	0.001151	8 <sub>24</sub> .2	4	14	1	2	<b>1.70012</b>
RP11-188E11	138500	1.486552	0.003659	8 <sub>24</sub> .2	10	13	2	1	<b>1.5418</b>
RP11-88M23	138731		0	8 <sub>24</sub> .2	10	13	4	2	×
RP11-21H16	142713	1.520416	0.011437	8 <sub>24</sub> .2	13	13	2	2	<b>1.53719</b>
RP11-28A04	143000	1.476169	0.004332	8 <sub>24</sub> .2	13	13	1	2	<b>1.45321</b>
RP11-44N11	143040	1.802164	0.019405	8 <sub>24</sub> .1	13	21	4	4	<b>1.80156</b>
RP11-45B19	144000	1.683446	0.01148	8 <sub>24</sub> .2	13	22	1	1	<b>1.68744</b>
RP11-122H07	144281	1.486672	0.03927	8 <sub>24</sub> .2	13	13	3	2	<b>1.51441</b>
RP11-17M08	146000	1.488447	0.027548	8 <sub>24</sub> .2	7	22	4	1	<b>1.48725</b>
RP11-642A01	146287	1.635294	0.014417	8 <sub>24</sub> .3	13	22	2	1	<b>1.63514</b>
RP11-13A18	146319	1.529817	0.021448	8 <sub>24</sub> .3	10	13	3	2	<b>1.56431</b>
GS1-26111	147000	0.029512	0.022589	8 <sub>01</sub> tel	10	9	4	2	<b>0.03121</b>





## **CHAPTER 9**

### **Conclusions and Future work**

---

*Brief summary of the research work conducted and the important conclusions there on are highlighted in this chapter. The scope for further work in this field as an extension of the present study has also discussed.*

---

## 9.1 Conclusions

In this thesis, different techniques for image analysis of high density microarrays have been investigated. Most of the existing image analysis techniques require prior knowledge of image specific parameters and direct user intervention for microarray image quantification. The objective of this research work was to develop of a fully automated image analysis method capable of accurately quantifying the intensity information from high density microarrays images. The method should be robust against noise and contaminations that commonly occur in different stages of microarray development.

The research work concentrates on three main areas of microarray image analysis such as gridding, segmentation and spot quality control analysis. Various two channel cDNA and array CGH microarray images available from Stanford microarray data base were used for the analysis. First, a novel method for automatic gridding has been developed. Each subarrays and individual spots were addressed using this method. The method identifies an optimum subimage with regular profile within each subarray using a moving window approach. Using the intensity projection profile of the identified subimage the parameters necessary for gridding are estimated. The method has been validated with different real microarray images with irregular spot size shape and contamination level. Performance of the system has been evaluated in terms of gridding accuracy, robustness against noise and computation time. The morphological filters used in the preprocessing steps make the method robust with respect to different types of noises. The new gridding technique can tolerate a high percentage of missing spots make it a suitable for gridding high density microarray images. Existing method

for automatic gridding based on intensity projection profile of the whole microarray has been developed for comparison. When compared with this method, the new method is found superior in gridding of microarrays with comet tails, doughnut and dilated spots as well as images with large coefficient of variation.

In the second stage of the work a novel segmentation technique (AASRG) has been developed for extracting the foreground and background regions. The method uses the principle of region growing for image segmentation. The seed and threshold values were selected automatically depending on the spot characteristics. The AASRG algorithm was applied on each spot within the array, using block processing technique. Monte Carlo simulations were conducted to study the segmentation accuracy of the method. Various real microarray spots with different morphology, Intensity variations, different levels of contaminations were use to test the segmentation accuracy. The performance of the algorithm has been compared with existing segmentation methods such as fixed circle, adaptive circle and conventional seed region growing method used in MAGIC software. It was observed that the new method is capable of segmenting spots with low SNR levels. The background extraction was carried out considering the characteristics of both global and local background pixels. This method greatly improves the segmentation accuracy of spot with high local background intensity especially for high density microarrays. The ratio of intensity between the two channels in evaluated and log transformation was performed. Different data visualization tools like Scatter plot, MA plots and Box plots were used to examine the results obtained. Normalization techniques based on linear and lowess regression were applied to remove the systematic errors that occurred within the arrays.

Spot quality analysis is an essential part of microarray image analysis. This problem is difficult to formalize due to the diversity of instrumental and biological

factors that can influence the result. Third part of this research work concentrates on the development of spot quality control measures to analyze the quality of microarray spots to filter out low quality ones. Six quality measures are defined for this purpose. These parameters characterize different features of the spot. These parameters are scaled between 0(bad spot) and 1 (good spot) to facilitate further quality analysis. The overall quality of the spot is defined by a composite score. Using thresholding low quality spots were identified and discard them from further analysis. The developed procedure provides an automatic tool to quantify the microarray spot quality for high density microarray images.

arrayCGH  $\log_2$ -based intensity ratios provide useful information about genome-wide CNAs. The developed new automatic image analysis algorithm has been implemented on arrayCGH microarray to study the chromosomal abnormalities related to human colon carcinoma. Chromosome 8 and 2 in the human colon carcinoma cell lines HT29 vs. normal female HT 29 cell line microarray were considered for the study.  $\log_2$  ratio of the intensity values results were validated with already known data. The developed new automatic image analysis algorithm has been implemented in MATLAB. The method is fully automatic, robust and can aid high throughput microarray image analysis. It can process microarray images with different spot size, with broad range of experimental distortions such as non uniform background level, intensive dust spots and large bubbles.

## **9.2 Scope for Further Investigations**

Arrays have become an increasingly diverse set of tools in various scientific fields and wide range of cutting edge research is being conducted using them. DNA / Protein expression profiling and genotyping, tissue arrays for histological analysis and biomarker discovery are some of these areas. The

technology, formats and protocols of microarray are continuing to evolve. Investigators can choose from the growing range of options, when selecting an array technology that is appropriate for reaching their research objectives. A common image analysis platform for the analysis of these microarrays is a mandatory requirement in such situations. With the rapid development of microarray fabrication technology, analysis of huge amount of data is also a major challenge. A more precise quality measures is necessary that uses a set of parameters corresponding to the common quality problem in microarray and set a threshold for each, rather than defining composite quality score.



## REFERENCES

- Abdel-Rahman, W. M., Katsura, K., Rens, W., Gorman, P. A., Sheer, D., Bicknell, D., & Edwards, P. A. (2001). Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement. *Proceedings of the National Academy of Sciences*, 98(5), 2538-2543.
- Adams, R., & Bischof, L. (1994). Seeded region growing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6), 641-647.
- Angulo, J., & Serra, J. (2003). Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics*, 19(5), 553-562.
- Athanasiadis, E., Cavouras, D., Spyridonos, P., Kalatzis, I., & Nikiforidis, G. (2007). An automatic microarray image gridding technique based on continuous wavelet transform. *In Computer Analysis of Images and Patterns* (pp. 864-870). Springer Berlin/Heidelberg.
- Bajcsy, P. (2004). Gridline: automatic grid alignment DNA microarray scans. *Image Processing, IEEE Transactions on*, 13(1), 15-25.
- Bajcsy, P. (2006). An overview of DNA microarray grid alignment and foreground separation approaches. *EURASIP Journal on Advances in Signal Processing*, 2006.
- Bariamis, D., Maroulis, D., & Iakovidis, D. K. (2010). Unsupervised SVM-based gridding for DNA microarray images. *Computerized Medical Imaging and Graphics*, 34(6), 418-425.

## References

---

- Bengtsson, A., & Bengtsson, H. (2006). Microarray image analysis: background estimation using quantile and morphological filters. *BMC bioinformatics*, 7(1), 96.
- Bergemann, T. L. (2010). Use of signal quality measurements to gain efficiency in the analysis of cDNA microarray data. *Journal of Genetics and Genomics*, 37(4), 265-279.
- Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., Olenych, S., Bonifacino, J. S. & Hess, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793), 1642-1645.
- Bowtell, D., & Sambrook, J. (2003). *DNA microarrays: a molecular cloning manual*. CSHL press.
- Brown, C. S., Goodwin, P. C., & Sorger, P. K. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences*, 98(16), 8944-8949.
- Brownstein M J and A. B.Khodursky A B. *Methods in Molecular Biology series*, Humana Press, Totowa, NJ
- Bruchez, M. P. (2009). State of the Art and Beyond: Fluorescent Probes for Living Cells. *The Change We Nseed: New Frontiers in Live-Cell Imaging*, 31-40.
- Buckley, M. J. (2000). The Spot user's guide.
- Buhler, J., Ideker, T., & Haynor, D. (2000). Dapple: improved techniques for finding spots on DNA microarrays. *University of Washington CSE Technical Report UWTR*, Brueck, Chad, Sunny Song, and Jim Collins. "Oligonucleotide array CGH analysis of a robust whole genome amplification method." *Biotechniques* 42.2 (2007): 230-233.08-05.



- 
- Bylesjö, M., Eriksson, D., Sjödin, A., Sjöström, M., Jansson, S., Antti, H., & Trygg, J. (2005). MASQOT: a method for cDNA microarray spot quality control. *BMC bioinformatics*, 6(1), 250.
- Canny, J. (1986). A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 679-698.
- Ceccarelli, M., & Antoniol, G. (2006). A deformable grid-matching approach for microarray images. *Image Processing, IEEE Transactions on*, 15(10), 3178-3188.
- Chen, Y., Dougherty, E. R., & Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical optics*, 2(4), 364-374.
- Demirkaya, O., Asyali, M. H., & Shoukri, M. M. (2005). Segmentation of cDNA microarray spots using markov random field modeling. *Bioinformatics*, 21(13), 2994-3000.
- DeRisi, Joseph L., Vishwanath R. Iyer, and Patrick O. Brown. "Exploring the metabolic and genetic control of gene expression on a genomic scale." *Science* 278.5338 (1997): 680-686.
- Doi, K. (2006). Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. *Physics in Medicine and Biology*, 51(13), R5.
- Draghici, S. (2003). *Data analysis tools for DNA microarrays* (Vol. 4), Chapman & Hall/CRC.

## References

---

- Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 12(1), 111-140.
- Eisen, M. ScanAlyze user manual. (1999). *Stanford University*, USA.
- Fadiel, A., & Naftolin, F. (2003). Microarray applications and challenges: a vast array of possibilities. *Int Arch Biosci*, 2003, 1111-1121.
- Frank Y.S, Shouxian C (2005), "Automatic seeded region growing for color image segmentation", *Image and Vision Computing*, vol.23, pp.877-886.
- GenePix Pro, <http://www.axon.com/gn>
- Gonzalez, R. C., & Richard, E. (2002). Woods, digital image processing. *ed: Prentice Hall Press, ISBN 0-201-18075-8*.
- Hautaniemi, S.; Edgren, H.; Vesanen, P.; Wolf, M.; Järvinen, A.K.; Yli-Harja, O.; Astola, J.; Kallioniemi, O. & Monni, O. (2003). A novel strategy for microarray quality control using Bayesian networks. *Bioinformatics*, Vol. 19, 2031-2038.
- Heyer, L. J., Moskowitz, D. Z., Abele, J. A., Karnik, P., Choi, D., Campbell, A. M., ... & Akin, B. K. (2005). MAGIC Tool: integrated microarray data analysis. *Bioinformatics*, 21(9), 2114-2115.
- Ho, J., Hwang, W. L., Lu, H. H. S., & Lee, D. T. (2006). Gridding spot centers of smoothly distorted microarray images. *Image Processing, IEEE Transactions on*, 15(2), 342-353. Identification for High Throughput Microarray Analysis, *Bioengineering & Biomedical Science?*
- <http://www.bio.davidson.edu/projects/magic/magic.html>
- <http://www.biodiscovery.com/software/imagene/>

- 
- <http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm>
- <http://www1.cse.wustl.edu/~jbuhler/dapple/>
- <http://ccr.coriell.org/>
- <http://www.ncbi.nlm.nih.gov/>
- <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28>.
- Jain, A. N., Tokuyasu, T. A., Snijders, A. M., Segraves, R., Albertson, D. G., & Pinkel, D. (2002). Fully automatic quantification of microarray image data. *Genome research*, 12(2), 325-332.
- Jie Wu., Poehlman, S., Noseworthy, M. D., & Kamath, M. V. (2008). Texture feature based automated seeded region growing in abdominal MRI segmentation. In *BioMedical Engineering and Informatics Vol. 2*, pp. 263-267.
- Katzer, M., Kummert, F., & Sagerer, G. (2003, March). A Markov Random Field model of microarray gridding. In *Proceedings of the 2003 ACM symposium on Applied computing* (pp. 72-77). ACM.
- Kherlopian, A. R., Song, T., Duan, Q., Neimark, M. A., Po, M. J., Gohagan, J. K., & Laine, A. F. (2008). A review of imaging techniques for systems biology. *BMC systems biology*, 2(1), 74.
- Kim, P. G., Park, K., & Cho, H. G. (2005). A Quality Measure Model for Microarray Images. *International Journal of Information Technology*, 11(8).
- Kreil, D. P., Karp, N. A., & Lilley, K. S. (2004). DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics*, 20(13), 2026-2034.

## References

---

- Kuklin, A. A., & Shams, S. (2001). Quality control in microarray image analysis. *Microscopy Research*.
- Labib, F. E. Z., Fouad, I., Mabrouk, M., & Sharawy, A. (2012). An Efficient Fully Automated Method for Gridding Microarray Images. *American Journal of Biomedical Engineering*, 2(3), 115-119.
- Lawrence, N. D., Milo, M., Niranjana, M., Rashbass, P., & Soullier, S. (2004). Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics*, 20(4), 518-526.
- Lehmussola, A., Ruusuvaara, P., & Yli-Harja, O. (2006). Evaluating the performance of microarray segmentation algorithms. *Bioinformatics*, 22(23), 2910-2917.
- Li, Q., Fraley, C., Bumgarner, R. E., Yeung, K. Y., & Raftery, A. E. (2005). Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, 21(12), 2875-2882.
- Lichtenbelt, K. D., Knoers, N. V. A. M., & Schuring-Blom, G. H. (2011). From karyotyping to array-CGH in prenatal diagnosis. *Cytogenetic and Genome Research*, 135(3-4), 241-250.
- Lim, Y. W., & Lee, S. U. (1990). On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. *Pattern Recognition*, 23(9), 935-952.
- Lipori, G. (2005). Efficient gridding of real microarray images. *In Proceedings of the Workshop on Biosignal Processing and Classification of the International Conference on Informatics in Control, Automation and Robotics*.

- 
- Lukac, R., Plataniotis, K. N., Smolka, B., & Venetsanopoulos, A. N. (2004). A multichannel order-statistic technique for cDNA microarray image processing. *NanoBioscience, IEEE Transactions on*, 3(4), 272-285.
- Ma, M. Q., Zhang, K., Wang, H. Y., & Shih, F. Y. (2007). ELB-Q: A New Method for Improving the Robustness in DNA Microarray Image Quantification. *Information Technology in Biomedicine, IEEE Transactions on*, 11(5), 574-582.
- Massoud, T. F., & Gambhir, S. S. (2003). Molecular imaging in living subjects: seeing fundamental biological processes in a new light. *Genes & development*, 17(5), 545-580.
- Médigue, C., Rose, M., Viari, A., & Danchin, A. (1999). Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence. *Genome research*, 9(11), 1116-1127.
- Morris, D., Wang, Z., & Liu, X. (2007). Microarray subgrid detection: a novel algorithm. *International Journal of Computer Mathematics*, 84(5), 669-678.
- Ni, S. H., Wang, P., PAUN, M., DAI, W., & PAUN, A. (2009). Spotted cDNA microarray image segmentation using ACWE. *Romanian Journal of Information Science and Technology*, 12(2), 249.
- Novikov E, Barillot E: (2005) An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments. *BMC Bioinformatics* 6(1),293.
- Novikov, E., & Barillot, E. (2006). A noise-resistant algorithm for grid finding in microarray image analysis. *Machine Vision and Applications*, 17(5), 337-345.

## References

---

- Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern recognition*, 26(9), 1277-1294.
- Qi F, Luo Y, Hu D(2006): Recognition of Perspectively Distorted Planar Grids, *Pattern Recognition Letters*, 27(14),1725-1731.
- Rahmenführer, J., & Bozinov, D. (2004). Hybrid clustering for microarray image analysis combining intensity and shape features. *BMC bioinformatics*, 5(1), 47.
- Robins, G., Robinson, B. L., & Sethi, B. S. (1999). On detecting spatial regularity in noisy images. *Information Processing Letters*, 69(4), 189-195.
- Rueda, L. (2007). Sub-grid detection in DNA microarray images. *Advances in Image and Video Technology*, 248-259.
- Rueda, L., & Rezaeian, I. (2011). A fully automatic gridding method for cDNA microarray images. *BMC bioinformatics*, 12(1), 113.
- Rueda, L., & Vidyadharan, V. (2006). A hill-climbing approach for automatic gridding of cDNA microarray images. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(1), 72-83.
- Russell, S., Meadows, L. A., & Russell, R. R. (2009). *Microarray technology in practice*. Academic Press.
- Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A. & Quackenbush, J. (2006). [9] TM4 Microarray Software Suite. *Methods in enzymology*, 411, 134-193.
- Sahoo, P. K., Soltani, S., & Wong, A. K. C. (1988). A survey of thresholding techniques. *Computer vision, graphics, and image processing*, 41(2), 233-260.

- 
- Sauer, U., Preininger, C., & Hany-Schmatzberger, R. (2005). Quick and simple: quality control of microarray data. *Bioinformatics*, 21(8), 1572-1578.
- Schena, M. (2003). *Microarray analysis* (pp. 1-630). Hoboken, NJ: Wiley-Liss.
- Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., & Cherry, J. M. (2001). The Stanford microarray database. *Nucleic Acids Research*, 29(1), 152-155.
- Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., ... & Croner, L. J. (2006). The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9), 1151-1161.
- Shih, F. Y., & Cheng, S. (2005). Automatic seeded region growing for color image segmentation. *Image and Vision Computing*, 23(10), 877-886.
- Smyth, G. K., Yang, Y. H., & Speed, T. (2003). Statistical issues in cDNA microarray data analysis. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-*, 224, 111-136.
- Snijders, A. M., Nowak, N., Segreaves, R., Blackwood, S., Brown, N., Conroy, J., ... & Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number by CGH. *Nature genetics*, 29, 263-264.
- Soille, P. (2002). On morphological operators based on rank filters. *Pattern recognition*, 35(2), 527-535.
- Southern, E. M. (1992). Detection of specific sequences among DNA fragments separated by gel electrophoresis. 1975. *Biotechnology (Reading, Mass.)*, 24, 122.
- Speed, T. P. (2002). John W. Tukey's contributions to analysis of variance. *The Annals of Statistics*, 30(6), 1649-1665.

## References

---

- Stefano Lonardi, Yu Luo (2004) "Gridding of microarray images", *Proceedings of IEEE Computational Systems Bioinformatics conferences (CSB 2004)* 0-7695-2194-0/04.
- Stefano Lonardi, Yu Luo, (2004) "Gridding and Compression of Microarray Images", *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE Volume , Issue , 16-19 Page(s): 122 – 130.*
- Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U., & O'Brien, J. (2001). Automated image analysis for array hybridization experiments. *Bioinformatics, 17(7)*, 634-641.
- Stekel, D. (2003). *Microarray bioinformatics*. Cambridge University Press.
- Subramanian, G., Adams, M. D., Venter, J. C., & Broder, S. (2001). Implications of the human genome for understanding human biology and medicine. *JAMA: the journal of the American Medical Association, 286(18)*, 2296-2307.
- Tang, T., François, N., Glatigny, A., Agier, N., Mucchielli, M. H., Aggerbeck, L., & Delacroix, H. (2007). Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment. *Bioinformatics, 23(20)*, 2686-2691.
- Tran, P. H., Peiffer, D. A., Shin, Y., Meek, L. M., Brody, J. P., & Cho, K. W. (2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Research, 30(12)*, e54-e54.
- Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., & Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic acids research, 29(12)*, 2549-2557.



---

UCSF spot <http://www.jainlab.org/downloads.html>

- Wang, X. H., Istepanian, R. S., & Song, Y. H. (2003). Application of wavelet modulus maxima in microarray spots recognition. *NanoBioscience, IEEE Transactions on*, 2(4), 190-192.
- Wang, X., Ghosh, S., & Guo, S. W. (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15), e75-e75.
- Wang, X., Hessner, M. J., Wu, Y., Pati, N., & Ghosh, S. (2003). Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics*, 19(11), 1341-1347.
- Wang, Y., Ma, M. Q., Zhang, K., & Shih, F. Y. (2007). A hierarchical refinement algorithm for fully automatic gridding in spotted DNA microarray image processing. *Information Sciences*, 177(4), 1123-1135.
- Wang, Y., Shih, F. Y., & Ma, M. Q. (2005). Precise gridding of microarray images by detecting and correcting rotations in subarrays. In *Proceedings of the 8th Joint Conference on Information Sciences* (pp. 1195-1198).
- Wee-Chung Liew, A., Yan, H., & Yang, M. (2003). Robust adaptive spot segmentation of DNA microarray images. *Pattern Recognition*, 36(5), 1251-1254.
- Wu, J., & Chimka, J. R. (2012). Similarity of Multivariate Methods to Establish Microarray Quality Control Standards. *Quality Engineering*, 24(3), 381-385.

## References

---

- Wu, J., Poehlman, S., Noseworthy, M. D., & Kamath, M. V. (2008, May). Texture feature based automated seeded region growing in abdominal MRI segmentation. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on* (Vol. 2, pp. 263-267). IEEE.
- Xiang, C. C., & Chen, Y. (2000). cDNA microarray technology and its applications. *Biotechnology Advances*, 18(1), 35-46.
- Yang, Y. H., Buckley, M. J., & Speed, T. P. (2001). Analysis of cDNA microarray images. *Briefings in bioinformatics*, 2(4), 341-349.
- Yang, Y. H., Buckley, M. J., Dudoit, S., & Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of computational and graphical statistics*, 11(1), 108-136.
- Yee, Hwa, Yang, Michel J. Buckley, Sandrine Dudoit, and Trence P. Speed "Comparison of methods image analysis on cDNA microarray data", *Journal of computational and graphical statistics*, volume, 11, Number 1, pages 1-29.
- Zacharia, E., & Maroulis, D. (2008, June). A Precise and Automatic Gridding Approach to Noise-Affected and Distorted Microarray Images. In *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on* (pp. 605-607). IEEE.
- Zacharia, E., & Maroulis, D. (2008, October). Microarray image gridding via an evolutionary algorithm. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* (pp. 1444-1447).
- Zacharia, E., & Maroulis, D. E. (2010). 3-D Spot Modeling for Automatic Segmentation of cDNA Microarray Images. *NanoBioscience, IEEE Transactions on*, 9(3), 181-192.

- Zhang A. (2006). *Advanced analysis of gene expression microarray data*. World Scientific Publishing Company Incorporated.
- Zhang, J., Campbell, R. E., Ting, A. Y., & Tsien, R. Y. (2002). Creating new fluorescent probes for cell biology. *Nature Reviews Molecular Cell Biology*, 3(12), 906-918.



## LIST OF PUBLICATIONS

### International conferences

1. Deepa J, Tessamma Thomas, “ Automation Segmentation and Background Correction of DNA Microarray Spots-A New Approach” ,in Proceedings of *International Conference on Modelling and Simulation(MS09),AMSE*, Dec 09,pp.188-192.
2. Deepa J, Tessamma Thomas “Automatic Segmentation of DNA Microarray Images using an improved Seeded Region Growing Method” *Proc.International symposium on Innovations in Natural Computing INC'09pp.1469-1474*.
3. Deepa, J. Thomas, T. Automatic Segmentation of DNA Microarray Images Using an improved Seeded Region Growing Method” in *IEEE conference proceedings of Nature & Biologically Inspired Computing, 2009.NaBIC 2009*.pp.1469-1474
4. Deepa J, Tessamma Thomas “Image adaptive automatic gridding of microarray images”, in *International conference on Advances in neurosciences IAN 2008,Dec2008,CUSAT*.

### **International Journals**

1. J Deepa, Tessamma Thomas, Automatic Gridding of DNA Microarray Images using Optimum Subimage, *International Journal in Recent Trends in Electrical & Electronics (IJRTEE)*, ISSN 1797-9617, Academy Publishers, Finland ,volume1,pp.37-40
2. J Deepa, Tessamma Thomas Image adaptive automatic gridding of Microarray images, *Annals of Neurosciences, Official journal of Indian Academy of Neurosciences*, ISSN 0972-7531, Volume15, pp.59
3. J Deepa, Tessamma Thomas, A Robust Method for Extracting Features from Noisy Microarray Images, *International Journal of Research and Reviews in Computer Science (IJRRCS)* Vol. 2, No. 5, October 2011, ISSN: 2079-2557
4. Deepa J, Tessamma Thomas, A New Gridding Technique for High Density Microarray Images Using Intensity Projection Profile of Best Sub Image , *Computer Engineering and Intelligent Systems*, IISTE , Computer Engineering and Intelligent Systems [www.iiste.org](http://www.iiste.org) ISSN 2222-1719 ,vol.4 No1,2014.
5. Deepa J, Tessamma Thomas, A Fully automated image analysis technique for arrayCGH images, *IEEE transaction on Nano Bioscience* (Submitted)

## Resume



### Personal profile

Name : Deepa J

Sex : Female

Date of Birth : 13-05-1973

Area of Interest : Digital signal /image processing  
Bioinformatics

Languages known : English, Hindi, Malayalam

Computer Skills : MATLAB ,VHDL, MASM,

Professional Experience : 17 years teaching experience at Engineering college. At present working as Associate Professor, Dept. of Electronics, College of Engg, Chengannur.

**Professional Memberships** : Life member, ISTE,

### Publications

International/ National Conferences : 7

Journal Publication : 5

Qualifications	Branch/Specialization	Year	University	Percentage/CGPA
B. Tech	Electronics & Communication Engineering	1994	Cochin University of science &Technology	72.7 %
M. Tech	Digital Electronics	2002	Cochin University of science &Technology	7.4
Ph.D	Microarray Image Analysis	Doing	Cochin University of science &Technology	NA

Email Address : deepaj@ceconline.edu, jdeeparegi@gmail.com

Permanent Address : "Harisree",  
Pattathimukku, Perunna PO,  
Changanacherry, Kerala ,PIN: 686101

Phone No : 0481-2427404(Res), 9447660410(Mob)