ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΣΧΟΛΕΣ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ - ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑΤΑ ΙΑΤΡΙΚΗΣ - ΦΥΣΙΚΗΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ ΣΤΗΝ ΙΑΤΡΙΚΗ ΦΥΣΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΩΝ cDNA ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ
ΒΑΣΙΣΜΕΝΗ ΣΕ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥΣ
ΚΥΜΜΑΤΙΔΙΩΝ ΚΑΙ ΤΥΧΑΙΩΝ ΠΕΔΙΩΝ ΜΑΡΚΟΥ

ΘΕΟΔΩΡΟΣ Α. ΜΑΝΤΟΥΚΑΣ
B. Sc. Τεχνολογία Ιατρικών Οργάνων,
Τεχνολογικού Εκπαιδευτικού Ιδρύματος Αθήνας

Πάτρα, Νοέμβριος 2008, Ελλάδα
COMPENSATORY DNA MICROARRAY IMAGE PROCESSING BASED ON WAVELETS AND MARKOV RANDOM FIELDS MODELS

THEODORE A. MANTOUKAS
B. Sc. Medical Instruments Technology
Technological Educational Institute of Athens

Patras, November 2008, Greece
To my family

With great love
Supervising Committee:
Professor George Nikiforidis
Professor Dionisis Cavouras

Examining committee:
Professor George Nikiforidis
Professor Bezerianos Anastasios
Associate Professor George Sakellaropoulos
ACKNOWLEDGEMENTS

First of all, I would like to express my appreciation to my supervisor, Professor G. Nikiforidis, for the opportunity he gave me to attend and deliver with success this postgraduate program. Also, I would like to express my deep gratitude to my supervisor, Professor D. Cavouras, for his essential guidance in both graduate and post-graduate studies. Additionally, I would like to thank Mr. Manolis Athanasiadis for the essential guidance and help throughout of this thesis. I would also like to thank Dr. John Kalatzis for his assistance in programming in MATLAB. Many thanks go to all my friends in MIPA and MEDISP labs for their support and patience. Finally, I would like to express to my family, my deep appreciation for their love and support.
# TABLE OF CONTENTS

Acknowledgements .................................................................................................................. 5
Table of Contents ...................................................................................................................... 6
List Of Figures .......................................................................................................................... 8
List of Tables ............................................................................................................................ 10
Abbreviations ........................................................................................................................ 11
Περιληπνη ............................................................................................................................... 12
Abstract (in English) .............................................................................................................. 14

## CHAPTER 1 INTRODUCTION TO MICROARRAY TECHNOLOGY ...... 16

1.1. Introduction ......................................................................................................................... 16
1.2. Biological Background ....................................................................................................... 18
1.2.1. Deoxyribonucleic acid (DNA) ..................................................................................... 18
1.2.2. Ribonucleic Acid (RNA) ............................................................................................... 21
1.2.3. Proteins ........................................................................................................................ 22
1.2.4. Gene Expression ............................................................................................................. 23
1.2.5. Polymerase Chain Reduction (PCR) ............................................................................... 25
1.3. cDNA Microarray Images ................................................................................................. 27
1.3.1. Ideal cDNA image characteristics ............................................................................... 31
1.3.2. Real cDNA image characteristics ............................................................................... 32

## CHAPTER 2 MICROARRAY IMAGE PROCESSING ..................................... 36

2.1 Introduction ........................................................................................................................ 36
2.2 Griding or Addressing ........................................................................................................ 37
2.2.1 Manual Method ............................................................................................................... 37
2.2.2 Semi-Automatic method ................................................................................................. 38
2.2.3 Fully-automatic method ................................................................................................. 39
2.3 Segmentation ...................................................................................................................... 40
2.3.1 Circle shape segmentation ............................................................................................. 41
2.3.1.1 Fixed circle segmentation .......................................................................................... 41
2.3.1.2 Adaptive circle segmentation ................................................................................... 42
2.3.2 Adaptive shape segmentation ......................................................................................... 43
2.3.3 Histogram- based segmentation ..................................................................................... 44
2.4 Intensity Extraction ........................................................................................................... 45
2.5 Normalization .................................................................................................................... 46
CHAPTER 3 MATERIALS AND METHODS ................................................. 48

3.1 Addressing ................................................................................. 48
3.1.1 Continues Wavelet Transform (CWT) .................................... 48
3.1.2 Addressing Procedure ............................................................... 49
3.2 Segmentation .............................................................................. 50
3.2.1 A-trous Wavelet Transform (AWT) .......................................... 51
3.2.2 Markov Random Field Model .................................................. 53
3.2.3 Majority Vote Operator ............................................................. 55
3.2.4 Min and Product Operators ..................................................... 56
3.2.5 Probabilistic Product Operator ............................................... 57
3.3 Intensity Extraction .................................................................... 58
3.4 Materials ..................................................................................... 58
3.5 Performance Evaluation ............................................................... 60
3.5.1 Evaluation of Simulated images .............................................. 60
3.5.2 Evaluation of Real images ....................................................... 61

CHAPTER 4 RESULTS AND DISCUSSION ............................................. 62

4.1 Results and Discussion ............................................................... 62
4.2 Conclusions ............................................................................... 68
4.3 Future work .............................................................................. 68

References ....................................................................................... 69
List of publications ........................................................................... 73
LIST OF FIGURES

Figure 1.1: Schematically presentation of the involved tasks in biological discovery based on microarray technology \[2\] ................................................................. 17
Figure 1.2: DNA within cell nucleus \[4\] ............................................................... 18
Figure 1.3 DNA Base Pairs\[5\] .......................................................................... 19
Figure 1.4 DNA Structure\[5\] ............................................................................. 20
Figure 1.5: DNA Replication\[5\] ....................................................................... 20
Figure 1.6: Ribose chemical structure and Uracil chemical structure\[5\] .......... 21
Figure 1.7 Gene segment: interleaving exons and introns\[7\] ......................... 23
Figure 1.8 The central dogma of molecular biology, where R,T,C,T donate the Replication, Transcription and Translation process respectively.\[6\] .......... 23
Figure 1.9: Protein Synthesis \[8\] ..................................................................... 24
Figure 1.10 Exponential amplification of trace amount DNA based on PCR.\[9\] ... 25
Figure 1.11: Polymerase Chain Reduction Process \[9\] ................................. 26
Figure 1.12: Microarray Experiment \[10\] .......................................................... 28
Figure 1.13: Hybridization \[11\] ...................................................................... 28
Figure 1.14: Block diagram of cDNA microarray Scanner \[12\] ..................... 29
Figure 1.15: Schematically presentation of the involved tasks during scanning in a cDNA microarray experiment \[12\] ......................................................... 29
Figure 1.16: Ideal Microarray Image \[13\] .......................................................... 31
Figure 1.17: uncertainty and dynamic range dependencies in cDNA images \[13\] .... 32
Figure 1.18: Rotational grid of cDNA image \[13\] ............................................. 33
Figure 1.19: cDNA image with low discrimination power \[13\] ....................... 33
Figure 1.20: Morphological variations in cDNA images \[13\] ......................... 34
Figure 1.21: Comet tails in cDNA images \[14\] ................................................. 34
Figure 1.22: cDNA image with locally high background \[13\] ....................... 35
Figure 2.1 Semi-automatic gridding method as it is used in MAGIC Tool \cite{20}. ........38
Figure 2.2: Fully automatic representation of estimated grid based on: (a) fitted foreground grid and (b) fitted background grid. ..........................................................39
Figure 2.3: Representation of typical cell mask \cite{21} ........................................40
Figure 2.4: Fixed Circle implementation on a cDNA image \cite{22}. ................42
Figure 2.5: Implementation of Adaptive circle segmentation on a microarray image \cite{22}. ........................................................................................................43
Figure 2.6: Implementation of Seed Region Growing on a microarray image \cite{22} ....44
Figure 2.7: Histogram based segmentation \cite{22} ........................................45
Figure 2.8: Image illustrating different background estimation methods \cite{22}. ............46
Figure 3.1: Three-dimensional representation of CWT coefficients \cite{29}. .................49
Figure 3.2: (a) Original simulated microarray image and (b) gridding result. ..........49
Figure 3.3: Schematically approach of the proposed ensemble scheme based on à trous wavelet transform up to second scale and MRF model. ..................................50
Figure 3.4: Distance of central pixel and the adjacent ones during two levels of à-trous decompositions \cite{31} ..................................................................................................51
Figure 3.5: AWT implementation on a signal containing 1000 points \cite{29} ..........52
Figure 3.6: Hard-Threshold Function \cite{33} ..................................................................52
Figure 3.7: Example of Majority Vote Ensemble Scheme. ........................................55
Figure 3.8: Ensemble decision based on the aggressive rule \cite{37} .........................56
Figure 3.9: Binary Mask \cite{21} ..................................................................................57
Figure 3.10: Intensity Extraction Flow chart .............................................................58
Figure 3.11: Block diagram of the microarray simulation model \cite{40} ....................59
Figure 3.12: A sub-grid of the produced simulated cDNA microarray image. .........59
Figure 4.1: Graphical representation of SMF for all applied methods and cases ......63
Figure 4.2: Graphical representation of the calculated mean probability of error ......64
Figure 4.3: Graphical representation of the calculated coefficient of determination for all applied methods and cases. .........................................................................65
Figure 4.4: Box plots summarizing the normalized MAE values of the proposed ensemble schemes in comparison to Conventional (Simple) MRF model. .............67
LIST OF TABLES

Table 1.1 Classification of Proteins by Biological Activity[6] ........................................22
Table 2.1: Existed segmentation methods and their implementations on existed image processing software tools. ............................................................................................................. 41
Table 3.1: Steps of the MRF based segmentation method..............................................54
Table 4.1: Segmentation Matching Factor results for all applied methods in all cases.
........................................................................................................................................63
Table 4.2: Percentages of mean probability of error in all cases. .................................64
Table 4.3: Calculated results of the coefficient of determination for all applied methods and cases...................................................................................................................65
Table 4.4: Comparative results for four cells of the G channel......................................66
Table 4.5 The calculated mean MAE value of each algorithm........................................67
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cDNA</td>
<td>Complementary Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>A/D</td>
<td>Analog to Digital Converter</td>
</tr>
<tr>
<td>AC</td>
<td>Adaptive Circle</td>
</tr>
<tr>
<td>AWT</td>
<td>A trous Wavelet Transform</td>
</tr>
<tr>
<td>C</td>
<td>Cytocine</td>
</tr>
<tr>
<td>CWT</td>
<td>Continues Wavelet Transform</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DP</td>
<td>Decision Profile</td>
</tr>
<tr>
<td>FC</td>
<td>Fixed Circle</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>HTF</td>
<td>Hard Threshold Filter</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov Random Fields</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reduction</td>
</tr>
<tr>
<td>PE</td>
<td>Probability of Error</td>
</tr>
<tr>
<td>PMT</td>
<td>Photomultiplier</td>
</tr>
<tr>
<td>R</td>
<td>Repication</td>
</tr>
<tr>
<td>$R^2$</td>
<td>Coefficient of Determination</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse Transcription Polymerase Chain Reduction</td>
</tr>
<tr>
<td>SMF</td>
<td>Segmentation Matching Factor</td>
</tr>
<tr>
<td>SRG</td>
<td>Seed Region Growing</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>Tc</td>
<td>Transcription</td>
</tr>
<tr>
<td>TIFF</td>
<td>Tagged Image Format File</td>
</tr>
<tr>
<td>$T_L$</td>
<td>Translation</td>
</tr>
<tr>
<td>tRNA</td>
<td>transport RNA</td>
</tr>
<tr>
<td>U</td>
<td>Uracil</td>
</tr>
</tbody>
</table>
ΠΕΡΙΛΗΨΗ

Οι μικροσυστοιχίες συμπληρωματικού DNA (cDNA) αποτελούν αποδοτικά και αποτελεσματικά μέσα για την ταυτόχρονη ανάλυση της λειτουργίας δεκάδων χιλιάδων γονιδίων. Μια τυπική cDNA εικόνα μικροσυστοιχίων αποτελεί μια συλλογή πράσινων και κόκκινων κηλίδων (spots) που περιέχουν DNA. Κάθε κηλίδα καταλαμβάνει ένα μικρό τμήμα της εικόνας, με τη μέση τιμή της έντασης της κηλίδας να είναι στενά συνδεδεμένη με το επίπεδο έκφρασης του αντίστοιχου γονίδιου. Η κύρια διαδικασία υπολογισμού της έντασης περιλαμβάνει τρία στάδια: Διευθυνσιοδότηση η κατασκευή πλέγματος (gridding), κατάτμηση (Segmentation) και τέλος η διαδικασία εξαγωγής έντασης.

Στη παρούσα εργασία, η διευθυνσιοδότηση πραγματοποιήσεις χρησιμοποιώντας μια αυτόματη τεχνική κατασκευής πλέγματος στηρίζομενη στον συνεχή μετασχηματισμό κυματιδίων (CWT)\(^\text{[29]}\). Ποιο συγκεκριμένα, υπολογίστηκαν τα προφίλ του χ και γ άξονα της εικόνας. Δεύτερον, εφαρμόστηκε, σε κάθε προφίλ, ο CWT μετασχηματισμός εώς το 15 επίπεδο, χρησιμοποιώντας daubechies 4 (db4) ως μητρικό κυματίδιο\(^\text{[29]}\). Τρίτον, υπολογιστικά το άθροισμα των 15 επιπέδων για κάθε ένα από τα δύο σήματα x και y. Τέταρτον, εφαρμόστηκε στα δύο νέα σήματα τεχνική καταστολής θορύβου με χρήση μετασχηματισμού Wavelet. Τελικά, το κέντρο και όρια της κάθε κηλίδας καθορίστηκαν μέσω του υπολογισμού των τοπικών ελαχίστων και μέγιστων του κάθε σήματος.

Για την κατάτμηση της εικόνας, μια νέα μέθοδος προτάθηκε, η οποία διακρίνεται σε τρία βασικά βήματα: Πρώτου, ο άτρομος μετασχηματισμός κυματιδίων (AWT) εφαρμόστηκε έως το δεύτερο επίπεδο στην αρχική εικόνα. Δεύτερον, στις λεπτομέρειες (details coefficients) του κάθε επιπέδου εφαρμόστηκε φίλτρο καταστολής θορύβου, προκειμένου να υποβαθμιστεί ο θόρυβος. Τρίτον, η αρχική εικόνα μαζί με τις προσεγγίσεις (approximations) και τις λεπτομέρειες (details) του κάθε επιπέδου εφαρμόστηκαν σε ένα συλλογικό σχήμα (ensemble scheme) σηματοδοτών στο MRF μοντέλο κατάτμησης. Ως τελεστές του σχήματος χρησιμοποιήθηκαν οι: Majority Vote, Min, Product και Probabilistic Product.

Η αξιολόγηση των προτεινόμενων αλγορίθμων πραγματοποιήθηκε με τη χρήση υψηλής ποιότητας εικόνας προσομοίωσης αποτελούμενη από 1040 κηλίδες (spots) με ρεαλιστικά μορφολογικά χαρακτηριστικά, η οποία δημιουργήθηκε σύμφωνα με το μοντέλο προσημείωσης μικροσυστοιχίων του Matlab\(^\text{[40],[41]}\) καθώς και 14 πραγματικές εικόνες, επτά 16-bit grayscale TIFF εικόνες από κάθε κανάλι (κόκκινο και πράσινο), οι οποίες αποκτήθηκαν από την ευρέως διαδεδομένη βάση δεδομένων DERICI\(^\text{[20],[42]}\). Επιπλέον, προκειμένου να παρατηρηθεί η συμπεριφορά των αλγορίθμων στη παρουσία θορύβου, η απομονώμενη εικόνα υποβαθμιστήκε με τη προσθήκη λευκού Gaussian θορύβου.
Η ακρίβεια της ακολουθούμενης διαδικασίας κατάμηνης, στη περίπτωση της εικόνας προσομοίωσης, προσδιορίστηκε μέσω του segmentation matching factor (SMF)\(^{35}\), probability of error (PE)\(^{35}\) και coefficient of determination (CD)\(^{44}\) με σεβασμό στη πραγματική κλάση στην οποία ανήκουν (φόντο-υπόβαθρο). Στη περίπτωση των πραγματικών εικόνων η αξιοπιστία των αλγορίθμων προσδιορίστηκε έμμεσα μετρώντας την ένταση κάθε κηλίδας, μέσω του Mean Absolute Error (MAE)\(^{43}\).

Το σύνολο των αλγορίθμων, εφαρμοσμένο στην απομονωμένη εικόνα, κατάφερε να οδηγήσει σε καλύτερο προσδιορισμό των κηλίδων σε σχέση με το απλό MRF μοντέλο κατάμηνης. Επιπλέον, ο τελεσίτης Majority Vote επέτυχε το υψηλότερο ποσοστό σε όλες τις περιπτώσεις, ειδικά σε κελία (cells) με υψηλή παρουσία θορύβου (SMF: 82.69%, PE: 6.60% and CV:0.809 ), ενώ το απλό μοντέλο περιορίστηκε στο χαμηλότερο ποσοστό (SMF:94.87%-82.69%, PE:3.03%-9.85%, CV:0.961-0.729). Στη περίπτωση των πραγματικών εικόνων ο min τελεσίτης επέτυχε το χαμηλότερο ποσοστό (MAE: 803.96 and Normalized MAE: 0.0738), σε αντίθεση με τον τελεσίτη Majority Vote, ο οποίος κατάφερε να επιτύχει το υψηλότερο ποσοστό ανάμεσα στους χρησιμοποιούμενους τελεσίτες (MAE 990.49 and Normalized MAE 0.0738). Επιπλέον όλοι οι προτεινόμενοι αλγορίθμοι κατάφεραν να μειώσουν τη μέση τιμή MAE σε σχέση με το απλό μοντέλο MRF (MAE 1183.50 and Normalized MAE 0,0859).
Complementary DNA microarrays are a powerful and efficient tool that uses genome sequence information to analyze the structure and function of tens of thousands of genes simultaneously. A typical cDNA microarray image is a collection of green and red discrete spots containing DNA. Each spot occupies a small fraction on the image and its mean fluorescence intensity is closely related to the expression level of the genes. The main process for measuring spot intensity values involves three tasks: gridding, segmentation and data extraction.

In the present study, spot location was accomplished by using an automatic gridding method based on continues wavelet transform (CWT). Firstly, line-profiles for x and y axes were calculated. Secondly, the CWT was applied up to 15 scales to both profiles by using daubechies 4 (db4) as mother wavelet. Thirdly, a summation point by point of the signals of all the 15 scales was calculated. Fourthly, a hard-thresholding wavelet based technique was applied to each signal. Finally, spots centers and boundaries were defined by calculating the local maxima and the local minima on both signals.

The proposed segmentation method is divided into three major steps: Firstly, à trous wavelet transform was applied up to second scale on the initial cell. Secondly, on the details coefficients, a hard threshold filter was carried out in order to suppress the noise. Finally, the initial image among the approximations and details of each scale were implemented in an ensemble scheme based on MRF model. As operators of the ensemble scheme were chosen: Majority Vote, Min, Product and Probabilistic Product.

The validation of the proposed algorithms was accomplished by a high quality simulated microarray image of 1040 cells with realistic morphological characteristics generated by using the Matlab microarray simulation model and fourteen real cDNA microarray images, seven 16-bit grayscale TIFF images of both channels (green and red), collected from the DERICI public database. In order to investigate the performance of the algorithms in presence of noise, the simulated image was corrupted with additive white Gaussian noise.

In the case of simulated image, the segmentation accuracy was evaluated by means of segmentation matching factor, probability of error and coefficient of determination in respect to the pixel actual classes (foreground-background pixels). In the case of real images the evaluation was based on Mean Absolute error, in order to measure indirectly their reliability.

According to our results in simulated cells, the proposed ensemble schemes managed to lead to more accurate spot determination in comparison to conventional
MRF model. Additionally, the majority vote operator managed to accomplish the highest score in all cases, especially on cells with high noise (SMF: 82.69%, PE: 6.60% and CV: 0.809), while the conventional MRF managed to gather the lowest score in all cases (SMF: 94.87% - 82.69%, PE: 3.03% - 9.85%, CV: 0.961 - 0.729). In the case of real images, the min operator achieved the lowest score (MAE: 803.96 and Normalized MAE: 0.0738) in contrast to majority vote, which reached the highest score among the proposed evaluating methods (MAE 990.49 and Normalized MAE 0.0738). Additionally, all the proposed algorithms managed to suppress MAE value compared to the conventional MRF segmentation model (MAE 1183.50 and Normalized MAE 0.0859).
Chapter 1

Introduction to Microarray Technology

1.1. Introduction

The Human Genome Project was formally launched in 1990 to develop detailed genetic and physical maps of the human genome, to locate and identify virtually all human genes and to determine the sequence of most or all of the human genome. The approach of the human genome project is based on examining the order of the basic building blocks of DNA (A, C, G and T in abbreviation) along the human genome. The project was completed in 2003 and guided to the discovery of 20,000-25,000 estimated human genes [1].

The information provided by the human genome project is expected to contribute to molecular biology which aims to understand the genes functional role, the relationship between them and how the changes in genes affect their functions in all cell process. To fulfil this task, scientists have developed new techniques and tools for conducting research. Complementary DNA microarrays are an example of those tools that uses genome sequence information to analyze the structure and function of tens of thousands of genes simultaneously. This global analysis allows the determination of genes cellular function, the nature and regulation of biochemical pathways, and the regulatory mechanisms during certain conditions or diseases.
Complementary DNA microarray experiments have been designed and applied in many different contexts and have covered the following areas\[^2\]:

- Differential gene expression between two or more sample types.
- Similar gene expression across treatments.
- Tumour sub-class identification.
- Classification of malignancies into known classes.
- Identification of genes that characterize different tumour classes.
- Identification of genes associated with clinical outcomes.

The process of biological discovery based on microarray technology, as it is illustrated in Figure 1.1, involving five tasks: Biological Hypothesis, Experimental design, Microarray Experiment, image analysis, normalization, and data analysis.

Figure 1.1: Schematically presentation of the involved tasks in biological discovery based on microarray technology\[^2\].
1.2. Biological Background

1.2.1. Deoxyribonucleic acid (DNA)

A DNA molecule is a double stranded linear polymer of nucleic acids that contains the genetic instructions specifying the biological development of all cellular forms of life. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. Most DNA is located in the cell nucleus but a small amount of DNA can also be found in the mitochondria. Within cell nucleus, DNA is organized into structures called chromosomes (see Figure 1.2).

Figure 1.2: DNA within cell nucleus [4].
Chemically, DNA molecules are composed of three subunits called nucleotides. Each nucleotide is comprised by a phosphate group, a deoxyribose sugar and one of four complementary nitrogen bases which correspond to Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). In base pairing, A always pair with T with 2 hydrogen bounds and G always pair with C, with 3 hydrogen bounds (see Figure 1.3). The order, or sequence, of these bases determines the information available for building and maintaining an organism.

![DNA Base Pairs](image)

Figure 1.3: DNA Base Pairs \(^5\).

Nucleotides are arranged in two long strands that form a spiral called a double helix as it shown in figure 1.4. The two strands are held together by weak hydrogen bonds between the complimentary bases. Along each strand, nucleotides are
connected by covalent bonds between the sugar of one nucleotide and the phosphate group of the next nucleotide. Each full turn of the DNA double helix has 10 nucleotide pairs.

An important property of DNA is that it can be replicated by itself. Each strand of DNA of the double helix serves as *template* for the reproduction of the opposite strand. The base sequences of the new strands are complementary in sequence to the template strands. This means that G, A, C, and T in the template are replaced to new strand with C, T, G, and A, respectively as it is shown in Figure 1.5.
1.2.2. **Ribonucleic Acid (RNA)**

RNA molecules have a chemical structure similar to that of DNA and they are synthesized by using DNA as template. Three major differences distinguish a RNA molecule from DNA. Firstly, deoxyribose has been replaced by ribose. Secondly, the pyrimidine *uracil (U)* has replaced the thymine (Figure 1.6) and base pairing is accomplished by following the rule that *Adenine (A)* always pair with *Uracil (U)* and *Guanine (G)* always pair with *Cytocine (C)*. Finally, RNA molecule is usual single-stranded molecule.

![Ribose chemical structure and Uracil chemical structure](image)

Figure 1.6: Ribose chemical structure and Uracil chemical structure[^5].

There are three major types of RNA that play key roles in protein synthesis[^5]:

- **mRNA**: messenger RNA carries coding information to ribosomes and functions as blueprint.
- **tRNA**: transport RNA is RNA that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein biosynthesis during translation.
- **rRNA**: ribosomal RNA is the central component of ribosome that provide a mechanism for decoding mRNA into amino acids.
1.2.3. Proteins

*Proteins* play crucial role in virtually all biological processes. Their biological functional role is depended on the three dimensional arrangement of the atoms of the molecule. They are composed of one or more chains of amino acids and as building blocks, twenty amino acids with unique shape and chemical properties are used.

Their biological activities are involved in enzymatic catalysis, transport and storage, coordinate motion, mechanical support, immune protection, and control of cell growth and differentiation. Additionally, the production of energy and the ability to respond to intra and extracellular stimuli are all protein dependent. A classification of proteins with respect to their activities is illustrated in table 1.1.

<table>
<thead>
<tr>
<th>TYPE OF PROTEIN</th>
<th>BIOLOGICAL ACTIVITY</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Carrier proteins</strong></td>
<td>Move molecules within and between cells or between the bloodstream and the lymphatic system.</td>
<td>Albumin; hemoglobin; lipoproteins; transferring</td>
</tr>
<tr>
<td><strong>Enzymes</strong></td>
<td>Organic catalysts that increase rate of chemical reactions without themselves being consumed.</td>
<td>Alcohol dehydrogenase; hexokinase; hyaluronidase</td>
</tr>
<tr>
<td><strong>G-proteins</strong></td>
<td>Transduce signals from outside the cell to the inside by stimulating production of second messengers.</td>
<td>Transducin; G; Q</td>
</tr>
<tr>
<td><strong>Information molecules</strong></td>
<td>Hormones and some other proteins that, when functioning, cause a particular metabolic or physiological activity in their target cells.</td>
<td>Insulin; glucagon; luteinizing hormone; prolactin</td>
</tr>
<tr>
<td><strong>Muscle proteins</strong></td>
<td>Interaction of actin and myosin causes muscle contraction.</td>
<td>Actin; myosin</td>
</tr>
<tr>
<td><strong>Protection molecules</strong></td>
<td>Protect against microbial invasion and toxins.</td>
<td>Antibodies; complement; interferons; interleukins</td>
</tr>
<tr>
<td><strong>Receptor proteins</strong></td>
<td>Transmembrane proteins that relay information transmitted by hormones and neurotransmitters and that are needed for endocytosis.</td>
<td>Insulin and adrenalin receptors on cell surfaces</td>
</tr>
<tr>
<td><strong>Regulatory proteins</strong></td>
<td>Control gene and cellular activities.</td>
<td>Intercellular hormone-receptors; repressors and Corepressors</td>
</tr>
<tr>
<td><strong>Structural proteins</strong></td>
<td>Provide shape, internal organization, and movement to cells and their components.</td>
<td>Cytochromes; cytoskeleton; histones; ribosomes</td>
</tr>
<tr>
<td><strong>Miscellaneous proteins</strong></td>
<td>Channel proteins, which allow movement of ions across membranes, and special proteins involved in cell-to-cell communication and transport.</td>
<td>ion channel proteins; chloride adhesion proteins;</td>
</tr>
</tbody>
</table>

Table 1.1: Classification of Proteins by Biological Activity \[6\].
1.2.4. Gene Expression

Genes contain two types of sequence segments, *coding* segments and *non-coding* segments. Coding segments are called *exons* and carry the genetic information to synthesize proteins, while non-coding segments are called *introns* and do not carry genetic information for protein synthesis. Exons and introns are interleaving within genes as it is shown in Figure 1.7.

![Gene segment: interleaving exons and introns](image)

Figure 1.7: Gene segment: interleaving exons and introns \(^7\).

A gene *is expressed* when its genetic information is transferred to mRNA and then to protein. The process of genetic information flowing from DNA into RNA and from RNA into protein is also known as the central dogma of molecular biology. Two major tasks are involved in the gene expression process: transcription and translation.

![The central dogma of molecular biology](image)

Figure 1.8: The central dogma of molecular biology, where R, T\(_C\), T\(_L\) donate the Replication, Transcription and Translation process respectively \(^6\).
During transcription, that takes place in the nucleus, one strand of the DNA is served as template in order to be synthesized a single-stranded mRNA molecule. The synthesis is based on the complementary base paring mechanism of nucleotides.

The amount of produced mRNA molecules is closely related to the gene expression levels. It is determined mainly by two regulatory DNA sequences: the promoter and the enhancer. The promoter indicates the transcription start site and the enhancer regulates the transcriptional efficiency. Regulation is accomplished by the action of the activator and repressor proteins binding to enhancers. The binding of enhancer to activators makes the corresponding genes raise the production of mRNA molecules and the binding to repressors decreases the amount of produced mRNA molecules.

Figure 1.9: Protein Synthesis. [8]
During the translation phase, the mRNA molecule enters cytoplasm. In the cytoplasm, the ribosome recognize the starting codon (AUG) and start to add amino acids, carried by the tRNA molecule, in the growing protein chain. The process is terminated when one of the three stopping codons (UAA, UGA and UAG) in the mRNA sequence is encountered. Finally, the amino acids are bound one by one to form the final protein form and the mRNA molecule is released.

1.2.5. Polymerase Chain Reduction (PCR)

The polymerase chain reduction is a laboratory procedure for amplifying trace amount of DNA exponentially to a sufficient quantity (Figure 1.10). This continuous doubling is accomplished by a primer, a small fragment of single stranded DNA, DNA nucleotides and finally enzymes that link nucleotides to form a long chain of DNA, known as DNA polymerases.

![Exponential amplification diagram](image)

Figure 1.10: Exponential amplification of trace amount DNA based on PCR. 

\[ 2^3 = 8 \text{ copies} \]
\[ 2^4 = 16 \text{ copies} \]
\[ 2^5 = 32 \text{ copies} \]
\[ 2^{35} > 68 \text{ billion copies} \]
The PCR procedure, as it is shown in Figure 1.11, involves three major tasks: **Denaturation, Annealing or hybridisation, Extension.** Firstly, the reaction solution is heated at 95 °C and the double stranded DNA is separated to two complementary strands, a process known as **Denaturation.** Secondly, the temperature is lowered at 55 °C that causes the primers bound to each DNA strand (**hybridisation**) and additional complementary nucleotides are attached to form more stable bonds by the DNA polymerases enzymes. Finally, the reaction solution is heated at 72 °C and DNA polymerases adds further nucleotides to the developing DNA strand (**extension**). Simultaneously, any loose bonds that have formed between the primers and DNA segments that are not fully complementary are broken. By repeating the process, many copies of the same segment of DNA can be obtained.

**Figure 1.11: Polymerase Chain Reduction Process**[^9].
A RNA molecule can also be amplified by *Reverse Transcription Polymerase Chains Reduction (RT-PCR)* procedure. The RNA molecule has to be firstly transcribed into DNA and the resulting single complementary DNA (cDNA) strand is amplified by PCR. The transformation from RNA to DNA is called *reverse transcription.* TRT-PCR is widely used in the diagnosis of genetic diseases and in the determination of the abundance of specific different RNA molecules within a cell or tissue as a measure of gene expression.

1.3. cDNA Microarray Images

A cDNA microarray is a collection of green and red discrete spots that containing DNA. Each spot is deposited on the surface of a microscope glass slide and occupies a small fraction on the acquired digital image. The corresponding fluorescence intensity value is closely related to the expression level of the genes.

The whole procedure is initialized by the selection of a set of genes of interest (ESTs) by the Biologist. For the purpose to produce a sufficient EST amount that can be printed in a high density array on a glass microscope slide, the ESTs are amplified by PCR.

The printing task is obtained by a robotic arrayer. An arrayer is a device that under a certain number of programmable pins deposits EST aliquots in an array form. An *aliquot* is a small quantity of premade nucleic acid. The number of pins as well as pins configuration varies from arrayer to arrayer, thus the resulting layout of a microarray experiment highly depends on the using arrayer.

In a typical microarray experiment, as it is illustrated in Figure 1.12 , preparative to compare the genetic expression in two cell samples the mRNAs of both samples are firstly isolated. The control sample contains mRNAs of a cell that is developed under normal conditions and the test or experimental sample contains mRNAs of a cell that has received a specific treatment.
The two mRNAs samples or targets are then reverse-transcribed into cDNA followed by a labeling procedure, where they are tagged with two different fluorescent dyes. The mRNA samples are typically tagged with Cyanine (Cy3) and (Cy5), fluorescent dyes with different wavelengths and visualized with green and red pseudo-color respectively. Afterwards, the two samples are mixed and hybridized with the arrayed DNA sequences or probes (Figure 1.13).
Two lasers in the wavelength of each dye are used to produce two scanned images. These scanners can be either sequential or dual laser. The former scanners perform a scan of the glass slide with one wavelength first and then with the other. The latter laser scanners perform a glass slide scan with both wavelengths simultaneously. A block diagram of a typical sequential cDNA microarray scanner is illustrated in Figure 1.14.

![Block diagram of cDNA microarray Scanner](image)

Figure 1.14: Block diagram of cDNA microarray Scanner\textsuperscript{[12]}.

A schematically presentation of the involved tasks during scanning is illustrated in Figure 1.15.

![Schematically presentation of the involved tasks during scanning in a cDNA microarray experiment](image)

Figure 1.15: Schematically presentation of the involved tasks during scanning in a cDNA microarray experiment\textsuperscript{[12]}. 
The scanning procedure followed can be described in the following steps:\textsuperscript{[12]}

- The laser scanner excites the hybrids, with Cyanine dye Cy3 and Cy5 that emit at the range of 510-550 nm and 630-660 nm respectively.
- The excitation photons are absorbed by the fluorescence molecules, which leads to the emission of fluorescence photons.
- Objective lens gathers a fraction of the emitted photons, which travel at any direction.
- A dichroic beamsplitter and a band-pass filter are usually put in front of the detector in order to discriminate excitation and emission photons and to suppress the probability of excitation photons detection.
- A pinhole is placed in front of the detector in order only the glass slide emitted photons to be detected.
- The detected photons are then converted into electric current by using Photomultipliers tubes (PMT). The number of electrons that are produced per photon is controlled by the PMT’s voltage input.
- The electric current is converted into digital signal by an analog to digital converter (A/D).
- Finally, in the digitization step the produced signal, for each pixel, represents the total fluorescence of the dye molecules that corresponded to pixel glass region.

At the end of the scanning procedure, two images are generated (usually produced in 16 bit Tagged Image Format) reporting the fluorescence of hybrids on the slide. The fluorescence intensities notify the relative presence of a hybrid compound in one sample and not in the other or vice versa. Thus, the change and the level of change in mRNA expression between the two samples can be evaluated. The final hybridized image is obtained by merging the two generated images.
1.3.1. Ideal cDNA image characteristics

An ideal cDNA microarray image is a high resolution image with the number of pixels per spot reaching infinity in a theoretical level. An example of a theoretical ideal image consisted of four sub-grids and containing 64 spots in total is illustrated in Figure 1.16.

An ideal image is characterized by the followings \[^{13}\]:

- Deterministic grid geometry.
- Zero uncertainty about the background intensity.
- Defined spot shape.
- Constant spot intensity with zero uncertainty that is different from the background intensity and is directly proportional to the biological phenomenon.
- High statistical confidence in microarray measurements due to large number of pixels per spot.

![Ideal Microarray Image](image-url)
1.3.2. Real cDNA image characteristics

The generation of real cDNA images is based on a complex process that provokes several image variations that are mainly observed on the foreground and background, or even on the spot intensity value.

The most common variations are summarized in the followings \cite{10,11}:

- **Variations of microarray image channels:** The dynamic range, as it is illustrated in Figure 1.17, corresponds to maximum minus minimum amplitude. Any value outside the dynamic range is mapped with either the minimum or maximum value. The uncertainty of each intensity measurement increasing for a fixed number of bytes and expanded dynamic range.

![Figure 1.17: uncertainty and dynamic range dependencies in cDNA images](image)

Figure 1.17: uncertainty and dynamic range dependencies in cDNA images \cite{13}. 


Variations of grid geometry: Any rotational slide offset produces a rotated grid in microarray image. An example of rotational of grid is illustrated in Figure 1.18.

![Rotational grid of cDNA image](image1.png)

Figure 1.18: Rotational grid of cDNA image\textsuperscript{[13]}.

Variation of background and foreground intensities: low differences between the intensities of background and foreground affect their discrimination. An example of this case is illustrated in Figure 1.19.

![cDNA image with low discrimination power](image2.png)

Figure 1.19: cDNA image with low discrimination power\textsuperscript{[13]}.  

\[
\text{Variations of grid geometry: Any rotational slide offset produces a rotated grid in microarray image. An example of rotational of grid is illustrated in Figure 1.18.}
\]

\[
\text{Variation of background and foreground intensities: low differences between the intensities of background and foreground affect their discrimination. An example of this case is illustrated in Figure 1.19.}
\]
Variation of spot morphology: the most common spot shape is circular. However, a large number of shape deviations are existed. Six cases of the most common spot variation are presented in Figure 1.20.

Figure 1.20: Morphological variations in cDNA images\textsuperscript{[13]}.  

Variations of slide preparation: insufficient rapid immersion of the slides in the succinic anhydride blocking solution produces Comet tails. An example is illustrated in Figure 1.21.

Figure 1.21: Comet tails in cDNA images\textsuperscript{[14]}.  
Variations of background: Inappropriate following procedures (dust or dirty), and the instrumentation noise are few of the responsible factors for background variations. An example of cDNA image with locally high background is illustrated in Figure 1.22.

Figure 1.22: cDNA image with locally high background [13].
Chapter 2

Microarray Image Processing

2.1 Introduction

Image processing has a great impact on clustering and identification of differential expressed genes. The process is based on measuring the mean fluorescence intensity of each spot, which is closely related to the transcript abundance in the two labeled mRNA samples.

The main process for measuring spot intensity values is sub-divided into three major steps:  

- **Addressing or Gridding**: the process of assigning coordinates to each cell.
- **Segmentation**: where each cell pixel is classified to either as foreground or background.
- **Intensity extraction**: where for each cell of the array, intensity values for foreground and background are calculated.

In the past years, several efforts have been made and several software tools have been developed for the task of processing microarray images. The most commonly used are ScanAlyze [16], GenePix [17], ImaGene [18], Spot [19] and MAGIC Tool [20].
2.2 Gridding or Addressing

The gridding procedure is the first task of image analysis procedure. The objective of this step is an approximation of cell locations by determining spot center and its corresponding boundaries. Each cell is represented as a rectangular image which enclosing the spot. However in order to address the cells on a microarray image several parameters have to be estimated. These parameters must include information about:\footnote{13}

- Displacement of grids or spots from the expected position caused by slight variations in print-tip position.
- Small individual translation of spots.
- Separation between rows and columns of grids.
- Separation between rows and columns of spots within each grid.
- Overall position of the microarray in the scanned image.

The existing methods can be characterized as \textit{manual}, \textit{semi-automatic} and \textit{fully-automatic}. The reliability of these methods is mainly depended on the users’ level of interference.

2.2.1 Manual Method

Manual gridding methods are based on a computer inference and high level user intervention. The users, in most cases, have to specify the radius of each spot and the dimensions of a grid template. This template is then adjusted to spots layout, in order to achieve a match between them.

The major drawback of this method is that they are extremely time-consuming and tedious. Thus, a reproduction of manual gridding is practically impossible. However, this method is characterized with high reliability due to the fact that a perfect grid alignment could be obtained.
2.2.2 Semi-Automatic method

Semi-automatic methods are not exclusively depended on user interference. The process is firstly initialized by the user and carried out with an automatic gridding procedure. The automatic component is based on either grids template with the matching to be based on image correlation techniques or with data-driven techniques that assumes the intensity of foreground as heterogeneous and background as homogeneous. These methods are reliable with efficient repeatability for high-throughput image analysis.

An example of MAGIC Tool software using a semi-automatic method is illustrated in figure 2.1. Procedure is firstly initialized by user inputs about the first top left, right spot and with additional information about the number of lines and the coordinates of the last grid line. Based on these inputs a grid template is created and the adjustment is made by successive updates.

![Semi-automatic gridding method as it is used in MAGIC Tool](image.png)

Figure 2.1: Semi-automatic gridding method as it is used in MAGIC Tool [20].
2.2.3 Fully-automatic method

These methods need minimum human interference. All parameters, which were a prior knowledge in the previous methods, have to be calculated. There are two existed methods for the representation of these estimated grids (Figure 2.2). The first one consists of horizontal and vertical lines passing through the estimated centers of the spots, named as fitted foreground grids. The second one consists of horizontal and vertical lines passing through the estimated centers of the boundaries between the spots and it is named as fitted background grids.

A reliable fully-automatic gridding method has to take into account the following variations of cDNA microarray images:

- **Variation of grid alignment**: rotational grid
- **Variation of background and foreground intensities**: missing lines or spots in a microarray image layout.

![Figure 2.2: Fully automatic representation of estimated grid based on: (a) fitted foreground grid and (b) fitted background grid.](image)
2.3 Segmentation

Following the gridding procedure, the next task is the classification of produced cells pixels as *foreground*, spots area of interest, or *background*, so that fluorescence intensities can be calculated for each spotted cDNA sequence.

The accuracy of this task has a substantial impact on the accuracy and the effectiveness of gene expression analysis. Thus, a reliable segmentation process has to take into account the followings\textsuperscript{[13],[15]}:

a. the variations of block and spot positions,
b. the spot shape, which is not perfectly circular
c. the existence of non-expressed spots that have zero intensity
d. dust or other contamination on the slide generates artifacts in the image.

The outcome of segmentation task for each cell is a binary image, known as *cell mask*. Each pixel that is assigned to background is usually represented with zero value while pixels assigned to foreground to be represented with “one”. An example of the resulting spot mask is represented in Figure 2.3.

![Cell Mask](image)

Figure 2.3: Representation of typical cell mask \textsuperscript{[21]}.

Several methods have been proposed for the segmentation task. Some of these have already been adopted by several image processing (analysis) software tools. A summary of the existing segmentation methods with respect to their software
implementation is shown in table 2.1. The most recent segmentation methods involve clustering algorithms. These methods have not been, yet, employed by any existed image processing or analysis tool.

The existed segmentation methods can be categorized, in general, into the following categories [22]:

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

<table>
<thead>
<tr>
<th>Segmentation Methods</th>
<th>Image Processing tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle shape segmentation</td>
<td>ScanAlyze, GenePix, QuantArray, Dapple</td>
</tr>
<tr>
<td>Adaptive shape segmentation</td>
<td>Spot</td>
</tr>
<tr>
<td>Histogram based segmentation</td>
<td>Imagen, QuantArray, DeArray</td>
</tr>
</tbody>
</table>

Table 2.1: Existed segmentation methods and their implementations on existed image processing software tools.

2.3.1 Circle shape segmentation

2.3.1.1 Fixed circle segmentation

Fixed circle (FC) segmentation method is based on the assumption of circular shape with the same diameter for all spots. It is using a circle with constant diameter to separate the foreground from the background. The pixels inside the circle is assigned to foreground, whereas the rest is assigned to the background.
However, in real cDNA microarray images the shape and the size of each spot can vary. Thus, the accuracy of this method is adequate mainly for circular spots with approximately same size. An example of FC implantation on a cDNA microarray image is illustrated in Figure 2.4.

![Fixed Circle implementation on a cDNA image](image)

Figure 2.4: Fixed Circle implementation on a cDNA image [22].

### 2.3.1.2 Adaptive circle segmentation

Adaptive circle (AC) segmentation method is based again on the assumption of circular shape for all spots with the center and the diameter of each spot to be estimated. It is a two-step process [13].

- Firstly, the center of each spot has to be estimated.
- Secondly, the circle diameter has to be adjusted to each spot.
The diameter adjustment it could be accomplished by using a typical edge detection technique. The major drawbacks are summarized in the circular shape assumption and the adjustment that has to be made separately for every spot. An example of AC segmentation method implementation on cDNA microarray image is illustrated in Figure 2.5.

Figure 2.5: Implementation of Adaptive circle segmentation on a microarray image [22].

### 2.3.2 Adaptive shape segmentation

There are two methods for adaptive segmentation: watershed and seeded region growing (SRG) techniques (Figure 2.6). Both techniques are based on the specification of a starting point-seed for the foreground and background determination. The nearest neighbor pixel of a region is classified to that region according to the difference between the pixel intensity value and the mean intensity
value of the adjoining region. The process is repeated until all pixels have been assigned to one of the two regions. The main drawback of these techniques is the uncertainty of the way the number and the location of the starting seeded points are selected.

![Figure 2.6: Implementation of Seed Region Growing on a microarray image](image_url)

2.3.3 Histogram-based segmentation

There are two Histogram based segmentation methods. The first one uses a circular target mask that covers all spot area and a threshold derived from the pixels, within mask, intensity values. The determination of background and foreground regions is obtained by assigning the pixels with intensity value greater than the estimated threshold to the foreground region. The threshold estimation is accomplished by using statistical tests.

The second method uses a squared target mask that covers all spot area. The foreground and the background determination is obtained by using the histogram
percentile values that range from the 80\textsuperscript{th} to the 95\textsuperscript{th} value and the 5\textsuperscript{th} to the 20\textsuperscript{th} respectively.

The size of the used target mask has a subsequent impact on the accuracy of the segmentation methods. The basic disadvantage of these methods is summarized in that large target masks can overlap with neighbor spots and in contrary small target masks might not cover the hall spot.

![Image](image.png)

Figure 2.7: Histogram based segmentation \cite{22}.

### 2.4 Intensity Extraction

The last step in image processing procedure is the calculation of the foreground and background intensities with respect to the resulting binary cell mask. The total foreground intensity is estimated by using the either the mean or median fluorescence intensity within the estimated spot. In the histogram based segmentation methods the foreground and the background are obtained directly by the histogram of the cell.

Nevertheless, the estimated foreground is not totally independent from the background variations of a real cDNA microarray image. The correction is obtained by subtracting the estimated total foreground intensity from the mean or median fluorescence intensity of the background.
There are several proposed and employed methods for background estimation. In Scanlayze, the local background is considered as the median value of pixels that are within a square center at the spot centre. In QuantArray, the area between two concentric circles is assigned to the background. In Spot, four diagonal diamond-shaped areas donate the background. An example of these methods is illustrated in Figure 2.8, where the region inside the red circle indicates the spot mask. Green, blue and pink areas indicate the background estimation regions by using QuanArray, ScanAlyze and Spot package respectively.

![Image illustrating different background estimation methods](image.png)

Figure 2.8: Image illustrating different background estimation methods \[22\].

### 2.5 Normalization

The purpose of normalization is to minimize the systematic variations that effects the measured fluorescent intensities, so that the biological differences to be more easily distinguished and a comparison between them across slides of the same experiment to be possible. The existed normalization techniques involves global, local and scale techniques \[23,24\].
- **Global Normalization**: Assumes that the measured intensities of red and green channel are related by a constant factor.

- **Location Normalization**: The correction is based on a robust locally regression, where a linear or quadratic lowess function is locally fitted to data with respect to the distance between them.

- **Scale Normalization**: The logarithmic ratio of the red and green channel, across images of the same experiment, follow normal distribution with zero mean and different variance and thus an adjustment to the scale is essential. The within-print normalization will center log ratios around zero. The normalized values are computed as:

\[ S_i \times \log_2 R/G - L_i(A) \]  

(2.1)

Where \( S_i \) is the scale factor for print tip \( i \) using Median Absolute Deviation and calculated according to equation (2.2).

\[ S_i = \frac{MAD_i}{\sqrt{\prod_{i=1}^{t} MAD_i}} \]  

(2.2)

The Median Absolute Deviation (MAD) for each print tip is calculated according to equation (2.3).

\[ MAD_i = M_{ij} - \text{median} \ M_{ij} \]  

(2.3)

Where \( M_{ij} \), represents the \( j \)th log ratio in the \( i \)th print tip.
In the present study, spot location was accomplished by using an automatic gridding method based on continues wavelet transform. The proposed segmentation method is based on an ensemble scheme developed by using the à trous wavelet transform up to second scale in combination with Markov Random Field (MRF) model. As operators of the ensemble scheme were chosen: Majority Vote, Min, Product and Probabilistic Product.

3.1 Addressing

3.1.1 Continues Wavelet Transform (CWT)

The Continues Wavelet Transform is used for signal decomposition into small oscillations that are highly localized in time, named as wavelets, according to equation (3.1):

$$C(a,b) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{|a|}} \Psi\left(\frac{t-b}{a}\right) dt$$

(3.1)

Where, C is the corresponding wavelets, f(t) is the original signal, a is the scale, b refers to time or space and \( \Psi \) to chosen mother wavelet. A three dimensional representation of cwt coefficients is illustrated in Figure 3.1, with x axis to refer to time, y axis to calculated coefficients and z axis to scale.
3.1.2 Addressing Procedure

Spot localization was accomplished as followed \cite{30}: firstly, line-profiles for $x$ and $y$ axes were calculated by finding the mean intensity value of rows and columns. Secondly, the Continues Wavelet Transform (CWT) was applied up to 15 scales to both profiles by using daubechies 4 ($db4$) as mother wavelet \cite{26}. Thirdly, a summation point by point of the signals of all the 15 scales was performed for the purpose of increasing the actual signal of the spots and decreasing the noise. Fourthly, a hard-threshold filter was applied to each signal for the task of further noise suppression. Finally, spots centers and boundaries were defined by calculating the local maxima and the local minima on both signals. In Figure 3.2 is illustrated the result of the gridding step.

Figure 3.2: (a) Original simulated microarray image and (b) gridding result.
3.2 Segmentation

The proposed method is divided into three major steps. Firstly, the à trous wavelet transform (AWT) was applied up to second scale on the initial cell (spot with its background as it was determined in the gridding step). Secondly, on the details coefficients, a hard threshold filter was carried out for noise suppression. Finally, the initial cell along with the approximations and the details of each scale were implemented in an ensemble MRF scheme. As operators of the ensemble scheme, majority vote, min, product, probabilistic product were chosen. The whole procedure is schematically illustrated in Figure 3.3.

Figure 3.3: Schematically approach of the proposed ensemble scheme based on à trous wavelet transform up to second scale and MRF model.
3.2.1 À-trous Wavelet Transform (AWT)

The à-trous algorithm represents the discrete convolution between the input signal and a properly chosen mother wavelet, which refers to a low-pass filter. During discrete convolution, the distance between a central pixel and adjacent ones is $2^{j-1}$ [28] (Figure 3.4).

![Diagram of à-trous algorithm](image)

Figure 3.4: Distance of central pixel and the adjacent ones during two levels of à-trous decompositions [31].

In the present study, the B3-Spline scaling function was chosen as mother wavelet to process the sampled data $\{c_0(k)\}$. The approximations coefficients $\{C_j(k)\}$ (at resolution $j$ and position $k$) were calculated by using equation (3.2) [28]:

$$C_{j+1}(k) = \frac{1}{16} C_j(k - 2^{j+1}) + \frac{1}{4} C_j(k - 2^j) + \frac{3}{8} C_j(k) + \frac{1}{4} C_j(k + 2^j) + \frac{1}{16} C_j(k + 2^{j+1})$$ (3.2)

The details coefficients $\{w_j(k)\}$ were calculated by subtracting the approximations of two sequential resolution levels (equation 3.3):

$$w_j(k) = c_{j-1}(k) - c_j(k)$$ (3.3)

An example of AWT implementation on a signal containing 1000 points is illustrated in Figure 3.5. The corresponded wavelet coefficients of the first scale have the same size as the initial signal.
In order this algorithm to be extended to two dimensions, a row by row followed by a column by column convolution was performed. The extremities of the signal were handled by using the mirror effect technique \[^{33}\].

After applying the AWT, a hard-threshold wavelet based technique was performed according to equation (3.4).

\[
\begin{align*}
W_{\text{out}} &= W_{\text{in}} + T(G - 1) & W_{\text{in}} > T \\
W_{\text{out}} &= W_{\text{in}} - T * (G - 1) & W_{\text{in}} < -T \\
W_{\text{out}} &= 0 & \text{otherwise}
\end{align*}
\]

where \(W_{\text{out}}\) denotes the output and \(W_{\text{in}}\) the input coefficient values of the details. \(T\) and \(G\) are threshold and gain values respectively (Figure 3.6) \[^{33}\].

Figure 3.5: AWT implementation on a signal containing 1000 points \[^{29}\].

Figure 3.6: Hard-Threshold Function \[^{33}\].
### 3.2.2 Markov Random Field Model

The MRF segmentation model is based on the probability theory in order to detect and analyze spatial or contextual dependencies of physical phenomena. Furthermore, probabilistic distributions of interacting labels can also be established for visual labeling [34]. According to MRF, the probability of a pixel belonging to foreground or background (label) \( y \), given that it has a value (feature) \( f \), follows the Gibbs distribution as it is described in equation (3.5)\([35,36]\):

\[
P(label = y \mid feature = f) = \frac{1}{Z} e^{-\frac{E}{T}}
\]  

(3.5)

where \( T \) is a temperature constant, \( E \) is an energy function given by equation (3.6) and \( Z \) is a normalized constant determined by equation (3.7).

\[
E = E_{Labels} + aE_{Features}
\]  

(3.6)

\[
Z = \sum_{all\ pixels} e^{-\frac{E}{T}}
\]  

(3.7)

The determination of individual contributions of \( E_L \) (Labels energy) and \( E_F \) (Feature energy) is obtained by using a weighted parameter \( a \) (equation (3.8)), which is able to achieve a proper balance between the two components [30].

\[
a(t) = c_1 \cdot \gamma^t + c_2
\]  

(3.8)

where \( \gamma, c_1 \) and \( c_2 \) are constants and set to 0.9, 80 and 0.5 respectively. Finally, \( t \) represents the number of interactions.

In present study, the first order pairwise Multilevel Logistic (MLL) (equation 3.9) model [35] was utilized with the purpose of quantifying the label’s distribution, taking into consideration the neighboring pixels.

\[
E_L = \sum_{i \in \text{Neighbors}} \beta \cdot \delta(y_i, y_{center})
\]  

(3.9)
where $\beta$ is a constant set to 1, $\delta(y_i, y_{\text{center}})$ returns -1 if the labels have the same value or 1 otherwise:

$$
\delta(y_i, y_{\text{center}}) = \begin{cases} 
-1, & \text{if } y_i = y_{\text{center}} \\
+1, & \text{if } y_i \neq y_{\text{center}}
\end{cases}
$$

(3.10)

The feature energy for each class considering Gaussian intensity distribution is given by the following equations:

$$
E_{F1(\text{center})} = \sum_{i \in \text{neighbours}} \left( \frac{(f_i - \mu_1)^2}{2(\sigma_1)^2} + \log(\sqrt{2\pi}\sigma_1) \right)
$$

(3.11a)

$$
E_{F2(\text{center})} = \sum_{i \in \text{neighbours}} \left( \frac{(f_i - \mu_2)^2}{2(\sigma_2)^2} + \log(\sqrt{2\pi}\sigma_2) \right)
$$

(3.11b)

where $\mu_1/\mu_2$, $\sigma_1/\sigma_2$ are considered to be the mean and standard deviation value of the background and the foreground respectively.

A summary of involved steps in MRF based segmentation method is illustrated in Table 3.1:

1. Apply $k$-means clustering to the initial image to obtain 2 classes (Foreground and Background Binary image)
2. Calculate the Mean Value $\mu$ and the Standard Deviation $\sigma$ of the 2 classes by using as feature the pixel intensity.
3. Calculate the Energy of the labels ($E_L$) (3x3 moving window on the binary image).
4. Calculate 2 Feature Energies ($E_{F1}$, $E_{F2}$), of the 2 classes on the actual image.
5. Calculate the total Energies.

$$
E_L = E_L + a E_{F1} \quad E_L = E_L + a E_{F2}
$$

6. Calculate the posteriori probabilities (according to Gibbs distribution).
7. Update the labels of the binary image and repeat the process from step 2.

Table 3.1: Steps of the MRF based segmentation method.
3.2.3 Majority Vote Operator

Following the implementation of MRF segmentation model to the initial cell and its wavelet coefficients, the corresponding binary cells were applied into a majority vote ensemble scheme (see Figure 3.3). The ensemble decision for each pixel of the cell was made according to the plurality rule\[^{[37],[39]}\], where the most represented class label was assigned to the corresponding pixel on the resulting binary cell.

An example of the proposed ensemble scheme based on plurality rule is illustrated in Figure 3.7.

Figure 3.7: Example of Majority Vote Ensemble Scheme.
3.2.4 Min and Product Operators

After the implementation of initial cell and its corresponding wavelet coefficients to MRF segmentation model, the corresponding probabilistic outputs for each pixel were organized in a decision profile (DP) matrix as followed \cite{37,39}:

\[
DP(x) = \begin{pmatrix}
P_{1,1} & P_{2,1} \\
P_{1,2} & P_{2,2} \\
P_{1,3} & P_{2,3} \\
P_{1,4} & P_{2,4} \\
P_{1,5} & P_{2,5}
\end{pmatrix} \rightarrow \text{MRF output} \quad (3.12)
\]

Where, the columns of the DP matrix corresponded to the number of labels and the rows to the probabilistic output of the applied MRF.

The ensemble decision for each pixel was based on the maximum value of the minimum and product aggressive rule output (equation 3.13). The whole procedure is repeated until all the cells pixels have been assigned to one of the two classes (background or foreground). An example of the aggressive rule implementation is illustrated in Figure 3.8, with x and D_L to donate each cell pixel and the number of applied MRF respectively.

\[
\text{Label} = \max(\text{Aggressive Rule}(DP(x))) \quad (3.13)
\]

Figure 3.8: Ensemble decision based on the aggressive rule \cite{37}.
3.2.5 Probabilistic Product Operator

In the case of probabilistic operator the followed procedure is similar to the previous one. Firstly, the probabilistic outputs of the applied MRF to the initial cell and its corresponding wavelet coefficients were organized into DP matrix (equation 3.12). The ensemble decision for each pixel of the cell was obtained according to equation (3.14)\cite{37,38}:

\[
Label = \max_{j=1}^{c} \left( \prod_{i=1}^{L} \frac{DP(i,j)}{P(j)^{L-1}} \right) \tag{3.14}
\]

Where \( c \) and \( L \) are constants and correspond to the class total number and the number of applied MRF respectively. \( P(j) \) donates the prior probability for each class (equation 3.15).

\[
P(j) = \frac{N_j}{N} \tag{3.15}
\]

Where \( N_j \) and \( N \) are the number of pixel that belong to class \( j \) and the total number of cell pixels respectively.

For the calculation of prior probabilities a binary mask, representing the average segmented spot shape, was produced by calculating the average binary cell of the MRF outputs. An example of a belief mask is illustrated in Figure 3.9.

![Binary Mask][21]

Figure 3.9: Binary Mask [21].
3.3 Intensity Extraction

The final step of the process was to compute the representative spot intensity by subtracting the mean of the Foreground from the Background (equation 3.16).

\[ I = \mu_{FG} - \mu_{BG} \]  \hspace{1cm} (3.16)

Where \( \mu_{FG} \) and \( \mu_{BG} \) are the mean foreground and mean background respectively, both calculated from the ensemble decision labeled cell pixels. The flow chart of the followed procedure is illustrated in Figure 3.10.

![Intensity Extraction Flow chart](image)

Figure 3.10: Intensity Extraction Flow chart.

3.4 Materials

The validation of the described algorithms was accomplished on both simulated and real microarray images. In order to produce microarray images with realistic morphological characteristics the MATLAB microarray simulator toolbox \(^{35-36}\) was used. The simulator toolbox includes all the steps that affect the quality of real microarray images. The structure of the model is presented in figure 3.11.
The layout and the quality of the produced images are determined by the user input data. The user can specify information about the spot, size, location, and additional information about the type of implemented errors on the image during scanning and hybridization.

In the present study, one high quality microarray image of 1040 spots with low variability of spot size and shape was produced. The noise level was reasonably low. In figure 3.12 is illustrated a sub-grid of the produced cDNA microarray image. Additionally, in order to investigate the performance of the algorithms in presence of noise, the produced image was corrupted with additive white Gaussian noise (zero mean value and variance ranging from 0.001-0.012).
Finally, the proposed segmentation ensemble schemes were applied on fourteen real cDNA microarray images, seven 16-bit grayscale TIFF images of red channel and seven from the green channel, collected from public database of the MicroArray Genome Imaging & Clustering Tool website (MAGIC tool)\cite{20}. Each image contained 6400 spots investigating the diauxic shift of Saccharomyces cerevisiae\cite{42}. Images included spots of various shapes as well as artifacts (scratches and dust).

3.5 Performance Evaluation
3.5.1 Evaluation of Simulated images

The segmentation accuracy was numerically calculated by using the segmentation matching factor (SMF) (equation 3.17)\cite{35} for every binary cell, produced by the proposed algorithm and conventional MRF segmentation model.

\[
SMF = \frac{A_{\text{cal}} \cap A_{\text{real}}}{A_{\text{cal}} \cup A_{\text{real}}}
\] (3.17)

where \(A_{\text{cal}}\) denotes the spot area as calculated by each segmentation model and \(A_{\text{real}}\) is the actual spot area. A perfect match is obtained by a 100% score. Any score higher than 50% indicates reasonable segmentation whereas, a score of less than 50% indicates poor segmentation\cite{35}.

Additionally, the coefficient of determination in respect to the actual classes (foreground-background pixels) of the simulated spots, was calculated according to equation (3.18):

\[
R^2 = \frac{\sum_{i=1}^{\text{All spots}} (I_{\text{segment}(i)} - \bar{I}_{\text{actual}})^2}{\sum_{i=1}^{\text{All spots}} (I_{\text{actual}(i)} - \bar{I}_{\text{actual}})^2}
\] (3.18)
where $I_{\text{segment}}$ represent the calculated intensity derived from the segmentation model, $I_{\text{actual}}$ donates the actual spot Intensity value and $I_{\text{actual}}$ is the mean intensity derived from all spots. The coefficient of determination is ranging from 0-1. Score close to 1 denotes high performance of the applied algorithm \cite{44}.

Finally, the probability of error was calculated according to equation 3.19\cite{35}:

$$PE = P(F) \times P(B/F) + P(B) \times P(F/B)$$

(3.19)

Where, $P(B/F)$ is the probability of error in false classifying foreground as background, $P(F/B)$ is the probability of error in classifying background as foreground, $P(F)$ is the prior probability of the foreground and $P(B)$ is the prior probability of background.

### 3.5.2 Evaluation of Real images

The Mean Absolute Error (MAE) of each spot within the segmented cell was numerically calculated according to equation (3.20) \cite{43}:

$$\text{MAE}_{\text{spot}} = \frac{1}{n} \sum_{i=1}^{n} |I_i - \bar{I}|$$

(3.20)

Where, $n$ is the total number of images, $I_i$ donates the mean fluoresce intensity value of each spot and $\bar{I}$ is the mean intensity derived from all spots across the used images.
4.1 Results and Discussion

Following the gridding procedure, the conventional MRF along with the proposed ensemble schemes were applied on the determined cells. Thus, a set of six binary images, having all spot boundaries delineated, were produced. Each binary image corresponded to a respective applied algorithm.

Based on the produced binary images and the ground truth binary image of the gold standard microarray image, the segmentation matching factor was calculated for all cases, in order to estimate the accuracy of each method quantitatively. According to our results, the ensemble scheme based on majority vote operator managed to accomplish the highest score in all cases, especially on cells with high noise (SMF: 88.21%). The conventional MRF model achieved better segmentation results at low levels of noise (SMF: 94.87%) than in the present of strong noise (SMF: 82.69%), where its segmentation ability reached the lowest score. The remaining operators proved to lead to more accurate spot determination in comparison to conventional MRF model. Results obtained in the presence of noise are illustrated in Table 4.1 and figure 4.1.
Additionally, in order to evaluate the pixel level accuracy of the segmentations methods the probability of error, which measures the mis-segmented pixels according to the prior knowledge of the pixels actual labels, was also calculated (see equation 3.19). Regarding pixel-based segmentation accuracy on the simulated images, in terms of mean probability of error, the proposed majority vote ensemble scheme managed to achieve the lowest score (PE: 1.98%). Even though the image quality of the evaluated images varied significantly, due to the additional noise implementation,
the accuracy of the proposed methodology remained high (PE Majority: 6.60%, Min: 8.01%, Product: 7.27% and Probabilistic Product: 7.27%). Furthermore, the conventional MRF segmentation model, according to table 4.2, achieved the highest mean probability of error and thus the lowest pixel-based segmentation accuracy in all cases (PE: 3.03-9.85%).

<table>
<thead>
<tr>
<th>Applied Method</th>
<th>Original Spot</th>
<th>White Gaussian Noise (v=0.001)</th>
<th>White Gaussian Noise (v=0.005)</th>
<th>White Gaussian Noise (v=0.008)</th>
<th>White Gaussian Noise (v=0.012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional MRF</td>
<td>3.03</td>
<td>4.07</td>
<td>6.70</td>
<td>7.70</td>
<td>9.85</td>
</tr>
<tr>
<td>Majority Vote</td>
<td>1.98</td>
<td>2.71</td>
<td>4.46</td>
<td>5.03</td>
<td>6.60</td>
</tr>
<tr>
<td>Min</td>
<td>2.38</td>
<td>3.21</td>
<td>5.35</td>
<td>6.17</td>
<td>8.01</td>
</tr>
<tr>
<td>Product</td>
<td>2.22</td>
<td>2.96</td>
<td>4.93</td>
<td>5.55</td>
<td>7.27</td>
</tr>
<tr>
<td>Probabilistic Product</td>
<td>2.22</td>
<td>2.96</td>
<td>4.94</td>
<td>5.55</td>
<td>7.27</td>
</tr>
</tbody>
</table>

Table 4.2: Percentages of mean probability of error in all cases.

In figure 4.2 is illustrated the graphical representation of the calculated results of the mean probability of error for 1040 simulated spots in the presence of noise.

Figure 4.2: Graphical representation of the calculated mean probability of error.
In order to measure the correlation between actual intensity value and calculated intensity value for each of the applied segmentation methods on the simulated cells, the coefficient of determination was also calculated. According to table 4.3, the Majority vote operator resulted to the highest score among all cases (CV: 0.975-0.809) while conventional MRF segmentation method gathered the lowest score (CV: 0.961-0.729). Moreover, the remaining operators proved to be more reliable in all examined cases in comparison to conventional MRF model. In figure 4.3 is illustrated a graphical representation of the calculated coefficient of the determination.

<table>
<thead>
<tr>
<th>Applied Method</th>
<th>Original Spot</th>
<th>White Gaussian Noise (v=0.001)</th>
<th>White Gaussian Noise (v=0.005)</th>
<th>White Gaussian Noise (v=0.008)</th>
<th>White Gaussian Noise (v=0.012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional MRF</td>
<td>0.961</td>
<td>0.934</td>
<td>0.843</td>
<td>0.805</td>
<td>0.729</td>
</tr>
<tr>
<td>Majority Vote</td>
<td>0.975</td>
<td>0.956</td>
<td>0.892</td>
<td>0.865</td>
<td>0.809</td>
</tr>
<tr>
<td>Min</td>
<td>0.969</td>
<td>0.947</td>
<td>0.870</td>
<td>0.837</td>
<td>0.764</td>
</tr>
<tr>
<td>Product</td>
<td>0.970</td>
<td>0.947</td>
<td>0.873</td>
<td>0.843</td>
<td>0.773</td>
</tr>
<tr>
<td>Probabilistic Product</td>
<td>0.970</td>
<td>0.947</td>
<td>0.873</td>
<td>0.843</td>
<td>0.774</td>
</tr>
</tbody>
</table>

Table 4.3: Calculated results of the coefficient of determination for all applied methods and cases.

Figure 4.3: Graphical representation of the calculated coefficient of determination for all applied methods and cases.
In Table 4.4 are illustrated comparative results of four randomly chosen cells of the G channel. The first column indicates the simulated spot, the second column indicates the actual boundaries of the spot and the rest columns present the resulting binary cells of the applied segmentation methods with their corresponding matching factors and probability of error scores.

<table>
<thead>
<tr>
<th>Original Cell</th>
<th>Actual Boundaries</th>
<th>Conventional MRF</th>
<th>Majority Operator</th>
<th>Min Operator</th>
<th>Product Operator</th>
<th>Prob. Product operator</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
</tr>
<tr>
<td>SMF=84.81 PE=7.10</td>
<td>SMF=94.52 PE=2.37</td>
<td>SMF=39.33 PE=4.73</td>
<td>SMF=90.67 PE=4.14</td>
<td>SMF=90.67 PE=4.14</td>
<td>SMF=90.67 PE=4.14</td>
<td>SMF=90.67 PE=4.14</td>
</tr>
<tr>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
</tr>
<tr>
<td>SMF=87.67 PE=5.77</td>
<td>SMF=95.83 PE=1.92</td>
<td>SMF=93.15 PE=3.21</td>
<td>SMF=95.83 PE=1.92</td>
<td>SMF=95.83 PE=1.92</td>
<td>SMF=95.83 PE=1.92</td>
<td>SMF=95.83 PE=1.92</td>
</tr>
<tr>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
<td><img src="image21" alt="Image" /></td>
</tr>
<tr>
<td>SMF=98.44 PE=0.59</td>
<td>SMF=100 PE=0.0</td>
<td>SMF=98.46 PE=0.59</td>
<td>SMF=100 PE=0.0</td>
<td>SMF=100 PE=0.0</td>
<td>SMF=100 PE=0.0</td>
<td>SMF=100 PE=0.0</td>
</tr>
<tr>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
<td><img src="image25" alt="Image" /></td>
<td><img src="image26" alt="Image" /></td>
<td><img src="image27" alt="Image" /></td>
<td><img src="image28" alt="Image" /></td>
</tr>
<tr>
<td>SMF=82.89 PE=7.69</td>
<td>SMF=90.00 PE=4.14</td>
<td>SMF=86.49 PE=5.92</td>
<td>SMF=88.89 PE=4.73</td>
<td>SMF=88.89 PE=4.73</td>
<td>SMF=88.89 PE=4.73</td>
<td>SMF=88.89 PE=4.73</td>
</tr>
</tbody>
</table>

Table 4.4: Comparative results for four cells of the G channel.

Following the results obtained by simulation, the proposed ensemble segmentation schemes were applied on fourteen real cDNA microarray images containing 89600 spots in total. Due to the fact that the ground truth information about the images is not a prior knowledge, we indirectly measured the performance of the algorithms by calculating the pairwise mean absolute error (MAE) among the replicates for the common reference channel. The calculated intensity values were normalized according to LOWESS method, in order the comparison among the calculated MAE of all methods to be possible.
According to our results, among the ensemble schemes, the minimum mean absolute error was accomplished by min operator (MAE: 803.96 and Normalized MAE: 0.0570), while the maximum value was obtained by majority vote (MAE: 990.49 and Normalized MAE: 0.0738), which proved to be more depended on the image variations than the rest of the segmentation methods.

Apart from majority vote, the remaining operators of the proposed ensemble scheme produced similar results. Product (MAE: 810.17 and Normalized MAE: 0.0602) and Probabilistic product (MAE: 806.42 and Normalized MAE: 0.0598) proved to be more reliable in comparison to conventional MRF model, in terms of mean absolute error. Box plots summarizing the mean MAE values are presented in Figure 4.4. The calculated mean MAE value and its corresponded normalized mean value of each algorithm are illustrated in Table 4.5.

<table>
<thead>
<tr>
<th>Segmentation Method</th>
<th>mean MAE</th>
<th>Normalized mean MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple MRF</td>
<td>1183.50</td>
<td>0.0859</td>
</tr>
<tr>
<td>Majority Vote</td>
<td>990.49</td>
<td>0.0738</td>
</tr>
<tr>
<td>Min</td>
<td>803.96</td>
<td>0.0570</td>
</tr>
<tr>
<td>Product</td>
<td>810.17</td>
<td>0.0602</td>
</tr>
<tr>
<td>Probabilistic Product</td>
<td>806.42</td>
<td>0.0598</td>
</tr>
</tbody>
</table>

Table 4.5 The calculated mean MAE value of each algorithm.

Figure 4.4: Box plots summarizing the normalized MAE values of the proposed ensemble schemes in comparison to Conventional (Simple) MRF model.


4.2 Conclusions

In the present study, a new multiscale approach of MRF was introduced based on an MRF ensemble scheme. The resultant binary image was produced by combining the corresponding outputs of the implementation of the initial cell and its wavelet coefficients on the conventional MRF segmentation model. As combining operator majority vote, min, product and probabilistic product were chosen.

According to our results on the simulated image, the proposed ensemble scheme managed to lead to more accurate spot determination in comparison to conventional MRF model. Additionally, an ensemble scheme based on majority vote operator accomplished the highest score in all cases, especially on cells with high noise.

Following the results obtained by simulation, the proposed ensemble segmentation schemes were applied on real cDNA microarray images, in order to indirect measure their reliability. According to our results, all examined operators managed to suppress the MAE value in comparison to conventional MRF segmentation model. The min operator achieved the lowest score in contrast to majority vote, which scored the highest score among the proposed evaluating methods.

4.3 Future work

The most challengeable tasks in microarray image processing techniques are the addressing and the segmentation steps. The accuracy of those steps has a substantial impact on the accuracy and the effectiveness of the subsequent gene expression analysis. Thus, new gridding techniques, with efficient repeatability for high-throughput image analysis, should be introduced as an attempted to overcome the problems of current methods.

New segmentation algorithms could also be investigated. These methods could be based on either combining the information provided by different segmentation methods or hybrid clustering methods, or on development of new segmentation methods based on Hidden Markov Random Fields and wavelets.
REFERENCES


[2] Statistics for Microarray, http:\statwww.epfl.ch\davison\teaching\Microarrays\lec/


[22] http://binfo.ym.edu.tw/microarray/summer05/


[29] MatLab Tutorials Version 7


[34] Markov Random Fields and Gibbs Distributions: http://www.cbsr.ia.ac.cn/users/szli/MRF_Book/Chapter_1/node9.html


[41] Microarray simulation model. Matlab implementation of the microarray simulation model. http://www.biomedcentral.com/content/subplementary/1471-2105-7-349-S1.gz

