University of Crete
School of Sciences and Engineering
Computer Science Department

DNA Microarray Image Enhancement in a MultiResolution Framework

Hara Stefanou

Master of Science Thesis

Heraklion, July 2007

University of Crete
School of Sciences and Engineering
Computer Science Department

DNA MICROARRAY IMAGE ENHANCEMENT IN A MULTIRESOLUTION FRAMEWORK

HARA STEFANOU

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Author:

_____

Hara Stefanou, Computer Science Department

Supervisory Committee:

_____

Panagiotis Tsakalides, Associate Professor, Supervisor

_____

Konstantinos Marias, Researcher, Vice-Supervisor

_____

Ioannis Tollis, Professor, Member

Approved by:

_____

Panos Trahanias, Professor, Postgraduate Studies' Chairman

Heraklion, July 2007

Summary

DNA microarrays have demonstrated an excellent potential in correlating specific gene expression profiles to specific conditions (e.g., disease) as they allow the concurrent observation of all known genes. Because patterns of gene expression correlate strongly with function, microarrays are providing unprecedented information both on basic research, such as the expression profiles of different tissues and the effect of deletion of specific genes, as well as on applied research, such as human disease, aging, drug and hormone action, mental illness, diet and many other clinical matters.

Microarray experiments, however, involve a large number of error-prone procedures that lead to a high level of noise in the resulting data. The high level of the uncertainty associated with each microarray experiment originates by biological variations (corresponding to real differences between different cell types and tissues) and experimental noise. This uncertainty often obscures some of the important characteristics of the biological processes of interest. More specifically, changes in the measured transcript values in the samples render the clustering of genes into functional groups and the classification of samples difficult.

A major challenge in DNA microarray analysis is to eliminate the effect of the noise, which has an additive and a multiplicative component, and recover the gene expression measurements. A number of well-known image processing techniques, including soft and hard thresholding, Bayesian denoising based on Gaussian or Laplacian signal modeling and multiresolution methods that exploit the correlation between the representation coefficients of adjacent scales, have been applied to microarray images by ordinarily assuming the presence of either additive or multiplicative noise.

In this dissertation, we propose an image denoising method which accounts for both noise components and makes the microarray spot area more homogeneous and more distinctive from their local background. The proposed approach consists of two stages: one that processes the additive component of the noise and one that processes the multiplicative component. The method first performs a multiresolution decomposition of the image and then accounts for the heavy-tailed statistical behavior of the representation coefficients as well as for their strong statistical dependence across multiple scales. The utility of this framework is validated with real microarray data through visual evaluation and quantitative performance metrics.

Supervisor: Panagiotis Tsakalides
Associate Professor

Περίληψη

Οι μικροσυστοιχίες DNA έχουν επιδείξει μια εκπληκτική δυνατότητα στη συσχέτιση συγκεκριμένων προφίλ γονιδιακής έκφρασης με συγκεκριμένες καταστάσεις (π.χ. ασθένεια) αφού επιτρέπουν την ταυτόχρονη παρατήρηση όλων των γνωστών γονιδίων. Επειδή κάποια πρότυπα γονιδιακής έκφρασης συσχετίζονται σε μεγάλο βαθμό με συγκεκριμένες λειτουργίες, οι μικροσυστοιχίες προσφέρουν νέα πληροφορία τόσο στη βασική έρευνα, όπως στα προφίλ έκφρασης διαφορετικών ιστών και στο πώς επιδρά η απαλοιφή συγκεκριμένων γονιδίων, όσο και στην εφαρμοσμένη έρευνα, όπως στις ανθρώπινες ασθένειες, στη γήρανση, στη φαρμακευτική και ορμονική δράση, στις νοητικές ασθένειες και σε πολλά άλλα κλινικά θέματα.

Τα πειράματα μικροσυτοιχιών, όμως, συνιστώνται από έναν μεγάλο αριθμό διαδικασιών, οι οποίες είναι επιρρεπείς στα σφάλματα και έτσι τα τελικά αποτελέσματα περιέχουν μεγάλο ποσοστό θορύβου. Οι παρατηρούμενες διαφορές στη γονιδιακή έκφραση σε κάθε πείραμα μικροσυστοιχιών προέρχεται από βιολογικές διαφορές (που αντιστοιχούν σε πραγματικές διαφορές μεταξύ διαφορετικών τύπων κυττάρων και ιστών) και σε πειραματικό θόρυβο. Αυτή η αβεβαιότητα καθιστά συχνά μη εμφανή κάποια από τα πιο σημαντικά χαρακτηριστικά της βιολογικής διεργασίας. Πιο συγκεκριμένα, οι μεταβολές στις μετρούμενες τιμές μετάφρασης στα δείγματα καθιστούν δύσκολη την ομαδοποίηση των γονιδίων σε λειτουργικές ομάδες και την ταξινόμηση των δειγμάτων.

Μεγάλη πρόκληση στην ανάλυση των DNA μικροσυστοιχιών αποτελεί η εξάλειψη του θορύβου, ο οποίος έχει μια αθροιστική και μια πολλαπλασιαστική συνιστώσα, και η ανάκτηση των μετρήσεων της γονιδιακής ρύθμισης. Σε αυτήν την εργασία, προτείνουμε μια μέθοδο αποθορύβωσης εικόνας που λαμβάνει υπόψη και τις δύο συνιστώσες του θορύβου και κάνει τις περιοχές των spots πιο ομογενείς και πιο διακριτές από το τοπικό υπόβαθρο. Η προτεινόμενη μέθοδος αποτελείται από δύο στάδια: ένα για την επεξεργασία της αθροιστικής συνιστώσας του θορύβου κι ένα για αυτή της πολλαπλασιαστικής. Η μέθοδος πρώτα εκτελεί μια ανάλυση πολλαπλής διακριτικής ικανότητας στην εικόνα και στη συνέχεια αντιμετωπίζει τη στατιστική συμπεριφορά των συντελεστών αναπαράστασης που εμφανίζει βαριές ουρές, αλλά και τη συσχέτιση που υπάρχει ανάμεσα στους συντελεστές της αναπαράστασης σε διαδοχικά επίπεδα ανάλυσης. Η χρησιμότητα αυτής της μεθόδου αξιολογήθηκε με πραγματικά δεδομένα από μικροσυστοιχίες μέσω οπτικής αξιολόγησης και ποσοτικών μέτρων επίδοσης.

Επόπτης: Παναγιώτης Τσακαλίδης
Αναπληρωτής Καθηγητής

Ευχαριστίες

Αισθάνομαι την ανάγκη να εκφράσω ένα μεγάλο ευχαριστώ στον επόπτη μου, Καθ. Παναγιώτη Τσακαλίδη, για τη διαθεσιμότητά του κάθε στιγμή που τον χρειαζόμουν, για το οργανωτικό του πνεύμα που με βοήθησε να συντονίσω την προσπάθειά μου και να μάθω τον τρόπο διεξαγωγής έρευνας, για την ουσιαστική καθοδήγηση και συμβολή του στην ολοκλήρωση της παρούσας εργασίας, για την ώθηση που μου έδινε κάθε φορά που καταλάβαινε ότι έμενα πίσω.

Επίσης, θα ήθελα να ευχαριστήσω τον Καθ. Γιάννη Τόλλη για τη συμμετοχή του στην εξεταστική επιτροπή και για τις καίριες παρατηρήσεις του.

Ευχαριστώ θερμά τον Κώστα Μαριά, Ερευνητή του ΙΤΕ – ΙΠ, για τη υποστήριξη που μου παρείχε προτού ακόμα γίνω δεκτή ως μεταπτυχιακή φοιτήτρια στο Τμήμα Υπολογιστών του Πανεπιστημίου Κρήτης και καθόλη τη διάρκεια των σπουδών μου. Τον ευχαριστώ ακόμα για τις πολύτιμες συμβουλές και τα δημιουργικά σχόλια του κατά τη διεκπεραίωση αυτής της εργασίας, καθώς και τον πολύτιμο χρόνο που μου αφιέρωνε.

Ιδιαίτερες ευχαριστίες θέλω να δώσω στον Θανάση Μαργαρίτη, διδακτορικό φοιτητή του τμήματος Βιολογίας του Πανεπιστημίου Κρήτης και του ΙΤΕ – ΙΜΒΒ, για τη μετάδοση των βιολογικών γνώσεων, για τη συμβολή του στην εργασία αυτή με τη συνεχή τροφοδότηση νέων ιδεών και γιατί ήταν πάντα διαθέσιμος σε μια εποχή της ζωής του με λιγοστό ελεύθερο χρόνο.

Περισσότερο, όμως, από όλους ευχαριστώ την οικογένειά μου για την συνεχή στήριξη σε όλη τη διάρκεια των σπουδών μου και κυρίως τα δύο τελευταία χρόνια, που ήμουν μακριά τους. Τον πατέρα μου, Γιώργο, που αν και έλειπε συχνά, δεν άφηνε να γίνει αισθητή η απουσία του με τις συμβουλές του και τη συνεχή παρότρυνση να υλοποιώ τις επιθυμίες μου. Τη μητέρα μου, Καίτη, για την ανατροφή μου, την παρότρυνσή της για συνεχές διάβασμα, το ενδιαφέρον της για την εξέλιξή μου και την υποστήριξη ακόμα και σε εποχές δύσκολες για εκείνη. Την αδερφή μου, Μαίρη, για την ψυχολογική υποστήριξη και γιατί είναι πάντα δίπλα μου σε ό,τι και αν τη χρειαστώ.

Τέλος, ευχαριστώ τους φίλους μου και ιδιαίτερα το Δημήτρη Μανακανάτα που με βοήθησε ψυχολογικά όποτε το χρειαζόμουν, αλλά και πρακτικά με τις καίριες επισημάνσεις του.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Over the last decade, a revolution has been witnessed in the biosciences, medical sciences, biotechnology and pharmaceutical industry. High-throughput technologies are producing massive amounts of data of the -omes (genome, transcriptome, proteome, metabolome, etc.) instead of the traditional analysis of single genes. The revolution is driven mostly by microarray technology [69,78]. The technology has centered on providing a platform for determining, in a single experiment, the gene expression profiles of hundreds to tens of thousands of genes in tissue, tumors, cells or biological fluids. The rapid and global adoption of this technology has been predicated on its simplicity and success in providing large amounts of highly relevant data.

DNA microarray technology has a profound impact on research as it allows the concurrent observation of the expression of all known genes. Because patterns of gene expression correlate strongly with function [26,88], microarrays are providing unprecedented information both on basic research, such as the expression profiles of different tissues [93] and the effect of deletion of specific genes [38], as well as on applied research, such as human disease, aging, drug and hormone action, mental illness and many other clinical matters. Microarrays can also be used for the identification of alterations in gene sequences, paving the way for a new era of genetic screening, testing and diagnostics.

The correlation between gene mutation, altered protein, and disease was first made by Pauling and co-workers [66]. It was showed that hemoglobin from sickle cell patients differs from hemoglobin in healthy individuals in that it migrates abnormally in gel electrophoresis assays, a finding the authors attributed correctly to a change in the surface charge of hemoglobin. By examining normal individuals, earners and patients with sickle cell disease, Pauling and co-workers concluded that changes in the hemoglobin gene were responsible for the altered protein, which was verified later in gene sequencing studies. This remarkable paper paved the way for the molecular genetic analysis of human disease, and provided a conceptual foundation for the use of microarrays in genetic screening, testing, and diagnostics.

Maxam and Gilbert [59] at Harvard and Sanger and co-workers [76] at the Medical Research Council (MRC) developed DNA sequencing technology independently. Gilbert and Sanger shared half of the Nobel Prize in 1980 “for their contributions concerning the determination of base sequences in nucleic acids” [64]. Sanger chemistry was used to sequence the human genome and has provided much of the sequence database information used to manufacture DNA microarrays.

The origin of the microarray technique was evolved from E. Southern’s technique in the 1970s [90]. In the late 1980’s a team of scientists led by Stephen P.A. Fodor, Ph.D. made the first microarray protocol. The publication in 1994 [69] was the first reported description of microarray technology. This was the Affymetrix VLSIPS technology which was adapted for the production of the first two-color DNA microarray by Dr. Schena and his colleagues at Stanford University. Their first

publication in Science [78] is the first one on microarrays and the most highly cited one. The Scientist places Dr. Schena at positions 1 and 2 on "the microarray family tree", confirming his role as the founder of microarray technology and substantiating his "Father of Microarray Technology" status [96].

In the near future, it will be possible to profile the whole transcriptome of higher organisms with only a few DNA gene chips. This will allow us to obtain a global view of the genotypes corresponding to different cell phenotypes. Such capability will greatly accelerate and perhaps fundamentally change biomedical research and development in many areas, ranging from developing advanced diagnostics to unravelling complex biological pathways and networks, to eventually facilitating individual-based medicine [53,12]. The demand to conduct analysis on a genome-wide basis grows and so does the need for improved data extraction and analysis.

1.1 Motivation

DNA microarray experiments consist of procedures that are prone to errors yielding data with a high level of noise. This noisy nature makes deciphering high throughput gene expression experiments difficult. In general, the changes in the measured transcript values between different experiments are caused by both biological variations (corresponding to real differences between different cell types and tissues) and experimental noise. A major challenge in DNA microarray analysis is to effectively dissociate gene expression values from experimental noise. Elucidating the sources of noise may be of help for identifying the steps of the techniques that need to be modified to improve the signal-to-noise ratio.

Several authors have addressed the noise issue explicitly. Lee et al. [49] have noticed that results from repeated gene array experiments differed substantially, reaching in that way the conclusion that repetition can increase the significance of conclusions from gene array experiments. Chen et al. [14] and Ermolaeva et al. [29] used a ratio distribution to determine the statistical significance of an observed change in expression levels. Hughes et al. [39] suggested statistics for estimation and uncertainty from multiple repetitions. Unlike the purely additive model of Chen et al. [14] and Ermolaeva et al. [29], the model of Hughes et al. [39] incorporates both additive and multiplicative noise. Yet, these methods consider only the process statistics. Hartemink et al. [34] claimed that the Affymetrix chip data follow a log-normal distribution. Rocke and Durbin [74] introduced a model for measurement error in gene expression arrays as a function of the expression level. The Bayesian estimation of array measurements (BEAM) technique, proposed by Dror et al. [21], results in a noise model for Affymetrix chips that includes a heavy-tailed additive noise and a gene-specific bias term.

These techniques deal mainly with the measurement error and not the noise inherent in the microarray images. Currently, there are various techniques which deal with the image noise by increasing the accuracy and signal-to-noise ratio (SNR) of the estimated values [110]. Nonetheless, most techniques rely on mathematical algorithms which distinguish between noise and real signal, the well-known thresholding methods [2], [103]. Thresholding methods, however, have some drawbacks. Mastriani and Giraldez [58] made an attempt to overcome the disadvantages of thresholding by applying a bidimensional smoothing within each

highest subband. This method, as all smoothing methods, does not discriminate between spots and noise; therefore, spot information may be removed together with noise. Another technique [55] uses vector processing filters based on fuzzy logic concepts for the attenuation of noise in two-channel images, i.e. cDNA microarrays. But this technique deals only with the additive noise component. However, none of these studies actually quantify the benefit of image enhancement in facilitating segmentation and consequently gene quantification. Consequently, a new technique which removes both microarray image noise components and enhances the image has to be employed.

1.2 Contributions

The main contributions of this dissertation are:

- application of a noise removal algorithm on the microarray image,
- implementation of a noise removal method which deals with both noise components,
- account for the heavy-tailed statistical behavior of the representation coefficients,
- account for the coefficients' strong statistical dependence across multiple scales,
- enhancement of the dynamic range of existing microarray imaging technology in the resulting image,
- identification of the most significant spots with increased accuracy and robustness,
- better spot segmentation.

1.3 Organization

This dissertation is organized as follows. Section 2 is an introduction to the microarray technology containing an overview of microarray types, types of probes and the procedure for microarray production. The latter consists of the fabrication, label targeting, array scanning, typical image processing and information extraction. Section 3 accounts for the characteristics of the microarray signal and noise and presents most microarray noise sources. Section 4 is an analytical presentation of the decimated and undecimated wavelet transform and the multi-resolution analysis.

The proposed denoising method is described in Section 5, along with some other denoising techniques, including including soft and hard thresholding, Bayesian denoising based on Gaussian or Laplacian signal modeling, and multiresolution methods that exploit the correlation between the wavelet coefficients of adjacent scales, that were used for the comparison and the evaluation of our method. Section 6 describes the characteristics of the used microarray images and presents the processed images, the spot segmentation results and quantitative metrics that were used for the evaluation of the proposed method. The conclusions of this work are drawn in Section 7.

2 MICROARRAYS

2.1 What is a Microarray?

In their most generic form, microarrays are ordered arrays of microscopic elements (probes) on a planar substrate that allows the specific binding of genes or gene products [80]. Hundreds to tens of thousands DNA molecules are organized in a two dimensional array (matrix). The DNA molecules are typically either oligonucleotide (ranging from 35 base pairs to several hundred) or cDNAs[1]. The substrate to which the DNA molecule is attached is usually glass, silicon, or nylon.

2.2 Types of Microarrays

Microarrays are categorized into several groups regarding the technology used in their production and the type of probe they use. The latter parameter for discrimination is taken into consideration only in spotted microarrays where many types of material can be used for the probes.

2.2.1 Regarding the Technology

When DNA microarrays are used for measuring the concentration of mRNA in living cells, a probe of one DNA strand that matches a particular mRNA in the cell is used. The concentration of a particular mRNA is a result of expression of its corresponding gene, so this application is often referred to as expression analysis. For expression analysis, there are many technologies of microarrays production but the field has been dominated in the past by two major technologies, the *Affymetrix chips* and the *spotted microarrays*.

The Affymetrix, Inc. GeneChip system [52,69] uses prefabricated oligonucleotide chips made by light mask technology while custom-made chips use a robot to spot cDNA, oligonucleotides or PCR[2] products on a solid support. The first is easier to control and therefore the variation between these chips is smaller. On the other hand, spotted arrays [79] are more flexible. Moreover, *digital micromirror arrays* [86] of NimbleGen Systems Inc. and Febit Biotech Gmbh combine the flexibility of the spotted arrays with the speed of the prefabricated Affymetrix chips due to their ability to control light-directed synthesis of oligonucleotide microarrays. Agilent manufactures DNA microarrays on glass slides using the inkjet technology of Hewlett Packard. These arrays are known as *inkjet microarrays* [38]. The *bead-based array system* of Illumina Inc. allows small glass beads with covalently attached oligo probes self-assemble into etched substrates. The location of each bead on the array is then

[1] Complementary DNA. Single-stranded DNA that is complementary to messenger RNA or DNA that has been synthesized from messenger RNA by reverse transcriptase [95].
[2] Polymerase Chain Reaction. Revolutionary technology developed during the 1980's that allows massive amplification of any gene sequence of interest [80].

read by a decoder. *Serial analysis of gene expression* (SAGE) [101] measures the times a cDNA fragment occurs in the sequence of concatenated fragments of DNA and this number is proportional to the abundance of its corresponding mRNA [48]. A more detailed description of the two major technologies – Affymetrix and spotted arrays – follows.

2.2.1.1 Affymetrix Microarrays

The production of an Affymetrix microarray consists of six steps. Firstly, a photo-protected glass substrate is selectively illuminated by light passing through a photolithographic mask. After that, the de-protected areas are activated and chemical coupling occurs at the activated positions with the nucleoside incubation following up. Then the coupling step is repeated with a new mask pattern applied and this process is repeated until the desired set of probes is obtained.

In Affymetrix microarrays each gene or portion of a gene is represented by 11 to 20 oligonucleotides of 25 base-pairs and the probe is an oligonucleotide of those, i.e. a 25-mer. The *perfect-match* (PM) is a 25-mer complementary to a reference sequence of interest while a *mismatch* (MM) is the same as a PM but with a single homomeric base change for the middle – that is the 13^{th} – base. Both PM and MM constitute a probe-pair and a collection of these probe pairs (from 11 to 20) related to a common gene or fraction of a gene is a probe-pair set.

Affymetrix claims that the MM oligonucleotides will be able to detect non-specific and background hybridization, which is important for quantifying weakly expressed mRNAs. However, for weakly expressed mRNAs where the signal-to-noise ratio is smallest, subtracting mismatch from perfect match adds considerably to the noise in the data [77]. That is because subtracting one noisy signal from another noisy one yields a third signal with even more noise.

2.2.1.2 Spotted Microarrays

In spotted arrays a robot spotter is used to move small quantities of probe in solution from a microtiter plate to the surface of a glass plate. The probe can consist of cDNA, oligonucleotides, proteins and even a whole section from a human tumor. Each probe is complementary to a unique gene. Probes can be fixed to the surface in a number of ways. The classical way is by non-specific binding to slides. The slides are first coated with polylysine while the probes are prepared in microtiter plates. The robot spots the probes on the glass slides and when this procedure is over the remaining exposed amines of polylysine are blocked with succinic anhydride. The cDNA has to be denatured to produce single-stranded DNA for the hybridization step.

Extracted mRNA from cells is converted to cDNA and then every sample is labelled fluorescently with different dyes. The two most commonly used dyes – usually referred as fluorochromes – are rhodamine (Cy3) that is green and flourescein (Cy5) that is red. After mixing, the labelled samples are hybridized to the probes on the glass slides. Then, the unhybridized material is washed away and the slide is scanned with a confocal laser.

The advantage of spotted arrays compared to the Affymetrix GeneChips is that any probe for spotting on the array can be designed. However, the spotting will not be

nearly as uniform as the in-situ synthesized Affymetrix chips and the cost of oligonucleotides, for chips containing thousands of probes, increases. As far as the data analysis is concerned, in cDNA microarrays the two samples are hybridized to the same chip using different fluorochromes, whereas the Affymetrix chip can handle only one fluorochrome so two chips are required for the comparison between two samples.

In this dissertation we used a spotted cDNA microarray and later on we will further analyze the procedure of making such a microarray.

2.2.2 Regarding the Type of the Probe

The first microarray experiments were performed with cDNA microarrays and continue to find wide use in gene expression assays. A cDNA is a nucleic acid molecule derived from mRNA and its length is typically 500-2500 base pairs. Microarrays containing such molecules provide intense hybridization signals because of their extensive complementarity to fluorescent probe molecules in solution. Database analysis of 2000 microarray citations (arrayit.com/e-library) reveals that cDNA microarrays account for approximately 65% of all microarray publications.

Oligonucleotide microarray is another commonly used microarray, finding wide use in a variety of applications, including gene expression profiling and genotyping. Oligonucleotides are single-stranded 15- to 70-nucleotide molecules made by chemical synthesis, and these synthetic targets produce high specificity and good signal strength in hybridization reactions. More than one quarter of all microarray publications to date use oligonucleotides as the target molecules. Complementary DNA and oligonucleotide microarrays both exploit the chemical process of hybridization to generate microarray signals. They, also, fall into a broader category known as nucleic acid microarrays, which encompasses microarrays containing any type of DNA or RNA as the target material.

Tissue and protein microarrays are more recent than nucleic acid microarrays, but are being used with increasing frequency, combining for nearly 10% of the scientific publications to date. Tissue microarrays contain sections from human tumor specimens and oilier tissues of interest, and protein microarrays contain pure proteins or cell extracts at each microarray location. These new microarray formats are replacing many of the traditional histological and biochemical assays because the parallelism**,** miniaturization, and automation of microarray assays afford a precision, speed, and information content unattainable with the antecedent technologies.

2.3 The Microarray Procedure

The procedure of a spotted DNA microarray production consists of four steps. A short outline of these steps is going to be presented, though every step is going to be analyzed in the rest of this section. Firstly, the matrix is typically coated with chemicals to make the matrix reactive. Usually, the DNA is attached to it using UV radiation or covalent coupling to permanently link the DNA to the surface. In this form a cDNA or oligonucleotide, corresponding to a specific sequence of a gene, can be spotted onto the solid surface. That can be repeated for hundreds to tens of thousands of genes. With this set of DNA segments attached to the surface, RNA

from a specimen (e.g. tissue, cell line, tumor) can be labeled directly or indirectly (usually with a fluorescent nucleotide) and hybridized to the array of genes. The amount of fluorescence at each DNA spot corresponds to the transcript level of that particular gene. Therefore, the expression of thousands of genes can be analyzed in a single specimen by analyzing the microarray image.

2.3.1 Fabrication

Production of spotted arrays begins with the selection of the probes to be printed on the array. In many cases, these are chosen directly from databases including GenBank [6], dbEST [10] and UniGene [82]. Additionally, full−length cDNAs, collections of partially sequenced cDNAs (or ESTs[3]), or randomly chosen cDNAs from any library of interest can be used. Arrays for higher eukaryotes are typically based on the EST portions of these projects, whereas for yeast and prokaryotes, probes are usually generated by amplifying genomic DNA with gene−specific primers[4]. Given the expense of obtaining clones, producing DNA from them and printing them, it is usually preferable to produce arrays with a low redundancy of representation, so as to survey the broadest possible set of genes.

cDNA arrays are produced by spotting PCR products representing specific genes onto a matrix. These are usually generated from purified templates, so that cellular contaminants do not find their way onto the array. Typically, the PCR product is partially purified by precipitation, gel−filtration, or both — to remove unwanted salts, detergents, PCR primers and proteins present in the PCR cocktail. For both glass and membrane matrices, each array element is generated by the deposition of a few nanoliters of purified PCR product [16]. A robot spots a sample of each gene product onto a number of matrices in a serial operation performing in this way the printing. The first spotting robots relied on contact printing with a device like a fountain pen. Many variations on this design are now available [11], in addition to a “spotter” that is a capillary tube, to which a low but constant pressure is applied. Non−contact printing modes, using either piezo or ink−jet devices, are also being evaluated.

The types of membranes commonly used are nitrocellulose and charged nylon commercial varieties that are used for various blotting assays. Glass−based arrays are most often made on microscope slides, which have low inherent fluorescence. Glass has many of the advantages of nylon but it also has some unique merits:

i. DNA samples can be covalently attached onto a treated glass surface.
ii. Glass is a durable material that sustains high temperatures and washes of high ionic strength.
iii. It is non-porous so the hybridization volume can be kept to a minimum, thus enhancing the kinetics of annealing probes to targets.
iv. As a consequence of its low fluorescence, it does not significantly contribute to background noise.

[3] Expressed Sequence Tag. A short sub-sequence of a transcribed protein-coding or non-protein coding DNA sequence [105].
[4] A primer is a nucleic acid strand (or related molecule) that serves as a starting point for DNA replication [104].

v. Two different probes can be labelled with different fluors, and simultaneously incubated with a microarray in a single reaction [16]. The glass−based arrays are coated with some chemicals which enhance both the hydrophobicity of the slide and the adherence of the deposited DNA. They also limit the spread of the spotted DNA droplet on the slide.

In most cases, DNA is cross−linked to the matrix by ultraviolet irradiation. After fixation, residual amines on the slide surface are reacted with succinic anhydride to reduce the positive charge at the surface. As a final step, some percentage of the DNA deposited is rendered single−stranded by heat or alkali [16]. The state of bound DNA is ill−defined. It is deposited in double−stranded form, intra−strand cross−linked to some extent, and may well have multiple constraining contacts with the matrix along its length (induced by drying the DNA onto the matrix). It is therefore probably not the best hybridization probe. One can imagine that oligonucleotide matrices, with their short chains and single points of constraint at each chain end, may be a far better probe for hybridization. Against this advantage, however, the disadvantages of using short−chain detectors must be weighed. Chief among these are the variations in melting temperature due to AT−GC composition, and the reduction in specificity due to truncating the number of nucleotides from hundreds to as few as twenty. A format in which the accessibility of a simply tethered, single−stranded probe could be combined with the specificity of a long probe would provide a considerable improvement for the field [23].

2.3.2 Target labelling and hybridization

The targets for arrays are labelled representations of cellular mRNA pools. Typically, reverse transcription from an oligo−dT primer[5] is used. This has the virtue of producing a labelled product from the 3´ end of the gene, directly complementary to immobilized targets synthesized from ESTs. Frequently, total RNA pools (rather than mRNA selected on oligo−dT) are labelled, to maximize the amount of message that can be obtained from a given amount of tissue. The purity of RNA is a critical factor in hybridization performance, particularly when using fluorescence, as cellular protein, lipid and carbohydrate can mediate significant non−specific binding of fluorescently labelled cDNAs to slide surfaces. The fact that array elements are physically close to each other and strong hybridization with a radioactive target can easily interfere with detection of weak hybridization in surrounding targets has to be taken into consideration. As far as the fluorescent labels are concerned, Cy3 and Cy5 are frequently paired, as they have relatively high incorporation efficiencies with reverse transcriptase, good photostability and yield and are widely separated in their excitation and emission spectra, allowing highly discriminating optical filtration [23].

A clear limitation to the application of this technology is the large amount of RNA required per hybridization. If not a quite large amount of total RNA is used in the hybridization there will not be an adequate fluorescence, because a few fluors will be present in a scanned pixel from a specific probe. Such low levels of signal are at the lower limit of fluorescence detection, and could easily be rendered undetectable by assay noise. Although radioactive targets may have a higher intrinsic detectability,

[5] The oligo-dT primer consists of a string of deoxythymidylic acid residues and is designed to prime poly A+ RNA molecules for first-strand cDNA synthesis [40].

they too reach a level of dilution that prohibits effective detection, thus making experimentation on very small numbers of cells impossible [23].

A variety of means by which to improve signal from limited RNA has been proposed. Efficient mixing of the hybridization fluid should bring more molecules into contact with their cognate probe, increasing the number of productive events. This entails, however, a larger mixing volume, which might offset the potential gain. Methods that produce multiple copies of mRNA using highly efficient phage RNA polymerases have been developed [70]. A version of this approach, in which labelled target (cRNA) is made directly from a cDNA pool, having a T7 RNA polymerase promoter site at one end via in vitro transcription, has been applied to arrays [52]. Post−hybridization amplification methods have also been reported in which detectable molecules are precipitated at the target by the action of enzymes "sandwiched" to the cDNA target [15]. Detection of hybridized species using mass spectroscopy or local changes in electronic properties can also be imagined [98,57].

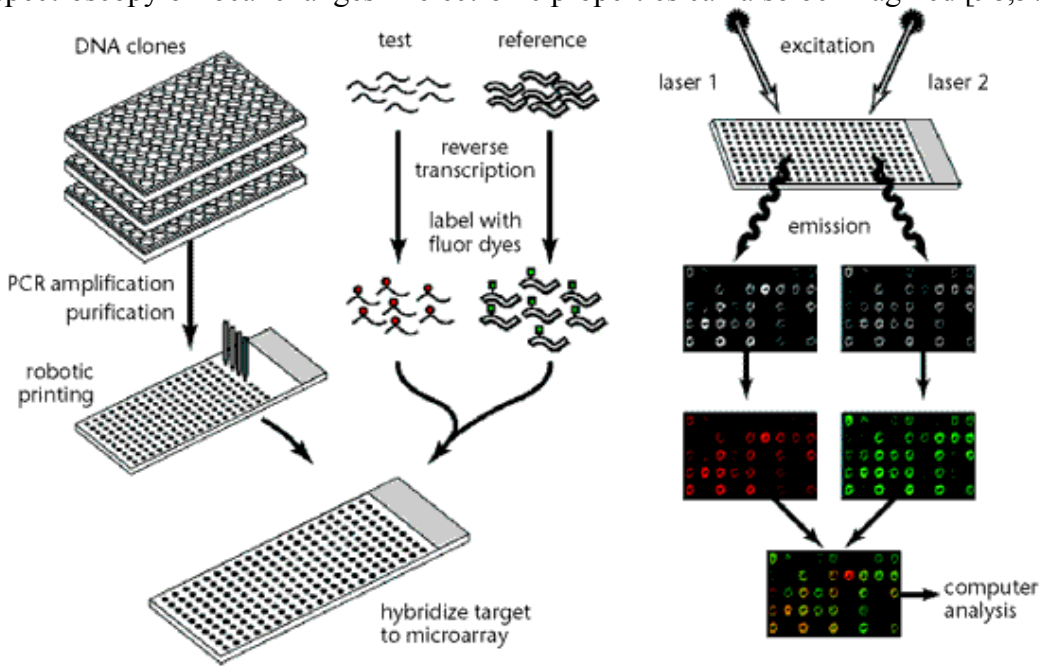


Figure 2.1 – The Microarray Experiment [23].

2.3.3 Image analysis and data extraction

Image data is prone to extraction by highly developed, digital image processing procedures due to the highly regular arrangement of detector elements and crisply delineated signals that result from robotic printing and confocal imaging of fluor−detected arrays. Grids specifying target locations can be readily overlaid on the images. Local sampling of background can be used to specify a threshold which true signal must exceed and mathematical morphology methods to predict the likely shape and placement of the hybridization signal. By applying these methods it is possible to accurately detect even weak signals [14] and extract a mean intensity above background for the target. In contrast, extraction of data from film or phosphor−image representations of radioactive hybridizations presents many difficulties for image analysis. If the array is on a membrane, there is frequently non−linear warping of the matrix, which means that the observed array will not have the strict geometric regularity of an array printed to a stiff matrix, such as glass. This

introduces difficulty in developing highly accurate grids to specify target locations. The spread of detectable particles from a disintegrating nuclide to the detector is highly sensitive to variations in distance between source and detector, and produces a smooth transition from the highest levels of intensity to background. This ensures that the image produced by radioactive exposure is composed of sections at many focal planes, and renders impossible the application of single, simple, point−spread functions to reconstitute a "focused" representation of the data. The smoothness of the transition from maximum signal intensity to background signal intensity makes consideration of local background for each signal a difficult proposition as one does not observe an abrupt, readily discerned transition between signal and background, but a smooth curve without a sharp derivative [23].

In carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. All experiments are carried out under conditions of a large excess of immobilized probe relative to labelled target. The kinetics of hybridization is therefore pseudo−first order, and inter−probe competition is not a factor. Under these conditions, the linear differences arising from exact amount of applied target, extent of target labelling, efficiencies of fluor excitation and emission, and detector efficiency can be compounded into a single variable and the information from each detection channel normalized. It is best to achieve normalization by adjusting the sensitivity of detection (photomultiplier voltage with fluorescence or exposure time with radioactivity) so that the measurements occupy the same dynamic range in the detector. There are essentially two strategies that can be followed in carrying out the normalization. One is based on a consideration of all of the genes in the sample, and the other, on a designated subset expected to be unchanging over most circumstances. In either case, variance of the normalizing set can be used to generate estimates of expected variance, leading to predicted confidence intervals. In instances of closely related samples, the transcript level of many genes will remain unchanged, making global normalization a useful tool. As samples become more divergent, the fraction of genes showing altered transcript levels increases, and global normalization yields a poorer estimate of normalization than would be achieved using a subset of constantly expressed genes. Explicit methods have been developed which make use of a subset of genes for normalization, and extract from this subset's variance statistics for evaluating the significance of observed changes in the complete dataset [14].

A common aspect is the extent of reliability and variance in measurements. So far, most array methods have been validated by probing northern blots [3] of the biological samples. As with sequencing, the best comparisons and measures of reliability can be made only when large data sets contain significant repetitions and overlapping data are freely available. One can, however, clearly envisage strengths and weaknesses. The simple and highly determined nature of immobilized hybridization probes in oligonucleotide arrays make them likely to yield the highest level of reproducibility of absolute measurement for a given element. The ability of cDNA arrays to achieve element−by−element normalization with two−colour fluorescence detection and to use a single, highly specific immobilized probe could provide the most accurate measurements of relative expression levels. All methods should readily disclose large changes in transcript levels among those readily detected genes [23].

The main objectives in microarray image analysis are:

- Noise suppression.
- Spot localization and detection, including the extraction of the background intensity, the spot position and the spot boundary and size.
- Data quantification and quality assessment.

All these processing steps are going to further analyzed in Sections 2.5 and 5.

2.3.4 Data management and mining

All array methods require the construction of databases for the management of information on the genes represented on the array, the primary results of hybridization and the construction of algorithms to make it possible to examine the outputs from single and multiple array experiments. Correlation−based approaches that apply methods developed for the analysis of data which are more highly constrained (such as protein or amino acid sequence comparisons) than at the transcript level have been applied to microarray data analysis. This level of analysis on large data sets provides new perspectives of the operation of genetic networks. Comparison of expression profiles undoubtedly provides useful insights into the molecular pathogenesis of a variety of diseases [47,17]. It cannot, however, deliver the kind of intimate understanding of the highly interrelated control circuitry that is necessary to achieve true understanding of genome function. A number of publications [60,109,30] suggest that to achieve this objective, we should reconsider our perception of transcriptional control as a simple on−off switch to a model whereby control is analogous to a highly gated logic circuit, where numerous, often contradictory, inputs are summed to produce a response. To reach these goals, biologists must expand the arsenal of tools they use to analyse expression data — recruiting statisticians and mathematicians to consider multivariant problems of a size never before attempted.

2.4 Microarray Scanning

After competitive hybridisation with the control and the query labelled DNAs, the array is scanned. Many scanners for microarray analysis exist on the market today differing in technology, size, speed, sensitivity, capacity and many more. Though, it is quite difficult to discriminate which scanner is the best one due to the quick evolution of the field of biochip microarrays.

As array technology transitioned from nylon-based arrays to glass substrate, new instrumentation became necessary to measure the fluorophores used in target labeling. To that end, all scanners were designed to measure the amount of fluorescence emitted from a microarray slide. However, because it is assumed that the quantity of fluorophores in a given location (feature or DNA spot) correlates to the level of endogenous gene transcription, it is important to obtain an accurate measurement of this fluorescent signal. Consequently, scanner specifications (regarding sensitivity, uniformity, resolution, throughput, dynamic range and cross-talk) are important for selecting a quality optical instrument. In addition, engineering improvements to reduce noise and image-extraction algorithms to estimate noise have important implications for detecting subtle differences in gene expression with confidence based on statistical analysis [44].

2.4.1 Parameters for a Microarray Scanner Designing

A microarray scanner designing should take into consideration several key points in order to produce an image that suffers from noise as less as possible. The most important parameters along with the ones responsible for a noisy result are analytically described in the following sections.

2.4.1.1 Resolution

A frequently-used rule of thumb is that the scanning resolution of each pixel should be set as 1/10 the diameter of the spot being scanned. Since not all genes in the human genome are needed in a microarray experiment, the number of spots per slide is reduced so the diameter of each spot can be larger. However, having only one replicate of each gene feature per chip may not be sufficient to generate reliable data. In order to calculate statistics, such as confidence intervals, it is necessary to include replicates, positive and negative controls, and introduce quality control features. So, in theory, a well-designed complete human genome microarray chip should contain 100000 features (while the human genome consists of 30000 to 40000 genes), many of which are replicates and controls [61].

2.4.1.2 Sensitivity

Generating a good image from microarray scanners can be challenging because the quantity of fluorescently-labeled DNA hybridized to the probes on the microarray is not very large. Additionally, the glass substrate for the microarray generates a low, but significant, level of background fluorescence. Therefore, it is necessary to use extremely sensitive scanning and detection strategies that enable the scanner to detect the faint signals and block out as much of the undesirable background fluorescence as possible.

A simple way to increase sensitivity is to increase the power of the excitation light source in order to attempt to make the fluorophores emit greater quantities of fluorescent light. However, it is known that organic fluorophores can only be excited a finite number of times before they are photobleached, which means the dye undergoes a permanent chemical change and is destroyed. It is also possible to increase sensitivity by increasing the detector gain. Depending on the type of detector, there is usually an optimal setting for the detector gain, and further increase of the gain often results in increased noise, which is unwanted. Therefore, diminishing returns are encountered when attempting to increase sensitivity by increasing the power of the light source or increasing the detector gain beyond certain limits. However, if the optimal combination of components is chosen, then excellent sensitivity can be achieved [61].

2.4.1.3 Multiple Dyes

The absorption and emission spectral curves of the fluorescent dyes, used to label the samples, must be well separated to minimize cross-talk. Cross-talk results from either the simultaneous excitation of multiple dyes by one light source, the simultaneous detection of multiple dyes in one detector channel or both. More than two dyes in one experiment would be desirable so that additional controls are added or more data

from one slide are generated. However, the risk of cross-talk increases with the increase of the number of dyes, because most organic fluorophores have rather broad absorption and emission spectral curves. Cross-talk can be reduced by selecting filters that have a narrow spectral range, at the expense of reduced sensitivity [61].

2.4.1.4 Substrates

Many different types of microarray substrates are available such as glass, gold-plated, silicon, plastic or membrane-coated materials. The reasons for choosing a specific type of substrate vary. Some users may want to experiment with different adhesion chemistries that are only available on these different materials. Others may want to eliminate background fluorescence commonly caused by glass substrates. Often benefits of using new substrate materials come with trade-offs. Many of these materials are highly reflective and will direct the excitation light back into the detection optics and reduce the signal-to-background ratio, unless the scanner uses the proper optical design. Some substrates may have wells, which often have reflective walls and bottoms, and the biological materials of interest may be present at the bottom of the wells, or in the bulk solution. Coverslips present an interface where multiple reflections occur. Some provision for reducing reflections and the ability to scan into wells or beneath coverslips should be considered [61].

2.4.1.5 High Quality of Data

Above all, the quality of the image generated by the microarray scanner must satisfy certain criteria for producing valid data. For instance the signal-to-noise ratio of the spot should be sufficiently high so automatic spot finding algorithms can readily detect the spot signal from its surrounding background signal. Equally important is that the pixel value is accurately registered. This means that each data pixel directly corresponds to the actual feature on the microarray that is being imaged. The resolution of the scanner must be sufficiently high to provide enough pixels per spot so that image analysis algorithms can accurately analyze the spot intensities. Misregistered pixels, especially around the edges of the spot, can lead to difficulties in defining the spot location, shape and size, which will impact the spot intensity calculations. Of course, there are also other causes of low quality data so a more general approach of improving data quality should be adopted [61].

2.4.2 Types of Microarray Scanners

In general, microarray scanners fall into two main categories, those that use charged-coupled devices (CCDs) and others that use laser light with photo-multiplier tube (PMT) detection. CCDs commonly use flood-illumination to simultaneously acquire a microarray image that is divided into pixels by the detector. Because a CCD detector cannot acquire the entire dimension of a microarray slide with high resolution, smaller regions are typically scanned and “stitched” together by software to improve image quality [4]. To improve resolution and subsequent software manipulation the slide has to be longer exposed, but this may cause implications on array photobleaching and image integrity. In addition, the use of flood-illumination typically results in a higher background illumination from the glass substrate and the opposite slide surface, as it prevents rejection of these undesired sources of background. Point-source illumination (a laser spot of essentially constant power

scanned across the sample) typically results in better spatial uniformity of the measurement. Although a CCD-based system may compensate for signal magnitude variations by post-processing the data, it cannot compensate for the resulting variation in shot-noise (see section 3.2.2.2.3).

The more commonly used scanner technology involves laser excitation and PMT detection to build microarray images pixel by pixel via raster scanning. Here, well-defined wavelengths of laser light, corresponding to the excitation peaks of incorporated fluorophores, are directed to the microarray slide in a simultaneous or sequential fashion. As the fluorophore-labeled target is excited, the emitted photons impinge upon a photocathode material (PC) to cause photoelectron (PE) emission. The PE charge is then amplified by multiple dynodes to produce a current pulse that is proportional to the amount of incident light [67]. The fraction of impinging photons that result in photoelectrons and contribute to the output signal is referred to as the quantum efficiency Q_E of the PMT (not to be confused with the quantum efficiency Q_E of the dye). Often, confocal or other depth-discriminating optical designs are combined with PMT detection to ensure that only photons emitted from a defined plane of focus (i.e. the microarray features) are quantified. This increases scanner sensitivity by eliminating out-of-focus light not originating from the targets of interest. For state-of-the-art systems this results in a lower limit of detection and – in other words – higher sensitivity [99].

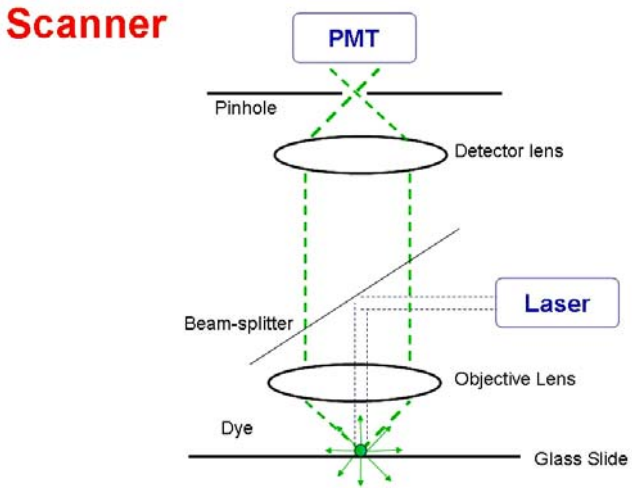


Figure 2.2 – A PMT detector.

2.5 Microarray Image Processing

Image processing has a potentially large impact on the subsequent analysis. The processing of scanned microarray images can generally be separated into four tasks.

- *Denoising* is the process of image noise removal. This step is going to be further analyzed in Section 5, where some denoising techniques are going to be presented.
- *Addressing* or *gridding* is the process of locating the signal spots in images and estimate the size of each spot. Spots are features in the image that are small compared to the whole image, but indeed relatively large when analysed locally and they consist the major amount of the image information. Automating this part of the procedure permits high-throughput analysis while it

reduces the human effort, minimizes the potential for human error and offers high consistency in the quality of the data.
- *Segmentation* allows the classification of pixels either as foreground – that is, within printed DNA spot – or as background.
- The *intensity extraction* step includes calculating, for each spot on the array, red and green foreground fluorescence intensity pairs (R, G), background intensities, and, possibly, quality measures [107].

Estimation of background intensity is generally considered necessary for the purpose of performing background correction. The motivation for background correction is that a spot's measured fluorescence intensity includes a contribution which is not specifically due to the hybridization of the mRNA samples to the spotted DNA. Background correction of the spot intensities is usually performed by subtracting background estimates from the red and green foreground values, with the aim of improving accuracy, that is, reducing bias. Spot quality scores may include measures of spot size and shape, or measures of background intensity relative to foreground intensity [108].

2.5.1 Addressing

The basic structure of a microarray image is determined by the arrayer and is therefore known. For example, it is known in advance that there are four rows and four columns of grids, and that within each grid there are 23 rows and 24 columns. However, to address the spots in an image – that is, to match an idealized model of the array with the scanned image data – a number of parameters need to be estimated. These parameters include separation between rows and columns of grids, individual translation of grids (caused by slight variations in print-tip positions), separation between rows and columns of spots within each grid, small individual translations of spots, and overall position of the array in the image. Within a batch of microarray images produced together, the last of these is usually the most highly variable. Other parameters that may in some cases need to be estimated are misregistration of the red and green channels, rotation of the array in the image, and skewness in the array. The last two parameters are important issues for automated gridding algorithms, but a minor problem if manual gridding techniques are used. In addition, with the improvement of printing and scanning technologies, some of these parameters such as misregistration between the two channels and small individual translations of spots are likely to decrease in importance [107].

To achieve higher levels of accuracy in the measurement process, it is desirable for the addressing procedure to be as reliable as possible. Reliability of the addressing stage can be enhanced by allowing user intervention. However, this can potentially make the process very slow. Ideally we seek reliability while attempting to minimize user intervention to maximize efficiency. The addressing steps are often referred to as "gridding" in the microarray literature. Most software systems now provide for both manual and automatic gridding procedures.

2.5.2 Segmentation

Generally, segmentation of an image can be defined as the process of partitioning an image into different regions, each having certain properties. In a microarray experiment, after the location of the spot is determined, a small area around the spot

(target region) is used to quantify the spot expression level; in other words, the signal and background pixel values are measured. The next step is to determine which of the pixels in the target region are due to the actual spot signal and which of them are considered background. Therefore, segmentation allows the classification of pixels as foreground or background, so that fluorescence intensities can be calculated for each spotted DNA sequence as measures of transcript abundance. Any segmentation method produces a spot mask, which consists of the set of foreground pixels for each given spot. Existing segmentation methods for microarray images can be categorized into four groups, according to the geometry of the spots they produce:

- fixed circle segmentation,
- adaptive circle segmentation,
- adaptive shape segmentation, and
- histogram segmentation.

Table 2.1 lists different segmentation approaches and examples of software implementations. In general, most software packages implement a number of segmentation methods.

Segmentation Methods	Software	Algorithms
Fixed circle	ScanAlyze, GenePix, QuantArray	
Adaptive circle	GenePix, Dapple	
Adaptive shape	Spot	Region Growing and Watershed
Histogram method	ImaGene [106], QuantArray, DeArray, SpotSegmentation	Adaptive Thresholding

Table 2.1 – Segmentation methods and examples of algorithms and software implementation

2.5.2.1 Fixed circle segmentation

Fixed circle segmentation fits a circle with a constant diameter to all the spots in the image. This method is easy to implement and works nicely when all the spots are circular and of the same size. It was probably first implemented in the ScanAlyze software written by Eisen [28] and it is usually provided as an option in most software. A fixed diameter segmentation may not be satisfactory to detect the exact shape for spots varying in diameter.

Theoretically, if the background affects the foreground values additively and the background value can be reliably estimated, one could use a very large fixed diameter for segmentation such that the entire spot is covered for all spots. That is, any segmentation that is too large can yield perfectly good (unbiased) estimates if the background contribution can be removed. On the other hand, an ability to detect the exact shape for all spots limits the amount of irregular noise within the spot mask (for example, bright pixels due to dust, scratch or contribution from neighboring spots) [107].

2.5.2.2 Adaptive circle segmentation

In this kind of segmentation, the circle's diameter is estimated separately for each spot. The software GenePix for the Axon scanner [31] implements such an algorithm. Note that GenePix and other software provide the user with the option to adjust the circle diameter spot by spot. This practice can be very time consuming, since each array contains thousands of spots. The software Dapple [13] finds spots by detecting edges of spots. Briefly, Dapple calculates the negative second derivative (Laplacian) of the image. Pixels with high values in the Laplacian image correspond to edges of a spot. In addition, Dapple enforces a circularity constraint by finding the brightest ring (circle) in the Laplacian images.

Adaptive circle segmentation methods will work rather well as circular spots are probably typical of most commercially produced arrays. However, spots printed from non-commercial arrayers are rarely perfectly circular and can exhibit oval or doughnut shapes [27]. A circular spot mask can thus provide a poor fit for a non-circular shaped spot. Sources of non-circularity include the printing process (e.g. features of the print-tips, uneven solute deposition) or the post-processing of the slides after printing (e.g. insufficient time of rehydration). Again, segmentation algorithms that do not place restrictions on the shape of the spots are thus more desirable if one is attempting to determine the exact spot shape.

2.5.2.3 Adaptive shape segmentation

Two commonly used methods for adaptive segmentation in image analysis are the watershed [7,102] and seeded region growing (SRG) [1]. These methods are beginning to be applied in microarray analysis, although not in the most widely-used software packages.

Both watershed and SRG segmentation require the specification of starting points, or seeds. A weakness of segmentation procedures using these methods can be the selection of the number and location of the seed points. In microarray image analysis, however, we are in the rather unusual situation where the number of features (spots) is known exactly a priori, and the approximate locations of the spot centres are determined at the addressing stage. Microarray images are therefore well suited to such methods. The SRG algorithm is implemented in Spot. Details regarding the placement of foreground and background seeds can be found in Yang et al [108].

2.5.2.3.1 The Watershed Segmentation

Every grayscale image can be interpreted as a topographic surface where the gray-levels of the image (or gradient image) represent altitudes. If the gradient image is used, region edges correspond to high watersheds, while low-gradient region interiors correspond to catchment basins. Watershed segmentation is a region-growing method [7,102]. In watershed segmentation, catchment basins represent the regions of the segmentation. In an image, every grayscale minimum represents a catchment basin and the idea lies in flooding with water every basin starting from the bottom. When floodings from two basins are about to merge a high dam is needed to be built in order to prevent this merging. When all the basins have been flooded the dams constructed represent the watershed lines. In a microarray image the catchment basins

correspond to the spots while the watershed lines are the lines separating two adjacent spots.

However, there are some problems when using the watershed algorithm. Firstly, thick watersheds are often produced when dealing with discrete space, but this can be solved with careful distance computing and queue handling. Secondly, watershed segmentation often leads to over-segmented results, i.e., too many regions. There are various ways in solving this, such as pre-processing (e.g., smoothing to remove small local minima), seeding, in which a marker (seed) is put inside each region and only marked regions are allowed to be found, and post-processing by merging either on edge strength or on valley depth.

2.5.2.3.2 Seeded Region Growing (SRG)

In the seeded region growing (SRG) algorithm of Adams and Bischof [1], a number of seeds are provided as input to the algorithm. These are groups of pixels which serve as starting points for a region growing process. Often seeds consist of only a single pixel, but they can be of any size and do not need to form a connected set. We describe below how we construct the seeds in this application of SRG.

After specification of seeds, the algorithm proceeds by growing all the foreground and background regions simultaneously until all pixels in the image have been allocated to one of the regions. At each stage, all pixels which are still unallocated, but which have at least one neighbor which has already been allocated, are considered for allocation. Out of all these region-neighboring pixels, the algorithm selects the one whose pixel value is nearest (in terms of absolute grey-level difference) to the average of the pixel values in the neighboring region. The process is repeated until all pixels have been allocated. Pixel queues are used to optimize the efficiency of the procedure.

For microarray segmentation using SRG, the foreground and background seeds are chosen using the grids calculated in the addressing stage. An obvious way to choose a seed for each spot is to choose a single pixel from the intersections of the horizontal and vertical grid lines of the fitted foreground grid. However, it is possible, particularly when the spot is small, that this intersection pixel may not be inside the spot because of local irregularities or small errors in the grid estimation. To overcome this problem, a point is chosen by finding the maximum of the combined intensity surface over a small region centered at the intersection pixel. The foreground seed is then set to be an n-by-n square of pixels centered on this point. The integer n is specified by the user.

Background seeds need to be computed also. A very simple approach would be to use the intersection points from the fitted background grids as background seeds, or indeed to use all of the grids together as one large background seed covering most of the image. Such a procedure has the advantage of separating the foreground seeds from each other and therefore ensuring that the segmented spots cannot merge or bleed into one another. There are, however, two reasons why the use of such large background seeds is undesirable. The first is that background intensity is often locally varying and poor performance is expected for SRG if regions are not homogeneous in intensity. A second reason is that we require local estimation of background intensity and this can be obtained by having smaller, more local background regions.

2.5.2.4 Histogram segmentation

This method uses a target mask chosen to be larger than any other spot. For each spot, foreground and background intensity estimates are determined in some fashion from the histogram of pixel values for pixels within the masked area. For example, QuantArray software for the GSI Lumonics scanner [71] uses a square target mask and defines foreground and background as the mean intensities between some predefined percentile values. By default, these are the 5th and 20th percentiles for the background and the 80th and 95th percentiles for the foreground. These methods therefore do not use any local spatial information.

Another example of this class of methods is described by Chen et al [14]. This method uses a circular target mask and computes a threshold value based on a Mann-Whitney test. Pixels are classified as foreground if their value is greater than the threshold and as background otherwise. This method is also implemented in the QuantArray and DeArray by Scanalytics. Simplicity is the main advantage of this method. However, a major disadvantage is that quantification is unstable when a large target mask is set to compensate for variation in spot size. Furthermore, the resulting spot masks are not necessarily connected.

SpotSegmentation software [51] employes a model-based clustering of pixels, which actually is a histogram-based method. This model-based clustering allows the estimation of the number of groups in the target area, and hence provides a formal basis for determining whether or not a spot is present. Typically, background pixels would be one group and pixels in the spot or foreground would be another. In addition, if an artifact is present, or if the spot is donut-shaped and has an inner hole, the corresponding pixels would form a third group. Artifacts often take the form of small disconnected groups, and so a thresholding on the size of the connected components in the spot cluster can identify a third cluster formed by artifacts.

2.5.3 Information Extraction

Information extraction deals with the measuring of the spot signal and background values. In microarrays, the key information that needs to be recorded is the expression strength of each target. When studying gene expression techniques we are typically interested in the difference in expression levels between the test and reference mRNA populations. This translates into differences in the intensities on the two images. Under idealized conditions, the total fluorescent intensity from a spot is proportional to the expression strength. To have this idealized situation we should prepare the probe cDNA so as its concentration in the solution is proportional to that in the tissue. Secondly, the hybridization experiment must be done in such a way that the amount of cDNA binding in the spots must be proportional to the probe cDNA concentration in the solution. Moreover, the amount of cDNA deposited on each spot during the chip fabrication should be constant, the spots must be uncontaminated and the signal pixels must be correctly identified by the image analysis software [44].

2.5.3.1 Spot intensity

Each pixel value in a scanned image represents the level of hybridization at a specific location on the slide. The total amount of hybridization for a particular spotted DNA sequence is proportional to the total fluorescence at the spot. The natural measure of

spot intensity is therefore the sum of pixel intensities within the spot mask. Since later calculations are based on the ratio of fluorescence intensities, we compute the average pixel value over the spot mask. This yields identical results, as the ratio of averages is equal to the ratio of sums. An alternative measure used is ratio of medians, where the median pixel value over the spot mask is computed. This measure is not associated with any biological meaning but can be seen as a robust variant of the ratio of means.

2.5.3.2 Background intensity

The motivation for background adjustment is the belief that a spot's measured intensity includes a contribution not specifically due to the hybridization of the target to the probe, for example, nonspecific hybridization and other chemicals on the glass. If such a contribution is indeed present, we would like to measure and remove it to obtain a more accurate quantification of hybridization. The glass slides are treated chemically so that the spotted cDNA fragments will bind to them. After the cDNA spots are printed, the slides are treated again so that target DNA does not bind to them. Nevertheless, some binding of the target to the slide may occur. Furthermore, there may be some fluorescence away from the spots due to the slide's surface treatment and the glass. It seems likely that the fluorescence from regions of the slide not occupied by DNA is different from that from regions occupied by DNA. It follows that measuring the intensity in some region near a spot and subtracting it may not be the best way to correct for this extra contribution, even though this is what many people are doing. It would be interesting to compare the morphological and local background estimates to ones based on local negative controls (i.e. nearby spotted cDNA sequences which should have no hybridization signal).

Apart from histogram-based methods, the rest segmentation procedures described above produce local background regions, as well. We can broadly classify the various background methods implemented in software packages into four categories.

2.5.3.2.1 Local background

Background intensities are estimated by focusing on small regions surrounding the spot mask. Usually, the background estimate is the median of pixel values within these specific regions. Most software packages we have encountered implement such an approach.

The ScanAlyze package considers as background all pixels that are not within the spot mask but are within a square centred at the spot centre. This is represented by the dotted square in Figure 2.3. The median value of these pixels is used as an estimate of the local background intensity. One of the background adjustment methods implemented in QuantArray, ArrayVision and ImaGene considers the area between two concentric circles, such as the area between the two larger circles in Figure 2.3. By not considering the pixels immediately surrounding the spots, the background estimate is less sensitive to the performance of the segmentation procedure. An alternate set of pixels to be considered as background (implemented in Spot) is shown as the four dashed diamond-shaped areas in Figure 2.3. These regions are referred to as the valleys of the array and have the furthest distance from all four surrounding spots. The local background for each spot can be estimated by the median of values from the four surrounding valleys. Depending on the software, the local valley

regions are different, but this method of background estimation is somewhat independent of the segmentation results. The background method implemented by GenePix effectively calculates the median intensity from local valley regions.

Using valley pixels which are very distant from all spots ensures to a large degree that the background estimate is not corrupted by pixels belonging to a spot. Such corruption by bright pixels may occur in the other methods, particularly in the ScanAlyze method, introducing an upward bias into the background estimate. Using remote pixels reduces this bias effectively but entails the use of a smaller number of pixels and therefore increases the variance of the estimate. This is an example of the bias—variance trade-off. Most software packages allow users to choose their preferred version of local background method.

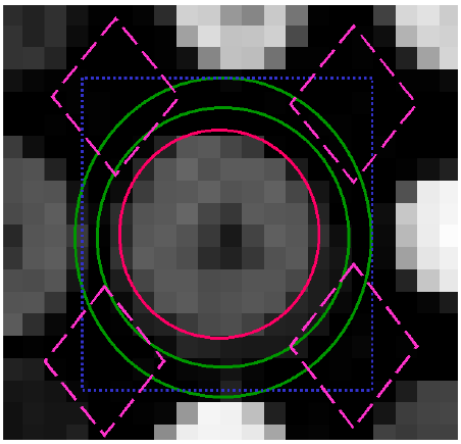


Figure 2.3 – Different local background approaches.
The spot is limited by the red circle. The other colored lines bound the regions used for local background calculations by different methods. Green ring: QuantArray, ImaGene, Blue rectangle: ScanALyze, Purple diamonds: Spot.

2.5.3.2.2 Morphological opening

This approach to background adjustment relies on a non-linear filter called morphological opening [89]. This filter is obtained by computing a form of local minimum filter (an erosion) followed by a form of local maximum filter (a dilation) with the same window. In a microarray image, the effect of such non-linear filtering using a window that is larger than any of the spots is to remove all spots, replacing them by nearby background values.

In Spot, morphological opening is applied to the original images R and G using a square structuring element with side length at least twice as large as the spot separation distance. This operation removes all the spots and generates an image that is an estimate of the background for the entire slide. For individual spots, the background is estimated by sampling this background image at the nominal centre of the spot. We simply chose to sample this image rather than take an average over a "background region" because very similar results are expected from both methods. A large window was used to create the morphological background image; hence it is expected to have slow spatial variation.

Morphological opening results in lower background estimates than other simpler methods. More importantly, though, morphological background estimation is

expected to be less variable than the other methods, because spot background estimates are based on pixel values in a large local window, and yet are not corrupted (i.e. biased upwards) by brighter pixels belonging to or on the edge of the spots.

2.5.3.2.3 Constant background

This is a global method which subtracts a constant background for all spots. The approaches previously described assume that the non-specific binding to a spot can be estimated by the surrounding area. However, some findings [54] suggest that the binding of fluorescent dyes to "negative control spots" (e.g. spots corresponding to plant genes that should not hybridize with human mRNA samples) is lower than the binding to the glass slide. If this is the case, it may be more meaningful to estimate background based on a set of negative control spots. When there are no negative control spots, one could approximate the average background by, for example, the third percentile of all the spot foreground values.

2.5.3.2.4 No adjustment

Finally, we also consider the possibility of no background adjustment at all.

2.5.3.3 Quality measures

In addition to the actual spot foreground and background intensities, it is also desirable to collect statistics describing the quality of these measurements. Examples of quality measures provided in most software include variability measures in pixel values within each spot mask, spot size (area in pixels), a circularity measure and relative signal to background intensity. Most software packages provide a reject and accept assessment on spot quality. Dapple defines two measures: b-score measures the fraction of background intensities less than the median foreground intensity while p-score measures the extent to which the position of a spot deviates from a rigid rectangular grid. A classifier is built based on these two measures to accept, reject or flag any spots. Flagged spots need to be manually accepted or rejected.

Most programs have yet to make fuller use of these measures in their analysis, as relating them to more common statistical concepts such as reproducibility seems to be difficult. Research along these lines is being carried out.

3 SIGNAL AND NOISE IN MICROARRAY EXPERIMENTS

Microarray experiments involve a large number of error-prone procedures that lead to a high level of noise in the resulting data. The high level of the uncertainty associated with each microarray experiment originates by biological variations (corresponding to real differences between different cell types and tissues) and experimental noise. This uncertainty often obscures some of the important characteristics of the biological processes of interest. More specifically, changes in the measured transcript values in the samples render the clustering of genes into functional groups [26,72], and the classification of samples difficult [88,100].

The main objective of this dissertation is to eliminate the effect of the noise and recover the gene expression measurements, which is a major challenge in microarray analysis. Before all, it is essential to discriminate signal from noise in a microarray experiment. Signal in a microarray experiment is defined as the desired output, whereas noise is defined as the sum of unwanted contributions to the instrument readings. Due to the fact that most microarray assays are based on fluorescent signals, the signal in nearly all microarray experiments derives from emission of fluorescent light from tags attached to the probe molecules. Noise, on the other hand, has many different origins. The quotient of signal and total noise is known as the signal-to-noise ratio. Because the information in microarray assays in contained within the signal, one goal in all microarray experiments is to maximize the ratio of signal to noise [80].

3.1 SIGNAL DETERMINANTS

Though the source of signal is much simpler to understand than the vast number of contributors to noise, total signal in a microarray analysis actually has three different determinants: intrinsic, extrinsic, and quantity. The intrinsic signal determinants are those that are inherent to the labels used on the probe molecules. Because nearly all microarray techniques use fluorescent labels, intrinsic signal determinants include the physical properties of the fluorescent dyes, such as molar extinction coefficient and quantum yield. To maximize microarray signals, prudent selection of dyes and other types of labels with superior intrinsic properties is the key.

Extrinsic signal determinants are external contributors to signal, the most pertinent in microarray experiments being instrument determinants and environmental determinants. With fluorescent detection instruments, some of the key instrument determinants of signal include the power of the light source, excitation wavelength, detection dwell time, and efficiency of the light path, detector and arid analog-to-digital converter. The main environmental determinants include the polarity of the solvent, pH of the buffer, presence of quenching species, and extent of energy transfer between adjacent dye molecules. Because most microarray substrates are detected in the dry state, solvent polarity, pH and presence of quenching species are fairly minor environmental signal determinants. Energy transfer or self-quenching, on the other hand, can exert a rather major effect on signal. Because energy transfer increases with the increasing proximity of dye molecules to each other, there can be a nonlinear

relationship between probe quantity and signal output. To obtain maximum microarray signals, all of the main extrinsic signal determinants, including instrument and environmental contributions, need to be optimized.

Microarray signal is also determined by the quantity of the label present at a given position on the microarray. Quantity determinants include the number of probe molecules bound and the number of labels present per bound probe molecule. The number of labels attached to a given probe molecule is sometimes referred to as the specific activity of the probe. Because the specific activity can vary greatly depending on the labeling scheme, there is not a 1:1 relationship between the number of molecules bound and the number of labels present at a given microarray location. Under conditions of target excess, a greater probe concentration results in a larger number of probe molecules binding to the surface. Stronger microarray signals are always observed with greater probe concentrations and higher specific activities and the researcher should always endeavor to maximize the quantity determinants in order to achieve maximum signal. With fluorescent labeling schemes, nonlinearity can be observed at high dye concentrations due to energy transfer between adjacent dye molecules [80].

3.2 NOISE IN MICROARRAY EXPERIMENTS

One of the major difficulties in decoding gene expression experiments comes from the noisy nature of the data. All undesirable features causing discrepancies in the digital image are considered noise. Noise is caused by both biological variations and experimental noise. To correctly interpret these data, it is crucial to understand the sources of the experimental noise. The noise sources are external (due to random nature of light, dust in the air, scratches on the objects being observed) or internal (due to the operation of the video sensor itself). In this dissertation, we are most interested in the noise which is caused by the microarray image generation process. Its origin, usually, involves the collection of fluorescence of the labeled samples, the amplification of the analog signal and the conversion to digital through dedicated imaging devices. In this section, most sources of noise in DNA microarray process are going to be presented [44, 99, 80, 81].

3.2.1 Systematic Noise vs. Random Noise

In microarray analysis the most damaging noise is background reflection, array misalignment, or scratches and dust on the film surface. As far as where the discrepancies appear, the noise may be geometric (spatial discrepancies in the image) or radiometric (discrepancies in the pixel value), while regarding its effects noise is either systematic or random [35]. Systematic noise affects the accuracy of the measurements made from the images, and random noise affects also the precision. Random noise includes randomness in the biological process, the camera noise, as well as random variation in the spot size and shape. Notice that the systematic and random noise are combined in the image, so treating the systematic noise, by default requires understanding the random noise as well.

Systematic noise is defined as the unwanted deviations from the intended detection protocol. If these errors are accurately evaluated, in theory, they can be compensated by post experiment data processing. If not, they result in a particular type of

measurement uncertainty, typically referred to as systematic noise. Systematic noise can be minimized by proper use of sensors, calibration procedures and proper set up of the experiment.

We define random or inherent noise of the detection system as the unavoidable uncertainties even with ideal detection where no systematic error exists. Inherent noise is basically inevitable since it originates from the stochastic nature of molecular-level interactions. Random noise, therefore, is not reproducible because there is variability in the pixel values each time images are taken.

3.2.2 Noise Sources

In general, the expression level uncertainty in microarray systems, fundamentally originates from the probabilistic characteristics of the detection process, all the way from sample extraction and mRNA purification to hybridization and imaging. We are at the greatest extend interested in the noise that originates from the microarray process because, as foresaid, this is the one that can help us in correctly interpret the gene expression microarray data.

3.2.2.1 Biological Noise

Noise enters very early in the microarray process due to biological variations. When the original volume, from which the sample is going to be extracted, noise due to randomness is introduced, because the position of the volume is not exactly the same in each repetition of the experiment. In the same way, the extraction of the desired fraction is also a random process. From the biological point of view, noise due to randomness occurs in three steps; first, in mRNA preparation, where probes may look very different from sample to sample depending on tissue and sensitivity to RNA degradation, second, in the reverse transcription to cDNA, which will result in DNA species of varying lengths, and finally in the fact that the clones of cDNA are subjected to PCR amplification, which is difficult to quantify and may fail completely.

3.2.2.2 Experimental Noise

Total noise in microarray detection is defined as the sum of all unwanted contributions to the instrument readings. There are many different sources of noise, with the two types being instrument noise and microarray noise. Instrument noise includes – among others – dark current, electronic noise, shot noise and optical noise. Microarray noise includes all of the noninstrument noise components of the system and consists of the chip-based sources that include substrate noise and sample noise. In some of the early detection instruments, instrument noise was a major source of noise in the system. In most modem systems, the chief component of total noise derives from microarray noise, placing great importance on high-quality surfaces and reaction chemistries. Each of these sources of noise is explained below.

3.2.2.2.1 Dark Current

Dark current, as the term implies, is instrument noise that originates in the absence of light. Dark current, or dark count, originates from the instrument detector and derives mainly from thermal PE emissions from the dynodes, thermally excited PE's leaving

the PC, or leakage currents between the electrodes [68], usually measured in electrons (e^-) per pixel per second at a given temperature. All instrument detectors including photomultiplier tubes (PMTs), CCD cameras, and CMOS cameras exhibit measurable dark current. PMTs and cameras that have very low dark current ratings are obviously the devices of choice for microarray detection systems, which endeavor to provide the greatest detectivity possible. Because scanning systems acquire data over a given area very rapidly, dark current is a minor consideration for most modern scanners and can be nearly negated through the proper choice of PMTs. Dark current can be a major source of noise for imaging systems, which require up to 60 seconds to read a given area. Typically, dark current doubles with every increase of 8 degrees Celsius [5]. For this reason, most CCD- and CMOS-based imaging systems used cameras cooled down to as low as -50^o C to reduce dark current noise, with high-quality cooled cameras providing ratings in the range of 0.5-2.0 e^-/pixel/s [80]. This is necessary in applications where the signal level is weak compared to the noise level such as medical applications.

3.2.2.2.2 Electronic Noise

A second source of instrument noise is electronic noise, which arises from the nondetector electrical components of the detection system, notably the amplifiers, circuitry, and analog-to-digital converter. For most microarray detection systems, electronic noise contributes less instrument noise than dark current.

3.2.2.2.3 Shot Noise

Shot noise is unwanted signal that derives from the fundamental process of electrical current flow, which corresponds to the discrete movement of electrons rather than a. continuous flow process. Because microarray detection systems are light based, shot noise, more precisely, derives from the fact that electrical flow is determined by the emission of photons from fluorescent sources, which fundamentally consist of particles rather than continuous beams causing fluctuations of the photon levels in the incoming light. As signal intensity increases, so does the level of shot noise, albeit proportional to the square root of signal [44]. Although this type of noise cannot be eliminated from any scanner, it can be estimated and accounted for in the data extraction model. In well-designed optical systems, all other sources of instrumental noise are minimized such that shot noise is the major contributor [67]. It should be noted, that the square-root dependence of noise on signal described above has its limitations: if signal is increased by integrating over more photons, the overall signal-to-noise ratio changes from being limited by photon statistics to being limited by molecular statistics [20]. For example, if only one photon is detected on average per hybridized molecule, then the resulting signal-to-noise ratio is already close to 70% of the limit set by the number of molecules present [99].

3.2.2.2.4 PMT Noise

As described earlier, PMTs detect fluorophore-emitted photons and amplify the signal through a series of dynodes to produce a current pulse. Fluctuations in this signal amplification that are not reflective of the initial photon emission are considered PMT excess noise. This is a multiplicative noise. Dark current of a PMT (or other detector) may cause additional noise. This dark current noise and electronic noise is sometimes

referred to as the additive noise in the detector [67]. Furthermore, poor signal amplification and digitizing circuitry can contribute additional noise in poorly designed systems. High quality components, precision engineering and low noise design can minimize the contribution of PMT and electronic noise to the overall measurement.

Because PMT gain/sensitivity is partially dependent upon an applied voltage, many commercial microarray scanners enable users to adjust this voltage with each slide. This allows users to increase the intensity of dim features on a microarray, or to decrease the intensity of saturated features. Although researchers may prefer visibly bright arrays, it should be noted that both signal and background intensity increase proportionally with increased PMT voltage. This means that researchers will typically not improve the signal-to-noise ratio of an array by increasing the PMT voltage, unless system noise was dominated by additive electronic noise or digitization noise which typically occurs for poorly designed systems only [4]. At very low PMT gain, the PMT excess noise may become noticeable too. In addition, adjusting PMT voltages often becomes a process of trial and error, as there is a highly non-linear relationship between PMT voltage and the resulting signal levels that a user sees in the image file. Even different PMTs from the same manufacturing lot can produce widely different results with the same applied voltage.

3.2.2.2.5 Laser Noise

In quantifying the fluorescent emission of microarray features, the assumption that the level of fluorescence is proportional to the amount of endogenous gene transcription is made. Because the level of fluorescence is also proportional to the amount of laser light falling on the array, it is important to compensate for laser drift over time. Noise in the laser can contribute to noise in the image, all other things being equal. Without laser monitoring and control, users may detect decreased microarray intensity over the life of the laser or even intensity fluctuations over the course of a single scan. Although some lasers can self-correct for temperature fluctuations or compensate for long-term laser light drift, these internal sensors respond in minutes and cannot compensate for real-time fluctuations. To maintain data integrity over the course of a single scan and consistent intensities over the life of the laser, users should consider scanners with external laser power modulation as well. This ensures that a uniform intensity of laser light will be applied to all features on the same microarray. In addition, external laser modulation virtually eliminates the long-term signal drift due to laser aging and enables calibration across scanners. This is critical for high throughput facilities where obtaining comparable results across multiple microarray scans and scanners is important.

3.2.2.2.6 Non-Uniformity

While scanning microarray slides in the X- and Y-direction, microarray scanners must also track the slides' surface in the Z-direction. This maximizes scanner sensitivity by restricting measurements to the light emitted from DNA features on or close to the microarray surface, rather than out-of-focus light. In the absence of autofocus, spatial non-uniformity of sensitivity across the scanned image can occur. This may result in artifacts that add to signal noise or, worse still, cause a bias that may be mistaken for a biological change. This source of noise is important to consider because the glass

surface of microarrays varies in thickness, surface roughness, and curvature. If the scanner cannot accurately measure DNA features in the focal plane, then sensitivity, uniformity, and data integrity will be compromised.

Some scanner manufacturers address this issue by widening the field of focus (depth discrimination). Although this approach addresses the surface variability, it decreases the overall scanner sensitivity by measuring light that does not originate from the DNA features of interest. Other manufacturers optimize the slide positioning in order to narrow the depth discrimination. Although this ensures maximal sensitivity for features within the set plane of focus, it does not account for curvature or variability in the glass surface. As a result, signal intensity will decrease for out-of-focus features. This reflects poorly on the scanner's field uniformity and data integrity because the X-Y position of DNA features now becomes important in the resulting signal intensity.

Poor scanner uniformity can be detected by scanning a microarray slide in one direction, turning the slide 180 degrees for re-scanning in the other direction, and comparing the data. Because variability in log ratios resulting from instrumentation can compromise the statistical confidence with which scientists measure differential gene expression, field uniformity specifications are important in selecting a quality microarray scanner.

3.2.2.2.7 Optical Noise

Optical noise refers to all components of instrument noise that require light, excluding shot noise. The most common sources of optical noise include reflected light from the substrate holder, spurious reflections from instrument enclosures, light leaks impinging on the detectors, and cosmic rays. In properly designed systems, optical noise can be greatly minimized but not eliminated.

3.2.2.2.8 Fixed-pattern noise

The non-uniformity in the physical characteristics of the individual sels is manifested in fixed pattern noise in dark images and flat fields. Dark images are images obtained with no presence of light (with lens caps on) and flat fields are obtained under uniform illumination (with an integrating sphere or defuse filters). Impulse noise is due to pixels whose responses differ significantly from their neighbors. A very high level of impulse noise is manifested as salt and pepper, or speckle noise, and an extremely high over-saturation of pixels, as blooming.

3.2.2.2.9 Substrate Noise

One of the two components of microarray noise is known as substrate noise. Noise from substrates derives either from the substrate material itself or from the surface treatment or surface coating that is applied to the substrate. Because most microarray substrates are made of glass, the inherent properties of transparency and low intrinsic fluorescence render glass substrates minimal contributors to substrate noise. Other substrate materials, including plastics and reflective metals, may present a considerable source of substrate noise in the system.

The noise contributed by surface treatments and surface coatings are generally the main source of substrate noise. The intrinsic fluorescence of different organic treatments can vary by more than three orders of magnitude, resulting in major research and developments efforts to reduce the noise contributed by the surface treatment. The high-quality organoamine and organoaldehyde surfaces generally contribute less than twofold noise above and beyond the contribution of glass itself and therefore enable remarkable detectivity when used with an appropriate detection instrument. Some of the older organic surface treatments produced noise levels that were 1000-fold greater than glass and thus reduced detectivity greatly. Gel and nitrocellulose coatings generally produce somewhat greater noise than the organic surface treatments, though microarray analysis with these surfaces can be successfully implemented by making adjustments to the instrument settings [80].

3.2.2.2.10 Sample Noise

Sample noise represents the second component of microarray noise. Noise from samples is introduced by the targets, probes, or solutions used to dissolve these components. Because most target molecules and target buffers are non-fluorescent, microarray targets generally contribute little to the sample noise. The main component of sample noise far and away, is attributable to the fluorescent probe molecules. Labelled probe solutions can react in a non-specific manner with the surface. This non-specific sticking of probe molecules to the surface can mask the productive interactions between targets and probes, obscuring the microarray signal. The noise attributed to non-specific interactions between probe molecules and the microarray surface is known commonly as background. The background noise is due to thermally generated dark current and internal luminance in the camera. In all cases, the background noise is temperature dependent, so cameras must be warmed up before use to allow background noise to stabilize [45]. Background fluorescence reduces the signal-to-noise ratio by elevating the noise and, therefore, compromises microarray assay detectivity. Of all of the sources of noise in microarray systems, background noise contributed by non-specific probe molecule interactions with the surface generally constitutes the main component. Vast resources have been devoted over the past few years to reducing background noise, with major successes in new blocking schemes, labelling procedures, and reaction and wash chemistries. At present, it is possible to perform microarray analyses with instrument controls adjusted to the highest settings of lasers and detectors. At these instrument settings, it is possible to detect a few dozen molecules bound to a single microarray spot [80].

3.2.2.2.11 Quantization noise

Quantization noise occurs from errors in assigning the pixel gray levels. When the number of quantization levels is not sufficient to represent faithfully the continuous signal, false contours may appear in the digital image. In addition, there is random quantization noise [91]. Geometric (spatial) noise is due mainly to the sampling process. The sels spacing puts a limit on the highest spatial frequency that could be recorded in the digital image and the area of the CCD chip on the lowest one. Geometric distortions in the images are in the center of the digital signal processing literature. In any case, violation of the sampling theorem leads to severe geometric or radiometric distortions [50,8]. For high quality digital cameras, these distortions are minimal.

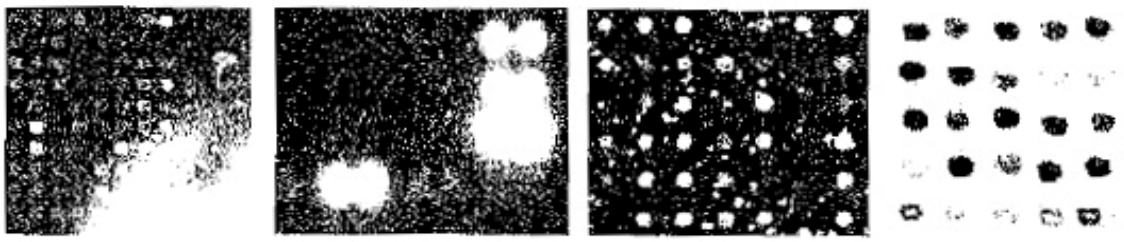


Figure 3.1 – Noisy Images.
From left to right: background illumination, blooming, dust, irregular spots locations and shapes (BioDiscovery, Inc.).

4 SIGNAL TRANSFORMATION

In general, the expression level uncertainty in microarray systems, fundamentally originates from the probabilistic characteristics of the detection process, from sample extraction and mRNA purification to hybridization and imaging. In the following, we formulate the microarray image noise removal with a brief essential overview of the signal model.

Denote by *I(x,y)* a noisy observation (i.e., the microarray image) of the two-dimensional (2-D) function *S(x,y)* (i.e., the noise-free image that has to be recovered) and by $n_m(x,y)$ and $n_a(x,y)$ the corrupting multiplicative and additive noise components, respectively. One can write

$$I(x,y) = S(x,y) \cdot n_m(x,y) + n_a(x,y) \tag{4.1}$$

The importance of including both additive and multiplicative measurement-specific noise in an error model for gene arrays is already established in the literature [74]. The omission of the measurement-specific additive noise term leads to exaggerated ratio estimates, false identification of significant differences, and understated uncertainty measures when the observations are small. The omission of the multiplicative noise term leads to similar problems when the observations are large.

To estimate the multiplicative noise component we have to ignore the additive component $n_a(x,y)$, and then (4.1) becomes

$$I(x,y) = S(x,y) \cdot n_m(x,y) \tag{4.2}$$

To transform the multiplicative noise model into an additive one, we apply the logarithmic function on both sides of (4.2)

$$\log I(x,y) = \log S(x,y) + \log n_m(x,y) \tag{4.3}$$

Expression (4.3) can be rewritten as

$$f(x,y) = g(x,y) + e(x,y) \tag{4.4}$$

where $f(\bullet)$, $g(\bullet)$, and $e(\bullet)$ are the logarithms of $I(\bullet)$, $S(\bullet)$, and $n_m(\bullet)$, respectively.

At this stage, one can consider to be white noise and subsequently apply any conventional additive noise suppression technique, such as Wiener filtering. However, it is recognized that standard noise filtering methods often result in blurred image features. Indeed, single-scale representations of signals, either in time or in frequency, are often inadequate when attempting to separate signals from noisy data. The wavelet transform has been proposed as a useful processing tool for signal recovery [2], [103], and is going to be analyzed straight forward.

The wavelet transform is a linear operation. Consequently, after applying the DWT to (4.4) we get, in each of the three directions (horizontal, vertical, diagonal), sets of noisy wavelet coefficients written as the sum of the transformations of the signal and of the noise

$$d^{i}_{j,k} = s^{i}_{j,k} + n^{i}_{j,k} \tag{4.5}$$

where k = 0, …, 2^{J+j} -1 and -1<j<-J refer to the decomposition level or scale and i = 1, 2, 3 refers to the three spatial orientations.

4.1 WAVELETS

Microarray images appear to have areas with both high and small frequency. Therefore, Fourier transform is not the proper tool for the image analysis as it is a stationary transform. Thus we use the wavelet transform which is known to be a superior approach to other time-frequency analysis tools due to its window of varying time scale width. Window width can be increased in time domain (thus decreased in frequency domain) when small frequency attributes are analyzed and decreased in time domain (increased in frequency domain) when high frequency attributes are analyzed. Therefore, it can match identically the original signal.

The wavelet transform provides an appropriate basis for separating noisy signal from the image signal. The motivation is that as the wavelet transform is good at energy compaction, the small coefficients are more likely due to noise and large coefficients due to important signal features [94,33,19]. These small coefficients can be processed in order to denoise the image without affecting the significant features of the image.

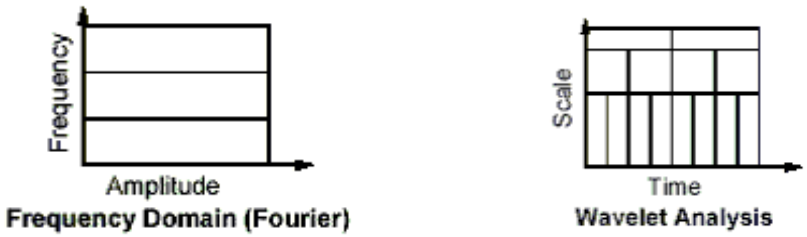


Figure 4.1 – Frequency and Wavelet based signal views.

4.1.1 Continuous Wavelet Transform (CWT)

We begin with a window function $\varphi(t)$ which is called a *mother wavelet* or *basic wavelet* [43]. This function introduces a scale in the analysis and since we want the transform to be scale-independent we will use every possible scaling of φ. To accomplish this, we arbitrarily fix $p \geq 0$ and for any real, non-zero number a, which we shall call scale factor, we define

$$\psi_a(t) \equiv |a|^{-p}\, \psi(t/a) \tag{4.6}$$

For various scale factors we observe the following relations between $\varphi_a(t)$ and φ:

- for $a > 1$, $\varphi_a(t)$ is a version of φ stretched by a in the horizontal direction.
- for $0 < a < 1$, $\varphi_a(t)$ is a version of φ compressed by a in the horizontal direction.
- for $a = -1$, $\varphi_a(t)$ is the reflection of φ.
- for $-1 < a < 0$, $\varphi_a(t)$ is reflected and compressed version of φ.
- for $a < -1$, $\varphi_a(t)$ is a reflected and stretched version of φ.

The factor $|a|^{-p}$ has a similar effect on the vertical direction. If p is positive then ψ is compressed along the vertical direction whenever stretched along the horizontal and it is stretched along the vertical whenever compressed along the horizontal. Usually p is set equal to ½ and so we get:

$$\psi_a(t) \equiv \frac{1}{\sqrt{a}}\psi(t/a) \tag{4.7}$$

Time localization of signals is achieved by looking at them through translated version of ψ_a. If $\psi(t)$ is supported on an interval of length T near t = 0, then $\psi_a(t)$ is supported on an interval of length $|a|$T near t = 0 and the function

$$\psi_{a,b}(t) \equiv \psi_a(t-b) \equiv \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{4.8}$$

is supported on an interval of length $|a|$T near t = b. The functions given by (4.7) are called *wavelets* and b is known as *translation parameter*.

Continuous wavelet transform of a signal $f(t)$ is defined as:

$$\tilde{f}(a,b) \equiv \int_{-\infty}^{\infty} dt\overline{\psi}_{a,b}(t)f(t) = \left\langle \psi_{a,b}, f \right\rangle = \psi_{a,b}^{*} f \tag{4.9}$$

where $a \in R^{+}$-{0}, $b \in R$

and as a function of b for a given a it represents the details contained in the signal $f(t)$ at the scale a. The result of CWT is a lot of wavelet coefficients, which are functions of time and frequency.

By reducing a, the support of $\psi_{a,b}$ is reduced in time and hence covers a larger frequency range. That is why, we consider the factor 1/a to be a frequency measure. On the other hand, b indicates the location of the wavelet window along the time axis. Thus, by altering a and b, CWT can be computed on the entire time-frequency plane [32].

For ψ to be a window function and to recover $f(t)$ from the inverse wavelet transform, $\psi(t)$ has to satisfy the following condition:

$$\widehat{\psi}(0) = \int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{4.10}$$

which states that the zeroeth Fourier coefficient must be 0. For the satisfaction of (4.10), the wavelet has to be constructed so that it has a higher order of vanishing moments [32]. A wavelet is said to have vanishing moments of order m if

$$\int_{-\infty}^{\infty} t^{p}\psi(t)dt = 0 \quad p = 0, \ldots, m\text{-}1 \tag{4.11}$$

Moreover, when the wavelet's k + 1 moments are equal to zero all the polynomial signals $s(t) = \sum_{0 \le p \le m-1} a_p t^p$ have zero wavelet coefficients. As a consequence, the details are also zero. This property ensures the suppression of signals that are polynomials of a degree lower or equal to k [62].

In addition, equation (4.10) declares that all wavelets must oscillate, giving them the nature of small waves hence the name wavelets [32].

CWT has the following property:
If we scale the signal by a factor σ as we have done to the wavelet, that is

$$f_{\sigma}(t) \equiv \sigma^{-\frac{1}{2}} f(t/\sigma) \tag{4.12}$$

then we have

$$\widetilde{f}_{\sigma}(\sigma a, \sigma b) = \widetilde{f}(a,b) \tag{4.13}$$

A more practical way to compute the CWT coefficients consists of the five following steps (Figure 4.2) [62]:

1. Take a wavelet and compare it to a section at the start of the signal.
2. Compute a number C ($=\widetilde{f}$), that represents how closely correlated the wavelet is with this specific section of the signal. The larger C is, the higher the correlation. More precisely, if signal's and wavelet's energies are equal to 1, C can be considered as a correlation coefficient.
3. Shift the wavelet to the right and repeat steps 1 and 2 until you cover the whole signal.
4. Stretch the wavelet and repeat steps 1 to 3.
5. Repeat steps 1 to 4 for all scales.

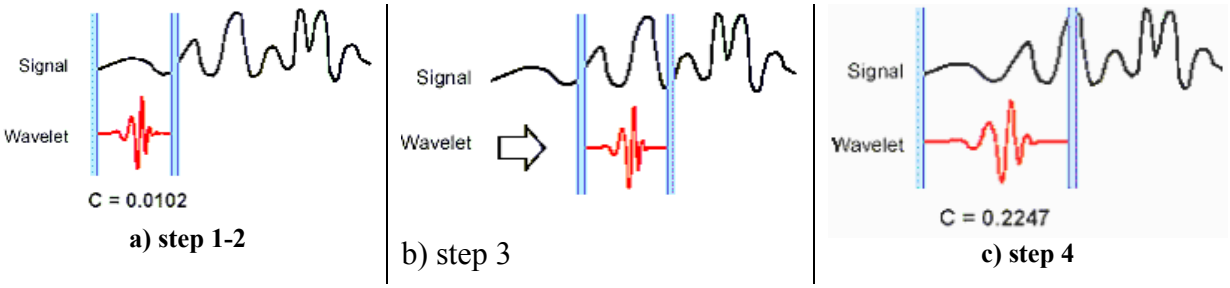


Figure 4.2 – Steps 1– 4 of the CWT coefficients' computation algorithm.

When you finish, you will have all coefficients produced at different scales by different sections of the signal. These coefficients can be presented as in Figure 4.4 and Figure 4.3.

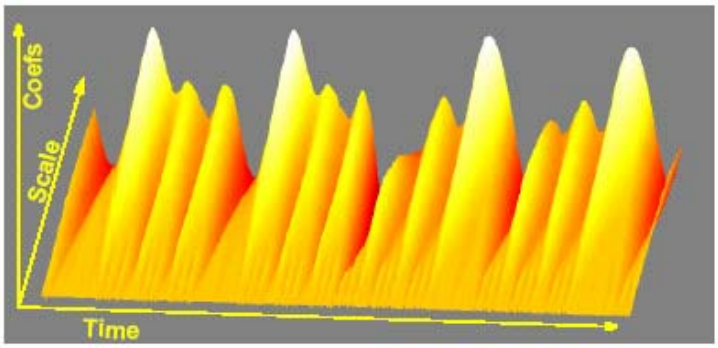


Figure 4.3 – Wavelet coefficients' presentation.

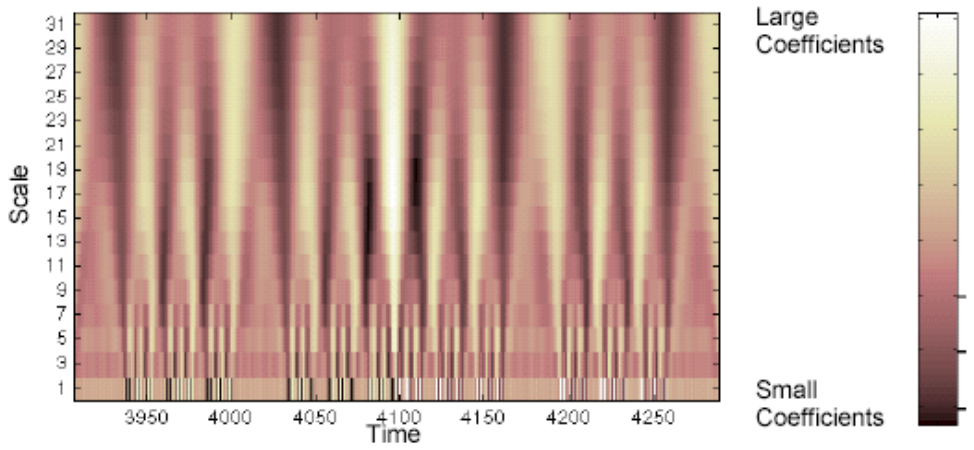


Figure 4.4 – Another presentation of Figure 4.3.
The horizontal axis represents time, the vertical axis represents scale and the color at each point represents the magnitude of the wavelet coefficient.

4.1.2 Discrete Wavelet Transform (DWT)

Continuous Wavelet Transform has a serious disadvantage; the coefficients have to be computed for each possible scale, resulting in a great computational cost and a massive amount of data. If the scales and positions are power of two – thus called dyadic scales and positions – and Discrete Wavelet Transform is used for the signal processing, the analysis would be as precise as CWT and even more efficient. DWT keeps enough information of the signal such that it reconstructs the signal perfectly from the wavelet coefficients. This process is known as critical sampling [32].

The discrete wavelet transform is defined as:

$$\widetilde{f}(a,b) \equiv \int_{-\infty}^{\infty} dt\, \overline{\psi}_{a,b}(t) f(t) = \left\langle \psi_{a,b}, f \right\rangle = \psi_{a,b}^{*} f \qquad (4.14)$$

where $a = 2^{-s}$, $b = k2^{-s}$, $s, k \in \mathbb{Z}$.

If we discretize the function $f(t)$ with sampling rate equal to 1, the above integral can be written as:

$$\widetilde{f}(2^{-s}, k2^{-s}) \equiv 2^{s/2} \sum_{n} f(n) \overline{\psi}(2^{s} n - k) \qquad (4.15)$$

To compute the wavelet transform of a function at some point in the time-scale plane, we do not need to know the function values for the entire time axis. All we need is the function at those values of time at which the wavelet is non-zero [32].

Figure 4.5 shows the differences arisen in a signal analysis when using DWT and CWT respectively. At DWT, time lies on the abscissa, scale α lies on the ordinate and is dyadic: 2^1, 2^2, 2^3, 2^4, 2^5, levels are between 1 and 5 and each coefficient at level k is repeated 2k times. At CWT, time lies on the abscissa, scale α lies on the ordinate and its value changes all the time from 2^1 to 2^5 with step equal to 1 [62].

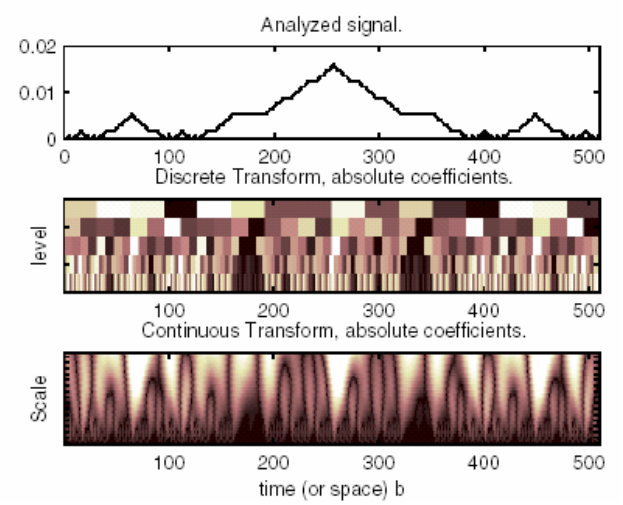


Figure 4.5 – Discrete versus Continuous Wavelet Transform.

For most signals, low frequency content is very important because it gives the signal's identity. High frequency components only give the fine differences between different signals. In wavelet analysis, low frequency – thus high scale – coefficients are called approximations while high frequency – thus low scale – ones are called details. Therefore, the filters that are used are a high-pass, which gives the details, and a low-pass, which gives the approximations, which are complementary to each other. The procedure is illustrated in Figure 4.6.

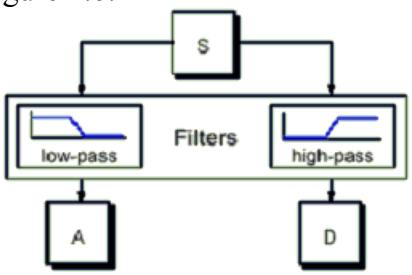


Figure 4.6 – Filtering Procedure.

Unfortunately, if we actually perform this operation on a real digital signal, we wind up with twice as much data as we started with. This problem is confronted with *decimation*. Decimation restores the number of the samples by keeping only one every two samples. Two sequences that give the discrete wavelet transform coefficients are, thus, produced; cA (for approximations) which contains less noise than the initial signal and whose coefficients have large values, and cD (for the details) that contains a great amount of high frequency noise and whose coefficients have small values. This procedure is shown in Figure 4.7 and the device which implements it is called *two-channel subband coder* [56].

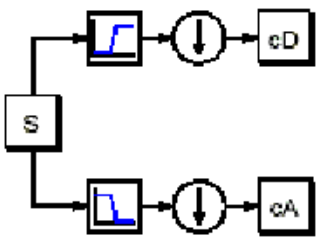


Figure 4.7 – DWT coefficients' production.

The discrete wavelet transform coefficients are computed with the use of multi-resolution analysis (MRA), which is further analyzed in section 4.1.3. At this point we accept that the following relations are true for MRA [32]

$$x_s(t) \in A_s \Leftrightarrow x_s(t) = \sum_k \alpha_{k,s}\phi_{k,s}, \tag{4.16}$$

$$x_{s+1}(t) \in A_{s+1} \Leftrightarrow x_{s+1}(t) = \sum_k \alpha_{k,s+1}\phi_{k,s+1}, \tag{4.17}$$

$$y_s(t) \in W_s \Leftrightarrow y_s(t) = \sum_k w_{k,s}\psi_{k,s}, \tag{4.18}$$

where A_s stands for the approximations and are generated by ϕ and $\mathcal{W}_s$ stands for the details and are generated by ψ.

Therefore, $A_s \oplus W_s = A_{s+1}$, which applies for this analysis, becomes

$$x_{s+1}(t) = x_s(t) + y_s(t) \tag{4.19}$$

and thus

$$\sum_k \alpha_{k,s+1}\phi_{k,s+1} = \sum_k \alpha_{k,s}\phi_{k,s} + \sum_k w_{k,s}\psi_{k,s}, \tag{4.20}$$

Let us define the decomposition relation

$$\phi(2^{s+1}t - \ell) = \sum_k \left\{ h_0[2k - \ell]\phi(2^s t - k) + h_1[2k - l]\psi(2^s t - k) \right\} \tag{4.21}$$

where h_0 and h_1 are the low-pass and high-pass filters that are used in the algorithm, respectively. If we combine it with (4.20) we get that

$$\alpha_{k,s} = \sum_k h_0[2k - \ell]\alpha_{\ell,s+1} \quad \text{and} \tag{4.22}$$

$$w_{k,s} = \sum_k h_1[2k - \ell]\alpha_{\ell,s+1} \tag{4.23}$$

where the right terms of the equations correspond to decimation every two samples after convolution, as former described.

As foresaid, DWT is a sampled version of CWT. The salient feature of the former is that the sampling rate is automatically adjusted to the scale. This means that a given signal is sampled by first dividing its frequency spectrum into bands, and then the signal in each band is sampled at a rate proportional to the ratio of the frequency scale of that band to the total frequency spectrum [43].

For the validity of the proportional to (4.13) relation, in going from scale $\alpha_m = \sigma^m$ to the next larger one $\alpha_{m+1} = \sigma \cdot \alpha_m$, we must increase the time-sampling interval Δt by a factor σ. So, we choose $\Delta t = \sigma^m \tau$ where τ is a positive, non-zero number which is equal to the time-sampling interval at the unit scale $\alpha = 1$. The signal is sampled only at times $t_{m,n} = n\sigma^m\tau$, where n is an integer, within the scale σ^m, which means that the time-sampling rate is automatically adjusted to the scale.

If we take a closer look at (4.15) we shall notice its time-variant nature. The DWT of a function shifted in time is not the same to DWT of the original function. If assumed that $f_m(t) = f(t - t_m)$ it gives

$$\begin{aligned} \tilde{f}(2^{-s}, k2^{-s}) &= 2^{s/2}\int_{-\infty}^{\infty} f_m(t)\overline{\psi}(2^s n - k)dt \\ &\approx 2^{s/2}\sum_n f(n-m)\overline{\psi}(2^s n - k) \\ &= 2^{s/2}\sum_n f(n)\overline{\psi}\left[2^s n - (k - m2^s)\right] \\ &\approx \tilde{f}\left[2^{-s}, (k - m2^s)2^{-s}\right] \end{aligned} \tag{4.24}$$

From which we can conclude that for DWT, a shift in time of a function manifests itself in a rather complicated way [32].

4.1.3 Multi-Resolution Analysis (MRA)

Let us consider a function consisting of slowly varying and rapidly varying segments and we would like to represent it at a single level of approximation, we should descritize it using a step determined by the rapidly varying segment. This will result – by no means – to a huge number of data points [32].

The wavelet transform provides the means of analyzing the input signal into a number of different resolution levels in a hierarchical fashion. This is known as *Multi-Resolution Analysis* [32], [43], [97], [62]. Thus, signal components corresponding to different physical activities can be best represented at different resolution levels: short high-frequency activities at the finer resolution and long low-frequency ones at the coarser resolution levels [97].

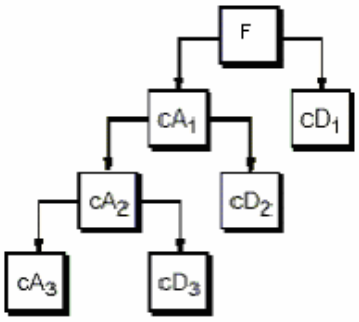


Figure 4.8 – Multi-resolution analysis representation.

At the first phase signal function is analyzed to approximation (cA1) and details (cD1). Then, the approximation is further analyzed to approximation (cA2) and details (cD2) and this procedure is recursively executed for all levels defined by the user or the system. Figure 4.8 illustrates this analysis and Figure 4.9 the application on a signal. As it can be now realized, the words approximation and details are justified by the fact that the approximation of one level (cA_1) arises from the approximation of the previous level (cA_2) taking into account the low frequencies of cA_2, whereas the details (cD_2) corresponds to the high frequency correction.

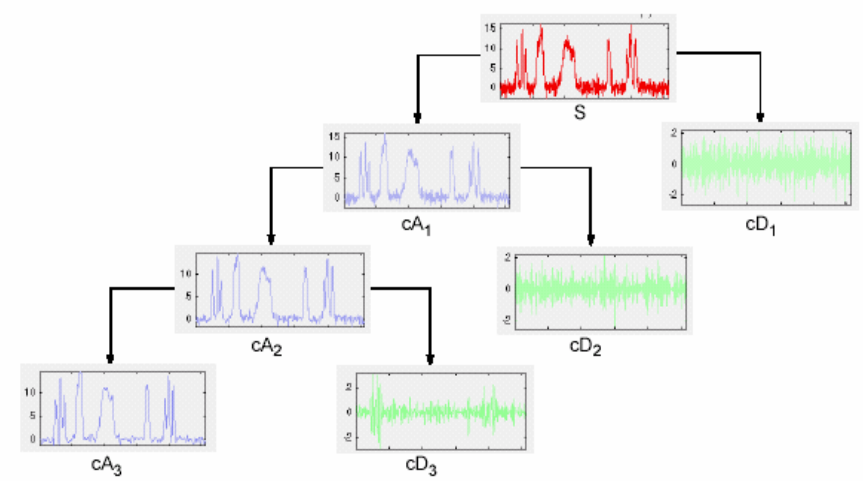


Figure 4.9 – Multi-resolution analysis of a signal

Instead of beginning with a mother wavelet, multi-resolution analysis begins with a basic function $\phi(t)$ called the *scaling function* [43]. This function will generate the wavelet ψ. The translated and dilated versions of ϕ are defined as:

$$\phi_{k,s} = 2^{s/2}\phi(2^s t - k)\,,\ k \in Z \tag{4.25}$$

and are used to sample signals at various times and scales. Unlike wavelet samples, which only provide details of the signal, the samples $\langle \phi_{k,s}, f \rangle \equiv \phi^*_{k,s} f$ are supposed to represent the values of the signal itself, averaged over a neighborhood of width $2^s W$ around $t = 2^s k$.

To achieve a multi-resolution analysis the scaling function has to satisfy certain conditions. *Orthonormality* within the scale $s = 0$ is one of them

$$\left\langle \phi_k,\ \phi_j \right\rangle \equiv \delta_k^j \tag{4.26}$$

It is proved that the operator which determines the time shift is unitary, so it is $\langle \phi_k, \phi_j \rangle = \langle \phi_{k-j}, \phi \rangle$ and the above relation becomes

$$\left\langle \phi_k, \phi \right\rangle \equiv \phi_k^* \cdot \phi = \delta_k^0\,,\ k \in Z \tag{4.27}$$

Furthermore, we get

$$\left\langle \phi_{k,s},\ \phi_{j,s} \right\rangle = \left\langle \phi_k,\ \phi_j \right\rangle. \tag{4.28}$$

Hence equation (4.28) implies orthonormality at every scale. Notice that $\phi_{k,s}$ at different scales need not be orthogonal.

If f is considered to be the constant function $f(t) = 1$ and ϕ is integrable then

$$\phi_k^* f = \int_{-\infty}^{\infty} dt\,\bar{\phi}(t-k) = \int_{-\infty}^{\infty} dt\,\bar{\phi}(t) = \hat{\bar{\phi}}(0) = 1 \tag{4.29}$$

Relation $\hat{\phi}(0) = 1$ is the second condition which we will call the *averaging property*.

Function $\phi(t)$ produces a sequence $\{A_s\}$ which consists of the approximations and is defined as

$$A_{s+1} \subset A_s \quad \text{for each } s \in Z. \tag{4.30}$$

If a signal is sampled at $\Delta t = 2^s$, the detail at scales less than 2^s are expected to be lost. That is why, A_s has to be regarded as containing signal information only down to the time scale $\Delta t = 2^s$ and for this idea to be precise we require (4.30). It is obvious that $\{\phi_{k,s} : k \in Z\}$ constitute an orthonormal basis for A_s.

Moreover, scaling function satisfies the dilation equation or two-scale relation for φ

$$\phi(t) = \sum_k h_0[k]\,\phi(\alpha t - k) \tag{4.31}$$

for some positive α and coefficients $\{h_0[k]\} \in \ell^2$. Function $\phi(t)$ is a translated and scaled version of itself, hence the name scaling function. $h_0[k]$ is known as the *two-scale sequence* or the set of filter coefficients for ϕ.

A_0 is generated by $\{\varphi(-k) : k \in Z\}$ and, in general, A_s by $\{\varphi k,\ s : k, s \in Z\}$. As a result we get:

$$x(t) \in A_s \Leftrightarrow x(2t) \in A_{s+1}, \tag{4.32}$$

$$x(t) \in A_s \Leftrightarrow x(t+2^{-s}) \in A_s, \tag{4.33}$$

These equations and the dilation equation define the functions that perform multi-resolution analysis [32].

Details are generated by $\psi_{k,s}(t) = 2^{s/2}\psi(2^s t - k)$ in the same way as the approximations are generated by $\phi(t)$. Thus:

$$x_s(t) \in A_s \Leftrightarrow x_s(t) = \sum_k \alpha_{k,s}\phi(2^s t - k) \quad \text{and} \tag{4.34}$$

$$y_s(t) \in W_s \Leftrightarrow y_s(t) = \sum_k w_{k,s}\psi(2^s t - k) \tag{4.35}$$

where A_s and W_s are the subspaces which contain the approximation and details respectively.

For A_s and W_s, which is called wavelet subspace [32], we have that:

$$A_s \oplus W_s = A_{s+1}, \tag{4.36}$$

$$A_s \cap W_s = \{0\}, \quad s \in Z. \tag{4.37}$$

Since $A_{s+1} = A_s \oplus W_s$ we have

$$A_{s+1} = W_s \oplus W_{s-1} \oplus W_{s-2} \oplus \ldots \tag{4.38}$$

$$\text{or} \quad A_s = \bigoplus_{l=-\infty}^{s-1} W_l \tag{4.39}$$

For some coefficients $\{\alpha_{k,s}\}_{k \in Z}$, $\{w_{k,s}\}_{k \in Z} \in \ell^2$.

Let us recall equations (4.22) and (4.23) which were proved in section 4.1.2

$$\alpha_{k,s} = \sum_k h_0[2k - l]\,\alpha_{l,s+1} \text{ and } w_{k,s} = \sum_k h_1[2k - l]\,\alpha_{l,s+1} \tag{4.40}$$

These relations give the scaling function at any scale in terms of the scaling function and the wavelet at the next-lower scale.

We notice that the wavelets constructed in this way form an orthonormal basis with as much locality and smoothness as desired. The unexpected existence of such bases is one of the reasons why wavelet analysis has gained such widespread popularity.

4.1.4 Decimation

Wavelet coefficients decimation is a complicated property of DWT which is performed during the decomposition. Decimation by two eliminates every other coefficient of the specific level. Hence, the calculation of the wavelet coefficients is quicker and needs less storage space. However, the important thing is that the original signal can be perfectly reconstructed from the coefficients left. As foresaid, decimation makes DWT a time-variant transform. Variance in translation means that the DWTs of a signal and its time-translated version are different because signal translations generate different wavelet coefficients. If we assume that sequence y is produced by decimation by two of the sequence x then we have:

$$[y] = [x]_{\downarrow 2} \Leftrightarrow \begin{bmatrix} . \\ . \\ y(-2) \\ y(-1) \\ y(0) \\ y(1) \\ y(2) \\ y(3) \\ . \\ . \\ . \end{bmatrix} = \begin{bmatrix} . \\ . \\ x(-4) \\ x(-2) \\ x(0) \\ x(2) \\ x(4) \\ x(6) \\ . \\ . \\ . \end{bmatrix} \Leftrightarrow \tag{4.41}$$

$$\Leftrightarrow \begin{bmatrix} . \\ . \\ y(-2) \\ y(-1) \\ y(0) \\ y(1) \\ y(2) \\ y(3) \\ . \\ . \\ . \end{bmatrix} = \begin{bmatrix} & & & & & & & & & & \\ 1 & 0 & 0 & . & . & & & & & & \\ 0 & 0 & 1 & 0 & 0 & . & . & & & & \\ & . & 0 & 0 & 1 & 0 & 0 & . & . & & \\ & . & . & . & 0 & 0 & 1 & 0 & 0 & . & \\ & & & . & . & . & 0 & 0 & 1 & 0 & 0 \\ & & & & & . & . & . & 0 & 0 & 1 \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \end{bmatrix} \begin{bmatrix} . \\ . \\ x(-2) \\ x(-1) \\ x(0) \\ x(1) \\ x(2) \\ x(3) \\ . \\ . \\ . \end{bmatrix} \Leftrightarrow [y] = [DEC_{\downarrow 2}][x].$$

The shift-variant property is evident if we shift the input column either up or down by a given number of position. It is noticeable that $[DEC_{\downarrow 2}]^{-1} = [DEC_{\downarrow 2}]^{t}$ that is decimation is an orthogonal transformation [32].

In order to ensure the shift invariance many algorithms for wavelet construction have been introduced. These algorithms are widely known as *Undecimated Wavelet Transforms (UWT*).

Moreover, the undecimated wavelet transforms increase the information amount regarding the transformed signal by comparing it with that of DWT. The number of wavelet coefficients does not alter within the levels; it remains the same and equal to the image pixels number at each level. This added information is very useful to better analyze and comprehend the signal attributes. For instance, in image denoising applications the resolution between data and noise can be increased. Large data amount is essential when statistical methods are used for the wavelet coefficients decomposition. However, undeniable disadvantages of UWTs are the big computational and storage cost together with the coefficient redundancy.

The most impulsive approach to the calculation of an undecimated wavelet transform is to omit the decimation step in DWT. The idea of the à trous algorithm, which is going to be further analyzed, is to double the filter coefficient number, which values are obtained by interpolation, and let the decimated sequence generated by DWT pass through this filter.

4.1.4.1 À Trous Algorithm

The difficulty in implementing a discrete wavelet series like the one below

$$\tilde{f}(2^i, 2^i k) \equiv \frac{1}{\sqrt{2^i}} \sum_n \overline{\psi}(n2^{-i} - k) f(n), \quad \text{where } i = -s \tag{4.42}$$

is that even for ψ(t) of finite support, as τo i increases, $\overline{\psi}(t)$ must be sampled at progressively more points, creating a large computational burden. Hence the à trous algorithm is applied as the next logical step [37], [82], [24]. This algorithm alters the filters at each step of the wavelet decomposition. While in DWT the filtered signal is downsampled, in the à trous algorithm the low-pass filter g is upsampled by inserting zeros between its coefficients, and then the discrete wavelet series of (4.42) passes through the filter. Thereafter, the even points' values are approximated by interpolation while the odd points are left fixed via a finite filter h. These kinds of filters are related with non-orthogonal wavelet analysis. The detail coefficients are computed from the difference between two low-passed images in adjacent levels. This procedure is presented in Figure 4.10. The invert transform is produced by the aggregation of the detail coefficients that have arisen at all levels and the last low-resolution image.

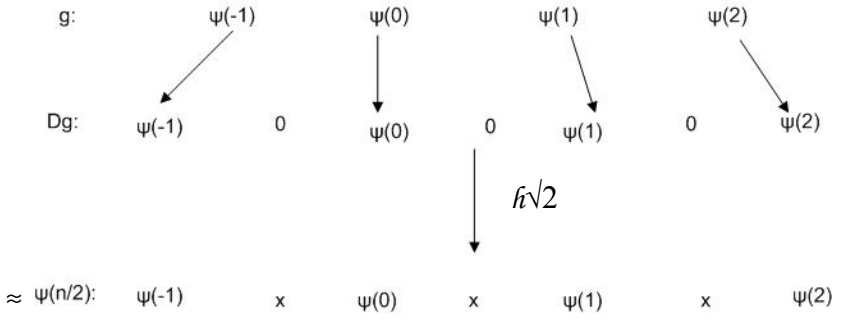


Figure 4.10 - Dilation and interpolation of a function ψ(t).

Definition: The low-pass filter h is said to be an à trous filter if it satisfies

$$h_{2n} = \delta(n) / \sqrt{2} \tag{4.43}$$

The result of the entire interpolation operation is

$$[h * Dg]_k = [H \cdot Dg]_k = \sum_n h_{k-2n} \varphi(n)$$

$$\approx \frac{1}{\sqrt{2}} \psi(k/2) \tag{4.44}$$

Thus, inserting the (4.44) into (4.43) and noting that $\varphi((n/2)-k) = \varphi((n-2k)/2)$ we obtain

$$\begin{aligned} \tilde{f}(2,2k) &= \sum_{n,m} h_{n-2k-2m} \bar{g}_m f_n \\ &= \sum_{n,m'} g_{k-m'} \bar{h}_{2m'-n} f_n \\ &= [g * (\Lambda(\bar{h} * f))]_k \end{aligned} \tag{4.45}$$

where $\Lambda_{k,m} = \delta(2k-m) = \delta_{2k,m}$ is the decimation factor.

The former equation is simply $\tilde{f}_k^i = \sum_j g_{k-j} [f^i]^j$ with $i = 1$. Continuing inductively by replacing f with f^{i-1} we result in $\tilde{f}\ (2^i, 2^i k) \approx \tilde{f}_k^i$ for all i. For real h we get

$$f^{i+1} = \Lambda\left(h * f^i\right) \text{ and} \tag{4.46}$$

$$\tilde{f}^i = g * h \tag{4.47}$$

Equations (4.46) and (4.47) contribute à trous algorithm which is not, in general, tranlation invariant. That is why the concept of undecimation is introduced which promises invariance. Therefore, the undecimated à trous algorithm is now presented.

We define T_m the operation of translation by m, i.e.:

$$(T_m f)_k \equiv f_{k-m} \tag{4.48}$$

and due to the dependency of f^0 on $\tilde{f}^i$ equations (4.46) and (4.47) become

$$\tilde{f}_i(f^0) = G(\Lambda H)^i f^0 \tag{4.49}$$

Moreover,

$$[(\Lambda H)^i]_{nk} = [(\Lambda H)^i]_{0,n-2^i k} \quad \text{and} \tag{4.50}$$

$$\sum_n [(\Lambda H)^i]_{nk} e^{jn\omega} = e^{j 2^i k \omega} \prod_{r=0}^{i-1} f_z(2^r \omega) \tag{4.51}$$

and it can be proved that $\tilde{f}^{i}$ is not translation invariant because $[\tilde{f}^{i}(T_m f^0)]_k \neq \sum_n [G(\Lambda H)^i]_{k-m,n} f_n^0$. If we replace m with 2^im and use (4.50) then we have

$$[f^i(T_{2^i m} f^0)]_k = [\tilde{f}^i(f^0)]_{k-m} \tag{4.52}$$

Thus, translating f^0 by 2^im translates octave i by m.

Note that the zeroth element of a series is invariant under decimation so that $\tilde{f}_k^i$ and $\bar{f}_k^i$ should coincide at $k = 0$. Utilizing this fact, we obtain the k–th output of the undecimated discrete wavelet transform by translating the signal back by k samples and taking the decimated transform at time zero.

Definition: The undecimated discrete wavelet transform $\bar{f}$ in terms of the decimated transform $\tilde{f}$ by

$$\tilde{f}_k^i \equiv [\bar{f}^i(f^0)]_k \equiv [\tilde{f}^i(T_{-k} f^0)]_0 \tag{4.53}$$

It is clear that the desired invariance is achieved and also that sampling $\bar{f}_k^i$ every 2^i points produces exactly $\tilde{f}_k^i$.

By taking z transforms we can prove that $\bar{f}$ may be computed by the filter sequence pictured in Figure 4.11, accepting also that $D^i h$ is filter h with $2^i - 1$ zeros inserted between every pair of filter coefficients. That is,

$$f^{i+1} = (D^i h) * f^i \text{ and} \tag{4.54}$$

$$\bar{f}^i = (D^i g) * f^i \tag{4.55}$$

which is the original (undecimated) à trous algorithm.

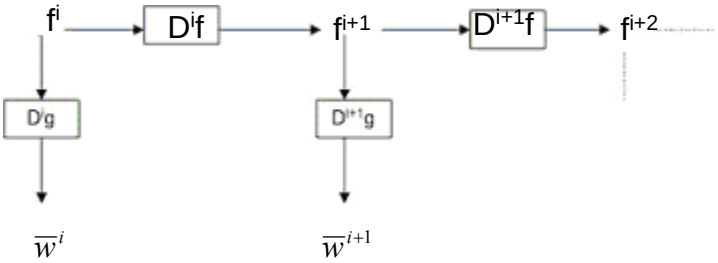


Figure 4.11 - Undecimated Discrete Wavelet Transform.

4.1.5 Signal Synthesis or Reconstruction

Up to now we analyzed how the signal is decomposed with the use of the wavelet transform either in the continuous or the discrete form. It is time we examined the other side of the coin; the *signal synthesis* or *reconstruction* from the coefficients that occurred during the transformation with no information loss.

For CWT, signal reconstruction is given by [62]:

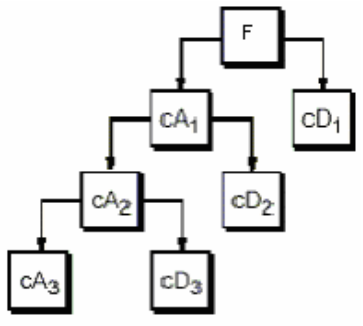
$$f(t) = \frac{1}{K_\psi} \int_{R^+} \int_R \tilde{f}(a,b) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{b}\right) \frac{da \quad db}{a^2} \tag{4.56}$$

where K_ψ is a constant depending on ψ.

For DWT, signal reconstruction is given by [62]:

$$f(t) = \sum_{s \in Z} \sum_{k \in Z} \tilde{f}(2^{-s}, k2^{-s}) \psi_{s,k}(t) \tag{4.57}$$

For multi-resolution analysis, the signal is reconstructed from the sum of the coarsest component and all the details components that have arose resulting in this way in a successively finer approximation [97]. From Figure 4.8 – which is displayed again for convenience – we get:



F = cA1 + cD1
= cA2 + cD2 + cD1
= cA3 + cD3 + cD2 + cD1.

While decomposition uses decimation after filtering, at synthesis we have upsampling before the coefficients pass the filters. Upsampling is a process reciprocal to decimation which inserts zeros every M samples and increases the signal length. Then, by interpolation we obtain the values of the added samples [32]. Interpolation function is:

$$x'(n) = \begin{cases} y\left(\frac{n}{M}\right) & for \quad n = kM, k \in Z \\ 0 & otherwise \end{cases} \tag{4.58}$$

or

$$x'(n) = \sum_k y(k)\delta(n - kM), \quad \text{k} \in \text{Z} \tag{4.59}$$

Yet we can write interpolation by 2 in a matrix form:

$$[\,\text{x}\,] = [\,\text{y}\,]_{\uparrow 2} \Leftrightarrow \begin{bmatrix} . \\ . \\ x'(-2) \\ x'(-1) \\ x'(0) \\ x'(1) \\ x'(2) \\ x'(3) \\ . \\ . \\ . \end{bmatrix} = \begin{bmatrix} . \\ . \\ y(-2) \\ 0 \\ y(0) \\ 0 \\ y(2) \\ 0 \\ . \\ . \\ . \end{bmatrix} \Leftrightarrow \tag{4.60}$$

$$\Leftrightarrow \begin{bmatrix} . \\ . \\ x'(-2) \\ x'(-1) \\ x'(0) \\ x'(1) \\ x'(2) \\ x'(3) \\ . \\ . \\ . \end{bmatrix} = \begin{bmatrix} . & . & 1 & 0 & & & & & \\ . & . & 0 & 0 & & & & & \\ . & . & 0 & 1 & 0 & & & & \\ . & . & 0 & 0 & 0 & 0 & . & & \\ . & . & 0 & 0 & 1 & 0 & & & \\ . & . & 0 & 0 & 0 & 0 & 0 & & \\ . & . & 0 & 0 & 0 & 1 & 0 & & \\ . & . & 0 & 0 & 0 & 0 & 0 & & \\ & & & & & & 1 & 0 & \\ & & & & & & 0 & 0 & \\ & & & & & & 0 & 1 & \end{bmatrix} \begin{bmatrix} . \\ . \\ y(-2) \\ y(-1) \\ y(0) \\ y(1) \\ y(2) \\ y(3) \\ . \\ . \\ . \end{bmatrix} \Leftrightarrow [x] = [INT_{\uparrow 2}][y].$$

By carefully choosing filters for the decomposition and reconstruction phases that are closely related (but not identical) we can cancel out the effects of aliasing, a distortion which is introduced by decimation performed during the decomposition phase. The low- and high- pass decomposition filters (H, L) together with their associated reconstruction filters (H ', L ') form a system of what is called *Quadrature Mirror Filters* (QMF) [62]. Figure 4.12 illustrates the reconstruction process if there is one decomposition level.

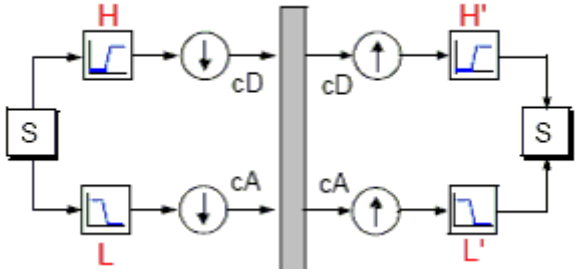


Figure 4.12 – Decomposition and Reconstruction processes.

If we want to reconstruct a signal from the sequences that contain the approximation (cA) and details (cD) coefficients we should follow the next steps: upsample the approximation sequence, then low-pass filter it and finally combine it with a vector of zeros with the same length as cA which has been similarly processed except the fact that it has been high-pass filtered. The final sequence has double length, which is the initial signal length. The same procedure is applied on the details only the filters are reciprocal – the details are high-pass filtered and the zero vector low-pass filtered. In order to retrieve the initial signal we add the two generated sequences. Notice that the sequences' summation has to be done always after their reconstruction so as to have obtained their initial length (Figure 4.13, Figure 4.14).

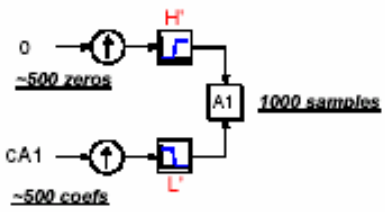


Figure 4.13 – Approximation Reconstruction

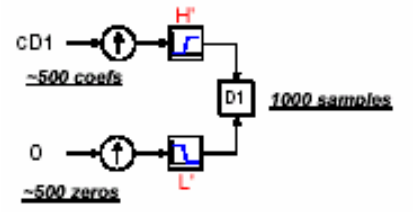


Figure 4.14 – Details Reconstruction

Signal reconstruction can be described by the relation [32]:

$$x_s(t) + y_s(t) = \sum_k \alpha_{k,s}\phi_{k,s}(t) + \sum_k w_{k,s}\psi_{k,s}(t) = x_{s+1}(t) \tag{4.61}$$

and substituting the two-scale relation for ϕ and ψ in (4.61) we get:

$$\sum_k \alpha_{k,s}\sum_l h_0[l]\phi(2^{s+1}t\text{-}2k\text{-}l) + \sum_k w_{k,s}\sum_l h_1[l]\phi(2^{s+1}t\text{-}2k\text{-}l) = \sum_l \alpha_{l,s+1}\phi(2^{s+1}t\text{-}l) \tag{4.62}$$

Comparing the coefficients of $\phi(2^{s+1}t - \ell)$ with both parts of the former equation we result in

$$\alpha_{l,s+1} = \sum_k \left\{ h_0[l-2k]\alpha_{k,s} + h_1[l-2k]w_{k,s} \right\} \tag{4.63}$$

where the right part of (4.63) represents interpolation followed by convolution as illustrated in Figure 4.15.

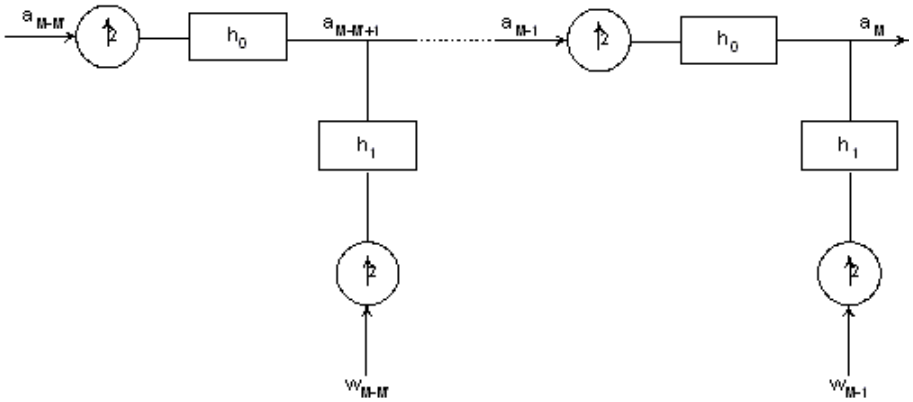


Figure 4.15 – Signal reconstruction from the approximation and details.

5 DENOISING STEP

Microarray images consist mostly of low-intensity features, which are not well distinguishable from the background. This is a result of the great differences on transcript abundance in all eukaryotic cells, which is a common knowledge for over 30 years now [9]. These differences cover over six orders of magnitude, even in relative simple eukaryotes, such as yeast [36].

The most interesting genes, including cell cycle and transcriptional regulators have an extremely low average expression level of one copy per cell [46]. Microarray images can accurately represent around three orders of magnitude. To increase the dynamic range of the measurements and better determine the most interesting genes, researchers produce multiple images of the same microarray at increasing detection settings [22,87] and transform the intensity values of the individual genes into one "true" measurement. By increasing the detection settings, the source noise (mostly additive), which includes photon noise and dust on the slides, remains unaffected. On the other hand, the detector noise (mostly multiplicative), which includes features of the amplification and digitization process, is increased [108].

After having decomposed the signal, the coefficients arose will be further processed in order to get an image with better resolution and more distinctive attributes. We, therefore have to denoise the image in order to discard any useless information. A number of well-known image processing techniques, including soft and hard thresholding, Bayesian denoising based on Gaussian or Laplacian signal modeling, and multiresolution methods that exploit the correlation between the wavelet coefficients of adjacent scales have been applied to microarray images. The proposed method consists of a Bayesian denoising based on Laplacian signal modelling stage and a correlation based stage.

5.1 Denoising via Thresholding

The simplest method of denoising is thresholding. With this method all wavelet coefficients of the detail subbands whose amplitude is below a given value – the threshold – are set to zero while the approximation coefficients are left unaltered.

It is clear that all coefficients subjected to thresholding that are smaller than the threshold is replaced by zero. Depending on how the coefficients are processed when their absolute value is larger than the threshold one can define different thresholding policies. The two most common thresholding polices are *hard* and *soft*. Hard thresholding can be described as the usual process of setting to zero the coefficients whose absolute values are lower than the threshold and left the rest of the coefficients unaltered. Soft thresholding is an extension of hard thresholding, first setting to zero the elements whose absolute values are lower than the threshold, and then shrinking the nonzero coefficients towards 0. By soft thresholding we do not confront the problem of discontinuities among the coefficients that are near the threshold values.

Some thresholding techniques try to compromise between hard and soft thresholding. The *hyperbolic* thresholding is an almost hard thresholder with the continuity property. The *semisoft (firm)* thresholding depends on two thresholds, $0 \le \tau_1 \le \tau_2$ being, therefore, a generalization of soft and hard thresholding; when τ_2 approaches infinity the semisoft rule transforms into soft thresholding with threshold τ_1, and when $\tau_2 \rightarrow \tau_1$ it transforms into hard thresholding. The *non-negative garrotte* shrinkage function is a continuous function which approaches the identity line as the absolute of the signal coefficient gets large, which provides a smaller bias than the soft shrinkage for large coefficient.

The following equations are the relations that define all these types of thresholding (Figure 5.1):

$$T^{hard}(x,\tau) = x \cdot \mathbf{1}(|x| > \tau) \tag{5.1}$$

$$T^{soft}(x,\tau) = \left(x - sgn(x)\tau\right) \cdot \mathbf{1}(|x| > \tau)^{6} \tag{5.2}$$

$$T^{hyper}(x,\tau) = sgn(x)\sqrt{x^2 - \tau^2} \cdot \mathbf{1}(|x| > \tau) \tag{5.3}$$

$$T^{semisoft}(x,\tau) = \begin{cases} 0 & |x| \le \tau_1 \\ sgn(x)\dfrac{\tau_2(|x| - \tau_1)}{\tau_2 - \tau_1} & \tau_1 < |x| \le \tau_2 \\ x & |x| > \tau_2 \end{cases} \tag{5.4}$$

$$T^{garrotte}(x,\tau) = \left(x - \frac{t^2}{x}\right) \cdot \mathbf{1}(|x| > \tau) \tag{5.5}$$

where x is the signal coefficient, τ_i are the thresholds value and $\mathbf{1}(g)$ is defined as

$$\mathbf{1}(g) = \begin{cases} 1, & if \;\; g \;\; is \;\; true \\ 0, & otherwise \end{cases} \tag{5.6}$$

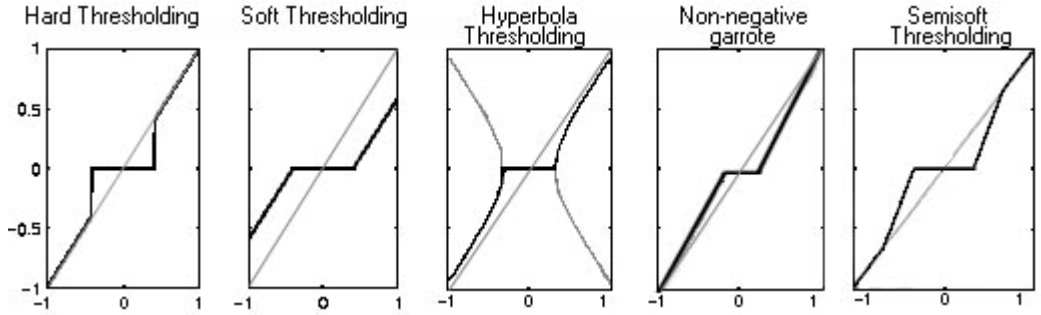


Figure 5.1 – Thresholding rules.

It is obvious that the selection of the threshold value is a critical and important step. If we choose a very small value then most detail coefficients are going to be left unaltered while by picking up a high one, most detail coefficients will be set to zero. In the first case, the image will not be thresholded therefore no noise is going to be discarded. On the other hand, when the threshold has a high value the final image is going to be an almost black canvas. Thus we should be very careful when selecting the threshold.

[6] sgn(x) is called signum function and is defined as $sgn(x) = \begin{cases} \dfrac{x}{|x|} & , x \neq 0 \\ 0 & , otherwise \end{cases}$

In this work, the threshold is selected using the principle of Stein's Unbiased Risk Estimate (SURE) [18]. In this method you get an estimate of the risk for a particular threshold value t. Minimizing the risks in t gives a selection of the threshold value. Let $\xi_i \overset{iid}{\sim} N(\mu_i,1)$, i = 1,…,k be the gaussian white noise and $\hat{\mu}$ be an estimator of $\underset{\sim}{\mu}=(\mu_1, \ldots, \mu_k)$. If the function $g = \{g_i\}_{i=1}^k$ in representation $\hat{\underset{\sim}{\mu}}(\underset{\sim}{\xi}) = \underset{\sim}{\xi} + g(\underset{\sim}{\xi})$ is weakly differentiable, then

$$E^{\mu} \| \hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu} \|^2 = k + \mathrm{E}^{\mu} \| g(\underset{\sim}{\xi}) \|^2 + 2\nabla g(\underset{\sim}{\xi}) \tag{5.7}$$

where $\nabla g = \frac{\partial}{\partial \xi_i} g_i$. It is interesting that estimator $\hat{\underset{\sim}{\mu}}$ can be nearly arbitrary; for instance, biased and non-linear. The application of (5.7) to $T^{soft}(\underset{\sim}{\xi},\tau)$ gives

$$\mathrm{SURE}(\underset{\sim}{\xi},\tau) = \mathrm{k} - 2 \sum_{i=1}^{k} \mathbf{1}(|\xi_i| \leq \tau) + \sum_{i=1}^{k} (|\xi_i| \wedge \tau)^2 \tag{5.8}$$

The SURE is an unbiased estimator of risk, i.e.,

$$E \| T^{soft}(\underset{\sim}{\xi},\tau) - \underset{\sim}{\mu} \|^2 = E\ \mathrm{SURE}(\underset{\sim}{\xi},\tau) \tag{5.9}$$

The LLN argument motivates the following threshold selection:

$$\tau^{\mathrm{SURE}} = \arg \min_{0 \leq \mathrm{T} \leq \mathrm{T}^U} \mathrm{SURE}(\underset{\sim}{\xi},\tau) \tag{5.10}$$

where $\tau^{\mathrm{U}} = \sqrt{2 \log J}\sigma$, J is the highest level of the decomposition and σ is the noise standard deviation. τ^{U} is called *universal threshold.*

It is possible to derive a SURE-type threshold for T^{hard} and T^{hyper} but the simplicity of the representation (5.8) is lost.

Because ξ is supposed to be a Gaussian white noise, we expect that the thresholding method kills roughly all the coefficients and returns the result *f(x)* = 0. For Stein's Unbiased Risk Estimate threshold, roughly 3% of coefficients are saved. So SURE threshold selection rule is more conservative and would be more convenient when small details of function *f* lie near the noise range, which is the case here.

5.2 Coring Suppresion

Thresholding has the following drawbacks: 1) it depends on the correct election of the type of thresholding, 2) the choice of the threshold, arguably the most important design parameter, is done in an ad hoc manner, 3) the threshold cannot be finely adjust after its calculation, 4) it should be applied at each level of decomposition, needed several levels, and 5) the specific distributions of the signal and noise may not be well matched at different scales. Therefore, methods without these constraints will represent an upgrade.

This denoising method makes use of some models (Gaussian, Laplacian) for the subband statistics of the signal and develops a noise-removal algorithm, which performs a "coring" operation to the data. The "coring" non linear noise suppression preserves high-amplitude observations while suppressing low-amplitude values from the high-pass bands of a signal decomposition [86, 85].

In a Bayesian framework, referring to (4.5), $d^i_{j,k}$, $s^i_{j,k}$, and $\xi^i_{j,k}$ are considered as samples of the random variables d, s, and ξ, respectively. The signal component is modeled according to a specific distribution (Gaussian, Laplacian), while the noise component is modeled as a zero-mean Gaussian random variable. Our goal is to find the Bayes risk estimator that minimizes the conditional risk, which is the loss averaged over the conditional distribution of s, given the set of wavelet coefficients,

$$\hat{s}(d) = \arg\min_{s} \int L[s, \hat{s}(d)] P_{s|d}(s \mid d) ds \tag{5.11}$$

The Bayes risk estimator under a quadratic cost function minimizes the mean-square error (MSE) and is given by the conditional mean of s, given d

$$\hat{s}(d) = \int s P_{s|d}(s \mid d) ds \tag{5.12}$$

Bayes' theorem gives the a posteriori probability density function of *s* based on the measured set of wavelet coefficients

$$P_{s|d}(s \mid d) = \frac{P_{d|s}(d \mid s) P_s(s)}{\int P_{d|s}(d \mid s) P_s(s) ds} \tag{5.13}$$

where $P_s(s)$ is the prior PDF of the signal component of the wavelet coefficients of the microarray image and $P_{d|s}(d|s)$ is the likelihood function. Substituting (5.13) into(5.12), we get:

$$\hat{s}(d) = \frac{\int P_{\xi}(\xi) P_s(s) s ds}{\int P_{\xi}(\xi) P_s(s) ds} \tag{5.14}$$

and because ξ = d-s we have:

$$\hat{s}(d) = \frac{\int P_{\xi}(d-s) P_s(s) s ds}{\int P_{\xi}(d-s) P_s(s) ds} \tag{5.15}$$

Let $A = \frac{P_{\xi}(d-s) P_s(s) ds}{\int P_{\xi}(d-s) P_s(s) ds}$ and from (5.15) we have

$$\hat{s}(d) = \int A \cdot s ds \tag{5.16}$$

Then, for the case of Gaussian signal (with standard deviation σ_s and zero mean) in Gaussian noise (with standard deviation σ and zero mean):

$$A = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(d-s)^2} \cdot \frac{1}{\sqrt{2\pi\sigma_s^2}} \cdot e^{-\frac{1}{2\sigma_s^2}s^2}}{\frac{1}{2\pi\sqrt{\sigma^2\sigma_s^2}} \cdot \int e^{-\frac{1}{2\sigma^2}(d-s)^2 - \frac{1}{2\sigma_s^2}s^2} ds} \tag{5.17}$$

We know that

$$\int e^{-\frac{1}{2}s^2} ds = \sqrt{2\pi} \text{ and } k_1 \int e^{-\frac{1}{2k_2}(s-k_3)^2} ds = k_1\sqrt{2\pi k_2} \tag{5.18}$$

thus, $A = \frac{\frac{1}{2\pi\sqrt{}}}{\frac{1}{\sqrt{2\pi(\sigma^2+\sigma_s{}^2)}}} \cdot \frac{e^{-\frac{1}{2}\left(\frac{(d-s)^2}{\sigma^2}+\frac{s^2}{\sigma_s^2}\right)}}{e^{-\frac{1}{2}\cdot\frac{d^2}{\sigma^2+\sigma_s{}^2}}} = \frac{1}{\sqrt{2\pi\frac{\sigma^2\sigma_s{}^2}{(\sigma^2+\sigma_s{}^2)}}} \cdot e^{-\frac{1}{2}\left[\frac{(d-s)^2}{\sigma^2}+\frac{s^2}{\sigma_s^2}-\frac{d^2}{\sigma^2+\sigma_s{}^2}\right]} \Leftrightarrow$

$$A = \frac{1}{\sqrt{2\pi\frac{\sigma^2\sigma_s{}^2}{(\sigma^2+\sigma_s{}^2)}}} \cdot e^{-\frac{1}{2}\left(s-\frac{d\sigma_s{}^2}{\sigma^2+\sigma_s{}^2}\right)\left(\frac{\sigma_s{}^2\sigma^2}{\sigma^2+\sigma_s{}^2}\right)^{-1}} \tag{5.19}$$

Now let $k = \frac{\sigma_s{}^2\sigma^2}{\sigma^2+\sigma_s{}^2}$. Substituting k,(5.19) into (5.16) we get:

$$\hat{s}(d) = \frac{1}{\sqrt{2\pi k}} \int e^{-\frac{1}{2k}\left(s-\frac{d\sigma_s{}^2}{\sigma^2+\sigma_s{}^2}\right)} s\,ds \tag{5.20}$$

Moreover, for Gaussian signals we have that

$$\frac{1}{\sqrt{2\pi\sigma_X^2}} \int e^{-\frac{1}{2\sigma_X^2}\left(s-\mu_X\right)^2} x\,dx = \mu_X \tag{5.21}$$

Therefore, for the case of Gaussian signal in Gaussian noise a well-known closed-form solution exists

$$\boxed{\hat{s}(d) = \frac{\sigma_s{}^2}{\sigma^2+\sigma_s{}^2} \cdot d} \tag{5.22}$$

It is a simple linear operation as illustrated in Figure 5.2.

As proposed in [19], a robust estimate of the noise standard deviation, σ, is obtained in the finest decomposition scale by the measured wavelet coefficients as

$$\sigma = \frac{1}{0.6745} MAD(\{d_{j,k}, 0 \le k \le 2^J\}) \tag{5.23}$$

where the operator MAD [75] signifies the median absolute deviation and J denotes the highest level of wavelet decomposition. A less robust estimation of σ is

$$\sigma = \frac{1}{2^J - 1} \sum_{k=0}^{J} \left(d_{J,k} - \bar{d}\right)^2 \tag{5.24}$$

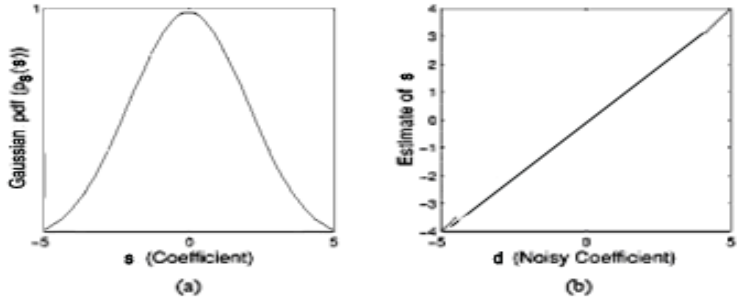


Figure 5.2 - (a) Gaussian pdf. (b) Corresponding shrinkage function.

Let now consider a signal that follows the Laplacian distribution (with standard deviation σ_s) in zero-mean Gaussian noise (with standard deviation σ):

$$P_s(s) = \frac{1}{\sqrt{2}\sigma_s} \mathrm{e}^{-\frac{\sqrt{2}|s|}{\sigma_s}} \tag{5.25}$$

To estimate s from the noisy observation d for this case we will use the maximum a-posteriori (MAP) estimator which is

$$\hat{s}(d) = \arg\max_s p_{s|d}(s \mid d) \tag{5.26}$$

Using Bayes rule[7] we get:

$$\hat{s}(d) = \arg\max_s \left[p_\xi(d-s) \cdot p_s(s) \right] \tag{5.27}$$

which is equivalent to

$$\hat{s}(d) = \arg\max_s \left[\log\left(p_\xi(d-s) \right) + \log\left(p_s(s) \right) \right] \tag{5.28}$$

Let $\log(p_s(s)) = f(s)$ which due to (5.25) is

$$f(s) = \log\left(\frac{1}{\sqrt{2}\sigma_s} e^{-\frac{\sqrt{2}|s|}{\sigma_s}} \right) = -\log(\sigma_s\sqrt{2}) - \sqrt{2}\frac{|s|}{\sigma_s} \tag{5.29}$$

and $\log(p_\xi(d-s)) = g(d-s)$ which in the same way becomes:

$$g(d-s) = -\log(\sigma\sqrt{2\pi}) - \frac{(d-s)^2}{2\sigma^2} \tag{5.30}$$

Thus, $\hat{s}(d) = \arg\max_s \left[\frac{(d-s)^2}{2\sigma^2} - \sqrt{2}\frac{|s|}{\sigma_s} \right]$ which is equivalent to solving the following equation for $\hat{s}$ is $p_s(s)$ is assumed to be strictly convex and differentiable:

$$\left(\frac{(d-s)^2}{2\sigma^2} \right)' - \left(\sqrt{2}\frac{|s|}{\sigma_s} \right)' = 0 \Leftrightarrow \frac{d-s}{\sigma^2} - \frac{\sqrt{2}}{\sigma_s} = 0 \tag{5.31}$$

Then the estimator will be

$$\boxed{\hat{s}(d) = sgn(d)\left(|d| - \frac{\sqrt{2}\sigma^2}{\sigma_s} \right)_+} \tag{5.32}$$

where $(g)_+ = \begin{cases} 0, & if \quad g < 0 \\ g, & otherwise \end{cases}$

[7] Bayes rule: $p_{s|d}(s \mid d) = p_{d|s}(d \mid s)p_s(s) = p_\xi(d-s)p_s(s)$

Equation (5.32) is the classical soft shrinkage function with $\frac{\sqrt{2}\sigma^2}{\sigma_s}$ as the threshold. The Laplacian pdf and corresponding shrinkage function are illustrated in Figure 5.3. Note that this thresholding method depends on the noise variance which is found by (5.23). Recall that the soft operator is defined as

$$\text{soft}(\text{d},\tau)= \text{sign}(\text{d})\cdot(|\text{d}|-\tau)_+ \tag{5.33}$$

The soft shrinkage function (5.32) can be written as

$$\hat{s}(d) = soft\left(d, \frac{\sqrt{2}\sigma^2}{\sigma_s}\right) \tag{5.34}$$

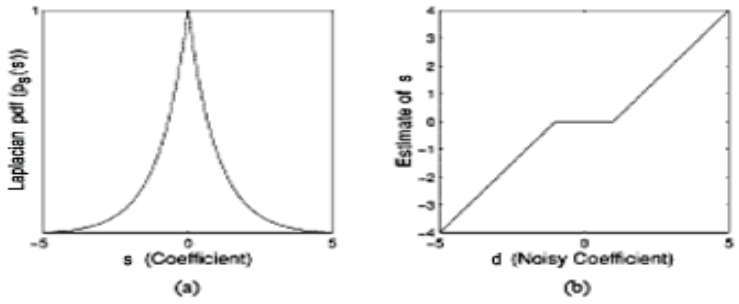


Figure 5.3 - (a) Laplacian pdf (b) Corresponding shrinkage function

5.3 Denoising based on Coefficient Correlation

The presented à trous wavelet transform gives a multiresolution representation of images consisting of approximation images which display the image with increasingly coarser resolution as the scale increases, and of detail planes which show the objects whose size is adapted to the resolution of the filter at each scale. There is an inherent adaptiveness of the analysis to the object size since, with the support of the convolution filter increasing with the scale of analysis (see Table 5.1), the filter smoothes out the response of too narrow objects at a given scale. At the first level of analysis, the support of the filter is such that the detail image has significant coefficients at those locations where pixel-sized significant features are present. When going down in resolution, the filter support increases in size and significant coefficients in the detail images correspond more and more to significant features of increasing spatial dimension. However, it is very difficult to pick up the interesting features from the analysis of one detail image only. This is because relevant coefficients are embedded into non-specific background detail coefficients.

Scale	1	2	3	4	5
Support	5	9	17	33	65

Table 5.1 - Length of wavelet filter support

To overcome the limitation of data coming from a single image and to distinguish important wavelet coefficients from non-relevant ones, we take advantage of the multiresolution representation provided by the à trous wavelet transform. As aforementioned, spots are features in the image that are small compared to the global image, but indeed relatively large when analysed locally. We assume that spots are features of interest represented by a small number of coefficients which are large and

correlated across levels. Following results first demonstrated in the case of additive Gaussian white noise [56], it has been shown that in the case of images contaminated by additive correlated Gaussian noise, local maxima in wavelet planes tend to propagate across scales when they are due to significant discontinuities in the image, while they do not if caused by noise [41,42]. We therefore design a multiscale spatial filtering scheme that results in wavelet coefficients that have high values in the presence of a spot and characterise it unambiguously, whereas they have non-significant values for the background or for large structures. To that goal, we compute a correlation image $P_J(x,y)$ which is defined at each location (x,y) by the direct spatial multiscale product of the wavelet coefficient images at adjacent scales in the à trous representation:

$$P_J(x, y) = \prod_{i=1}^{J} W_i(x, y), \tag{5.35}$$

where J is the deepest level at which the correlation is computed.

We subsequently use the fact that the product of significant coefficients across scales at the location (x,y) results in a significant value of $P_J(x,y)$ only if the local maxima propagate down to the considered scale. Obviously, if the local maxima die at some intermediate scale, this one small coefficient in the product will be sufficient to decrease the value of $P_J(x,y)$ significantly. The key point here is that the wavelet coefficients at large scales are significant only in the vicinity of an important feature while they are close to zero elsewhere. On the other hand, for a given feature, the support of its interval of relevance decreases at small scales. The spatial filtering method can therefore be interpreted as a process by which wavelet coefficient images at large scales are used to give a coarse estimation of possible spots positions. This estimation is then refined by supplementing data coming from finer scales only at those spatially filtered locations [65].

To increase further the efficiency of the method, we have found that before computing the multiscale correlation image, it is desirable to select the most significant wavelet coefficients and to reduce the influence of non-significant noisy coefficients by applying a threshold-based denoising to the wavelet coefficients. Given an input image of the form

$$Y = f + n, \tag{5.36}$$

where Y is the observation, f the noise-free data and n an additive Gaussian noise, we want to compute, from the wavelet transformation of Y, $W^Y = W^f + W^n$, an estimate $\hat{W}^Y$ where coefficients due to noise are replaced by zero. Assuming that noise is stationary and that the correlation between two noise realisations depends on their relative distance only, we have the following result [41,42] that, for a given resolution level i, the variance of the wavelet coefficients of a correlated noise W_i^n depends only on that resolution level i:

$$E(W_i^n)^2 = \sigma_i^2 \tag{5.37}$$

From this, we can define a thresholding strategy that makes use of the kσ hard thresholding technique to define a scale-dependent threshold t_i [63,92]. The wavelet coefficients W_i^Y are therefore transformed according to the following rule:

$$t_{hard}(W_i^Y, t_i) = \begin{cases} W_i^Y & W_i^Y \geq t_i \\ 0 & W_i^Y < t_i \end{cases} \tag{5.38}$$

with $t_i = k\sigma_i$, where σ_i is the standard deviation of the noisy wavelet coefficients at scale i and a usual choice is k =3 [63]. A robust estimation of σ_i is obtained from the MAD estimate [75], and is given by

$$\sigma_i = \bar{\sigma}/0.67 \tag{5.39}$$

where $\bar{\sigma}$ is the median absolute deviation of the wavelet coefficients at scale i.

5.4 Two-stage Multiresolution Technique

All former techniques assume the presence of either additive or multiplicative noise. As foresaid, the omission of the measurement-specific additive noise term leads to exaggerated ratio estimates, false identification of significant differences, and understated uncertainty measures when the observations are small. The omission of the multiplicative noise term leads to similar problems when the observations are large.

The proposed image denoising method accounts for both noise components via its two constituent stages: one that processes the additive component of the noise and another that processes the multiplicative component. The additive component is processed by the denoising method based on coefficient correlation, henceforth referred as *correlation stage*, which was described in Section 5.3 and the multiplicative component is attacked by the coring suppression method under the assumption of a Laplacian signal, henceforth referred as the *coring stage*, which was presented in Section 5.2.

Subband decompositions of an image have significantly high-order statistics that are eluded by the simple thresholding methods. Nonetheless, a Bayesian denoising method, like the coring suppression, exploits these higher-order statistics rendering it a more reasonable choice for image processing. Moreover, the correlation between the wavelet coefficients of adjacent scales infers that there is a significant feature at the position that should be passed through the filter. Therefore, a method which exploits this dependence would be an essential denoising method. A combination of these two methods, as the proposed approach, is obvious that provides a powerful image analysis tool.

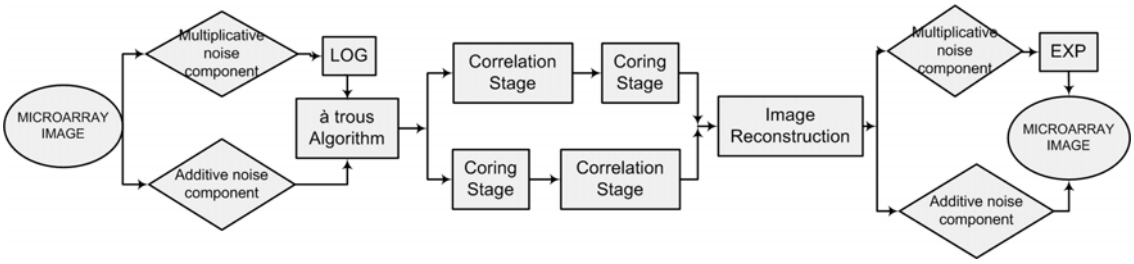


Figure 5.4 – Block diagram of the proposed method

6 RESULTS

6.1 Materials

The microarray images, which are processed by the proposed method, are 16-bit grayscale images. They come from the high detection settings of a homotypic hybridization of a leukemic cell line (KARPAS-231, human B cell leukemia) on microarrays containing oligos corresponding to both human genes (Operon Human Oligo Library v3) and external control genes from Bacillus subtilis, E. coli and phage P1.

4 μgr of total RNA were amplified, as described in [25] with the following modifications: first strand synthesis was perfomed using an anchored T7-oligo (dT) primer and Superscript III (Invitrogen) for 20min at 44^0C and 1h 45min at 50^0C. The amplified RNA (aRNA) in vitro transcription was done with Ampliscribe T7 transcription kit (Epicentre) at 42^0C following the manufacturer's protocol, using a ratio of UTP/aminoallyl-UTP (Epicentre) of 1. The aRNA was cleaned up using RNeasy columns (Qiagen) and quantified using Nanodrop. Equal quantities of aRNA, supplemented with aRNA of the externalcontrol genes at 1:1, 1:3 and 3:1 ratios, were labelled with Alexa 555 and 647 Succinimidyl ester (Molecular Probes, Invitrogen) according to the manufacturer's protocol at 50^0C.

The hybridization was performed in a Tecan HS4800 hybridization station. The slides were prehybridized for 1.5 hour at 42^0C with 5x SSC, 0.1% SDS, 1% BSA and then hybridized for 16h at 42^0C in a buffer with a final concentration of 5x SSC, 0.1 %SDS, 50% formamide and 1.5ìgr/ml fragmented salmon sperm DNA. The arrays were subsequently washed with 2x SSC, 0.1% SDS at 42^0C, followed by a wash in 0.1x SSC, 0.1% SDS at 23^0C and a third wash with 0.1x SSC at 23^0C. The slides were dried by nitrogen.

Arrays were read with a ScanArray 5000 scanner (GSI Lumonics) at 5 μm resolution at three different photomultiplier tube voltage settings (high, medium and low). The fluorescence intensity for each fluor and each element on the array was captured using spotSegmentation package [51] written in R [73] and ImaGene of Biodiscovery [106].

6.2 Evaluation Metrics

In order to evaluate the processed images and compare the results of the methods in a very objective manner we have used some quantitative performance metrics together with the qualitative visual evaluation.

6.2.1 Coefficient of Variation (CV)

The coefficient of variation (CV), or relative standard deviation, provides a quantitative measure of the homogeneity of both the background and the microarray

spot areas. The lower this metric is, the higher the homogeneity in the spot or background, respectively. It is defined as the standard deviation (std) to mean value:

$$CV_{spot} = \frac{std_{spot}}{mean_{spot}} \tag{6.1}$$

$$CV_{background} = \frac{std_{background}}{mean_{background}} \tag{6.2}$$

The main appeal of the CV is that the stds of microarray images generally increase or decrease proportionally as the mean increases or decreases, so that division by the mean removes it as a factor in the variability. The CV is therefore a standardization of the std that allows comparison of the variability inside the spots or local background regardless of the magnitude of the spots or local background respectively. As microarray images contain spots with different intensities and the background has not a constant intensity value, the standard deviation of the image cannot be a metric of homogeneity due to its intensity dependant nature. Therefore, the CV is a metric which demonstrates the spots and background homogeneity.

6.2.2 Confidence Interval (CI)

A confidence interval is an interval in which a measurement or trial falls corresponding to a given probability.

Definition: Given a random sample $X_1,X_2,...,X_n$ from a normal distribution $N(\mu,\sigma^2)$, consider the closeness of X, the unbiased estimator of μ, to the unknown μ. To do this, the error distribution of X, namely that X is $N(\mu,\sigma^2/n)$, is used in order to construct what is called a confidence interval for the unknown parameter μ, when the variance σ^2 is known.

$$CI = \mu \pm (Z_{a/2})\left(\frac{\sigma}{\sqrt{n}}\right) \tag{6.3}$$

where $Z_{\alpha/2}$ is the *confidence level*. The confidence level represents how willing we are to accidentally report a mistake. The Greek letter α actually represents something called the *alpha level*. Normally, we use values of $\alpha = 0.05$ or $\alpha = 0.01$, which mean we are willing to be wrong five out of every hundred times, or one out of every hundred times, respectively.

The *standard error of the mean*, $\sigma/\sqrt{n}$, is a measurement of how much error results from the size of our sample. The bigger our sample, the less likely we are to accidentally end up with an unrepresentative sample, and therefore the standard error of the mean will be smaller. Conversely, if we have a small sample, we expect the chances of us having a "bad" sample to be higher, so the standard error will be bigger.

The *confidence limits*, or *margin of error*, are simply the product of the standard error of the mean and the Z score for a given confidence level. More formally:

$$e = (Z_{\alpha/2})\left(\frac{\sigma}{\sqrt{n}}\right) \tag{6.4}$$

e represents the confidence limit for a given α, σ and *n*. As α decreases, the interval gets bigger.

In our case, the CIs are used as a measure of the spots homogeneity. An interval [0, t] is plotted versus the amount of spots that their CV lie in this interval. Therefore, the more rapid the increase of this amount while t remains small, the more homogeneous the spots.

6.2.3 Mahalanobis Distance

In statistics, Mahalanobis distance is a distance measure based on correlations between variables by which different patterns can be identified and analysed. It is a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements.

Formally, the Mahalanobis distance from a group of values with mean $\mu = (\mu_1, \mu_2, \mu_3, ..., \mu_p)^T$ and covariance matrix Σ for a multivariate vector $x = (x_1, x_2, x_3, ..., x_p)^T$ is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}. \tag{6.5}$$

Mahalanobis distance can also be defined as dissimilarity measure between two random vectors $\vec{x}$ and $\vec{y}$ of the same distribution with the covariance matrix Σ :

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \Sigma^{-1} (\vec{x}-\vec{y})}. \tag{6.6}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called the normalized Euclidean distance:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}}, \tag{6.7}$$

where σ_i is the standard deviation of the x_i over the sample set.

6.3 Results Presentation

In this subsection, the below qualitative and quantitative performance measures are going to be presented:

- visualization of the resulting image,
- visual comparison of the resulting images to the images resulting from other processing techniques,
- boxplots of the spots and background CVs, and the spots CI,
- plot of an approximation of the individual values of the mahalanobis distance between the background and spots,
- visualization of several spots that have been enhanced at a great extend by our method,

- visualization of the segmentation result for the selected spots, as it arises from the ImaGene package,
- amount of spots identified by SpotSegmentation.

For starters, an image dyed with Cy3 and scanned at the higher settings of the scanner is presented ('image1G'). Figure 6.1 shows the original and processed images for the two-stage method. The red box areas in Figure 6.1B show that the application of the correlation stage and then the coring stage significantly enhances low-intensity spots therefore discriminating them from the local noisy background. We also note that, in some cases (as highlighted by the yellow box area in Figure 6.1B) the resulting spots seem to be dilated. Application of the coring stage and then the correlation stage does not produce such an effect, as highlighted in Figure 6.1C. Comparison of Figure 6.1B and Figure 6.1C tends to indicate that the correlation stage applied before the coring stage has a better denoising effect for low-intensity microarray spots.

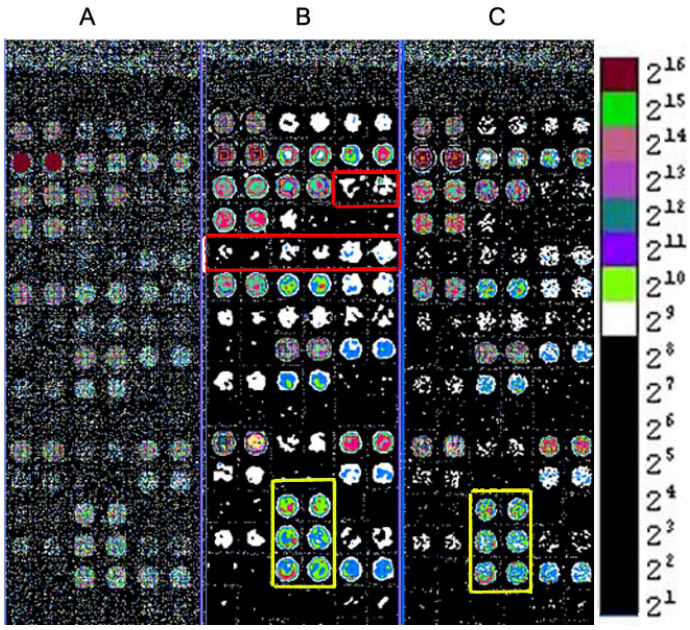


Figure 6.1 - Results of the proposed two-stage approach for image1G: correlation stage for additive noise removal and coring stage for multiplicative noise removal.
A: original image, B: image processed with the correlation stage and then with the coring stage, C: image processed with the coring stage and then with the correlation stage. Red box areas show low-intensity spots that are more enhanced by the application of the correlation stage and then the coring stage. Yellow areas show spots processed by the correlation stage and then the coring stage that have been dilated, while no such an effect is observed when applying first the coring stage and then the correlation one.

In Figure 6.2 and Figure 6.3, the proposed method is compared to other image processing techniques, which were described in Section 5. As foresaid, each of those methods account only for one noise component, either additive or multiplicative. Figure 6.2 demonstrates the results that arise when the methods attack only the additive component of the noise. In this case, it is obvious that when going up to analysis levels at the correlation method some low-intensity spots vanish (green areas in Figure 6.2O) while others are dilated (yellow areas in Figure 6.2O). We can clearly see that some low-intensity spots are enhanced after having been processed by this technique (red areas in Figure 6.2N) and all spots are much more homogeneous than the initial ones. Nonetheless, the application of the proposed two-stage approach (Figure 6.2P and Q) yields better results because it accounts for both noise

components. However, the conclusions for the correlation method confirm its selection for the additive noise component's removal.

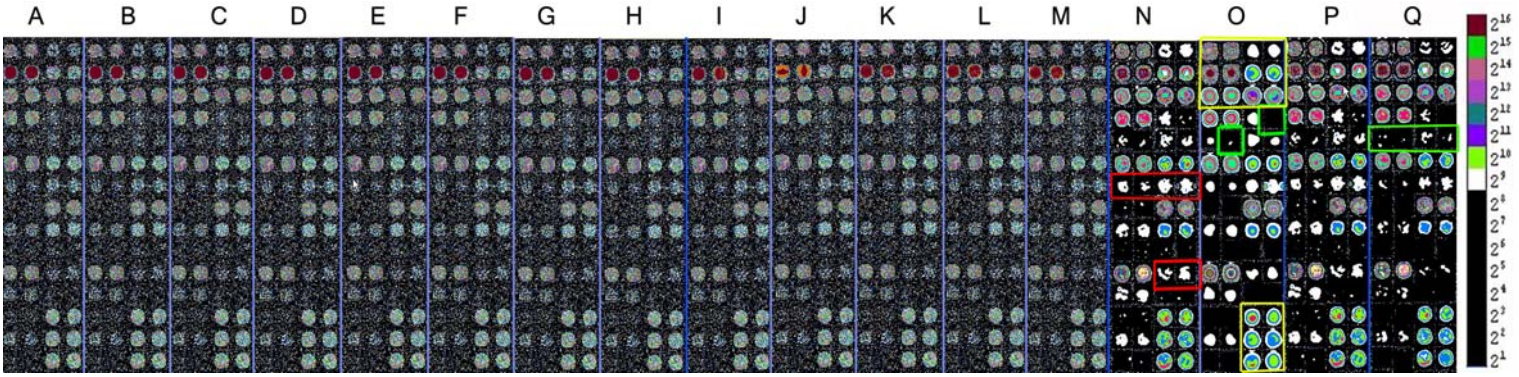


Figure 6.2 - Results when considering the additive component of noise for image1G. A:original image, B,C,D: images processed with the coring method assuming gaussian signal at 1,2,3 levels of analysis, respectively, E,F,G: images processed with hard thresholding at 1,2,3 levels of analysis, respectively, H,I,J: images processed with soft thresholding at 1,2,3 levels of analysis, respectively, K,L,M: images processed with the coring method assuming laplacian signal at 1,2,3 levels of analysis, respectively, N,O: images processed with the correlation technique at levels 1-2, 2-3 of analysis, respectively. P: image processed with the correlation stage and then with the coring stage, Q: image processed with the coring stage and then with the correlation stage. Red areas show low-intensity spots enhanced by the correlation method. Yellow areas show spots that have been dilated when processed by the correlation method. Green areas correspond to low-intensity spots that vanish when processed by the correlation technique in higher levels or by the coring stage and then with the correlation stage.

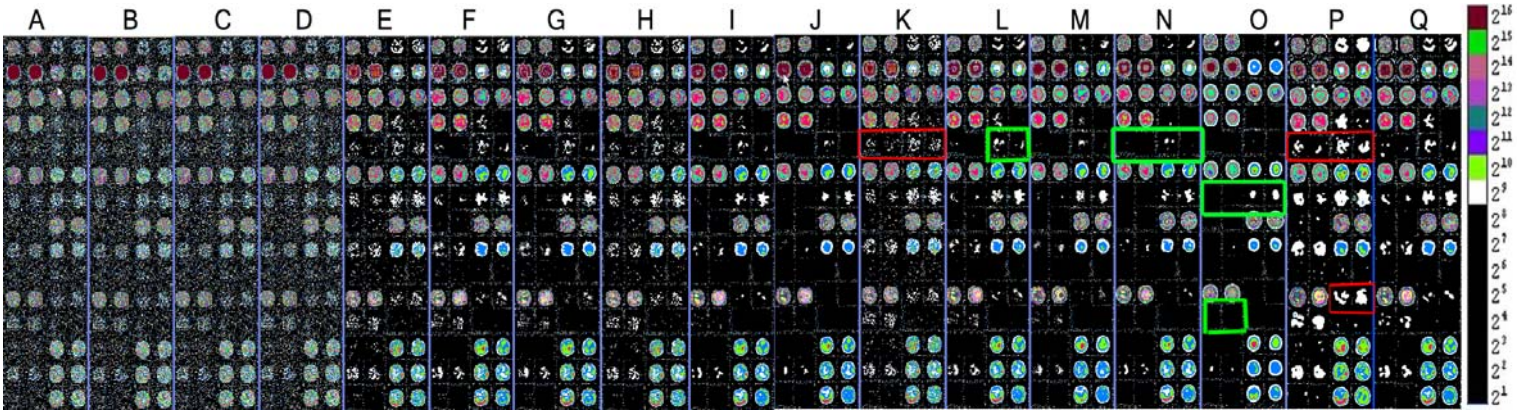
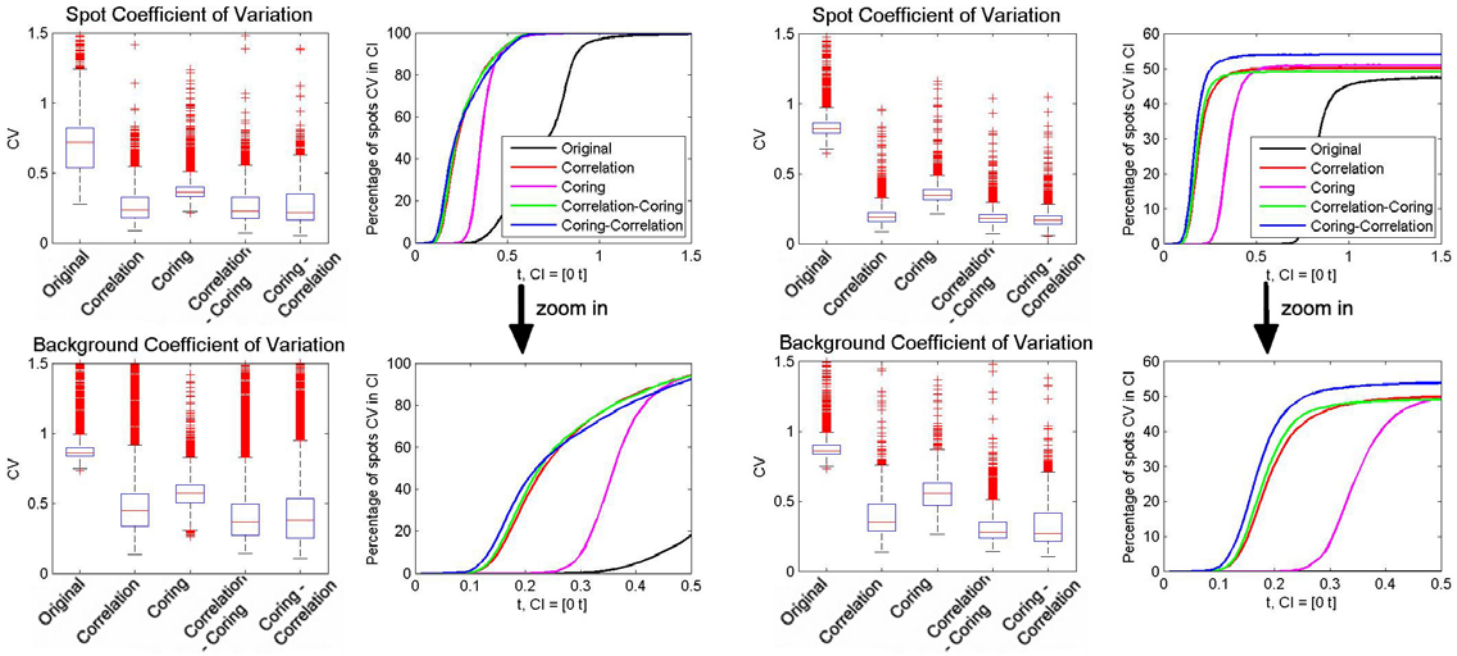


Figure 6.3 - Results when considering the multiplicative component of noise for image1G. A:original image, B,C,D: images processed with gaussian model at 1,2,3 levels of analysis, respectively, E,F,G: images processed with hard thresholding at 1,2,3 levels of analysis, respectively, H,I,J: images processed with soft thresholding at 1,2,3 levels of analysis, respectively, K,L,M: images processed with the coring method assuming laplacian signal at 1,2,3 levels of analysis, respectively, N,O: images processed with the correlation technique at levels 1-2, 2-3 of analysis, respectively. P: image processed with the correlation stage and then with the coring stage, Q: image processed with the coring stage and then with the correlation stage. Green areas correspond to some low-intensity spots that vanish. Red areas show that coring at the first level of analysis, as well as correlation and then coring stage, results into more enhanced spots.

Moreover, Figure 6.3 shows the results when accounting only for the multiplicative component of the noise. From this figure, it is obvious that the correlation method is not proper for this case due to the fact that some low-intensity spots vanish (green box areas in Figure 6.3N, and Figure 6.3O). We can clearly see (red areas in Figure 6.3K) that coring method based on Laplacian modeling at the first level, henceforth referred as coring method for simplicity, of analysis renders low-intensity spots more

enhanced. It is obvious, though, that the two-stage method we presented in this dissertation renders the low-intensity spots even more homogeneous (red areas in Figure 6.3P). On the other hand, the coring method yields the best results of all the methods that attack only one noise component, and this argues that it is the best method for the multiplicative noise component's removal. From now on, for space saving, our two-stage method will be compared only to the coring stage and the correlation stage, independently.



(a) **(b)**

Figure 6.4 - Effect of the two-stage approach on the homogeneity of the microarray spot and background areas for image1G.
(a) CV that correspond to all spots
(b) CV that correspond to low-intensity spots
The upper set of boxplots represents the signal $\log_2$CV. The lower set of boxplots represents the background $\log_2$CV. In each boxplot set, the first plot corresponds to the original image, the second corresponds to the image processed by the correlation stage, the third corresponds to the image processed by the coring stage, the fourth corresponds to the image processed by the two-stage correlation followed by coring method and the fifth corresponds to the image processed by the two-stage coring followed by correlation method.

An interesting issue for consideration is the order of applying the two processing stages. Figure 6.4 shows the coefficient of variance (CV), that demonstrates the effect of the two stages in the homogeneity of both the background and the microarray spot areas. More specifically, the figures provide boxplots of (i) the signal CV and (ii) the background CV and a plot which presents the percentage of spots whose CV falls into a confidence interval (CI). This CI plot indicates that the proposed method tends to assign a CV close to zero to the major percentage of spots. Figure 6.4b is the same boxplot as Figure 6.4a but it accounts only for the low-intensity spots. It illustrates that when we apply first the coring and then the correlation stage we get more low-intensity spots with CV close to zero which implies more homogeneous spot areas corresponding to low-intensity spots.

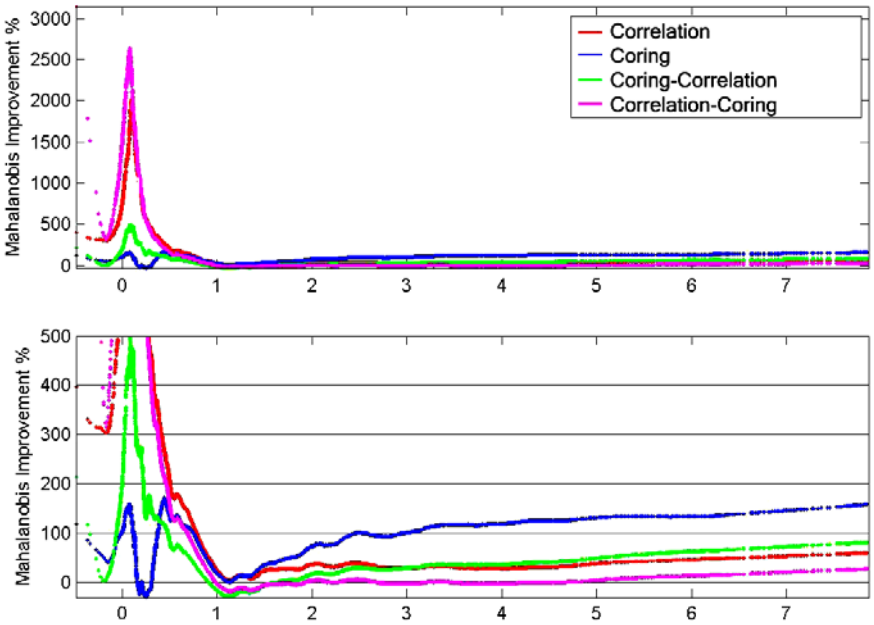
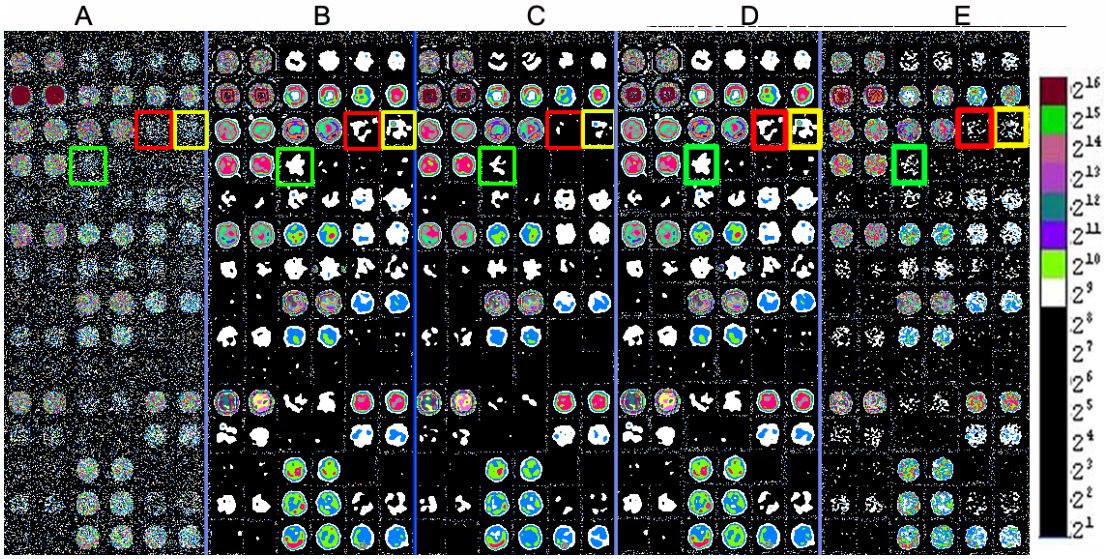


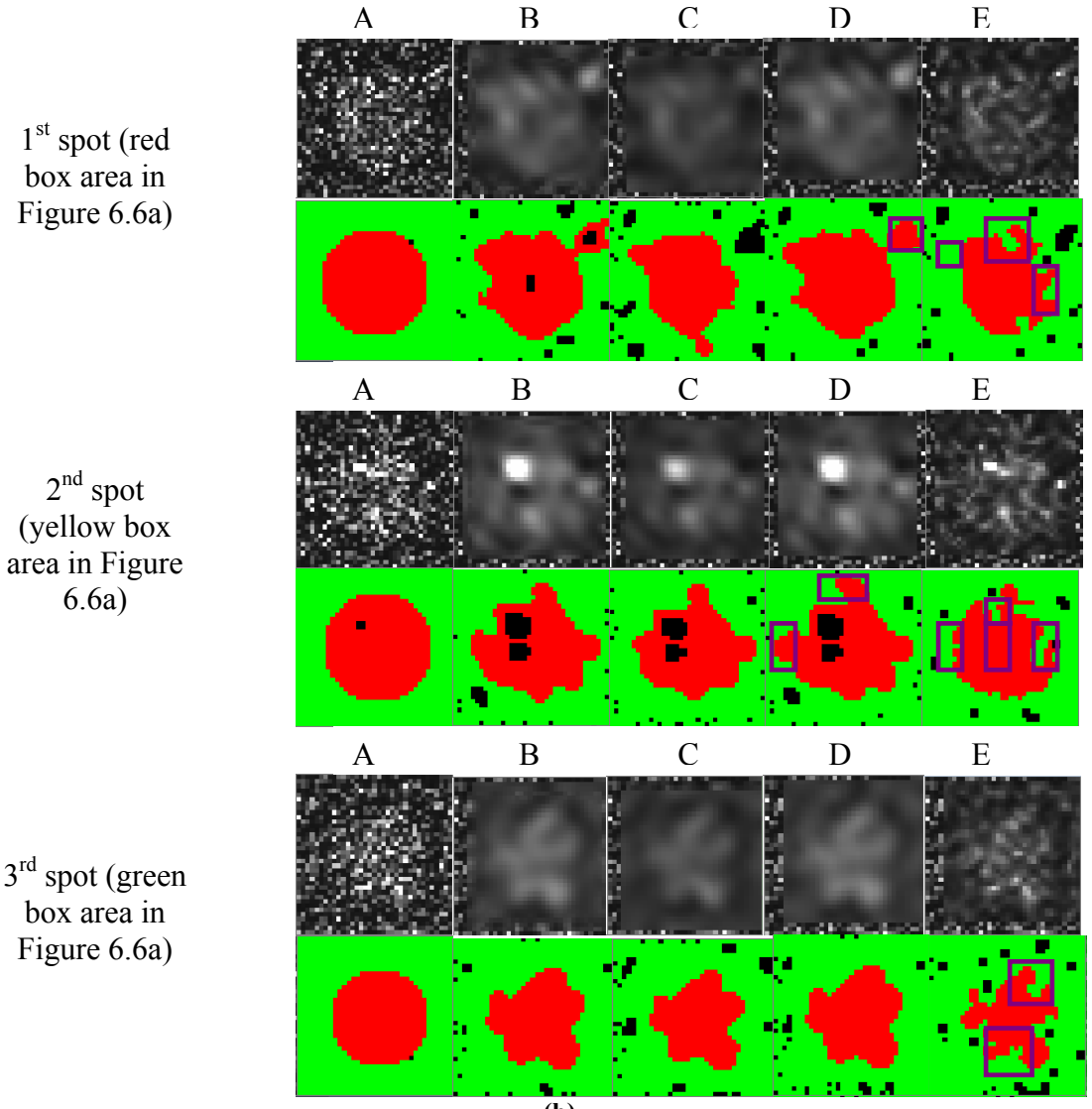
Figure 6.5 - Improvement of the Mahalanobis distance of spots and background between the original and the processed images as a function of the spot-to-background intensity ratio for image1G.

Figure 6.5 illustrates another quantitative performance metric, namely, the Mahalanobis distance improvement of spot and background areas between the processed and original images as a function of the spot-to-background intensity ratio. This Figure contains an approximation of the individual values. While Figure 6.4b illustrates that the application of the correlation stage followed by the coring stage tends to make the low-intensity spots more homogeneous than they were, Figure 6.5 shows that these spots become more distinctive from the local background. On the other hand, when combining the two stages vice versa we get a better discrimination for the high-intensity spots but not as good for the low-intensity ones (cf. Figure 6.5).

For this image, the three spots shown in Figure 6.6 have been enhanced at a great extent and the result of the ImaGene segmentation appears only for them. Regarding these spots (red, yellow and green box areas in Figure 6.6a), it is not possible for ImaGene to detect them in the original image (Figure 6.6b - A). Therefore, it defines a circular area to be the segmentation area for these spots. On the other hand, the proposed method results in better segmentation due to the fact that the spots have been enhanced before the segmentation step. Our two-stage method results in better results even than the application of each step on its own, as it was expected. For the first spot the segmentation of the correlation method's result does not discriminates that there are artifacts inside the spot (purple box area in Figure 6.6b – 1st spot – D). The segmentation of the coring method's result assumes a smaller spot area for all spots (purple box areas in Figure 6.6b – E).



(a)



(b)

Figure 6.6 – (a) Spot Selection. (b) Segmentation results from ImaGene for image1G. A: original image, B: image processed with the correlation stage and then with the coring stage, C: image processed with the coring stage and then with the correlation stage, D: image processed with the correlation method, E: image processed with the coring method.

We continue our results presentation with the results of the processing of the same image dyed with Cy5 and scanned at the higher settings of the scanner (‘image1R’). Figure 6.7, as Figure 6.1, shows the original image and those processed by the proposed method. The red box areas in Figure 6.7B show that the application of the correlation stage and then the coring stage significantly enhances low-intensity spots in this image as well, therefore discriminating them from the local noisy background. However, the application of the coring stage and then the correlation stage renders some spots, enhanced by the application of the stages in reverse order, vanish (as highlighted by the green box area in Figure 6.7C). From Figure 6.7B and Figure 6.7C along with Figure 6.7B and Figure 6.7D, we ensure the indication that the correlation stage applied before the coring stage has a better denoising effect for low-intensity microarray spots.

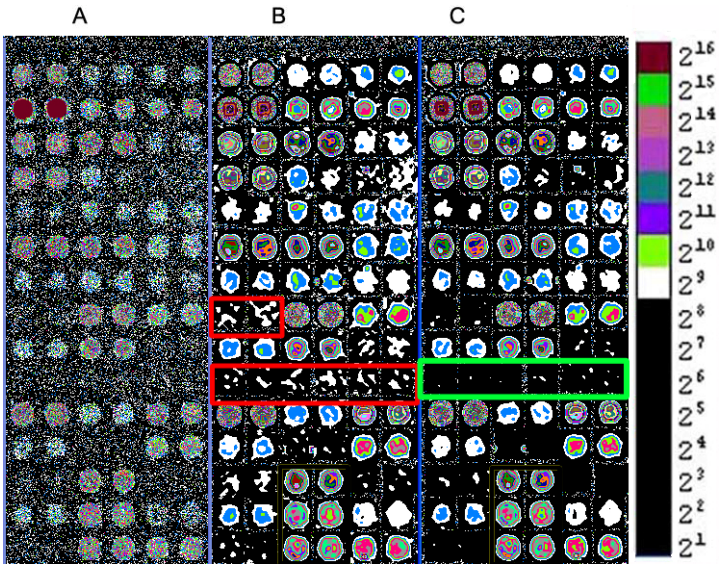


Figure 6.7 - Results of the proposed two-stage approach for image1R: correlation stage for additive noise removal and coring stage for multiplicative noise removal.
A: original image, B: image processed with the correlation stage and then with the coring stage, C: image processed with the coring stage and then with the correlation stage. Red box areas show low-intensity spots that are more enhanced by the application of the correlation stage and then the coring stage. Green areas show spots processed by the coring stage and then the correlation stage that have started vanishing.

In Figure 6.8 and Figure 6.9, the image processed by the proposed method is compared to images processed by the other described techniques. Figure 6.8 demonstrates the results that arise when the methods attack only the additive component of the noise. Also from this image, the correlation method appears to be the best choice for the additive noise component removal. Again, some low-intensity spots are enhanced when processed by this technique (red areas in Figure 6.8N) and all spots are much more homogeneous than the initial ones. However, when going up to analysis levels at the correlation method some low-intensity spots vanish (green areas in Figure 6.8O) while others are dilated (yellow areas in Figure 6.8O).

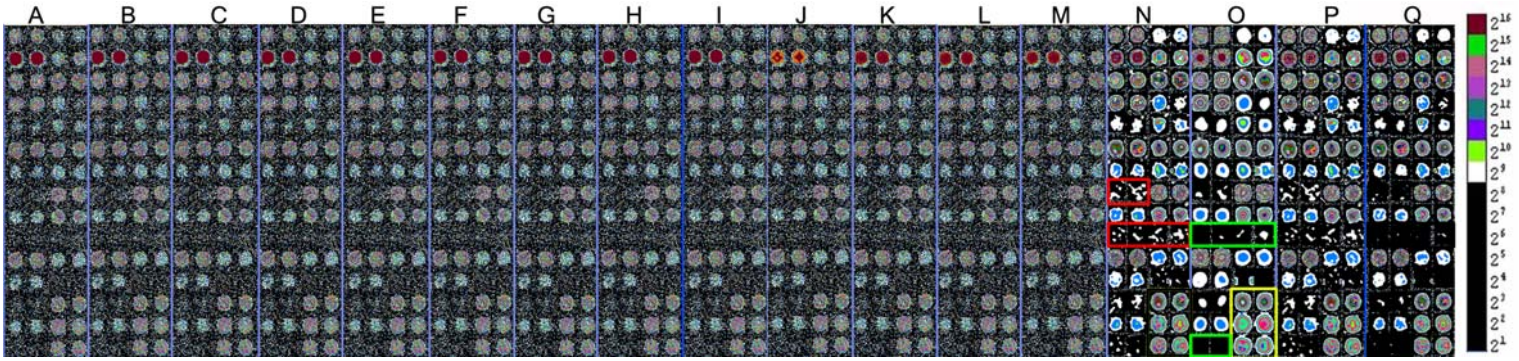


Figure 6.8 - Results when considering the additive component of noise for image1R.
A:original image,
B,C,D: images processed with the coring method assuming gaussian signal at 1,2,3 levels of analysis, respectively,
E,F,G: images processed with hard thresholding at 1,2,3 levels of analysis, respectively,
H,I,J: images processed with soft thresholding at 1,2,3 levels of analysis, respectively,
K,L,M: images processed with the coring method assuming laplacian signal at 1,2,3 levels of analysis, respectively,
N,O: images processed with the correlation technique at levels 1-2, 2-3 of analysis, respectively.
P: image processed with the correlation stage and then with the coring stage,
Q: image processed with the coring stage and then with the correlation stage.
Red areas show low-intensity spots enhanced by the correlation method. Yellow areas show spots that have been dilated when processed by the correlation method. Green areas correspond to low-intensity spots that vanish when processed by the correlation technique in higher levels or by the coring stage and then with the correlation stage.

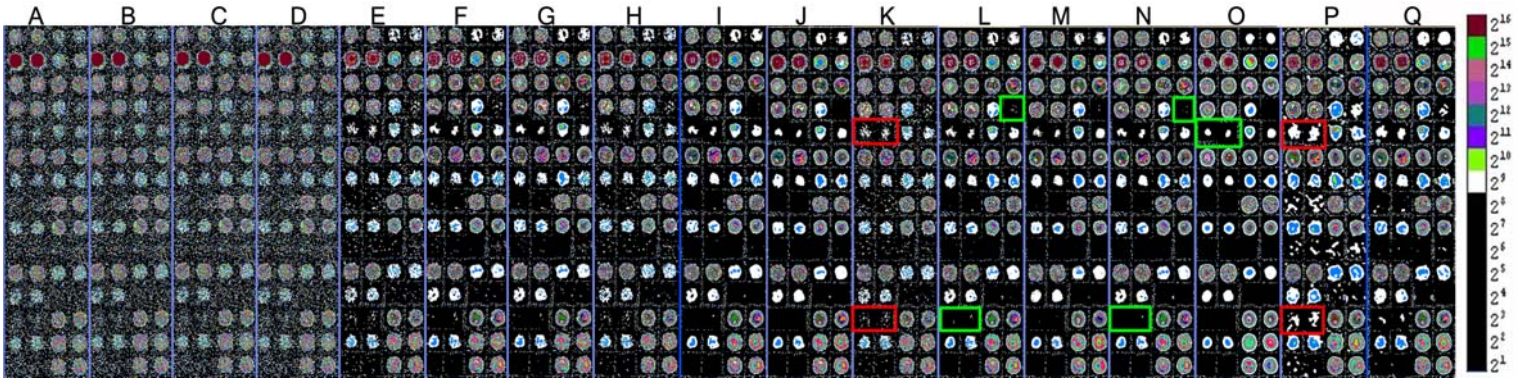


Figure 6.9 - Results when considering the multiplicative component of noise for image1R.
A:original image,
B,C,D: images processed with gaussian model at 1,2,3 levels of analysis, respectively,
E,F,G: images processed with hard thresholding at 1,2,3 levels of analysis, respectively,
H,I,J: images processed with soft thresholding at 1,2,3 levels of analysis, respectively,
K,L,M: images processed with the coring method assuming laplacian signal at 1,2,3 levels of analysis, respectively,
N,O: images processed with the correlation technique at levels 1-2, 2-3 of analysis, respectively.
P: image processed with the correlation stage and then with the coring stage,
Q: image processed with the coring stage and then with the correlation stage.
Green areas correspond to some low-intensity spots that vanish. Red areas show that coring at the first level of analysis, as well as correlation and then coring stage, results into more enhanced spots.

The results from the techniques when accounting only for the multiplicative component of the noise are shown in Figure 6.9. From this figure, the correlation method, again, appears not to be proper for this case because some low-intensity spots start vanishing (green box areas in Figure 6.9N, and Figure 6.9O). The coring method at the first level of analysis renders, again, low-intensity spots more enhanced (red areas in Figure 6.9K) but the low-intensity spots become even more enhanced and

homogeneous when processed by the two-stage proposed method (red areas in Figure 6.9P). However, the choice of the coring method for attacking the multiplicative noise component is once again confirmed to be the best one for this noise component's removal.

From Figure 6.10 we can see that our method yields – once again – more homogeneous background and microarray spot areas than in the original image. The same stands for the coring processing as well. From the zoomed version of the CI plot in Figure 6.10b, one can tell that the proposed method slightly outperforms the correlation method because there is a rapid and quick increase of the spot percentage while t, where CI = [0 t], remains close to zero.

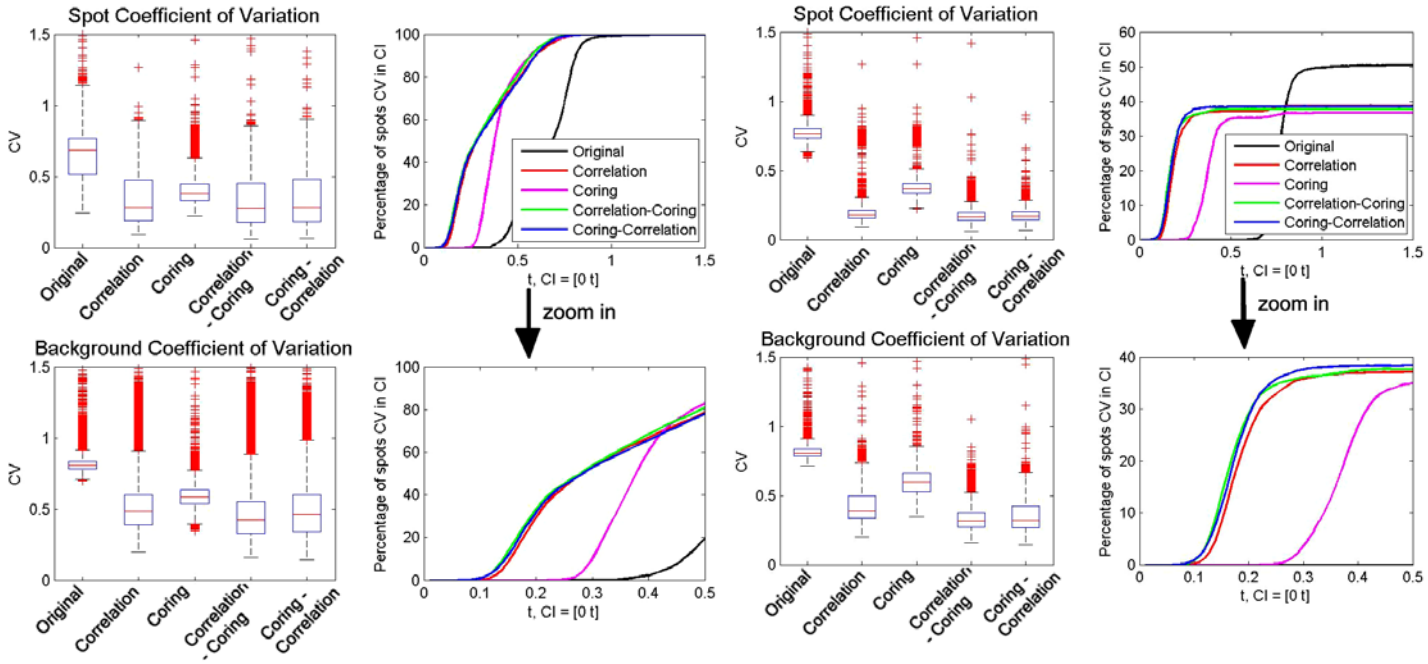


Figure 6.10 - Effect of the two-stage approach on the homogeneity of the microarray spot and background areas for image1R.
(a) CV that correspond to all spots
(b) CV that correspond to low-intensity spots
The upper set of boxplots represents the signal $\log_2$CV. The lower set of boxplots represents the background $\log_2$CV. In each boxplot set, the first plot corresponds to the original image, the second corresponds to the image processed by the correlation stage, the third corresponds to the image processed by the coring stage, the fourth corresponds to the image processed by the two-stage correlation followed by coring method and the fifth corresponds to the image processed by the two-stage coring followed by correlation method.

From Figure 6.11 we conclude that there is a significant improvement of the Mahalanobis distance of spot and background areas between the processed and original images for the low intensity spots. Moreover, all processing techniques perform a 100% increase in Mahalanobis distance for the high-intensity spots, whereas in the former image only the application of the coring stage and then the correlation resulted in such an improvement (see Figure 6.5).

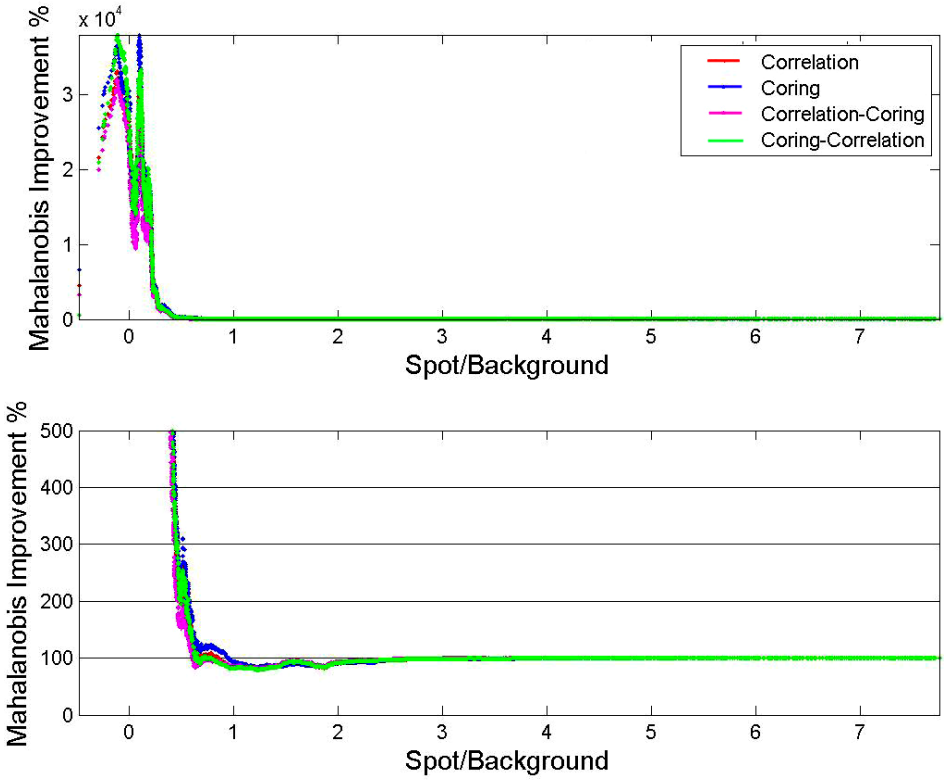
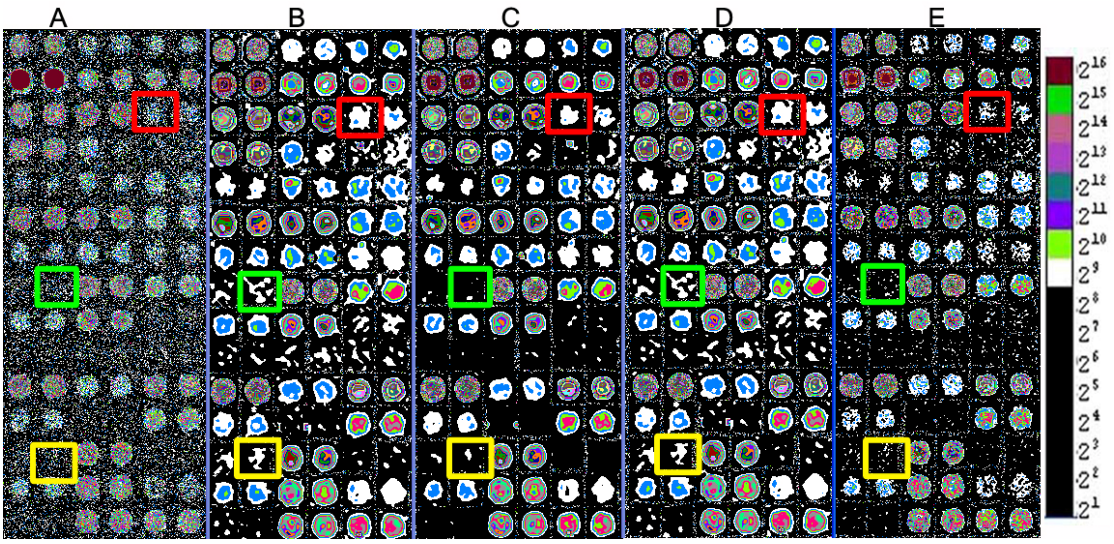
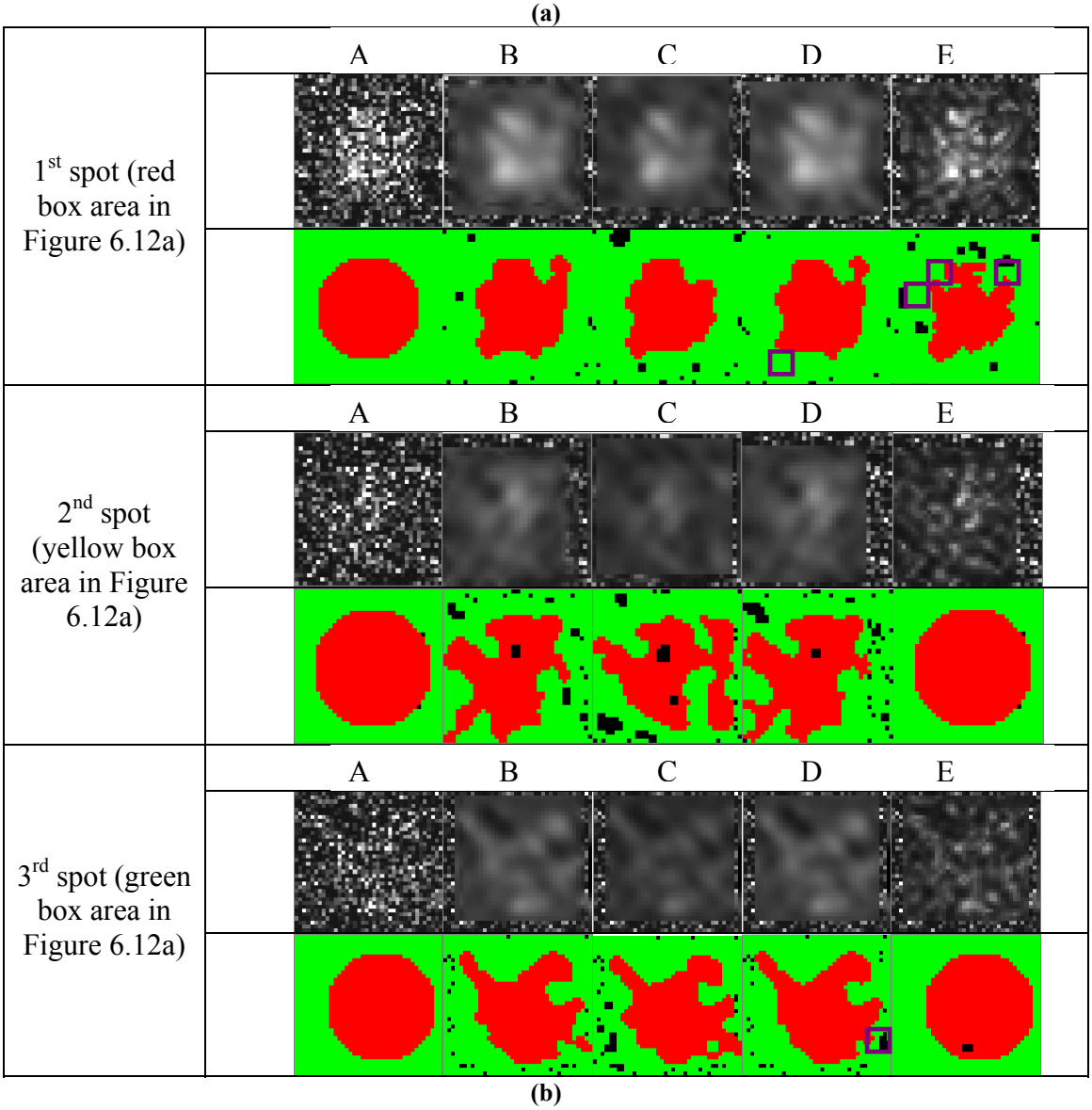


Figure 6.11 - Improvement of the Mahalanobis distance of spots and background between the original and the processed images as a function of the spot-to-background intensity ratio for image1R.

We, again, select some spots (red, yellow, green box areas in Figure 6.12a) that have been enhanced by our method and they are shown in Figure 6.12 along with their ImaGene segmentation result. ImaGene is still unable to detect these spots if they are not first processed (Figure 6.12b - A), consequently, assigning a circular area to be the segmentation area for these spots. However, processing these spots with our method allows ImaGene to perform better segmentation because the spots are more distinguishable from their local background. The coring method still results in bad segmentation especially for the second (Figure 6.12b – 2nd spot – E) and third spot (Figure 6.12b – 3rd spot – E) where ImaGene cannot identify a spot, but also for the first spot as it assumes smaller spot area (purple box areas in Figure 6.12b – 1st spot – E). Correlation method achieves better results than the coring method, however, not as good as the proposed approach. From the purple box area in Figure 6.12b – 1st spot – D it is obvious that ImaGene considers a smaller spot area and it reckons an artifact at the third spot (purple box area in Figure 6.12b – 3rd spot – D) although it should consider it as spot area.



(a)



(b)

Figure 6.12 – (a) Spot Selection. (b) Segmentation results from ImaGene for image1R. A: original image, B: image processed with the correlation stage and then with the coring stage, C: image processed with the coring stage and then with the correlation stage, D: image processed with the correlation method, E: image processed with the coring method.

Table 6.1 presents the amount of spots that SpotSegmentation is unable to detect. The results come from the composition of both red and green channel of the image that we have presented, so far. From this table, it appears that the coring stage finds most spots than the proposed approach. But when looking at the images provided by the software package (data not shown), we can discern that SpotSegmentation identifies "spots" in areas where no spot exists. The percentage of 8.3% consists of really detected spots along with a number of false positives. From the images of SpotSegmentation and from Table 6.1, it is obvious that the proposed method increases the spot detectability of the original image by approximately 40%. This result demonstrates the significance of the proposed two-stage approach.

	Original	Coring	Correlation	Coring - Correlation	Correlation - Coring
Spots NOT detected	42.41%	8.3%	21.2%	18.8%	20.8%

Table 6.1 – Results from SpotSegmentation for image1R and image1G.

Finally, we present an image that is dyed with Cy5 and is scanned at lower scanning settings ('image2R'). The result from the two-stage processing is demonstrated in Figure 6.13. It is obvious that scanning an image at lower settings results in a much more denoised image that has more low-intensity spots (green box areas in Figure 6.13A). The red box areas in Figure 6.13B, C shows spots that are enhanced after having been processed by our method. Processing results, once again, in more homogeneous spots (cyan box areas in Figure 6.13B, C) even if this colormap does not help us distinguish many spots (yellow areas in Figure 6.13B).

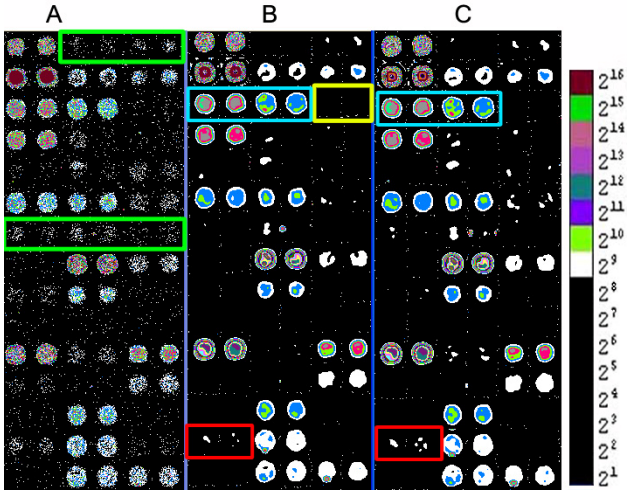


Figure 6.13 - Results of the proposed two-stage approach for image2R: correlation stage for additive noise removal and coring stage for multiplicative noise removal.
A: original image, B: image processed with the correlation stage and then with the coring stage, C: image processed with the coring stage and then with the correlation stage. Red box areas show low-intensity spots that are more enhanced by the proposed method. Cyan areas show spots which have become more homogeneous after being processed by the two-stage method. Green areas show spots that have a very low-intensity due to the low scanning settings. Yellow areas correspond to spots that are not obvious with this colormap.

For the third time in a row, the results in Figure 6.10 demonstrate the increase in homogeneity of the background and microarray spot areas after having the original image processed. This homogeneity is more significant for the low intensity spots, as illustrated in Figure 6.14b, and even more if the image is processed by the proposed

method as there is a rapid and quick increase of the spot percentage while *t*, where CI = [0 *t*], remains close to zero (cf. zoomed version of CI plot of Figure 6.14b).

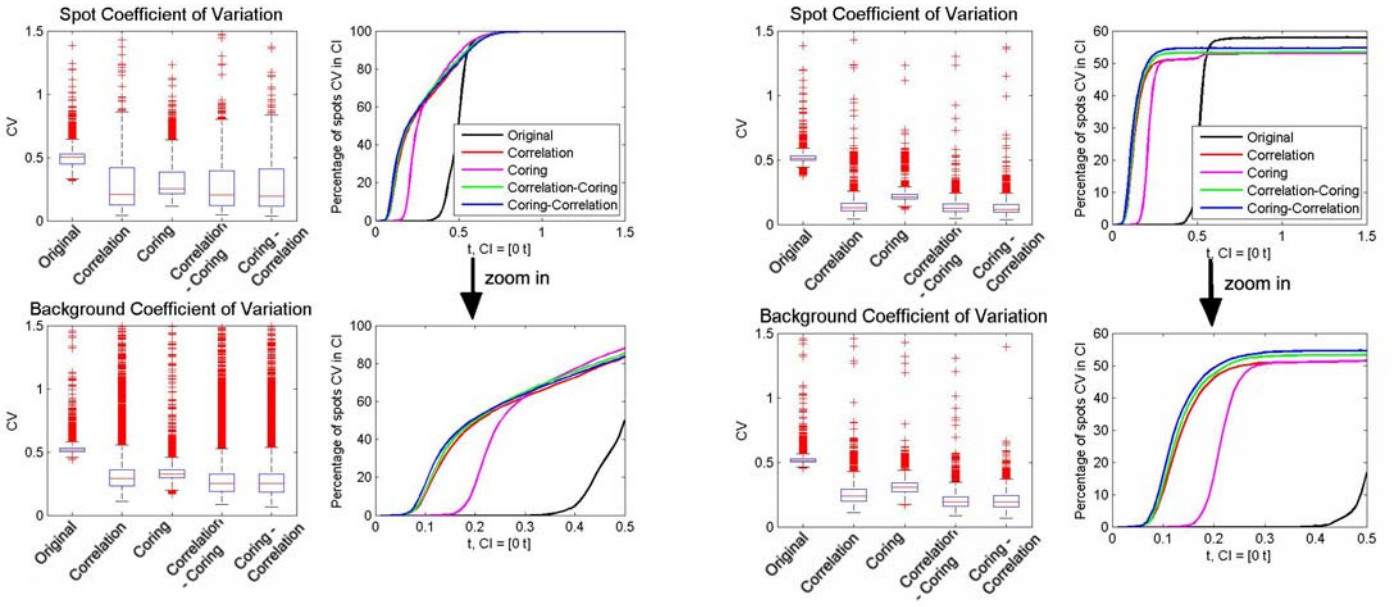


Figure 6.14 - Effect of the two-stage approach on the homogeneity of the microarray spot and background areas for image2R.
(a) CV that correspond to all spots
(b) CV that correspond to low-intensity spots
The upper set of boxplots represents the signal $\log_2$CV. The lower set of boxplots represents the background $\log_2$CV. In each boxplot set, the first plot corresponds to the original image, the second corresponds to the image processed by the correlation stage, the third corresponds to the image processed by the coring stage, the fourth corresponds to the image processed by the two-stage correlation followed by coring method and the fifth corresponds to the image processed by the two-stage coring followed by correlation method.

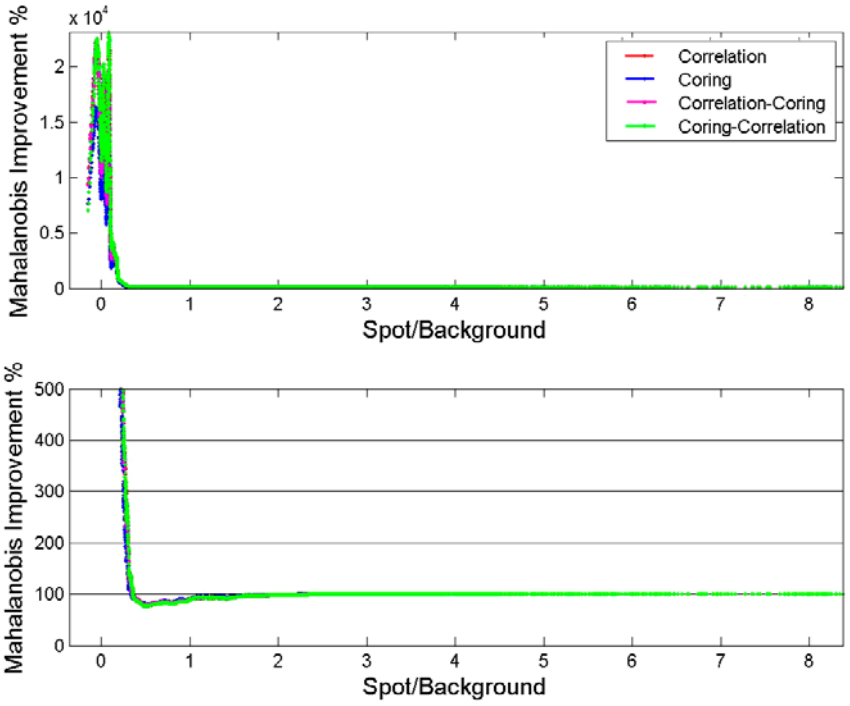
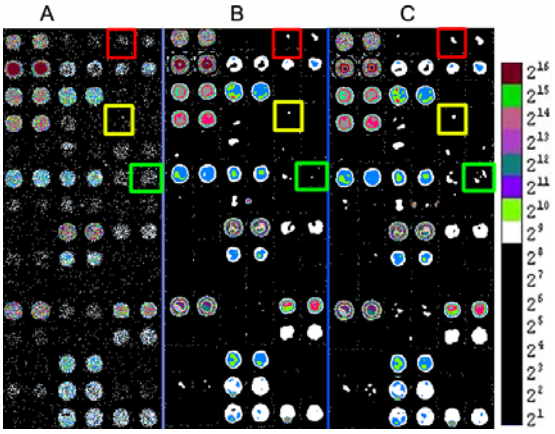
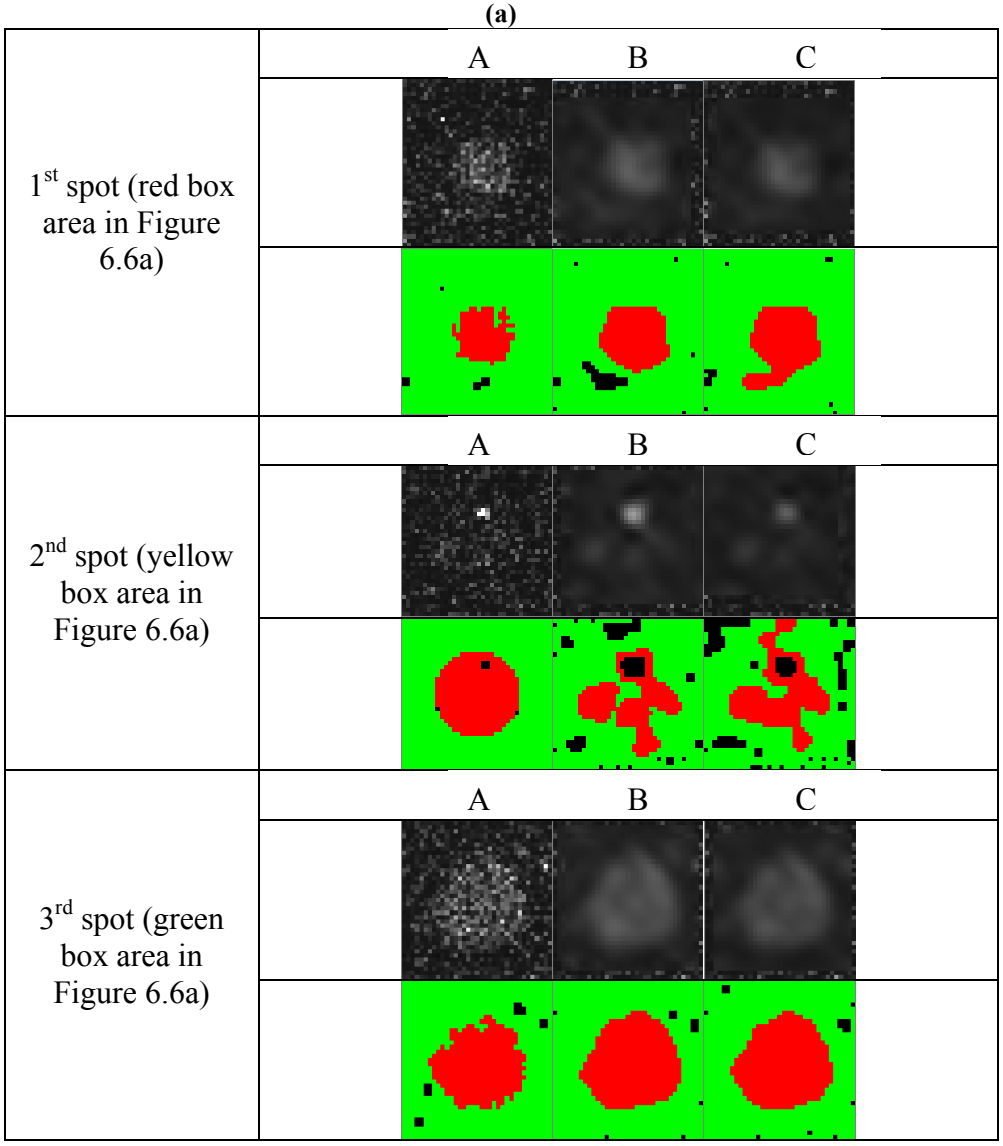


Figure 6.15 - Improvement of the Mahalanobis distance of spots and background between the original and the processed images as a function of the spot-to-background intensity ratio for image2R.

As in Figure 6.11, in Figure 6.15 we notice a significant improvement of the Mahalanobis distance of spot and background areas between the processed and original images for the low intensity spots. Again, all processing techniques perform a 100% increase in Mahalanobis distance for the high-intensity spots.



(a)



(b)

Figure 6.16 – (a) Spot Selection. (b) Segmentation results from ImaGene for image2R. A: original image, B: image processed with the correlation stage and then with the coring stage, C: image processed with the coring stage and then with the correlation stage.

In this image, the selection of individual spots from the resulting image in Figure 6.13 is very difficult because the colormap does not help for the discrimination of the spots. However, when the resulting images are processed with ImaGene we observe that the segmentation is better than that of the original image, even though the spots were not well discriminated. In Figure 6.16a the selected spots are presented as they appear in the images (red, yellow, green box areas) and the segmentation result form the ImaGene tool is presented in Figure 6.16b. It is evident that the spots are better segmented when having been processed by the proposed method.

Table 6.2 presents the amount of spots that SpotSegmentation is unable to detect. The results come from the composition of both red and green channel of the image scanned at lower settings. SpotSegmentation package is unable to detect more than half the spots at the unprocessed image while this percentage falls to 33 – 34 % for the processing images. It was expected that these percentages would be larger than those in Table 6.1 due to the large amount of low-intensity spots that exist in this image. It is inevitable that many spots are not going to be identified even if processed by the two-stage method.

	Original	**Coring - Correlation**	**Correlation - Coring**
Spots NOT detected	55.63%	33.31%	34.1%

Table 6.2 – Results from SpotSegmentation for image2R and image2G.

7 CONCLUSIONS

Microarray technology finds wide use, among others, in human disease, aging, drug and hormone action and mental illness. Exploiting the alterations in gene sequences, they pave the way for a new era of genetic screening, testing and diagnostics. It is, therefore, vital to identify every possible change in the gene sequences. This is almost never easy due to the noise inherent in the microarrays. Noise often obscures the spots that contain the information needed for identifying an alteration. Therefore, a denoising step, which enhances the image, plays a pivotal role in order not to omit any important information.

Our results, shown in Section 6, suggest that in high throughput whole-genome approaches, applying a two-stage approach enhances the dynamic range of existing microarray imaging technology, which is very important in order to identify the most significant genes with increased accuracy and robustness. This has been verified both by ImaGene and SpotSegmentation tools. As far as SpotSegmentation is concerned, it is obvious, from the relevant tables, that there is a major increase in the amount of identified spots after the image processing. In addition, the identification of more spots results in better segmentation by ImaGene.

Microarray images consist mostly of low-intensity spots that are not well distinguishable from the background. These low-intensity spots are affected by inherent additive and multiplicative noise components and are those which need enhancement. For these spots, our method achieves great homogeneity and discrimination from the local background. On the other hand, for high-intensity spots we confront the problem of spot dilation, especially in images scanned at higher settings. This artifact does not occur in the images at lower settings and this is due to the fact that the high-intensity spots at these images have smaller intensity than those in the higher settings images which are saturated. However, the spots with high intensity are well distinguishable from the background, consist a small percentage in the microarray images and there is no need to enhance them.

As we have seen, the proposed method is tested on the double-color DNA microarray technology. However, it is performed on the images separately and does not account for the correlation between the two channels. Consequently, we argue that it is applicable to the single-color (Affymetrix) microarrays, as well.

The basic concept of our method is that it accounts for both the additive and multiplicative noise component. Though, there are some microarray experimenters who, in order to overcome the effect of the noise without a denoising step, produce microarrays at the lowest scanning settings. These images suffer more from multiplicative noise and not that much from additive noise. In this case our method would, mistakenly, assume the existence of an additive noise component and try to remove it, consequently assuming an incorrect model for the signal and multiplicative noise component. The coring method applied on the image on its own would be the best choice for attacking this component. We realize that there is need of a metric which would tell if the image suffers from the multiplicative or the additive or both noise components. This metric for evaluating the presence of the noise components

would allow the application of the proper denoising technique for rendering the image more enhanced.

The proposed method is believed to be a novel technique for microarray image enhancement which achieves robust and accurate results. This method could constitute a part of a microarray data processing protocol as a pre-processing step prior to gridding. For example, such a protocol would apply our method on the image and the result would be processed by ImaGene in order to extract the gene information. If the whole process is to be automated, then even a normal user – who knows how to use the ImaGene tool – would be able to use it. This automated protocol would disencumber researchers of getting involved with the uphill task of image enhancement.

Microarray technology is one of the most important and promising research areas today. It finds wide use in many applications that will change the way we view health and disease. However, this technology has to confront some serious problems before it is possible for all its potentials to be in use. We hope that this approach would be a useful tool for the researchers, in order to eliminate the effect of noise in the gene expression measurements.

8 REFERENCES

1 Adams, R., Bischof, L., Seeded region growing, IEEE Trans. Pattern Anal. Machine Intell., 16, 641-647, 1994.

2 Ali, Q.M., Farooq, O., Wavelet transform for denoising and quantification of microarray data, Bioinformatics, Images, and Wavelets (eds. Aykroyd, R.G., Barber, S., & Mardia, K.V.), 81-84. Department of Statistics, University of Leeds.

3 Alwine, J.C., Kemp, D.J., Stark, G.R, Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes, Proc Natl Acad Sci U.S.A., 74(12), 5350-5354, 1977.

4 Basarsky, T., Verdnik, D., Zhai, J.Y, Wellis, D., Overview of a microarray scanner: Design essentials for an integrated acquisition and analysis platform, Microarray Biochip Technology (ed. Schena M.), Natick, 256-284, 2000.

5 Baynon, E., Lamb, D., Charged-coupled Devices and their Application, McGraw-Hill, 1980.

6 Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., GenBank, Nucleic Acids Res., 25, 1-6, 1997.

7 Beucher, S., Meyer, F., The morphological approach to segmentation: The watershed transformation, Mathematical Morphology in Image Processing (ed. E. Dougherty), Marcel Dekker, 1992.

8 Beyer, H., Linejitter and Geometric Calibration of CCD Cameras, ISPRS Journal of Photogrammetry and Remote Sensing, 45, 17-32, 1990.

9 Bishop, J.O., Morton, J.G., Rosbash, M., Richardson, M., Three abundance classes in HeLa cell messenger RNA, Nature, 250(463), 199–204, 1974.

10 Boguski, M.S., Lowe, T.M., Tolstoshev, C.M., dbEST – database for "expressed sequence tags", Nature Genet., 4, 332-333, 1993.

11 Bowtell, D.L., Options available – from start to finish – for obtaining expression data by microarray, Nature Genet., 21, 25-32, 1999.

12 Brown, P.O., Botstein, D., Nat. Genet., 21, Suppl., 33–371, 1999.

13 Buhler, J., Ideker, T., Haynor, D., Improved techniques for finding spots on cDNA microarrays, University of Washington, 2000.

14 Chen, Y, Dougherty, E.R, Bittner, M.L, Ratio-based decisions and the quantitative analysis of cDNA microarray images, J. Biomed. Optics, 2, 364–374, 1997.

15 Chen, J.J., Wu, R., Yang, P.C., Huang, J.Y., Sher, Y.P., Han, M.H., Kao, W.C., Lee, P.J., Chiu, T.F., Chang, F., Chu, Y.W., Wu, C.W., Peck, K., Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection, Genomics, 51, 313-324, 1998.

16 Cheung, V.G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., Making and reading microarrays, Nature Genet., 21, 15-19, 1999.

17 Debouck, C., Goodfellow, P., DNA microarrays in drug discovery and development, Nature Genet. 21, 48–50, 1999.

18 Donoho, D.L., Johnstone, I.M., Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81, 425–455, 1994.

19 Donoho, D.L, Johnstone, J.M., Adapting to Unknown Smoothness via Wavelet Shrinkage, Journal of American Stat.Assoc., 90, 432, 1200-1224, 1995.

20 Dorsel, A., Fundamental Performance Limitations of Hybridized Arrays, invited paper, Lake Tahoe Symposion on Microarray Algorithms and Statistical Analysis: Methods and Standards, 1999.

21 Dror, R.O., Murnick, J.G., Rinaldi, N.J., Marinescu, V.D., Rifkin, R.M., Young, R.A., Bayesian estimation of transcript levels using a general model of array measurement noise, J. Comp.Biology, 10(3/4), 433–452, 2003.

22 Dudley, A.M., Aach, J., Steffen, M.A., Church, G.M., Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range, Proc Nat Acad Sci U S A, 99, 7554–7559, 2002.

23 Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M., Expression profiling using cDNA microarrays, Nature Genet., 21, 10-4,1999.

24 Dutilleux P., An implementation of the algorithme à trous to compute the wavelet transform, in Wavelets, Time-Frequency Methods and Phase Space, (eds. Combes J.M., Grossmann A., Tchamitchian Ph.), Springer-Verlag, 298-304, 1987.

25 Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., Coleman, P., Analysis of gene expression in single live neurons, Proc Nat Acad Sci U S A, 89(7), 3010–3014, 1992.

26 Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D, Cluster analysis and display of genome-wide expression patterns Proc Nat Acad Sci U S A, 95(25), 14863–14868, 1998.

27 Eisen, M.B., Brown, P.O., DNA arrays for analysis of gene expression, Methods Enzymol, 303, 179-205, 1999.

28 Eisen, M, ScanAlyze User Manual, rana.lbl.gov/manuals/ScanAlyzeDoc.pdf, 1999 (22/11/2005).

29 Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M., Boguski, M.S., Data management and analysis for gene expression arrays, Nature Genet., 20, 19–23, 1998.

30 Evan, G., Littlewood, T., A matter of life and cell death, Science 281, 1317–1322, 1998.

31 GenePix Pro 4.0, User's Guide and Tutorial, 2001, Axon Instruments, Inc. http://files.axon.com/downloads/manuals/GenePix_Pro_4.0_User_Guide_Rev_E.pdf (22/11/2005)

32 Goswami, J.C., Chan, A.K., Fundamentals of Wavelets: Theory, Algorithms and Applications, J.Wiley & Sons, 57-107, 141-186, 1999.

33 Guo, H., Odegard, J.E., Lang, M., Gopinath, R.A., Selesnick, I., Burrus, C.S., Speckle reduction via wavelet shrinkage with application to SAR based ATD/R, Technical Report CML TR94-02, CML, Rice University, Houston, 1994.

34 Hartemink, A.J., Gifford, D.K., Jaakola, T.S., Young, R.A., Maximum likelihood estimation of optimal scaling factors for expression array normalization, In SPIE BiOS, San Jose, California, USA, 2001

35 Hassibi, A., Vikalo, H., A Probabilistic Model for Inherent Noise and Systematic Errors of Microarrays, Proc. of Workshop on Genomics Signal Processing and Statistics, 2005.

36 Holland, M.J., Transcript abundance in yeast varies over six orders of magnitude, J. Biol Chem, 277, 14363–14366, 2002.

37 Holschneider, M., Kronland-Martinet, R., Morlet, J., Tchamitchian, Ph., A real-time algorithm for signal analysis with the help of the wavelet transform, in Wavelets, Time-Frequency Methods and Phase Space, (eds. Combes, J.M., Grossmann, A., Tchamitchian, Ph.), Springer-Verlag, 286-297, 1987.

38 Hu, Z., Killion, P.J., Iyer, V. R., Genetic reconstruction of a functional transcriptional regulatory network, Nat Genet, 39, 683–687, 2007.

39 Hughes, T.R. *et al.*, Functional discovery via a compendium of expression profiles, Cell, 102(1), 109–126, 2000.

40 Invitrogen, http://www.invitrogen.com/content.cfm?pageid = 3433, (8/11/05)

41 Jansen, M., Bultheel, A., Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise, IEEE Transactions on Image Processing, 8 (7), 947–953, 1999.

42 Johstone, I.M., Silverman, B.W., Wavelet threshold estimators for data with correlated noise, J. Roy. Statist. Soc. Ser., B 59, 319–351, 1997.

43 Kaiser, G., A Friendly Guide to Wavelets, Birkhäuser,1994.

44 Kamberova, G., Shah, S., Microarray and Image Analysis: Introduction, DNA Array Image Analysis Nuts & Bolts (eds. Kamberova, G., Shah, S.), DNA Press, pp.7-16, 2002.

45 Kamberova, G., Introduction to Image Analysis, DNA Array Image Analysis Nuts & Bolts (eds. Kamberova, G., Shah, S.), DNA Press, pp.17-50, 2002.

46 Kang, J.J., Watson, R.M., Fisher, M.E., Higuchi, R., Gelfand, D.H., Holland, M.J., Transcript quantitation in total yeast cellular RNA using kinetic PCR, Nucleic Acids Res, 28(2), e2, 2000.

47 Khan, J., Simon, R., Bittner, M., Chen Y., Leighton, S.B., Pohida, T., Smith, P.D., Jiang, Y., Gooden, G.C., Trent, J.M., Meltzer, P.S., Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays, Cancer Res. 58, 5009–5013, 1998.

48 Knudsen, S., Guide to Analysis of DNA Microarray Data, Wiley-Liss, 2nd edition, 2004.

49 Lee, M.T., Kuo, F.C., Whitmore, G.A., Sklar, J., Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations, Proc Nat Acad Sci U S A, 97(9), 9834–9, 2000.

50 Lenz, R., Fritsch, D., Accuracy of Videometry with CCD Sensors, ISPRS Journal of Photogrammetry and Remote Sensing, 45, 90-110, 1990.

51 Li, Q., Fraley, C., Bumgarner, R.E., Yeung, K.Y., Raftery, A.E., Donuts, scratches and blanks: robust model-based segmentation of microarray images, Bioinformatics, 21(12), 2875–2882, 2005.

52 Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L., Expression monitoring by hybridization to high-density oligonucleotide arrays, Nature Biotechnol., 14, 1675-1680, 1996.

53 Lockhart, D.J., Winzeler, E. A., Nature, 405, 827–836, 2000.

54 Lou, X.J., Human primary cell gene expression monitoring using cDNA microarrays, poster, Lake Tahoe Center conference on Microarray Algorithms and Statistical Analysis: Methods and Standards, 1999.

55 Lukac, R., Plataniotis, K.N., Smolka, B., Venetsanopoulos, A.N., cDNA Microarray Image Processing Using Fuzzy Vector Filtering Framework, Journal of Fuzzy Sets and Systems: Special Issue on Fuzzy Sets and Systems in Bioinformatics, 152 (1), 17-35, 2005.

56 Mallat, S., Zhong, S., Characterization of signals from multiscale edges, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14 (7), 710–732, 1992.

57 Marshall, A., Hodgson, J., DNA chips: an array of possibilities, Nature Biotechnol., 16, 27-31, 1998.

58 Mastriani, M., Giraldez, A.E., Microarrays Denoising via Smoothing of Coefficients in Wavelet Domain, International Journal of Biomedical Sciences, 1, 1306-1216, 2006.

59 Maxam, A.M., Gilbert, W., A new method for sequencing DNA. Proc. Natl Acad. Sci. USA, 74, 560-564, 1977.

60 McAdams, H.H., Shapiro, L., Circuit simulation of genetic networks, Science 269, 650–656, 1995.

61 McGovern, M., Fayek, R., Advantages of Laser Confocal Microarray Scanning, DNA Array Image Analysis Nuts & Bolts (eds. Kamberova, G., Shah, S.), DNA Press, 51-67, 2002.

62 Misiti, M.,. Misiti, Y, Oppenheim, G., Poggi, J.M., Wavelet Toolbox For Use with MATLAB: User's Guide Version 2, The MathWorks.

63 Murtagh, F., Starck, J-L., Image processing through multiscale analysis and measurement noise modeling, Stat.Comput., 10, 95–103, 2000.

64 Nobel Prize Organization, The Nobel Prize in Chemistry 1980, http://nobelprize.org/chemistry/laureates/1980/index.html, (8/3/06).

65 Olivo-Marin, J-C., Extraction of spots in biological images using multiscale products, Pattern Recognition, 35, 1989–1996, 2002.

66 Pauling, L., Itano, H., Singer, S.J., Wells, I., Sickle cell anemia, a molecular disease, Science, 110, 543-548, 1949.

67 Pawley, J.B., Fundamental Limits in Confocal Microscopy, Handbook of Biological Confocal Microscopy (ed. Pawley, J.B.), 2nd edition, Plenum Press, 19-37, 1995.

68 Pawley, J.B., Sources of Noise in Three-Dimensional Microscopical Data Sets, Three-Dimensional Confocal Microscopy, Investigation of Biological Systems (eds. Stevens, J.K., Mills, L.R., Trogadis, J.E.), Academic Press, 47-69, 1996.

69 Pease, A.C., Solas, D.C., Sullivan, E.J., Cronin, M.T., Holmes, C.P., Fodor, S.P., Light-generated oligonucleotide arrays for rapid DNA sequence analysis, Proc. Natl Acad. Sci. USA, 91, 5022–5026, 1994.

70 Phillips, J., Eberwine J.H., Antisense RNA amplification: a linear amplification method for analyzing the mRNA population from single living cells, Methods 10, 283-288, 1996.

71 QuantArray Analysis Software, GSI Lumonics, www.bipl.ahc.umn.edu/quantarray.html (22/11/05).

72 Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., and Golub, T.R., Multiclass cancer diagnosis using tumor gene expression signatures, Proc Nat Acad Sci U S A, 98(26), 15149–15154, 2001.

73 R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2005.

74 Rocke, D., Durbin, B., A model for measurement error for gene expressionarrays, J Comput Biology, 8(6), 557–569, 2001.

75 Sadler, B.M., Swami, A., Analysis of multiscale products for step detection and estimation, IEEE Transactions on Information Theory, 45 (3), 1043–1051, 1999.

76 Sanger, F., Nicklen, S., Coulson, A.R., DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. USA, 74 (12), 5463-7, 1977.

77 Schadt, E.E., Li, C., Su, C., Wong, W.H., Analyzing high-density oligonucleotide gene expression array data, Journal of Cellular Biochemistry, 80, 192-202, 2000.

78 Schena, M., Shalon, D., Davis, R.W., Brown, P.O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, New York, N.Y 270, 1995.

79 Schena, M., DNA Microarrays: A Practical Approach, Oxford University Press, 1999.

80 Schena, M., Microarray Analysis, Wiley & Sons, 2003.

81 Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H., Normalization strategies for cDNA microarrays, Nucleic Acids Research, 28(10), e47, 2000.

82 Schuler, G.D., Boguski, M.S., Stewart, E.A., et al. (101 co-authors), A gene map of the human genome, Science, 274, 540-546, 1996.

83 Shensa, M.J., The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms, IEEE Transactions on Signal Processing, 40, 10, 2464-2482, 1992.

84 Simoncelli, E.P., Adelson, E.H., Noise removal via Bayesian wavelet coring, in Third Int. Conf. Image Processing, 1, 379–382, 1996.

85 Simoncelli, E.P., Bayesian denoising of visual images in the wavelet domain, Bayesian Inference in Wavelet Based Models, (eds. Muller, P., Vidakovic, B.) Springer-Verlag, 18, 291–308, 1999.

86 Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., Cerrina, F., Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array, Nat. Biotechnol., 17, 974–978, 1999.

87 Skibbe, D.S., Wang, X., Zhao, X., Borsuk, L.A., Nettleton, D., Schnable, P.S., Scanning microarrays at multiple intensities enhances discovery of differentially expressed genes. Bioinformatics, 22(15), 1863–1870, 2006.

88 Slonim, D., From patterns to pathways: gene expression data analysis comes of age, Nat Genet, 32 Suppl., 502–508, 2002.

89 Soille, P., MorphologicaI Image Analysis, Springer-Verlag, 2nd Edition, 2003.

90 Southern, E.M., Detection of specific sequences among DNA fragments separated by gel electrophoresis, J. Mol. Biol., 98, 503-517, 1975.

91 Sripad, A., Snyder, D., A Necessary and Sufficient Condition for the Quantization Noise to be Uniform and White, IEEE Trans. on Acoustic, Speech, and Signal Processing, 25(5), 442-448, 1977.

92 Starck, J-L., Murtagh, F., Bijaoui, A., Multiresolution support applied to image 2ltering and restoration, Graph. Models Image Process, 57 (5), 420–431, 1995.

93 Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R., Hogenesch, J.B., A gene atlas of the mouse and human protein-encoding transcriptomes, Proc Nat Acad Sci U S A, 101(16), 6062–6067, 2004.

94 Tan, H.S., Denoising of Noise Speckle in Radar Image, 2001, innovexpo.itee.uq.edu.au/2001/projects/s804298/thesis.pdf.

95 The Free Dictionary by Farlex, http://www.thefreedictionary.com/cDNA, (8/11/05)

96 The Microarray Family Tree: A historiograph of 13 influential papers, The Scientist, 17(16):29, Published 25th August 2003.

97 Theodoridis, S., Koutroubas, K., Pattern Recognition. Academic Press, San Diego, 1999.

98 Thorp, H.H., Cutting out the middleman: DNA biosensors based on electrochemical oxidation, Trends Biotechnol., 16, 117-121, 1998.

99 Vacha, S.J., McMillan, J., Dorsel, A., Considerations for a Quality Microarray Scanner, DNA Array Image Analysis Nuts & Bolts (eds. Kamberova G., Shah S.), DNA Press, 69-82, 2002.

100 van't Veer, L.J. et al, Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415, 530–536, 2002.

101 Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., Serial analysis of gene expression, Science, 270(5235), 484-7, 1995.

102 Vincent, L., Soille, P., Watersheds in digital spaces: An efficient algorithm based on immersion simulations, IEEE Trans. Pat. Anal. Machine Intell., 13, 583-598, 1991.

103 Wang, X.H., Istepanian, R.S.H., Song, Y.H., Microarray Image Enhancement by Denoising Using Stationary Wavelet Transform, IEEE Transactions on Nanobioscience, 2, 4, 184 – 189, 2003.

104 Wikipedia Encyclopedia, The Free Dictionary by Farlex, http://encyclopedia.thefreedictionary.com/Primer+(molecular+biology), (8/11/05).

105 Wikipedia Encyclopedia, The Free Dictionary by Farlex, http://encyclopedia.thefreedictionary.com/Expressed+sequence+tag, (8/11/05).

106 www.biodiscovery.com/index/imagene

107 Yang, Y.H., Buckley, M.J., Speed, T.P., Analysis of cDNA microarray images, Briefings in Bioinformatics,2 (4), 341-9, 2001.

108 Yang, Y.H., Buckley, M.J., Dudoit, S., Speed, T.P., Comparison of methods for image analysis of cDNA microarray data, Journal of Computational and Graphical Statistics, 11, 108-136, 2002.

109 Yuh, C.H., Bolouri, H., Davidson, E.H., Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene, Science 279, 1896–1902, 1998.

110 Zhang, W., Shmulevich, I. (editors), Computational and Statistical Approaches to Genomics, Kluwer Academic Publishers, 2002.