2014

Novel pattern recognition approaches for transcriptomics data analysis

Iman Rezaeian
University of Windsor

NOVEL PATTERN RECOGNITION APPROACHES FOR TRANSCRIPTOMICS DATA ANALYSIS

by
Iman Rezaeian

A Dissertation
Submitted to the Faculty of Graduate Studies
through School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
2014

NOVEL PATTERN RECOGNITION APPROACHES FOR TRANSCRIPTOMICS DATA ANALYSIS

by

IMAN REZAEIAN

APPROVED BY:

M. R. El-Sakka, External Examiner
Department of Computer Science, Western University

E. Abdel-Raheem
Department of Electrical and Computer Engineering

J. Lu
School of Computer Science

X. Yuan
School of Computer Science

A. Ngom, Co-Advisor
School of Computer Science

L. Rueda, Advisor
School of Computer Science

May 15, 2014

Declaration of Co-Authorship and Previous Publication

I. Co-Authorship Declaration:

I hereby declare that this Dissertation incorporates the outcome of a joint research undertaken in collaboration with Yifeng Li, Martin Crozier, Dr. Eran Andrechek, Dr. Alioune Ngom, Dr. Luis Rueda and Dr. Lisa Porter. The collaboration is covered in Chapter 6 of the Dissertation. In this research, experimental designs, applying and optimizing different machine learning methods for prediction, numerical and visual analysis were performed by the author.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my Dissertation, and have obtained written permission from each of the co-author(s) to include the above material(s) in my Dissertation. I certify that, with the above qualification, this Dissertation, and the research to which it refers, is the product of my own work.

II. Declaration of Previous Publications:

This Dissertation includes 5 original papers that have been previously published/submitted for publication in conferences and peer reviewed journals, as follows:

I certify that I have obtained a written permission from the copyright owner(s) to in-

Dissertation chapter	Publication title
Chapter 2	Luis Rueda, Iman Rezaeian: A Fully Automatic Gridding Method for cDNA Microarray Images. BMC Bioinformatics (2011) 12: 113.
Chapter 3	Luis Rueda, Iman Rezaeian: Applications of Multilevel Thresholding Algorithms to Transcriptomics Data. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 16th Iberoamerican Congress (CIARP), Chile, 2011: 26-37.
Chapter 4	Iman Rezaeian, Luis Rueda: A new algorithm for finding enriched regions in ChIP-Seq data. ACM International Conference on Bioinformatics, Computational Biology and Biomedicine (BCB), Chicago, USA, 2012: 282-288.
Chapter 5	Iman Rezaeian, Luis Rueda: CMT: A Constrained Multi-Level Thresholding Approach for ChIP-Seq Data Analysis. PLoS ONE 9(4): e93873, 2014.
Chapter 6	Iman Rezaeian, Yifeng Li, Martin Crozier, Eran Andrechek, Alioune Ngom, Luis Rueda, Lisa Porter: Identifying Informative Genes for Prediction of Breast Cancer Subtypes. Pattern Recognition in Bioinformatics - 8th IAPR International Conference (PRIB), France, 2013: 138-148.

clude the above published material(s) in my Dissertation. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my Dissertation does not infringe upon anyone copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my Dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my Dissertation. I declare that this is a true copy of my Dissertation, including any final

revisions, as approved by my Dissertation committee and the Graduate Studies office, and that this Dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

We proposed a family of methods for transcriptomics and genomics data analysis based on multi-level thresholding approach, such as OMTG for sub-grid and spot detection in DNA microarrays, and OMT for detecting significant regions based on next generation sequencing data. Extensive experiments on real-life datasets and a comparison to other methods show that the proposed methods perform these tasks fully automatically and with a very high degree of accuracy. Moreover, unlike previous methods, the proposed approaches can be used in various types of transcriptome analysis problems such as microarray image gridding with different resolutions and spot sizes as well as finding the interacting regions of DNA with a protein of interest using ChIP-Seq data without any need for parameter adjustment. We also developed constrained multi-level thresholding (CMT), an algorithm used to detect enriched regions on ChIP-Seq data with the ability of targeting regions within a specific range. We show that CMT has higher accuracy in detecting enriched regions (peaks) by objectively assessing its performance relative to other previously proposed peak finders. This is shown by testing three algorithms on the well-known FoxA1 Data set, four transcription factors (with a total of six antibodies) for *Drosophila melanogaster* and the H3K4ac antibody dataset. Finally, we propose a tree-based approach that conducts gene selection and builds a classifier simultaneously, in order to select the minimal number of genes that would reliably predict a given breast cancer subtype. Our results support that this

modified approach to gene selection yields a small subset of genes that can predict subtypes with greater than 95% overall accuracy. In addition to providing a valuable list of targets for diagnostic purposes, the gene ontologies of the selected genes suggest that these methods have isolated a number of potential genes involved in breast cancer biology, etiology and potentially novel therapeutics.

Dedication

to my love, Forough

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to Dr. Luis Rueda, my supervisor, for his steady encouragement, patient guidance and enlightening discussions throughout my graduate studies. Without his help, the work presented here could not have been possible.

I also want to express my gratitude to my co-advisor Dr. Alioune Ngom, who shared many works with me and gave me many useful suggestions during my research.

I also wish to express my appreciation to Dr. Mahmoud R. El-Sakka, Western University, Dr. Esam Abdel-Raheem, Department of Electrical and Computer Engineering, Dr. Jianguo Lu and Dr. Xiaobu Yuan, School of Computer Science for being in the committee and spending their valuable time. Moreover, I would like to express my gratitude to Dr. Lisa porter for her constructive suggestions and feedbacks.

Finally, I would like to thank all of my colleagues in pattern recognition and bioinformatics lab, especially, Forough Firoozbakht, Mina Maleki and Yifeng Li for their consistent support and help.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC: Area Under Curve, 112

cDNA: Complementary DNA, 22

ChIP: Chromatin Immunoprecipitation, 80

CMT: Constrained Multi-level Thresholding, 101

DNA: Deoxyribonucleic Acid, 60

FABLE: Fast Automated Biomedical Literature Extraction, 137

FDR: False Discovery Rate, 107

GEO: Gene Expression Omnibus, 24

MEDLINE: Medical Literature Analysis and Retrieval System Online, 137

MRF: Markov Random Field, 60

mRMR: minimum Redundancy Maximum Relevance, 126

OMT: Optimal Multi-level Thresholding, 81

OMTG: Optimal Multi-level Thresholding Gridding, 24

qPCR: quantitative Polymerase Chain Reaction, 94

Ribonucleic Acid, 5

RNA-Seq: RNA Sequencing, 81

ROC: Receiver Operating Characteristic, 109

SMD: Stanford Microarray Database, 3, 24

SVM: Support Vector Machine, 127

Chapter 1

Introduction

Pattern recognition and image processing and analysis approaches are some of the main streams for analysis of biological data, especially in transcriptomics. One of the main aims of pattern recognition techniques is to make the process of learning and detection of patterns explicit, in such a way that it can be implemented on computers. Automatic recognition, description and classification have become important problems in a variety of scientific disciplines such as biology, medicine and artificial intelligence. Classification, as one of the most well-known techniques in pattern recognition, is used to build models for identifying the correct class label corresponding to an unknown input sample. These methods are also very useful for analyzing biological data, identifying diseases and biomarkers. On the other hand, when there is no explicit class label corresponding to each sample, the model should figure out the appropriate label for each sample by analyzing the data. Clustering techniques are among these methods, which group similar samples together. Clustering methods has been used vigorously for analyzing multi-dimensional transcriptomics data. Clustering one dimensional data can be solved easily by using multi-level thresholding techniques, for which efficient algorithms are known. Multilevel thresholding has been applied to many

problems in signal and image processing and analysis. Examples are image segmentation, vector and scalar quantization, finding peaks in histograms, processing microarray images and finding enriched regions in next generation sequencing data [1–4].

1.1 Transcriptomics Data Analysis Using Machine Learning Methods

Transcriptomics data analysis is one of the areas that can benefit by using the computational methods. Clustering techniques are among those methods that can help scientists to detect patterns in biological data. Clustering one dimensional data can be efficiently solved using several techniques such as K-means, fuzzy K-means and multi-level thresholding. In particular, multilevel thresholding can solve this problem efficiently by using optimal, polynomial time algorithms. In this thesis, multilevel thresholding is used for finding subgrids and spots in microarray images in Chapters 2 and 3. This method is also used in Chapters 4 and 5 for finding enriched regions in ChIP-Seq data. Feature selection methods are among other machine learning techniques that can be used to select a subset of relevant features from a large set. There are different feature selection methods such as minimum Redundancy Maximum Relevance (mRMR) [5] and chi-squared [6]. In this thesis we use the chi-squared method in Chapter 6 to select a subset of genes that discriminate different subtypes of beast cancer. Classification techniques are other types of machine learning methods that can be used to train models for identifying unknown samples. There are different types of classification methods such as Decision tree [7], Bayes classifier [8], support vector machines (SVMs) [9], fuzzy rule-based classification methods [10] and neural networks [11] among others. In this thesis, we use SVM in Chapter 6 within a hierarchical

scheme to classify different subtypes of breast cancer.

1.2 Microarray Image Processing and Analysis

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Using Microarrays, scientists are able to measure the expression levels of large numbers of genes simultaneously. DNA microarray images are obtained by scanning DNA microarrays at high resolution and are composed of sub-grids of spots. There are different steps toward analyzing DNA microarray images such as gridding, segmentation and quantification among others. Gridding microarray images is one of the most important stages of microarray image analysis, since any error in this step is propagated to further steps and may reduce the integrity and accuracy of the analysis dramatically. DNA microarray images contain sub-grids, and each sub-grid contains a set of spots arranged in a grid with a certain number of rows and columns. Figure 1.1 depicts a real DNA microarray image downloaded from the Stanford Microarray Database (SMD) [12], which corresponds to a study of the global transcriptional factors for hormone treatment of *Arabidopsis thaliana* samples. The full image, Figure 1.1a, contains $12 \times 4 = 48$ sub-grids. Each sub-grid, contains $18 \times 18 = 324$ spots, which each has the resolution of 24×24 pixels. One of the sub-grids is shown in Figure 1.1b.

The aim of DNA microarray image processing and analysis is to find the positions of the spots and then identify the pixels that represent gene expression, separating them from the background and noise. The main steps involved in processing and analyzing a DNA microarray image are the following: spot addressing or gridding, segmentation, noise treatment and removal and background correction, which are discussed in more detail below.

When producing DNA microarrays, many parameters are specified, such as the number

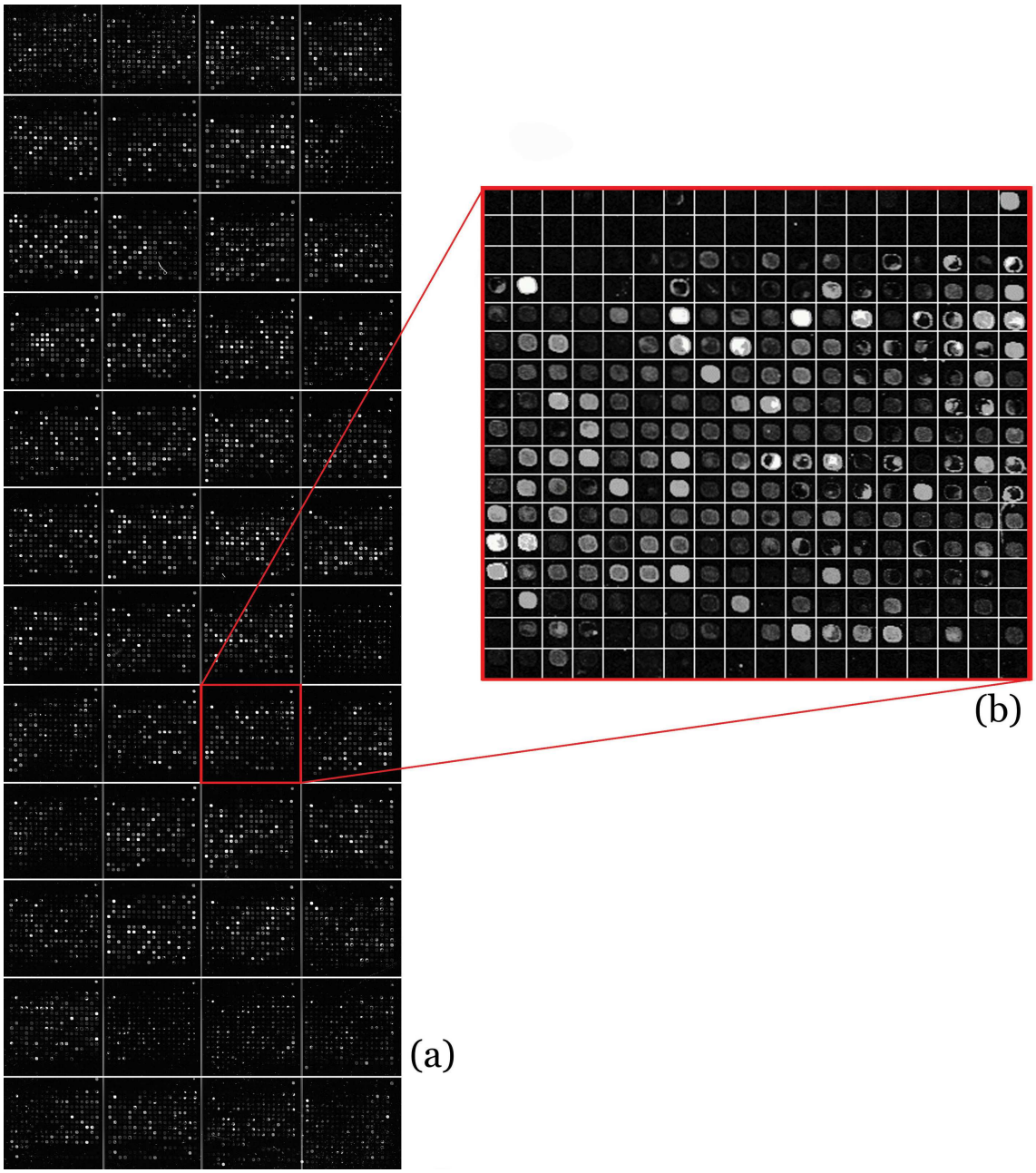

Figure 1.1: (a) Original DNA microarray image, 20391-ch1 (green channel), from the SMD; (b) sub-grid extracted from the 8^{th} column and 3^{rd} row.

and size of spots, number of sub-grids, and even their exact locations. However, many physicochemical factors produce noise, misalignment, and even deformations in the sub-grid template that it is virtually impossible to know the exact location of the spots after scanning, at least with the current technology, without performing complex procedures. Prior to applying the gridding process to find the locations of the spots, the sub-grids must be identified, a process that is also known as sub-gridding. Once the sub-grids are identified, the gridding step takes a sub-grid as input and aims to find the exact location of each spot. Depending on how complex the mechanisms are, the gridding method may or may not require some parameters about the sub-girds, namely the number of rows and columns of spots, the size of the spots in pixels, and others. Various methods have been proposed for solving this problem with some variations in terms of the amount of computer processing time, user intervention and parameters required [13–17].

1.3 ChIP-Seq Data Analysis

There are certain types of proteins that bind to some regions in DNA molecules, and these events are related to transcription and translation of RNA molecules into proteins. Protein-DNA binding has been studied using different biotechnological techniques such as ChIP-chip, ChIP-on-chip and ChIP-Seq [18–22]. They all use chromatin immunoprecipitation (ChIP), which precipitates a protein antigen out of the solution using a specific antibody designed to attach to that protein of interest. Of these, ChIP-Seq combines ChIP technology with high throughput, next generation sequencing, which allows one to investigate protein-DNA interactions more accurately. There are several advantages when using ChIP-Seq as an alternative technology to ChIP-chip, which combines the ChIP with microarray technology [23, 24]. Some of these are listed in Table 1.1, and include generating profiles

Table 1.1: Comparison of ChIP-Seq and ChIP-chip technology.

	ChIP-chip	**ChIP-Seq**
Resolution	30-100 bp	1 bp
Coverage	limited by sequence on the array	whole genome
Required amount of ChIP DNA	few micro grams	10-50 nano gram
cost	$400-$800 per array	$1000 per Illumina lane

with higher signal-to-noise ratios and a larger number of localized peaks. As observable from the table, ChIP-Seq has much higher resolution in comparison with ChIP-chip technology. Also, one of the main issues in ChIP-chip technology, which is noise pollution due to the hybridization step, does not exist in ChIP-Seq technology. Moreover, ChIP-Seq can cover the whole genome, whereas in ChIP-chip the coverage is limited to the amount of DNA attached to the array. Lastly, the amount of ChIP DNA required for performing the analysis is much higher in ChIP-chip technology in comparison with ChIP-Seq.

Figure 1.2 shows the work flow of ChIP-Seq data analysis. First, the DNA chromatin is sheared by sonication into small fragments (between 200-600 bp depends on the experiment). Then, using an antibody specific to the protein of interest, the DNA-protein complex is immunoprecipitated. Finally, after purifying DNA, the reads are sequenced and mapped to the reference genome. In the peak calling module, which is the step we focus on in this thesis, the locations in DNA that interact with certain proteins of interest are determined. After detecting those regions of interest, several analysis steps can be performed such as visualization, motif discovery, combining the results with gene expression data, and others.

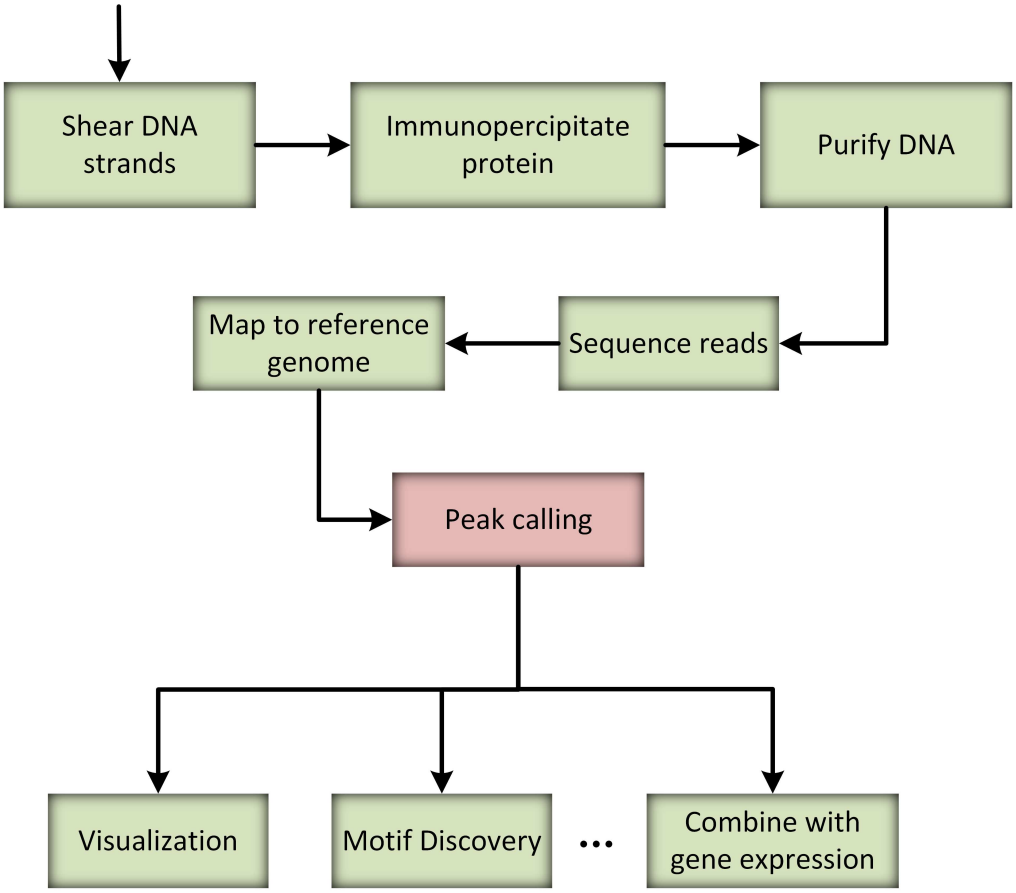

Figure 1.2: Diagrammatic view of the work flow of ChIP-Seq data analysis.

1.4 Finding Transcriptomics Biomarkers

Finding relevant transcriptome biomarkers corresponding to a certain disease is a key step toward efficient prediction and diagnosis of many diseases at early stages. Traditional gene selection approaches usually consider transcriptome of cancer cells for comparison to the patterns of normal cells in a cancer vs. non-cancer scenario for finding relevant transcriptome biomarkers. Here, we focus on a more challenging multi-class problem that consists of determining relevant and informative transcriptomics biomarkers in various subtypes of

a specific disease such as breast cancer.

While breast cancer is often thought of as a single disease, increasing evidence suggests that there are multiple subtypes of breast cancer that occur at different rates in different groups. They have their own specific treatment procedure, are more or less aggressive, and even have different survival rates. Having their own genetic and transcriptomics signatures makes the treatment procedure dramatically different from one subtype to another. The analysis in this case is more complicated, since each selected biomarker can be related to one or more classes with different possibilities or impact levels and it is essential to stratify patients into their relevant disease subtype prior to treatment. We address this problem by proposing a hierarchical method that finds an optimal subset of biomarkers for predicting a patient's subtype. It can be used for a wide range of diseases consisting a family of different subtypes with the ability of using different machine learning techniques to optimize the model based on the needs.

1.5 Motivation and Objectives

The first task in DNA microarray image processing is gridding, which, if done correctly, substantially improves the efficiency of the subsequent tasks that include segmentation, quantification, normalization and data mining [25]. Most of the proposed methods use one or more parameters to adjust their algorithms to the input image. Using more parameters can decrease the flexibility of the method, since these parameters are needed to be adjusted carefully based on the features of each microarray image before running the gridding algorithm. We introduced a parameterless and yet very powerful method for gridding microarray images that removes the burden of fine-tuning the parameters while providing a very high accuracy for finding the sub-grids within the microarray image as well as finding

the spots within each sub-grid.

As mentioned earlier, next generation sequencing offers higher resolution, less noise, and greater coverage in comparison with its microarray-based counterparts. Moreover, determining the interaction between a protein and DNA to regulate gene expression is a very important step toward understanding many biological processes and disease states. ChIP-Seq is one of the techniques used for finding regions of interest in a specific protein that interacts with DNA using next generation sequencing technology [26–32]. The growing popularity of ChIP-Seq has increased the need to develop new algorithms for peak finding. Due to mapping challenges and biases in various aspects of the existing protocols, identifying relevant peaks is not a straightforward task. One of the problems of the existing methods is that the locations of the detected peaks could be non-optimal. Moreover, for detecting these peaks all methods use a set of parameters that may cause variations of the results for different datasets. Thus, after some modifications, we proposed a new model for finding the interaction sites between a protein of interest and DNA using multi-level thresholding algorithm coupled with a model to find the best number of peaks based on clustering techniques for pattern recognition that addresses both of these issues.

Another downside of the existing methods is that they try to find all the enriched regions regardless of their length. These regions can be grouped by their length. For example, histone modification sites normally have a length of 50 to 60 kbp, while some other regions of interest like exons have a much smaller length of around 100 bp. Using these methods, there is no way to focus on regions with a specific length and all of the relevant peaks should be detected first. This is a time consuming task that forces the model to process all possible regions. We also proposed a modified version of multi-level thresholding to deal with this issue. Using the proposed method, we are able to search a specific region with a certain

length which consequently increases the accuracy and performance of the model.

On the other hand, as discussed in Section 1.4, another problem that is relevant in transcriptomics data analysis is finding the most informative genes associated with different subtypes of breast cancer, which is an important problem in breast cancer biomarker discovery. Finding relevant genes corresponding to each type of cancer is a key step toward efficient diagnosis and treatment of cancer. Machine learning approaches can be used to precisely determine the number of genes required to predict a patient subtype with a high degree of reliability. Moreover, modeling today's complex biological systems requires efficient computational models to extract the most valuable information from existing data. In this direction, pattern recognition techniques in machine learning provide a wealth of algorithms for feature extraction and selection, classification and clustering.

1.6 Contributions

The contributions of the thesis are based on using machine learning techniques for transcriptome data analysis. We propose various models and algorithms applicable on different transcriptome analysis technology. The main contributions of this thesis are summarized as follows:

- Proposing the optimal multi-level thresholding gridding (OMTG) method for finding sub-grids in microarray image and spots within each detected sub-grid. The proposed method is free of parameters (Chapter 2). We also proposed a new validity index (α) for finding the optimal number of sub-grids in microarray image and optimal number of spots within each sub-grid (Chapters 2 and 3). OMTG was originally proposed in 2011 for gridding microarray images. Since then, different articles have cited the

authors' proposed method for microarray image gridding. Table 1.2 shows the list of these publications.

- Adapting the proposed optimal multi-level thresholding model as a new framework (OMT) to find the interaction points between a protein of interest and DNA (Chapter 4) . Also, proposing a new high performance constrained based approach (CMT) used to find enriched regions in ChIP-Seq data (Chapter 5).

- Proposing a framework using *Chi2* feature selection [33] and a *support vector machine (SVM) classifier* [34] to obtain biologically meaningful genes, and to increase the accuracy for predicting breast tumor subtypes. The proposed model is flexible, in the sense that any feature selection and classifier can be embedded in it. The model can be used for prediction and diagnosis of various diseases with different subtypes (Chapter 6). We also discovered a new, compact set of biomarkers or genes useful for distinguishing among breast cancer types (Chapter 6).

1.7 Thesis Organization

The thesis is organized in seven chapters. Chapters 2 and 3 cover the topics related to the proposed optimal multilevel thresholding algorithm and its application in DNA microarray image analysis as follows:

Chapter 2: Luis Rueda, Iman Rezaeian: A Fully Automatic Gridding Method for cDNA Microarray Images. BMC Bioinformatics (2011) 12: 113.

Chapter 3: Luis Rueda, Iman Rezaeian: Applications of Multilevel Thresholding Algorithms to Transcriptomics Data. Progress in Pattern Recognition, Image Analysis,

Table 1.2: The list of papers that have cited the proposed method by the author.

Year	Title	Reference
2011	Automatic Spot Identification for High Throughput Microarray Analysis	[47]
2012	FPGA based system for automatic cDNA microarray image processing	[35]
2012	Denoising and block gridding of microarray image using mathematical morphology	[40]
2012	An improved automatic gridding based on mathematical morphology	[42]
2012	An improved automatic gridding method for cDNA microarray images	[43]
2013	Two dimensional barcode-inspired automatic analysis for arrayed microfluidic immunoassays	[48]
2013	A New Gridding Technique for High Density Microarray Images Using Intensity Projection Profile of Best Sub Image	[37]
2013	Recognition of cDNA micro-array image based on artificial neural network	[38]
2013	Using the Maximum Between-Class Variance for Automatic Gridding of cDNA Microarray Images	[41]
2013	An improved SVM method for cDNA microarray image segmentation	[44]
2013	A new method for gridding DNA microarrays	[36]
2014	gitter: A Robust and Accurate Method for Quantification of Colony Sizes from Plate Images	[46]
2014	Crossword: A fully automated algorithm for the image segmentation and quality control of protein microarrays	[39]
2014	An Effective Automated Method for the Detection of Grids in DNA Microarray	[45]

Computer Vision, and Applications - 16th Iberoamerican Congress (CIARP), Chile, 2011: 26-37.

Chapters 4 and 5 cover two proposed methods for analyzing ChIP-Seq data as follows:

Chapter 4: Iman Rezaeian, Luis Rueda: A new algorithm for finding enriched regions in ChIP-Seq data. ACM International Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB), Chicago, USA, 2012: 282-288.

Chapter 5: Iman Rezaeian, Luis Rueda: CMT: A Constrained Multi-Level Thresholding Approach for ChIP-Seq Data Analysis. PLoS ONE 9(4): e93873, 2014.

Similarly, a novel method for finding a subset of most informative genes to classify breast cancer subtypes is included in Chapter 6.

Chapter 6: Iman Rezaeian, Yifeng Li, Martin Crozier, Eran Andrechek, Alioune Ngom, Luis Rueda, Lisa Porter: Identifying Informative Genes for Prediction of Breast Cancer Subtypes. Pattern Recognition in Bioinformatics - 8th IAPR International Conference (PRIB), France, 2013: 138-148.

Finally, Chapter 7 concludes the thesis and identifies some problems arising from this work and relevant future work.

Bibliography

[1] Siddharth Arora, Jayadev Acharya, Amit Verma, and Prasanta K Panigrahi. Multilevel thresholding for image segmentation through a fast statistical recursive algorithm. *Pattern Recognition Letters*, 29(2):119–125, 2008.

[2] Hao Gao, Wenbo Xu, Jun Sun, and Yulan Tang. Multilevel thresholding for image segmentation through an improved quantum-behaved particle swarm algorithm. *Instrumentation and Measurement, IEEE Transactions on*, 59(4):934–946, 2010.

[3] Ming-Huwi Horng. Multilevel thresholding selection based on the artificial bee colony algorithm for image segmentation. *Expert Systems with Applications*, 38(11):13785–13791, 2011.

[4] Madhubanti Maitra and Amitava Chatterjee. A hybrid cooperative–comprehensive learning based pso algorithm for image segmentation using multilevel thresholding. *Expert Systems with Applications*, 34(2):1341–1350, 2008.

[5] Peng, Hanchuan and Long, Fulmi and Ding, Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[6] Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[7] Kingsford, Carl and Salzberg, Steven L. What are decision trees? *Nature*, 26(9):1011–1013, 2008.

[8] Ramoni, Marco, and Paola Sebastiani. Robust bayes classifiers. *Artificial Intelligence*, 125(1):209–226, 2001.

[9] Cortes, Corinna, and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[10] Ken Nozaki, Hisao Ishibuchi, and Hideo Tanaka. Adaptive fuzzy rule-based classification systems. *Fuzzy Systems, IEEE Transactions on*, 4(3):238–250, 1996.

[11] Hagan, M. T., Demuth, H. B. and Beale, M. H. Neural network design. *Boston: Pws Pub*, 1996.

[12] J. Hubble, J. Demeter, H. Jin, M. Mao, M. Nitzberg, T. Reddy, F. Wymore, Z. K. Zachariah, G. Sherlock, and C. A. Ball, "Implementation of genepattern within the stanford microarray database," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D898–D901, 2009.

[13] L. Rueda and I. Rezaeian, "A fully automatic gridding method for cdna microarray images," *BMC bioinformatics*, vol. 12, no. 1, p. 113, 2011.

[14] D. Bariamis, D. K. Iakovidis, and D. Maroulis, "M3g: maximum margin microarray gridding," *BMC bioinformatics*, vol. 11, no. 1, p. 49, 2010.

[15] G. Antoniol and M. Ceccarelli, "A markov random field approach to microarray image gridding," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 550–553, IEEE, 2004.

[16] M. Ceccarelli and G. Antoniol, "A deformable grid-matching approach for microarray images," *Image Processing, IEEE Transactions on*, vol. 15, no. 10, pp. 3178–3188, 2006.

[17] L. Rueda and V. Vidyadharan, "A hill-climbing approach for automatic gridding of cdna microarray images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 3, no. 1, p. 72, 2006.

[18] Michael J Buck and Jason D Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, 2004.

[19] W Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X Shirley Liu. Model-based analysis of tiling-arrays for chip-chip. *Proceedings of the National Academy of Sciences*, 103(33):12457–12462, 2006.

[20] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein–dna binding sites from chip-seq data. *Nucleic acids research*, 36(16):5221–5231, 2008.

[21] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[22] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide anal-

ysis of transcription factor binding sites based on chip-seq data. *Nature methods*, 5(9):829–834, 2008.

[23] Schones, Dustin E., and Keji Zhao. ”Genome-wide approaches to studying chromatin modifications.” Nature Reviews Genetics 9.3 (2008): 179–191.

[24] Ho, Joshua WK, et al. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC Genomics, 12.1(2011):134.

[25] L Qin, L Rueda, A Ali, and A Ngom: Spot Detection and Image Segmentation in DNA Microarray Data. *Applied Bioinformatics* 2005,4:1–12.

[26] Barski A, Zhao K (2009) Genomic location analysis by ChIP-Seq. Journal of Cellular Biochemistry 107: 11–18.

[27] Park P (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genetics 10: 669–680.

[28] Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics 13: 840–852.

[29] Micsinai M, Parisi F, Strino F, Asp P, Dynlacht BD, et al. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. Nucleic acids research 40: e70–e70.

[30] Jackman RW, Wu CL, Kandarian SC (2012) The ChIP-seq-Defined Networks of Bcl-3 Gene Binding Support Its Required Role in Skeletal Muscle Atrophy. PloS one 7: e51478.

[31] Auerbach RK, Chen B, Butte AJ (2013) Relating genes to function: identifying enriched transcription factors using the encode chip-seq significance tool. Bioinformatics 29: 1922–1924.

[32] Stower H (2013) DNA Replication: ChIP-seq for human replication origins. Nature Reviews Genetics 14: 78–78.

[33] Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. In: IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE Press, New York (1995)

[34] Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)

[35] Bogdan Belean, Monica Borda, Bertrand Le Gal, and Romulus Terebes. Fpga based system for automatic cdna microarray image processing. *Computerized Medical Imaging and Graphics*, 36(5):419–429, 2012.

[36] Christoforos C Charalambous and George K Matsopoulos. A new method for gridding dna microarrays. *Computers in biology and medicine*, 43(10):1303–1312, 2013.

[37] J Deepa and Tessamma Thomas. A new gridding technique for high density microarray images using intensity projection profile of best sub image. *Computer Engineering & Intelligent Systems*, 4(1), 2013.

[38] RM Farouk and EM Badr2and M SayedElahl. Recognition of cdna micro-array image based on artificial neural network. *Journal of Computer Vision Approaches Vol*, 1(1):1–12, 2013.

[39] Todd M Gierahn, Denis Loginov, and J Christopher Love. Crossword: A fully automated algorithm for the image segmentation and quality control of protein microarrays. *Journal of proteome research*, 2014.

[40] Nurnabilah Samsudin, Rathiah Hashim, and Noor Elaiza Abdul Khalid. Denoising and block gridding of microarray image using mathematical morphology. In *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on*, pages 230–235. IEEE, 2012.

[41] Gui-Fang Shao, Fan Yang, Qian Zhang, Qi-Feng Zhou, and Lin-Kai Luo. Using the maximum between-class variance for automatic gridding of cdna microarray images. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 10(1):181–192, 2013.

[42] Guifang Shao. An improved automatic gridding based on mathematical morphology. *Journal of Convergence Information Technology*, 7(1), 2012.

[43] Guifang Shao, Tingna Wang, Zhigang Chen, Yushu Huang, and Yuhua Wen. An improved automatic gridding method for cdna microarray images. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1615–1618. IEEE, 2012.

[44] Guifang Shao, Tingna Wang, Wupeng Hong, and Zhigang Chen. An improved svm method for cdna microarray image segmentation. In *Computer Science & Education (ICCSE), 2013 8th International Conference on*, pages 391–395. IEEE, 2013.

[45] PK Srimani and Shanthi Mahesh. An effective automated method for the detection of grids in dna microarray. In *ICT and Critical Infrastructure: Proceedings of the 48th*

Annual Convention of Computer Society of India-Vol II, pages 445–453. Springer, 2014.

[46] Omar Wagih and Leopold Parts. gitter: A robust and accurate method for quantification of colony sizes from plate images. *G3: Genes— Genomes— Genetics*, pages g3–113, 2014.

[47] Eunice Wu, Yan A Su, Eric Billings, Bernard R Brooks, and Xiongwu Wu. Automatic spot identification for high throughput microarray analysis. *Journal of bioengineering & biomedical science*, 2011.

[48] Yi Zhang, Lingbo Qiao, Yunke Ren, Xuwei Wang, Ming Gao, Yunfang Tang, Jianzhong Jeff Xi, Tzung-May Fu, and Xingyu Jiang. Two dimensional barcode-inspired automatic analysis for arrayed microfluidic immunoassays. *Biomicrofluidics*, 7(3):034110, 2013.

Chapter 2

A Fully Automatic Gridding Method for cDNA Microarray Images

2.1 Background

Microarrays are one of the most important technologies used in molecular biology to massively explore how the genes express themselves into proteins and other molecular machines responsible for the different functions in an organism. These expressions are monitored in cells and organisms under specific conditions, and have many applications in medical diagnosis, pharmacology, disease treatment, just to mention a few. We consider cDNA microarrays which are produced on a chip (slide) by hybridizing sample DNA on the slide, typically in two channels. Scanning the slides at a very high resolution produces images composed of sub-grids of spots. Image processing and analysis are two important aspects of microarrays, since the aim of the whole experimental procedure is to obtain meaningful biological conclusions, which depends on the accuracy of the different stages, mainly those at the beginning of the process. The first task in the sequence is gridding [1–5], which if

done correctly, substantially improves the efficiency of the subsequent tasks that include segmentation [6], quantification, normalization and data mining. When producing cDNA microarrays, many parameters are specified, such as the number and size of spots, number of sub-grids, and even their exact locations. However, many physicochemical factors produce noise, misalignment, and even deformations in the sub-grid template that it is virtually impossible to know the exact location of the spots after scanning, at least with the current technology, without performing complex procedures. Roughly speaking, gridding consists of determining the spot locations in a microarray image (typically, in a sub-grid). The gridding process requires the knowledge of the sub-girds in advance in order to proceed (sub-gridding).

Many approaches have been proposed for sub-gridding and spot detection. The Markov random field (MRF) is a well known approach that applies different constraints and heuristic criteria [1, 7]. Mathematical morphology is a technique used for analysis and processing geometric structures, based on set theory, topology, and random functions. It helps remove peaks and ridges from the topological surface of the images, and has been used for gridding the microarray images [8]. Jain's [9], Katzer's [10], and Stienfath's [11] models are integrated systems for microarray gridding and quantitative analysis. A method for detecting spot locations based on a Bayesian model has been recently proposed, and uses a deformable template to fit the grid of spots using a posterior probability model for which the parameters are learned by means of a simulated-annealing-based algorithm [1,3]. Another method for finding spot locations uses a hill-climbing approach to maximize the energy, seen as the intensities of the spots, which are fit to different probabilistic models [5]. Fitting the image to a mixture of Gaussians is another technique that has been applied to gridding microarray images by considering radial and perspective distortions [4]. A Radon-

transform-based method that separates the sub-grids in a cDNA microarray image has been proposed in [12]. That method applies the Radon transform to find possible rotations of the image and then finds the sub-grids by smoothing the row or column sums of pixel intensities; however, that method does not automatically find the correct number of sub-grids, and the process is subject to data-dependent parameters.

Another approach for cDNA microarray gridding is a gridding method that performs a series of steps including rotation detection and compares the row or column sums of the top-most and bottom-most parts of the image [13, 14]. This method, which detects rotation angles with respect to one of the axes, either x or y, has not been tested on images having regions with high noise (e.g., the bottom-most $\frac{1}{3}$ of the image is quite noisy).

Another method for gridding cDNA microarray images uses an evolutionary algorithm to separate sub-grids and detect the positions of the spots [15]. The approach is based on a genetic algorithm that discovers parallel and equidistant line segments, which constitute the grid structure. Thereafter, a refinement procedure is applied to further improve the existing grid structure, by slightly modifying the line segments.

Using maximum margin is another method for automatic gridding of cDNA microarray images based on maximizing the margin between rows and columns of spots [16]. Initially, a set of grid lines is placed on the image in order to separate each pair of consecutive rows and columns of the selected spots. Then, the optimal positions of the lines are obtained by maximizing the margin between these rows and columns using a maximum margin linear classifier. For this, a SVM-based gridding method was used in [17]. In that method, the positions of the spots on a cDNA microarray image are first detected using image analysis operations. A set of soft-margin linear SVM classifiers is used to find the optimal layout of the grid lines in the image. Each grid line corresponds to the separating line produced by

one of the SVM classifiers, which maximizes the margin between two consecutive rows or columns of spots.

2.2 Results and Discussion

For testing the proposed method (called Optimal Multi-level Thresholding Gridding or OMTG), three different kinds of cDNA microarray images have been used. The images have been selected from different sources, and have different scanning resolutions, in order to study the flexibility of the proposed method to detect sub-grids and spots with different sizes and features.

The first test suite consists of a set of images drawn from the Stanford Microarray Database (SMD), and corresponds to a study of the global transcriptional factors for hormone treatment of *Arabidopsis thaliana* samples. The images can be downloaded from smd.princeton.edu, by selecting "Hormone treatment" as category and "Transcription factors" as subcategory. Ten images were selected, which correspond to channels 1 and 2 for experiments IDs 20385, 20387, 20391, 20392 and 20395. The images have been named using AT (which stands for *Arabidopsis thaliana*), followed by the experiment ID and the channel number (1 or 2).

The second test suite consists of a set of images from Gene Expression Omnibus (GEO) and corresponds to an Atlantic salmon head kidney study. The images can be downloaded from ncbi.nlm.nih.gov, by selecting "GEO Datasets" as category and searching the name of the image. Eight images were selected, which correspond to channels 1 and 2 for experiments IDs GSM16101, GSM16389 and GSM16391, and also channel 1 of GSM15898 and channel 2 of GSM15898. The images have been named using GSM followed by the experiment ID, and the channel number (1 or 2).

The third test suite consists of two images, obtained from a dilution experiment (DILN) and correspond to channels experiments IDs Diln4-3.3942.01A and Diln4-3.3942.01B [18]. The specifications of the cDNA microarray images for each of these three test suites are summarized in Table 2.1.

Table 2.1: The specifications of the three datasets of cDNA microarray images used to evaluate the proposed method.

Suite Name	**SMD**	**GEO**	**DILN**
Database Name	Stanford Microarray Database	Gene Expression Omnibus	Dilution Experiment
Image Format	Tiff	Tiff	Tiff
No. of Images	10	8	2
Image Resolution	1910×5550	1900×5500	600×2300
Sub-grid Layout	12×4	12×4	5×2
Spot Layout	18×18	13×14	8×8
Spot Resolution	24×24	12×12	from 12×12 to 3×3

To assess the performance of the proposed method, we consider the percentage of the grid lines that separate sub-grids/spots incorrectly, marginally and perfectly. Each spot was evaluated as being perfectly, marginally or incorrectly gridded if the percentage of its pixels within the grid cell is 100%, between 80% and 100%, or less than 80% respectively [16]. These quantities were found by visually analyzing the result of the gridding produced by our method. For SMD and GEO, our gridding was not compared with the gridding currently available in these databases. For DILN, apart from the visual analysis, we also apply segmentation and quantification by computing the volume of log of intensity and relate these to the rate of dilution in the biological experiment. For the implementation, we used Matlab2010 on a Windows 7 platform and an Intel core i7 870 cpu with 8GB of memory. The average processing times for sub-grid and spot detections are shown in Table

2.2.

Table 2.2: Average processing times (in seconds) for detecting sub-grids within each cDNA microarray image and detecting spots within each detected sub-grid.

	Sub-grid Detection	Spot Detection
SMD	379.1	10.8
GEO	384.7	9.2
DILN	62.3	3.8

2.2.1 Sub-grid and Spot Detection Accuracy

Table 2.3 shows the results of applying the proposed method, OMTG, for spot detection on the SMD dataset. With the proposed method, spot locations can be detected very efficiently with an average accuracy of 98.06% for this dataset. The same sets of experiments were repeated for the GEO dataset and the results are shown in Table 2.4. Again, the spot locations are detected very efficiently with an average accuracy of 99.26%. The experiments were repeated for the DILN dataset and the results are shown in Table 2.5. Although the sizes of the spots in each sub-grid are different in this dataset, the spot locations are detected very efficiently with an average accuracy of 97.95%. In most of the images, the performance of the method is more than 98% and incorrectly and marginally aligned rates are less than 1%. Only in a few images with noticeable noise and defects, the accuracy of the method is less than 98%, while incorrectly aligned rates increase to more than 2%. This shows the flexibility and power of the proposed method. For all the images, in the sub-grid detection phase, the incorrect and marginal gridding rates are both 0%, yielding an accuracy of 100%. This means the proposed method works perfectly in sub-grid detection for this case.

One of the reasons for the lower accuracy in spot detection is that the distance between spots is smaller than the distance between sub-grids. In all three datasets, there are approxi-

Table 2.3: Accuracy of detected sub-grids and spots for each image of the SMD dataset and the corresponding incorrectly, marginally and perfectly aligned rates.

	Sub-grid Detection			Spot Detection		
Image	Incorrectly	Marginally	Perfectly	Incorrectly	Marginally	Perfectly
AT-20385-CH1	0.0%	0.0%	100%	4.30%	0.46%	95.24%
AT-20385-CH2	0.0%	0.0%	100%	2.83%	0.09%	97.08%
AT-20387-CH1	0.0%	0.0%	100%	2.90%	0.14%	96.96%
AT-20387-CH2	0.0%	0.0%	100%	0.52%	0.11%	99.37%
AT-20391-CH1	0.0%	0.0%	100%	0.64%	0.17%	99.19%
AT-20391-CH2	0.0%	0.0%	100%	0.32%	0.26%	99.42%
AT-20392-CH1	0.0%	0.0%	100%	4.10%	0.33%	95.57%
AT-20392-CH2	0.0%	0.0%	100%	0.21%	0.25%	99.54%
AT-20395-CH1	0.0%	0.0%	100%	0.41%	0.12%	99.47%
AT-20395-CH2	0.0%	0.0%	100%	0.98%	0.31%	98.71%

Table 2.4: Accuracy of detected sub-grids and spots for each image of the GEO dataset and the corresponding incorrectly, marginally and perfectly aligned rates.

	Sub-grid Detection			Spot Detection		
Image	Incorrectly	Marginally	Perfectly	Incorrectly	Marginally	Perfectly
GSM15898-CH1	0.0%	0.0%	100%	0.58%	0.16%	99.26%
GSM15899-CH2	0.0%	0.0%	100%	1.00%	0.21%	98.79%
GSM16101-CH1	0.0%	0.0%	100%	0.00%	0.32%	99.68%
GSM16101-CH2	0.0%	0.0%	100%	1.57%	0.06%	98.37%
GSM16389-CH1	0.0%	0.0%	100%	0.79%	0.12%	99.09%
GSM16389-CH2	0.0%	0.0%	100%	0.57%	0.04%	99.39%
GSM16391-CH1	0.0%	0.0%	100%	0.00%	0.24%	99.76%
GSM16391-CH2	0.0%	0.0%	100%	0.14%	0.13%	99.73%

Table 2.5: Accuracy of detected sub-grids and spots for each image of the DILN dataset and the corresponding incorrectly, marginally and perfectly aligned rates.

	Sub-grid Detection			Spot Detection		
Image	Incorrectly	Marginally	Perfectly	Incorrectly	Marginally	Perfectly
Diln4-3.3942.01A	0.0%	0.0%	100%	2.23%	0.05%	97.72%
Diln4-3.3942.01B	0.0%	0.0%	100%	1.71%	0.11%	98.18%

mately eight pixels between spots, and approximately 30 pixels horizontally and 100 pixels vertically between sub-grids in the SMD dataset, 200 pixels in the GEO dataset and 25 pixels horizontally, and 200 pixels vertically in the DILN dataset. Another possible reason for this behavior is that the number of pixels in each sub-grid is far lower than that of a microarray image (around $1/50$). Thus, the noise present in the image affects the spot detection phase much more than the sub-grid extraction stage. It is important to highlight, however, that because of the relatively large distance between sub-grids, the detection process is not affected by the presence of noise.

Additionally, to evaluate the effectiveness of the refinement procedure, we tested the accuracy of the proposed method with and without applying the refinement procedure. The results are shown in Table 2.6. For simplicity, we only include those images in which there is a change in accuracy. We observe that applying the refinement procedure slightly improve the efficiency of the method in all the images in the table.

Table 2.6: The accuracy of the proposed method with and without using the refinement procedure in the spot detection phase. Only images with changes in accuracy are listed.

	Without Refinement Procedure			With Refinement Procedure		
Image	Incorrectly	Marginally	Perfectly	Incorrectly	Marginally	Perfectly
AT-20385-CH1	4.73%	0.79%	94.48%	4.30%	0.46%	95.24%
AT-20387-CH2	0.93%	0.54%	98.53%	0.52%	0.11%	99.37%
AT-20391-CH2	0.71%	0.58%	98.71%	0.32%	0.26%	99.42%
AT-20395-CH2	1.37%	0.76%	97.87%	0.98%	0.31%	98.71%
GSM16101-CH2	2.13%	0.21%	97.66%	1.57%	0.06%	98.37%
GSM16389-CH1	0.93%	0.19%	98.88%	0.79%	0.12%	99.09%
GSM16391-CH2	0.47%	0.26%	99.27%	0.14%	0.13%	99.73%

To analyze the results from a different perspective, we have also performed a visual analysis. Figure 2.1 shows the detected sub-grids in the AT-20387-ch2 image (left) and the detected spots in one of the sub-grids (right). Also, Figure 2.2 shows the sub-grids detected

in the GSM16101-ch1 image (left) and the detected spots in one of the sub-grids (right), while Figure 2.3 shows the sub-grids detected in the Diln4-3.3942B image (left) and the detected spots in one of the sub-grids (right). As shown in the all three figures, the proposed method finely detects the sub-grid locations first, and in the next stage, each sub-grid is divided precisely into the corresponding spots with the same method. The robustness of OMTG is so high that spots in sub-grids can be detected very well even in noisy conditions, such as those observable in the selected sub-grid in Figure 2.1. The ability to detect sub-grids and spots in different microarray images with different resolutions and spacing is another important feature of the proposed method.

As mentioned earlier, deformations, noise and artifacts can affect the accuracy of the proposed method. Figure 2.4 shows an example in which the proposed method fails to detect some spot regions due to the extremely contaminated regions with noise and artifacts. In this particular sub-grid, noisy regions tend to be confused with spots. Also, most spots have low intensities that are confused with the background. After testing other methods on this image, we observed that they also fail to detect the correct gridding in these regions.

To further analyze the efficiency of the proposed method to automatically detect the correct number of spots and sub-grids, we show in Figures 2.5, 2.6 and 2.7 the plots for the indices of validity against the number of sub-grids for AT-20387-ch2 , GSM16101-ch1 and Diln4-3.3942B respectively. The plots on top of the figures represent the values of the index functions (y axis) for detecting the horizontal lines for the I, A and α indices respectively, while the plots of the indices for the vertical separating lines are shown at the bottom of the figures. We observe that it would be rather difficult to find the correct number of sub-grids using the I index or the A index, while the α index clearly reveals the correct number of horizontal and vertical sub-grids by producing an almost flat curve with

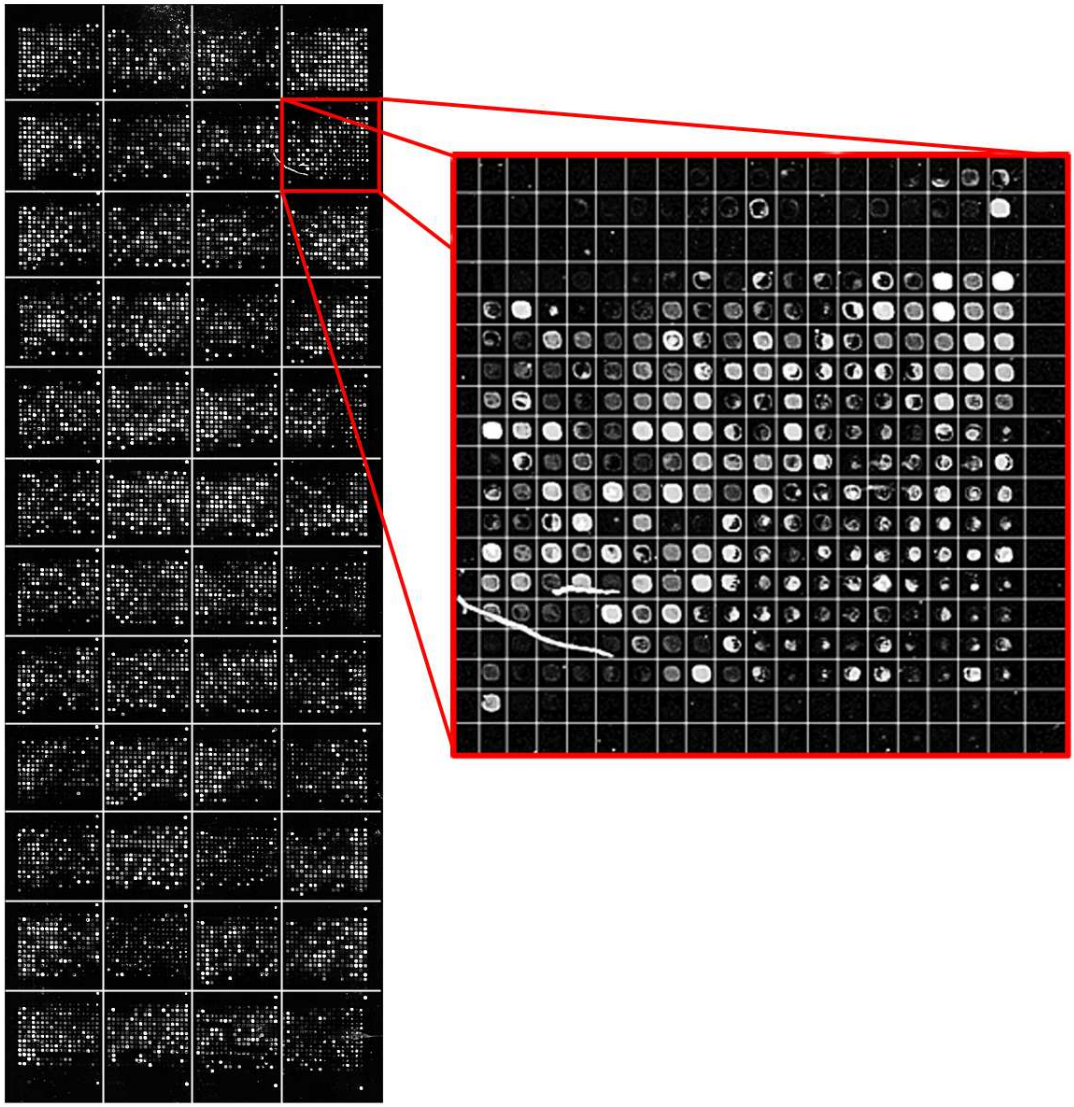

Figure 2.1: Sub-grid and spot detection in one of the SMD dataset images. Detected sub-grids in AT-20387-ch2 (left), and detected spots in one of the sub-grids (right).

pronounced peaks at 4 and 12 respectively for SMD and GEO images, and pronounced peaks at 2 and 5 respectively for DILN images. For example, it is clearly observable at the bottom plots in Figures 2.5 and 2.6 that the *I* index misses the correct number of sub-grids,

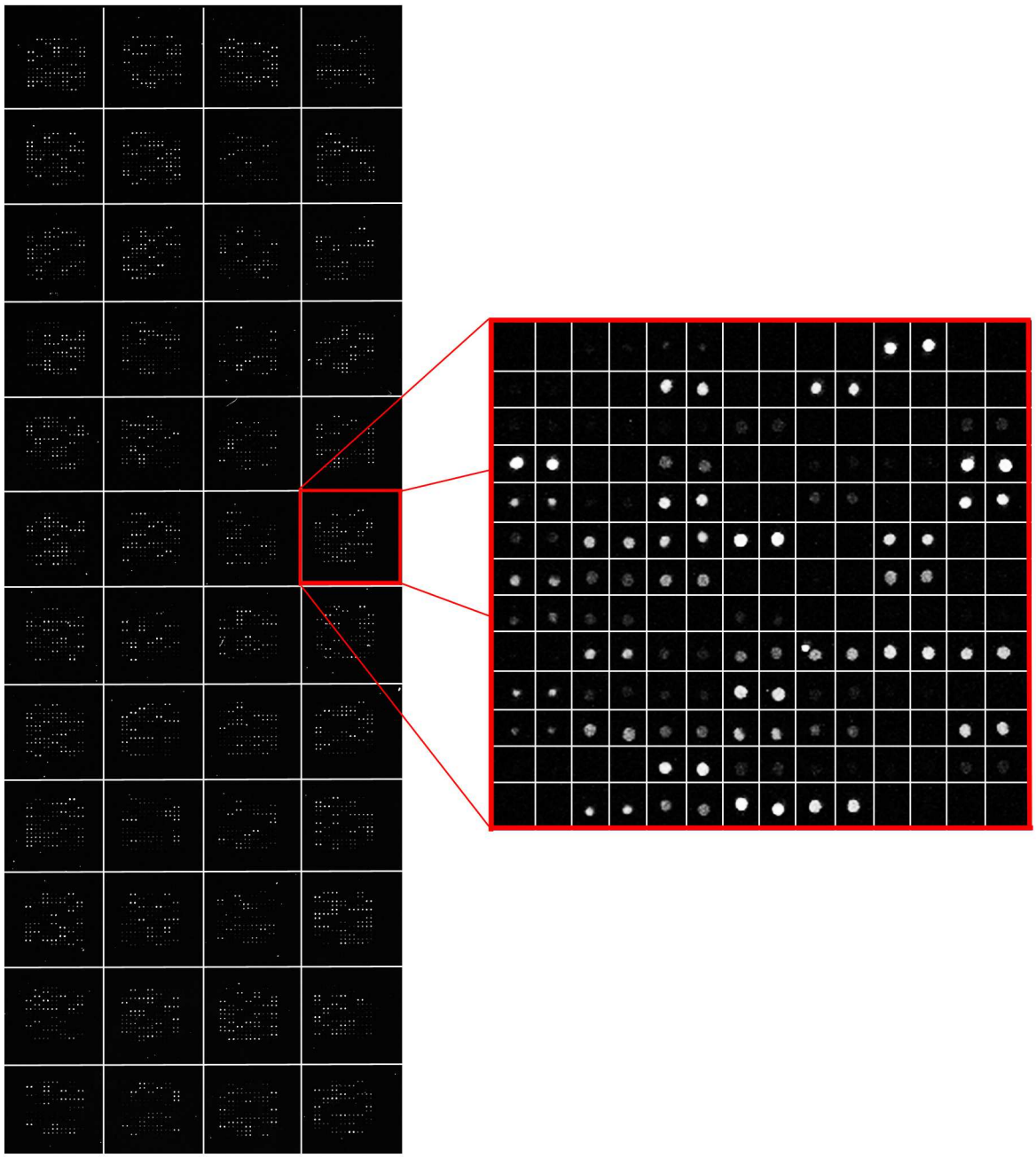

Figure 2.2: Sub-grid and spot detection in one of the GEO dataset images. Detected sub-grids in GSM16101-ch1 (left), and detected spots in one of the sub-grids (right).

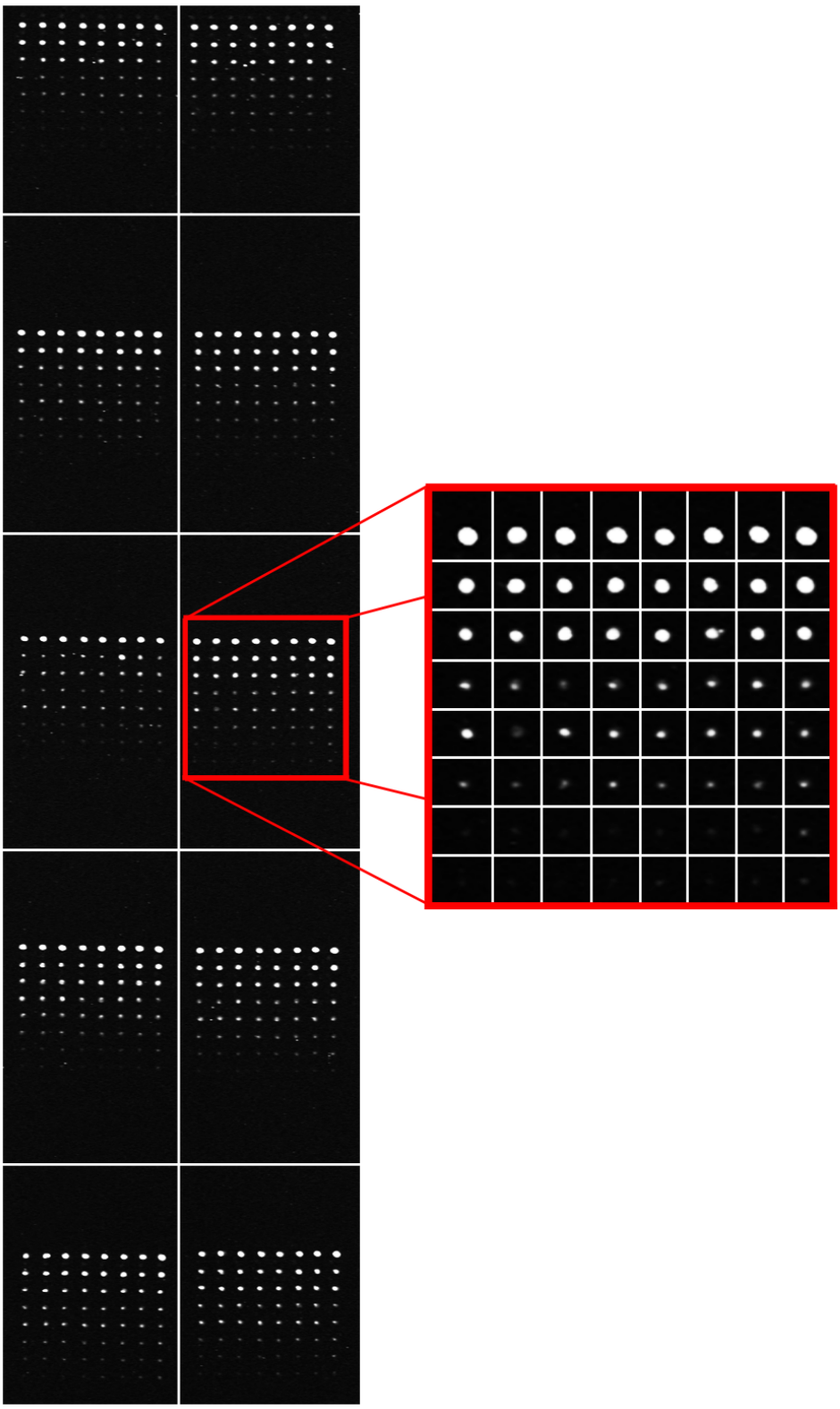

Figure 2.3: Sub-grid and spot detection in one of the DILN dataset images. Detected sub-grids in Diln4-3.3942B (left) and detected spots in one of the sub-grids (right).

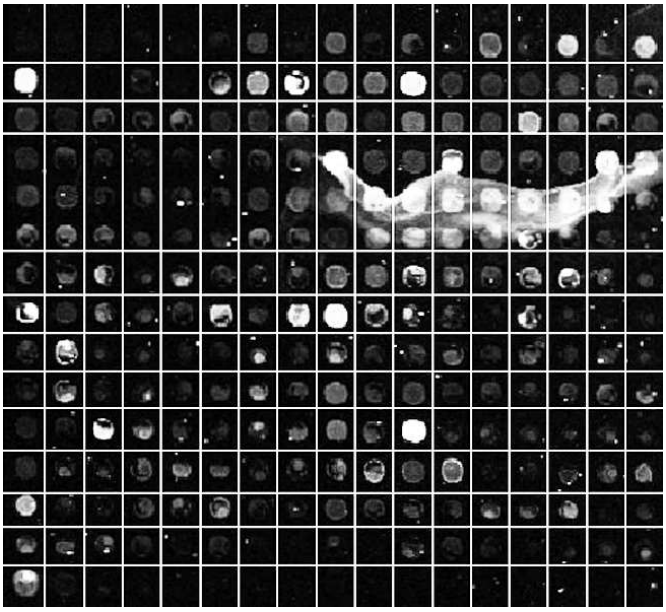

Figure 2.4: Failure to detect some spot regions due to the extremely contaminated images with artifacts in the sub-grid located in the first row and fourth column of AT-20392-ch1 from the SMD dataset.

12, by showing a higher peak at 13, while the α index finds the correct number of vertical sub-grids accurately.

2.2.2 Rotation Adjustment Accuracy

To test the effect of the Radon transform we rotate two of the images 5,10,15,20 and 25 degrees in both clockwise and counter-clockwise directions. Figure 2.8 shows the images rotated by -20, -10, 10 and 20 degrees (left) and the result of the adjustment after applying the Radon transform (right). Also, Table 2.7 shows the accuracy of the proposed method on two of the rotated images. In all cases, the adjustment method works accurately and corrects the rotations in both directions. Moreover, as shown in Table 2.7, the accuracy of the method remains nearly constant for all cases regardless of the degree of rotation.

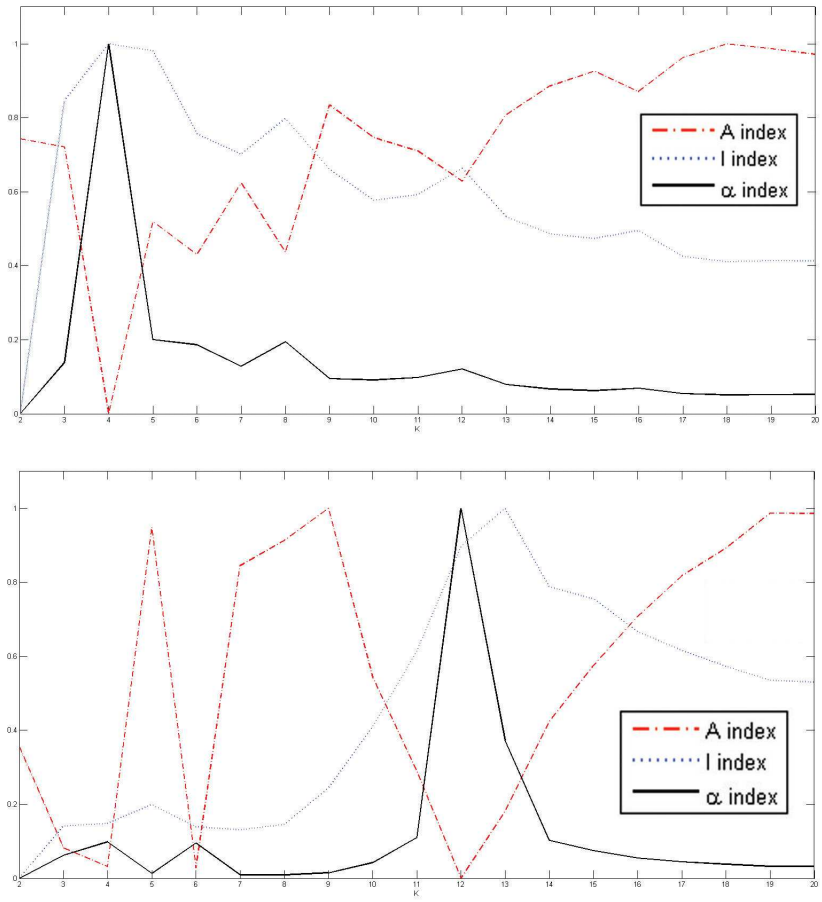

Figure 2.5: Plots of the index functions for AT-20387-ch2: (top) the values of the *I*, *A* and α indices for horizontal separating lines, and (bottom) the values of the *I*, *A* and α indices for vertical separating lines.

2.2.3 Comparison with other methods

A conceptual comparison between the proposed method, OMTG, and other microarray image griding methods based on their features is shown in Table 2.8. The methods included in the comparison are the following: (i) Radon transform sub-gridding (RTSG) [12], (ii) Bayesian simulated annealing gridding (BSAG) [3], (iii) genetic-algorithm-based gridding

(GABG) [15], (iv) hill-climbing gridding (HCG) [5], (v) maximum margin microarray gridding (M^3G) [16], and the proposed method, OMTG. As shown in the table, as opposed to other methods, OMTG does not need any number-based parameter, and hence making it much more powerful than the previous ones. One could argue, however, that the index or thresholding criterion can be considered as a "parameter". We have "fixed" these two on the α *index* and the *between class* criterion, and experimentally shown the efficiency of OMTG

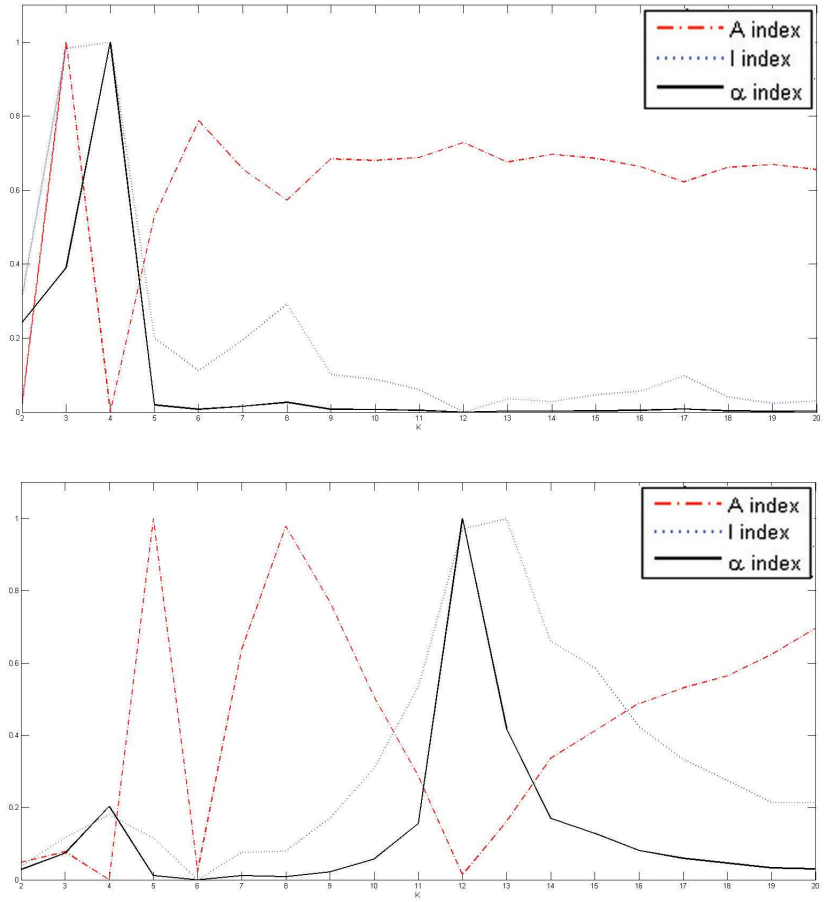

Figure 2.6: Plots of the index functions for the GSM16101-ch1: (top) the values of the *I*, *A* and α indices for horizontal separating lines, and (bottom) the values of the *I*, *A* and α indices for vertical separating lines.

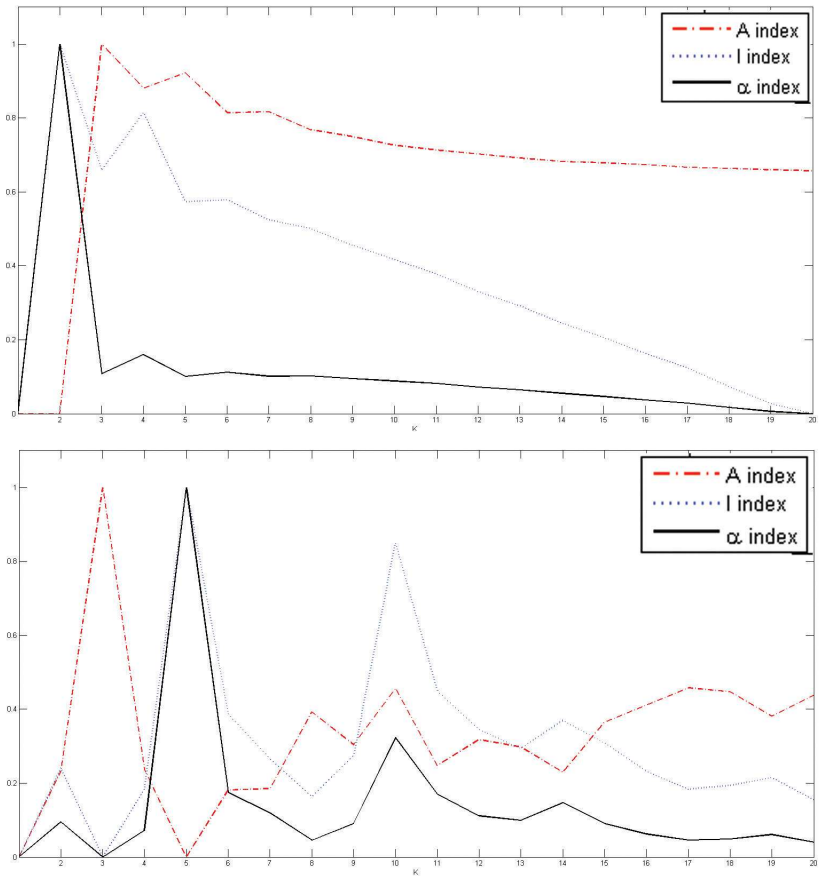

Figure 2.7: Plots of the index functions for the Diln4-3.3942B: (top) the values of the *I*, *A* and α indices for horizontal separating lines, and (bottom) the values of the *I*, *A* and α indices for vertical separating lines.

on various cDNA microarray images with different configurations.

An experimental comparison of the proposed method with GABG and HCG is shown in Table 2.9. As opposed to the proposed method that needs no parameters, GABG needs to set several parameters such as the mutation rate, μ, the crossover rate, c, the maximum threshold probability, p_{max}, the minimum threshold probability, p_{low}, the percentage of lines with low probability to be a part of the grid, f_{max} and the refinement threshold, T_p. Also,

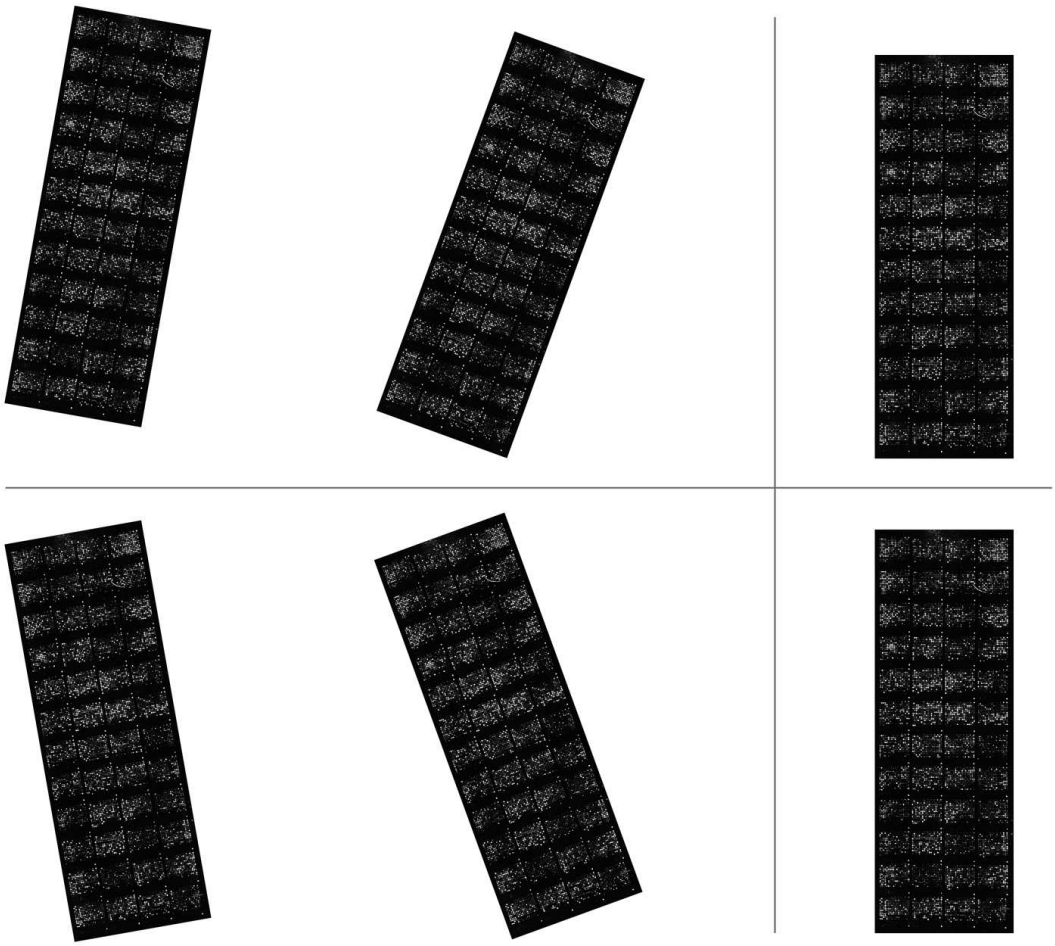

Figure 2.8: Rotation adjustment of AT-20387-ch2. Four different rotations from -20 to 20 degrees with steps of 10 degrees (left), and the adjusted image after applying the Radon transform (right).

HCG needs to set some parameters such as λ and σ. As shown in the table, the accuracy of our method is much higher than GABG and HCG. Since GABG and HCG use several parameters, to obtain good results for the SMD, GEO and DILN datasets, all the parameters must be set manually and separately for each dataset. If the same parameters for one of datasets were used for the others, unpredictable and poor results would be obtained – the

Table 2.7: Accuracy of detected spots for different rotations of AT-20395-CH1 and GSM16391-CH2, and the corresponding incorrectly, marginally and perfectly aligned rates.

	AT-20395-CH1			GSM16391-CH2		
Rotation	Incorrectly	Marginally	Perfectly	Incorrectly	Marginally	Perfectly
none	0.41%	0.12%	99.47%	0.14%	0.13%	99.73%
5°	0.41%	0.12%	99.47%	0.14%	0.13%	99.73%
10°	0.43%	0.12%	99.45%	0.15%	0.14%	99.71%
15°	0.41%	0.13%	99.46%	0.14%	0.13%	99.73%
20°	0.42%	0.13%	99.45%	0.15%	0.14%	99.71%
25°	0.42%	0.15%	99.43%	0.14%	0.15%	99.71%
-5°	0.41%	0.12%	99.47%	0.14%	0.13%	99.73%
-10°	0.41%	0.12%	99.47%	0.14%	0.13%	99.73%
-15°	0.42%	0.13%	99.45%	0.14%	0.14%	99.72%
-20°	0.42%	0.14%	99.44%	0.15%	0.13%	99.72%
-25°	0.42%	0.16%	99.42%	0.14%	0.15%	99.71%

accuracy of both methods could decrease to as low as 50%. This makes these methods fully dependent on the parameters, which have to be set manually and for specific datasets. The proposed method, however, does not need any parameter at all, and works exceptionally well in different kinds of images with different resolutions and noisy conditions.

2.2.4 Biological Analysis

In order to assess the proposed method on its suitability to perform in accordance with the biological problem, we analyze the quantification results and their relationships with the dilution experiment on the DILN dataset. To compute the volume intensity of each spot, first, we use *Sobel* method to detect the edge of each spot and then the region within the edge is defined as the primary region of each spot. The Sobel method finds edges of the spot using the Sobel approximation to the derivative and returns edges at those points where the gradient of image is maximum. In the next step, a set of morphological dilation and erosion operators are used to decrease the noise and artifacts in the region identified

Table 2.8: Conceptual comparison of our proposed method with other recently proposed methods based on the required number and type of input parameters and features.

Method	Parameters	Sub-grid Detection	Spot Detection	Automatic Detection No. of Spots	Rotation
RTSG	n: Number of sub-grids	$\surd$	$\times$	$\times$	$\surd$
BSAG	α ,β: Parameters for balancing prior and posterior probability rates	$\times$	$\surd$	$\surd$	$\surd$
GABG	μ , c :Mutation and Crossover rate, p_{max}: probability of maximum threshold, p_{low}: probability of minimum threshold, f_{max} : percentage of line with low probability to be a part of grid, T_p: Refinement threshold	$\surd$	$\surd$	$\surd$	$\surd$
HCG	λ , σ: Distribution parameters	$\times$	$\surd$	$\surd$	$\times$
M^3G	c: Cost parameter	$\times$	$\surd$	$\surd$	$\surd$
OMTG	None	$\surd$	$\surd$	$\surd$	$\surd$

Table 2.9: The results of the comparison between the proposed method (OMTG) and the GABG and HCG methods proposed in [5] and [15] respectively.

Dataset	Method	Incorrectly	Marginally	Perfectly
SMD	OMTG	1.72%	0.22%	98.06%
	GABG	5.37%	0.51%	94.12%
	HCG	2.12%	1.23%	96.65%
GEO	OMTG	0.58%	0.16%	99.26%
	GABG	4.49%	0.32%	95.19%
	HCG	2.55%	0.74%	96.71%
DILN	OMTG	1.97%	0.08%	97.95%
	GABG	4.35%	0.34%	95.31%
	HCG	3.78%	0.65%	95.57%

for each spot. Finally, the summation of all pixel intensities in the spot are used as the level of expression of the gene associated with that spot; this summation represents the *volume* of the spot. Table 2.10 shows the volume intensity of each dilution step for images A and B respectively. As shown in the table, the proposed method estimates the average intensities of dilution steps very well with near linear decreasing steps. Also, Figure 2.9 shows log-plots of the dilution steps for all 80 cases and the mean of them with a red line. The reference line with slope -1 is also shown in black. As shown in this figure, in most parts of the dilution experiment, the estimated intensities of each case follow a linear relationship. In step 4 of the dilution steps, there is an irregularity in the linearity of the red curve as shown in Table 2.10 and Figure 2.9. The reason for this irregularity is that, in some sub-grids of Diln4-3.3942.01A and Diln4-3.3942.01B, the intensities of the spots in step 4 are smaller than those of step 5. One example of this can be seen in the third and last rows of the sub-grids in Figure 2.10. As shown in Figure 2.10(b), this decrease in the intensity of the spots causes a slight nonlinearity in step 4 of the dilution steps. In general, we observe that the proposed method is able to capture the nonlinear relationships present in the dilution experiments. This is observable in the log-plots of Figure 2.9, as the black line follows the array of logs of spot volumes.

Table 2.10: Logs of volume intensities of each dilution step for images A and B from the DILN dataset.

Dilution steps	Diln4-3.3942.01A	Diln4-3.3942.01B
1	22.02	21.75
2	20.63	20.78
3	19.75	19.94
4	18.12	18.05
5	17.98	18.25
6	16.98	17.03
7	16.18	16.17
8	15.07	15.46

2.3 Conclusions

A new method for separating sub-grids and spot centers in cDNA microarray images is proposed. The method performs four main steps involving the Radon transform for detecting rotations with respect to the x and y axes, the use of polynomial-time optimal multilevel thresholding to find the correct positions of the lines separating sub-grids and spots, a new index for detecting the correct number of sub-grids and spots and, finally, a refinement procedure to increase the accuracy of the detection.

The proposed method has been tested on real-life, high-resolution microarray images drawn from three sources, the SMD, GEO and DILN. The results show that (i) the rotations are effectively detected and corrected by affine transformations, (ii) the sub-grids are accurately detected in all cases, even in abnormal conditions such as noisy areas present in the images, (iii) the spots in each sub-grid are accurately detected using the same method, (iv) using the refinement procedure increases the accuracy of the method, and (v) because of using an algorithm free of parameters, this method can be used for different microarray images in various situations, and also for images with various spot sizes and configurations effectively. The results have also been biologically validated on dilution experiments.

2.4 Methods

A cDNA microarray image typically contains a number of sub-grids, and each sub-grid contains a number of spots arranged in rows and columns. The aim is to perform a two-stage process in such a way that the sub-grid locations are found in the first stage, and then spots locations within a sub-grid can be found in the second stage. Consider an image (matrix) $A = \{a_{i,j}\}, i = 1,....,n$ and $j = 1,....,m$, where $a_{ij} \in \mathbb{Z}^+$, and A is a sub-grid of a

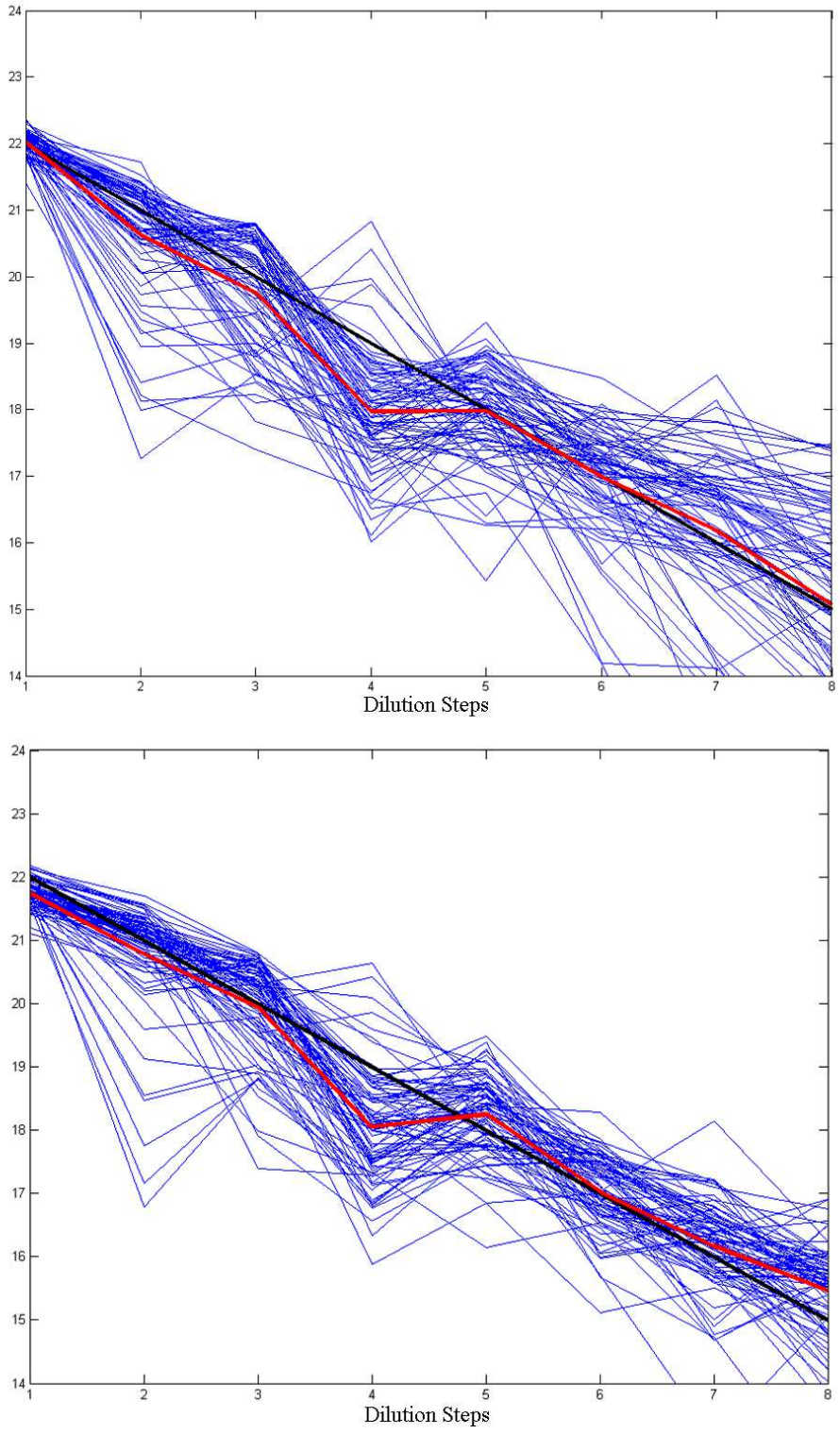

Figure 2.9: The logs of spot volumes that correspond to the dilution steps in Diln4-3.3942.01A (top) and Diln4-3.3942.01B (bottom). The red lines show the average of logs of spot volumes in different dilution steps. The black line corresponds to the reference line with slope equal to -1.

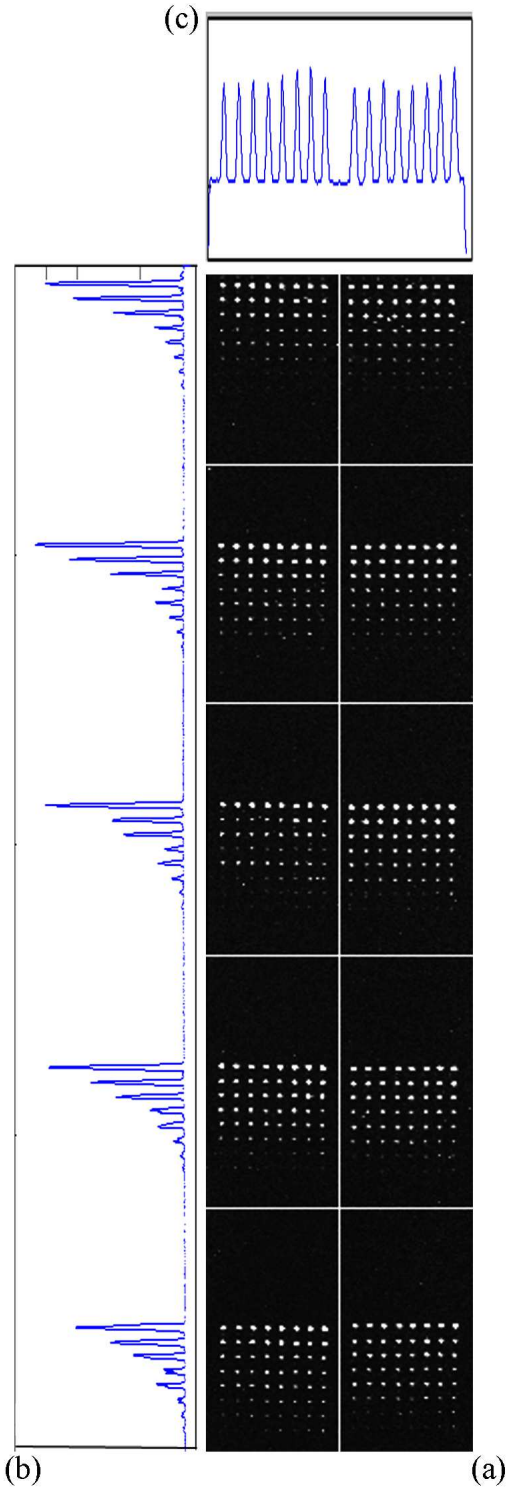

Figure 2.10: Detected sub-grids and the corresponding horizontal and vertical histogram. (a) detected sub-grids in Diln4-3.3942.01A, (b) vertical histogram (c) horizontal histogram.

cDNA microarray image. The method is first applied to a microarray image that contains a template of rows and columns of sub-grids (usually, a_{ij} is in the range [0..65,535] in a TIFF image). The aim of the first stage, sub-gridding, is to obtain vectors, $\mathbf{h} = [h_1, ...h_{p-1}]^t$ and $\mathbf{v} = [v_1, ...v_{q-1}]^t$, where $v_i \in [1, m]$, $h_j \in [1, n]$ and p and q are the number of horizontal and vertical sub-grids respectively. These horizontal and vertical vectors are used to separate the sub-grids.

Ones the sub-grids are obtained, the gridding process, namely finding the locations of the spots in a sub-grid, can be defined analogously. The rectangular area between two adjacent horizontal vectors h_j and h_{j+1}, and two adjacent vertical vectors v_i and v_{i+1} delimit the area corresponding to a spot (spot region). The aim of gridding is to find the corresponding spot locations given by the horizontal and vertical adjacent vectors. Post-processing or refinement allows us to find a spot region for each spot, which is enclosed by four lines.

To perform the gridding procedure our method may not need to know the number of sub-grids or spots. Although in many cases, based on the layout of the printer pins, the number of sub-grids or spots are known, due to misalignments, deformations, artifacts or noise during producing the microarray images, these numbers may not be accurate or unavailable. On the other hand, the optimal multi-level thresholding method needs the number of thresholds (sub-grids or spots) to be specified. Thus, we use an iterative approach to find the gridding for every possible number of thresholds, and then evaluate it with the proposed α index to find the best number of thresholds.

The sub-grids in a microarray image are detected by applying the Radon transform as a pre-processing phase and then using optimal multilevel thresholding in the next stage. By combining the optimal multilevel thresholding method and the α index (2.12) , the correct number of thresholds (sub-grids) can be found. Figure 2.11 depicts the process of finding

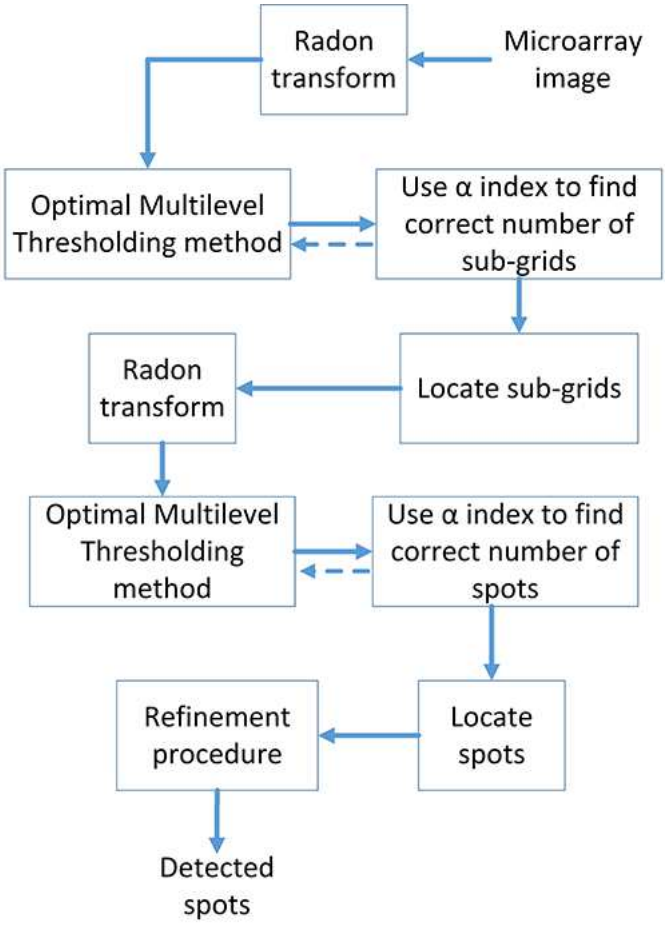

Figure 2.11: Schematic representation of the process for finding sub-grids (spots) in a cDNA microarray image.

the sub-grids in a microarray image and the spots in a sub-grid. The input to the Radon transform is a cDNA microarray image and the output of the whole process is the location (and partitioning) of the sub-grids. Analogously, the locations of the spots in each sub-grid are found by using optimal multilevel thresholding combined with the proposed α index to find the best number of rows and columns of spots. The input for this process is a sub-grid (already extracted from the sub-gridding step) and the output is the partitioning of the sub-grids into spots (spot regions).

2.4.1 Rotation Adjustment

Rotations of the images are seen in two different directions, with respect to the x and y axes. To find two independent angles of rotation for an affine transformation, the Radon transform is applied. Given an image $A = \{a_{x,y}\}$, the Radon transform performs the following transformation:

$$R(p,t) = \int_{-\infty}^{\infty} a_{x,t+px} dx\,, \tag{2.1}$$

where p is the slope and t its intercept. The rotation angle of the image with respect to the slope p is given by $\phi = \arctan p$. For the sake of the notation, $R(\phi,t)$ is used to denote the Radon transform of image A. Each rotation angle ϕ gives a different one-dimensional function, and the aim is to obtain the angle that gives the best alignment with the lines. This will occur when the lines are *parallel* to the y-axis. The best alignment will occur at the angle ϕ_{min} that minimizes the *entropy* as follows [1]:

$$H(\phi) = -\sum_{t=-\infty}^{\infty} R'(\phi,t) \log R'(\phi,t) dt\,. \tag{2.2}$$

$R(\phi,t)$ is normalized into $R'(\phi,t)$, such that $\sum_t R'(\phi,t) = 1$. The positions of the pixels in the new image, $[uv]$, are obtained as follows:

$$[u \quad v] \;=\; [x \quad y] \begin{bmatrix} \cos\phi_{\min_x} & \sin\phi_{\min_y} \\ -\sin\phi_{\min_y} & \cos\phi_{\min_x} \end{bmatrix}, \tag{2.3}$$

where $\phi_{\min_x}$ and $\phi_{\min_y}$ are the best angles of rotation found by the Radon transform.

2.4.2 Optimal Multilevel Thresholding

Image thresholding is one of the most widely-used techniques that has many applications in image processing, including segmentation, classification and object recognition. Given a sub-grid, we compute the row or column sums of pixel intensities, obtaining a discrete one dimensional function, where the domain is given by the positions of the rows/columns of pixels. In this work, that function is considered as a histogram or projection in which each bin represents one column (or row respectively), and the row or column sum of intensities corresponds to the frequency of that bin. We use the terms "histogram" or "sum" indistinctly. The frequencies are then normalized in order to be considered as probabilities of the corresponding bins. Figure 2.12 depicts a typical cDNA microarray image (AT-20387-ch2) that contains 12×4 sub-grids, along with the corresponding row or column sums. Also, Figure 2.13 depicts one of its sub-grids along with the corresponding row and column sums. Each row or column sum is then processed (see below) to obtain the optimal thresholding that will determine the locations of the sub-grids (spots).

Although various parametric and non-parametric thresholding methods and criteria have been proposed, the three most important streams are Otsu's method, which aims to maximize the separability of the classes measured by means of the sum of between-class variances [19], the one that uses information theoretic measures in order to maximize the separability of the classes [20], and the minimum error criterion [21]. In this work, we use the between-class variance criterion [19].

Consider a histogram H, an ordered set $\{1,2,\ldots,n-1,n\}$, where the ith value corresponds to the ith bin and has a probability, p_i. Given an image, $A = \{a_{i,j}\}$, as discussed earlier, H can be obtained by means of the horizontal (vertical) sum as follows: $p_i = \sum_{j=1}^{m} a_{i,j}$ ($p_j = \sum_{i=1}^{n} a_{i,j}$). We also consider a threshold set T, defined as an ordered

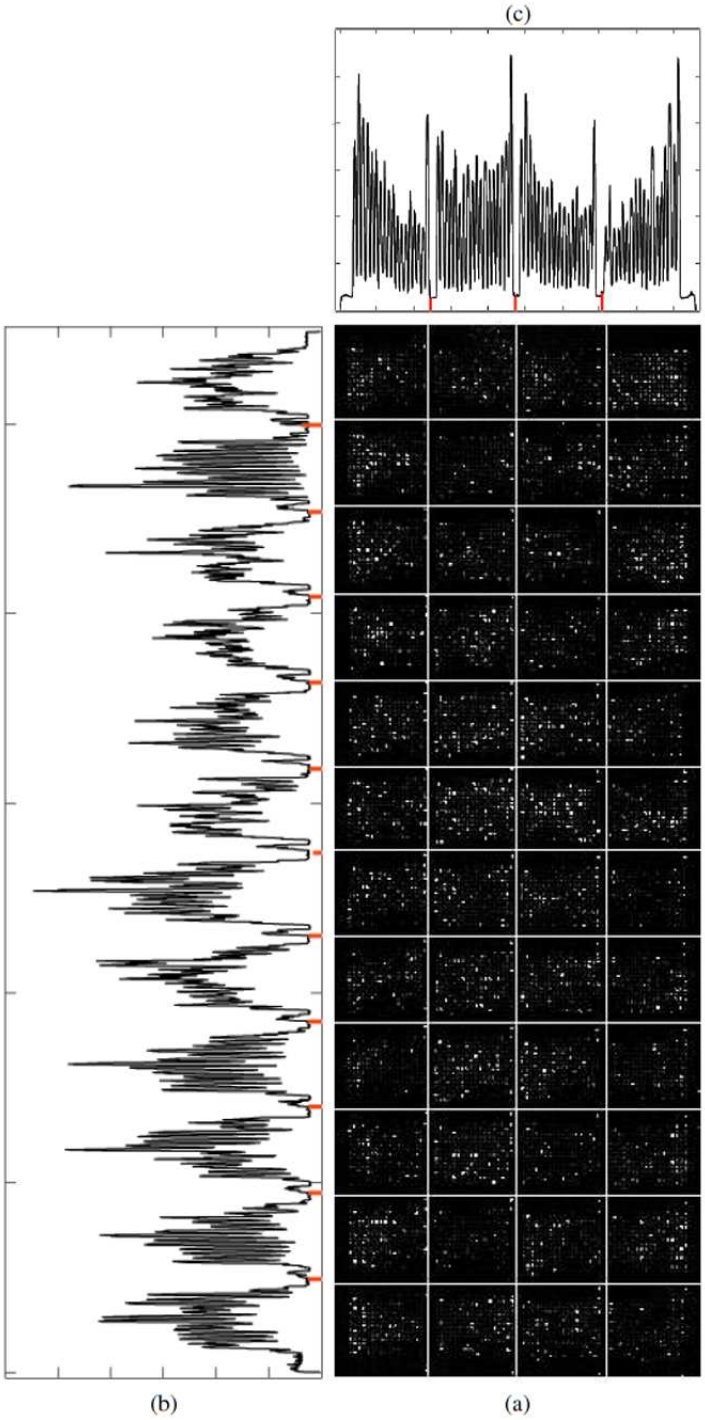

Figure 2.12: Sub-grid detection in a microarray image from the SMD dataset. (a) detected sub-grids in AT-20387-ch2 from the SMD dataset, (b) horizontal histogram and detected valleys corresponding to horizontal lines, (c) vertical histogram and detected valleys corresponding to vertical lines.

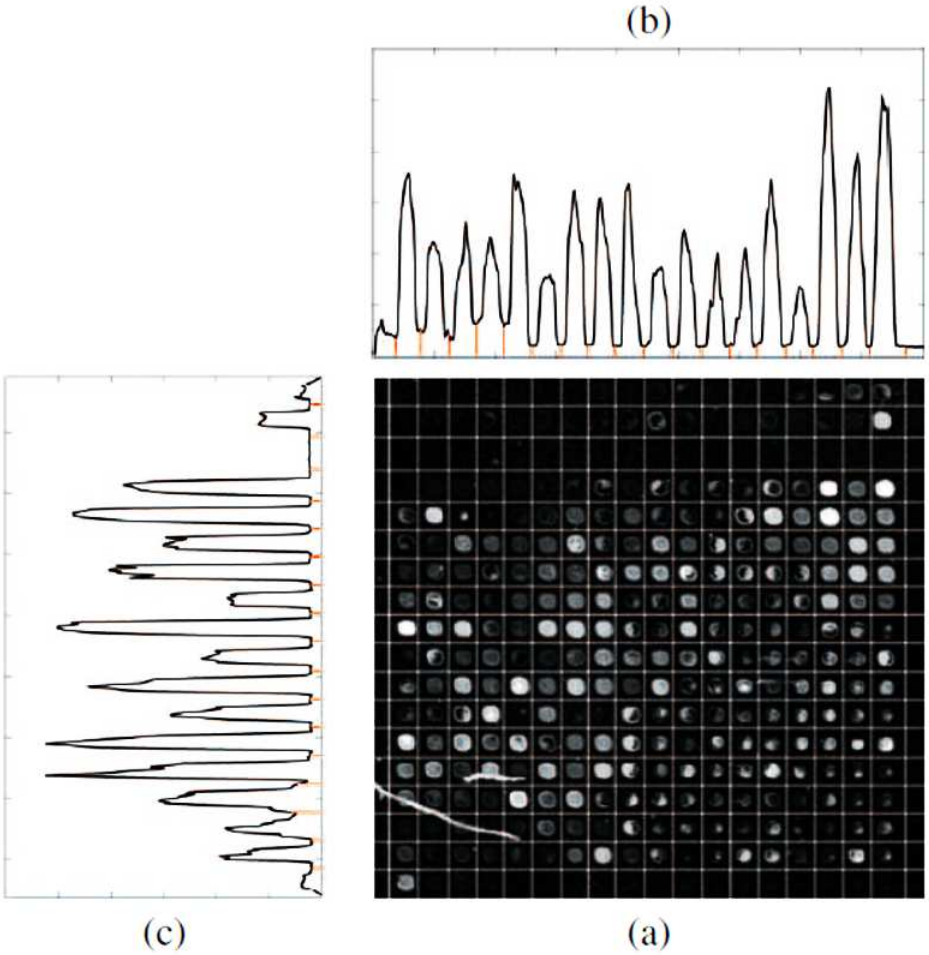

Figure 2.13: Spot detection in a sub-grid from AT-20387-ch2. (a) detected spots in one of the sub-grids in AT-20387-ch2, (b) horizontal histogram and detected valleys corresponding to horizontal lines, (c) vertical histogram and detected valleys corresponding to vertical lines.

set $T = \{t_0, t_1, \ldots, t_k, t_{k+1}\}$, where $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} = n$ and $t_i \in \{0\} \cup H$. The problem of multilevel thresholding consists of finding a threshold set, T^*, in such a way that a function $f : H^k \times [0,1]^n \rightarrow R^+$ is maximized/minimized. Using this threshold set, H is divided into $k+1$ classes: $\zeta_1 = \{1, 2, \ldots, t_1\}$, $\zeta_2 = \{t_1 + 1, t_1 + 2, \ldots, t_2\}$, ..., $\zeta_k = \{t_{k-1} + 1, t_{k-1} + 2, \ldots, t_k\}$, $\zeta_{k+1} = \{t_k + 1, t_k + 2, \ldots, n\}$. The between class variance

criterion is given by:

$$\Psi_{\mathrm{BC}}(T) = \sum_{j=1}^{k+1} \omega_j \mu_j^2 \,, \tag{2.4}$$

where $\omega_j = \sum_{i=t_{j-1}+1}^{t_j} p_i$, $\mu_j = \frac{1}{\omega_j} \sum_{i=t_{j-1}+1}^{t_j} i p_i$.

We use the dynamic programming algorithm for *optimal* multilevel thresholding proposed in [22], which is an extension for irregularly sampled histograms. To implement the between-class variance criterion, $\Psi_{\mathrm{BC}}(T)$ is expressed as follows: $\Psi_{\mathrm{BC}}(T) = \sum_{j=1}^{k+1} \omega_j \mu_j^2 = \sum_{j=1}^{k+1} \psi_{t_{j-1}+1,t_j}$, where $\psi_{t_j+1,t_{j+1}} = \omega_j \mu_j^2$. We consider the temporary variables a and b ,which are computed as follows:

$$a \leftarrow p_{t_{j-1}+1} + \sum_{i=t_{j-1}+2}^{t_j} p_i, \qquad \text{and} \tag{2.5}$$

$$b \leftarrow (t_{j-1}+1) p_{t_{j-1}+1} + \sum_{i=t_{j-1}+2}^{t_j} i p_i \,. \tag{2.6}$$

Since from (2.5) and (2.6), a and b are known, then $\psi_{t_{j-1}+2,t_j}$, for the next step, can be re-computed as follows in $\Theta(1)$ time:

$$a \leftarrow a - p_{t_{j-1}+1} \,, \tag{2.7}$$

$$b \leftarrow b - (t_{j-1}+1) p_{t_{j-1}+1} \,, \text{ and} \tag{2.8}$$

$$\psi_{t_{j-1}+2,t_j} \leftarrow \frac{b^2}{a} \,. \tag{2.9}$$

Full details of the algorithm, whose worst-case time complexity is $O(kn^2)$, can be found in [22].

Automatic Detection of the Number of Sub-grids and Spots

Finding the correct number of sub-grids and spots in each sub-grid is one of the most challenging issues in sub-grid and spot detection. This stage is crucial in order to fully automate the whole process. Multi-level thresholding uses the number of sub-grids (spots) as a single parameter. Thus, we need to determine the correct number of sub-grids (spots) prior to using multi-level thresholding methods. For this, we resort on validity indices used for clustering. By analyzing the traditional indices for clustering validity and their suitability to be combined with our measure, we propose a new index of validity for this specific problem. From the different indices of validity for clustering (cf. [23, 24]), we consider the I index as the basis of the proposed index. The I index is defined as follows:

$$I(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K \right)^2, \tag{2.10}$$

where $E_K = \Sigma_{i=1}^{K} \Sigma_{k=1}^{n_i} p_k ||k - z_i||$, $D_K = \underbrace{\max}_{i,j=1}^{K} ||z_i - z_j||$, n is the total number of points in the dataset (bins in the histogram), and z_k is the center of the kth cluster. We also consider the average frequency value of the thresholds in a histogram, which is computed as follows:

$$A(K) = \frac{1}{K} \sum_{i=1}^{K} p(t_i), \tag{2.11}$$

where t_i is the ith threshold found by optimal multilevel thresholding and $p(t_i)$ is the corresponding probability value in the histogram.

The proposed index, $\alpha(x)$, is the result of a combination of the I index, (2.10)and $A(K)$, (2.11), as follows:

$$\alpha(K) = \sqrt{K} \frac{I(K)}{A(K)} = \frac{\left(\frac{E_1}{E_K} \times D_K \right)^2}{\sqrt{K} \Sigma_{i=1}^{K} p(t_i)}. \tag{2.12}$$

For maximizing $I(K)$ and minimizing $A(K)$, the value of $\alpha(K)$ must be maximized. Thus, the best number of thresholds K^* based on the α index is given by:

$$K^* = \underset{1 \leq K \leq \delta}{\operatorname{argmax}} \alpha(K) = \underset{1 \leq K \leq \delta}{\operatorname{argmax}} \frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{2.13}$$

To find the best number of thresholds, K^*, we perform an exhaustive search on all positive values of K from 1 to δ and find the value of k that maximizes the α index. In our experiment we set δ to $\sqrt{n}$ (cf. [25]).

The Refinement Procedure

In some cases, the detected grid or sub-grid may not separate spots completely or may separate them marginally. In these cases, a refinement procedure can be used to boost the performance of method. For this, each horizontal or vertical line is replaced with a new line. Consider two horizontal lines h_j and h_{j+1} where $j \in [1, K^*]$ and a vertical line v_i where $i \in [1, K^*]$, and v_i is bounded between h_j and h_{j+1}. Given $A = \{a_{ij}\}$, line v_i can be moved to left and right in such way that $\Sigma_{i=h_j}^{h_{j+1}} a_{ik}$ is minimized. In other words, the vertical line v_i can be replaced with a new vertical line, v_r, in such a way that:

$$r = \underset{v_{i-1} \leq k \leq v_{i+1}}{\operatorname{argmin}} \Sigma_{i=h_j}^{h_{j+1}} a_{ik}. \tag{2.14}$$

Analogously, this procedure can be applied to each horizontal line. Figure 2.14 shows an example in which a vertical line is replaced by a new one during the refinement procedure. As shown in the figure, the vertical line v_i is originally located in the wrong place and does not separate two adjacent spots correctly. By moving it to left and right, the new line

v_r is found in such way that those adjacent spots are separated correctly.

Figure 2.15 shows the detected spots in one of the sub-grids of 20387-ch2 of SMD before and after using the refinement procedure. It is clear that there are some misalignments in separating the adjacent spots in the top part of the sub-grid before using the refinement procedure. After the refinement, all the spots are separated precisely as shown in the figure.

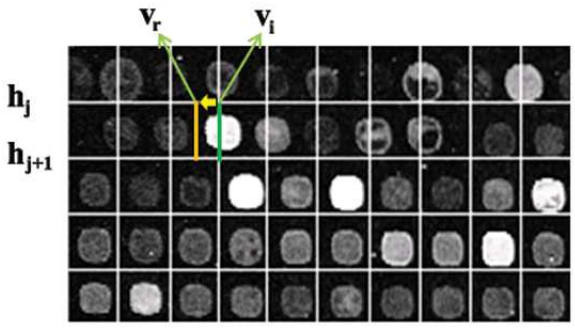

Figure 2.14: The refinement procedure. During the refinement procedure each line can be moved to left or right (for vertical lines) and up or down (for horizontal lines) to find the best location separating the spots. In this image, v_i is the sub-line before using the refinement procedure and v_r is the sub-line after adjusting it during refinement procedure.

Authors contributions

LR and IR conceived the gridding model, performed the data analysis, and elaborated the corresponding discussions. IR implemented the algorithms and conducted the experiments. Both authors have read and approved the final manuscript.

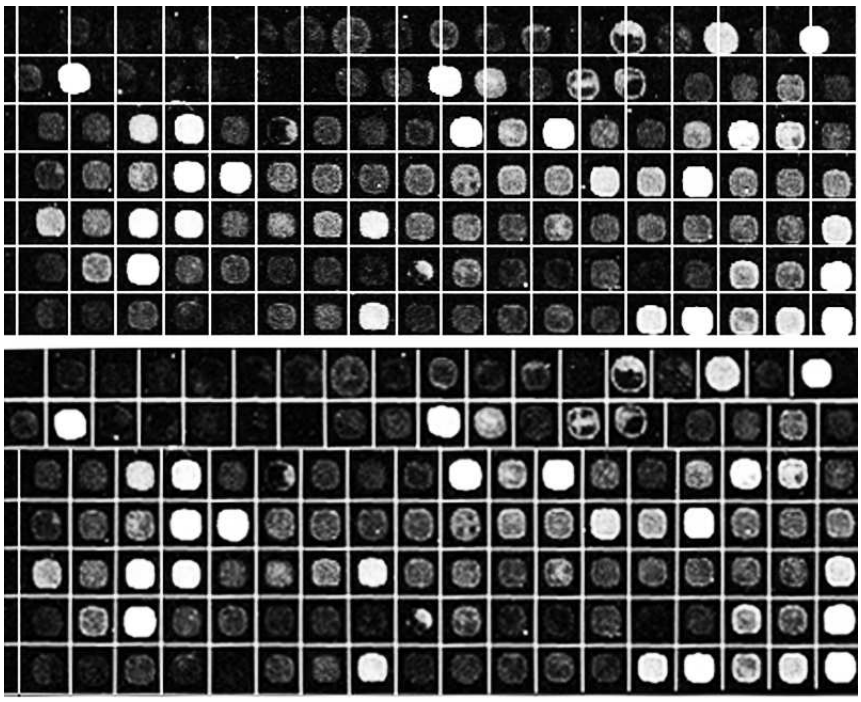

Figure 2.15: Effect of the refinement procedure to increase the accuracy of the proposed method. Detected spots in one of the sub-grids of AT-20387-ch1 from the SMD dataset before using the refinement procedure (top), and detected spots in the same part of the sub-grid after using the refinement procedure (bottom).

Bibliography

[1] G Antoniol and M Ceccarelli: A Markov Random Field Approach to Microarray Image Gridding. Proc. of the 17th International Conference on Pattern Recognition 2004, :550–553.

[2] N Brandle, H Bischof, and H Lapp: Robust DNA microarray image analysis. Machine Vision and Applications 2003, 15:11–28.

[3] B Ceccarelli and G Antoniol: A Deformable Grid-matching Approach for Microarray Images. IEEE Transactions on Image Processing 2006, 15(10):3178–3188.

[4] F Qi, Y Luo, and D Hu: Recognition of Perspectively Distorted Planar Grids. Pattern Recognition Letters 2006, 27(14):1725–1731.

[5] L Rueda and V Vidyadharan: A Hill-climbing Approach for Automatic Gridding of cDNA Microarray Images. IEEE Transactions on Computational Biology and Bioinformatics 2006, 3:72–83.

[6] L Qin, L Rueda, A Ali, and A Ngom: Spot Detection and Image Segmentation in DNA Microarray Data. Applied Bioinformatics 2005, 4:1–12.

[7] M Katzer, F Kummer, and G Sagerer: A Markov Random Field Model of Microarray Gridding. Proceeding of the 2003 ACM Symposium on Applied Computing 2003, :72–77.

[8] J Angulo and J Serra: Automatic Analysis of DNA Microarray Images Using Mathematical Morphology. Bioinformatics 2003, 19(5):553–562.

[9] A Jain, T Tokuyasu, A Snijders, R Segraves, D Albertson, and D Pinkel: Fully Automatic Quantification of Microarray Data. Genome Research 2002, 12(2):325–332.

[10] M Katzer, F Kummert, and G Sagerer: Automatische Auswertung von Mikroarraybildern. Proc. of Workshop Bildverarbeitung für die Medizin 2002.

[11] M Steinfath, W Wruck, and H Seidel: Automated Image Analysis for Array Hybridization Experiments. Bioinformatics 2001, 17(7):634–641.

[12] L Rueda: Sub-grid Detection in DNA Microarray Images. Proceedings of the IEEE Pacific-RIM Symposium on Image and Video Technology 2007, :248–259.

[13] Y Wang, M Ma, K Zhang, and F Shih: A Hierarchical Refinement Algorithm for Fully Automatic Gridding in Spotted DNA Microarray Image Processing. Information Sciences 2007, 177(4):1123–1135.

[14] Y Wang, F Shih, and M Ma: Precise Gridding of Microarray Images by Detecting and Correcting Rotations in Subarrays. Proceedings of the 8th Joint Conference on Information Sciences 2005, :1195–1198.

[15] E Zacharia and D Maroulis: Micoarray Image Gridding Via an Evolutionary Algorithm. IEEE International Conference on Image Processing 2008, :1444–1447.

[16] D Bariamis , D Maroulis and D Iakovidis: M^3G: Maximum Margin Microarray Gridding. BMC Bioinformatics 2010, 11:49.

[17] D Bariamis , D Maroulis and D Iakovidis: Unsupervised SVM-based gridding for DNA microarray images. Computerized Medical Imaging and Graphics 2010, 34(6):418–425.

[18] Ramdas L, Coombes KR, Baggerly K, Abruzzo L, Highsmith WE, Krogmann T, Hamilton SR, Zhang W: Sources of nonlinearity in cDNA microarray expression measurements. Genome Biology 2001, 2(11).

[19] N Otsu: A Threshold Selection Method from Gray-level Histograms. IEEE Trans. on Systems, Man and Cybernetics 1979, SMC-9:62–66.

[20] J Kapur , P Sahoo and A Wong: A New Method for Gray-level Picture Thresholding Using the Entropy of the Histogram. Computer Vision Graphics and Image Processing 1985, 29:273–285.

[21] J Kittler and J Illingworth: Minimum Error Thresholding. Pattern Recognition 1986, 19:41–47.

[22] L Rueda: An Efficient Algorithm for Optimal Multilevel Thresholding of Irregularly Sampled Histograms. Proceedings of the 7th International Workshop on Statistical Pattern Recognition 2008, :612–621.

[23] U Maulik and S Bandyopadhyay: Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Trans. on Pattern Analysis and Machine Intelligence 2002, 24(12):1650–1655.

[24] S. Theodoridis , K. Koutroumbas: Pattern Recognition. Elsevier Academic Press, 4th edition 2008.

[25] R. Duda, P. Hart, D. Stork: Pattern Classification. New York, NY: John Wiley and Sons, Inc., 2nd edition 2000.

Chapter 3

Applications of Multilevel Thresholding Algorithms to Transcriptomics Data

3.1 Introduction

Among other components, the genome contains a set of genes required for an organism to function and evolve. However, the genome is only a source of information and in order to function, the genes express themselves into proteins. The transcription of genes to produce RNA is the first stage of gene expression. The transcriptome can be seen as the complete set of RNA transcripts produced by the genome. Unlike the genome, the transcriptome is very dynamic. Despite having the same genome regardless of the type of cell or environmental conditions, the transcriptome varies considerably in differing circumstances because of the different ways the genes may express.

Transcriptomics, the field that studies the role of the transciptome, provides a rich source of data suitable for pattern discovery and analysis. The quantity and size of these data may vary based on the model and underlying methods used for analysis. In gene expression mi-

croarrays, the raw data are represented in terms of images, typically in TIFF format which are approximately 20-30MB per array. These TIFF files are processed and transformed into quantified data used for posterior analysis. In contrast, high throughput sequencing methods (e.g. ChIP-seq and RNA-seq) generate more than 1TB of data, while the sequence files (approximately 20-30GB) are typically used as a starting point for analysis [1]. Clearly, these sequence files are an order of magnitude larger than those from arrays.

3.1.1 DNA Microarray Image Gridding

Various technologies have been developed to measure the transcriptome, including hybridization or sequence-based approaches. Hybridization-based approaches typically involve processing fluorescently labeled DNA microarrays. Microarrays are one of the most important technologies used in molecular biology to massively explore the abilities of the genes to express themselves into proteins and other molecular machines responsible for different functions in an organism. These expressions are monitored in cells and organisms under specific conditions, and are present in many applications in medical diagnosis, pharmacology, disease treatment, among others. If we consider DNA microarrays, scanning the slides at a very high resolution produces images composed of sub-grids of spots. Image processing and analysis are two important aspects of microarrays, and involve various steps. The first task is gridding, which is quite important as errors are propagated to subsequent steps. Roughly speaking, gridding consists of determining the spot locations in a microarray image (typically, in a sub-grid). The gridding process requires the knowledge of the sub-girds in advance in order to proceed, which is not necessarily available in advance.

Many approaches have been proposed for microarray image gridding and spot detection, being the most widely known the following. The Markov random field (MRF) is

one of them, which applies specific constraints and heuristic criteria [2]. Other gridding methods used for gridding include mathematical morphology [3], Bayesian model-based algorithms [4, 5], the hill-climbing approach [6], a Gaussian mixture model approach [7], Radon-transform-based method [8], a genetic algorithm for separating sub-grids and spots [9], and the recently introduced maximum margin method [10]. A method that we have proposed and has been successfully used in microarray gridding is the multilevel thresholding algorithm [11], which is discussed in more detail later in the paper.

3.1.2 ChIP-Seq and RNA-Seq Peak Finding

Hybridization-based approaches are high throughput and relatively inexpensive, except for high-resolution tiling arrays that interrogate large genomes. However, these methods have several limitations, which include reliance upon existing knowledge about the genome, high background levels owing to cross-hybridization, and a limited dynamic range of detection owing to both background and saturation of signals [1, 12]. Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods.

Recently, the development of novel high-throughput DNA sequencing methods has provided a new method for both mapping and quantifying transcriptomes. These methods, termed ChIP-seq (ChIP sequencing) and RNA-seq (RNA sequencing), have clear advantages over existing approaches and are emerging in such a way that eukaryotic transcriptomes are to be analyzed in a high-throughput and more efficient manner [12].

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a technique that provides quantitative, genome-wide mapping of target protein binding events [13, 14]. In ChIP-seq, a protein is first cross-linked to DNA and the fragments sub-

sequently sheared. Following a size selection step that enriches for fragments of specified lengths, the fragments ends are sequenced, and the resulting reads are aligned to the genome. Detecting protein binding sites from massive sequence-based datasets with millions of short reads represents a truly bioinformatics challenge that has required considerable computational innovation in spite of the availability of programs for ChIP-chip analysis [7, 15–17].

With the increasing popularity of ChIP-seq technology, a demand for peak finding methods has emerged and it causes developing new algorithms. Although due to mapping challenges and biases in various aspects of existing protocols, identifying peaks is not a straightforward task.

Different approaches have been proposed for detecting peaks based ChIP-seq/RNA-seq mapped reads so far. Zhang et al. presents a Model-based Analysis of ChIP-seq data (MACS), which analyzes data generated by short read sequencers [18]. It models the shift size of ChIP-seq tags, and uses it to improve the spatial resolution of predicted binding sites. A two-pass strategy called PeakSeq has been presented in [19]. This strategy compensates for signal caused by open chromatin, as revealed by the inclusion of the controls. The first pass identifies putative binding sites and compensates for genomic variation in mapping the sequences. The second pass filters out sites not significantly enriched compared to the normalized control, computing precise enrichments and significance. A statistical approach for calling peaks has been recently proposed in [20], which is based on evaluating the significance of a robust statistical test that measures the extent of pile-up reads. Specifically, the shapes of putative peaks are defined and evaluated to differentiate between random and non-random fragment placements on the genome. Another algorithm for identification of binding sites is site identification from paired-end sequencing (SIPeS) [21], which can be

used for identification of binding sites from short reads generated from paired-end solexa ChIP-seq technology.

In this paper, we review the application of optimal multilevel thresholding (OMT) to gridding and peak finding problems in transcriptomics. Moreover, a conceptual and practical comparison between OMT and other state-of-the-art approaches is also presented.

3.2 Optimal Multilevel Thresholding

Multilevel thresholding is one of the most widely-used techniques in different aspects of signal and image processing, including segmentation, classification and object discrimination. Given a histogram with frequencies or probabilities for each bin, the aim of multilevel thresholding is to divide the histogram into a number of groups (or classes) of contiguous bins in such a way that a criterion is optimized. In microarray image gridding, we compute vertical (or horizontal) running sums of pixel intensities, obtaining histograms in which each bin represents one column (or row respectively), and the running sum of intensities corresponds to the frequency of that bin. The frequencies are then normalized in order to be considered as probabilities. Each histogram is then processed (see below) to obtain the optimal thresholding that will determine the locations of the separating lines.

Consider a histogram H, an ordered set $\{1,2,\ldots,n-1,n\}$, where the ith value corresponds to the ith bin and has a probability, p_i. Given an image, $A=\{a_{ij}\}$, H can be obtained by means of the horizontal (vertical) running sum as follows: $p_i=\sum_{j=1}^{m} a_{ij}$ ($p_j=\sum_{i=1}^{n} a_{ij}$). We also consider a threshold set T, defined as an ordered set $T=\{t_0,t_1,\ldots,t_k,t_{k+1}\}$, where $0=t_0<t_1<\ldots<t_k<t_{k+1}=n$ and $t_i\in\{0\}\cup H$. The problem of multilevel thresholding consists of finding a threshold set, T^*, in such a way that a function f : $H^k\times[0,1]^n\rightarrow\mathbb{R}^+$ is maximized/minimized. Using this threshold set, H is divided into $k+1$

classes: $\zeta_1 = \{1,2,\ldots,t_1\}$, $\zeta_2 = \{t_1+1, t_1+2,\ldots,t_2\}$, ..., $\zeta_k = \{t_{k-1}+1, t_{k-1}+2,\ldots,t_k\}$, $\zeta_{k+1} = \{t_k+1, t_k+2,\ldots,n\}$. The most important criteria for multilevel thresholding are the following [22]:

Between class variance:

$$\Psi_{\mathrm{BC}}(T) = \sum_{j=1}^{k+1} \omega_j \mu_j^2 \tag{3.1}$$

where $\omega_j = \sum_{i=t_{j-1}+1}^{t_j} p_i$, $\mu_j = \frac{1}{\omega_j} \sum_{i=t_{j-1}+1}^{t_j} i p_i$;

Entropy-based:

$$\Psi_{\mathrm{H}}(T) = \sum_{j=1}^{k+1} H_j \tag{3.2}$$

where $H_j = -\sum_{i=t_{j-1}+1}^{t_j} \frac{p_i}{\omega_j} \log \frac{p_i}{\omega_j}$;

Minimum error:

$$\Psi_{\mathrm{ME}}(T) = 1 + 2 \sum_{j=1}^{k+1} \omega_j (\log \sigma_j - \log \omega_j) \tag{3.3}$$

where $\sigma_j^2 = \sum_{i=t_{j-1}+1}^{t_j} \frac{p_i (i-\mu_j)^2}{\omega_j}$.

A dynamic programming algorithm for *optimal* multilevel thresholding was proposed in a previous work [22], which is an extension for irregularly sampled histograms. For this, the criterion has to be decomposed as a sum of terms as follows:

$$\Psi(T_{0,m}) = \Psi(\{t_0, t_1, \ldots, t_m\}) \triangleq \sum_{j=1}^{m} \psi_{t_{j-1}+1, t_j}\,, \tag{3.4}$$

where $1 \leq m \leq k+1$ and the function $\psi_{l,r}$, where $l \leq r$, is a real, positive function of $p_l, p_{l+1}, \ldots, p_r$, $\psi_{l,r} : H^2 \times [0,1]^{l-r+1} \rightarrow \mathbb{R}^+ \cup \{0\}$. If $m = 0$, then $\Psi(\{t_0\}) = \psi_{t_0,t_0} = \psi_{0,0} = 0$. The thresholding algorithm can be found in [22]. In the algorithm, a table C is filled in, where $C(t_j, j)$ contains the optimal solution for $T_{0,j} = t_0, t_1, \ldots, t_j$, $\Psi^*(T_{0,j})$, which is found from $\min\{t_j\} \leq t_j \leq \max\{t_j\}$. Another table, $D(t_j, j)$, contains the value of t_{j-1} for

which $\Psi^*(T_{0,j})$ is optimal. The algorithm runs in $O(kn^2)$, and has been further improved to achieve linear complexity, i.e. $O(kn)$, by following the approach of [23].

3.2.1 Using Multi-level Thresholding for Gridding DNA Microarray Images

A DNA microarray image contains spots arranged into sub-grids. The image contains various sub-grids as well, which are found in the first stage. Once the sub-grids are found, the spots centers are to be identified. A microarray image can be considered as a matrix $A = \{a_{i,j}\}, i = 1,....,n$ and $j = 1,....,m$, where $a_{ij} \in \mathbb{Z}^+$, and A is a sub-grid of a DNA microarray image. The aim of sub-gridding is to obtain vectors, namely $\mathbf{h} = [h_1,...h_{p-1}]^t$ and $\mathbf{v} = [v_1,...v_{q-1}]^t$, that separate the sub-grids. Finding the spot locations is done analogously – more details of this, as well as those of the whole process can be found in [11]. The aim of gridding is to find the corresponding spot locations given by the horizontal and vertical adjacent vectors. Post-processing or refinement allows us to find a spot region for each spot, which is enclosed by four lines.

When producing the microarrays, based on the layout of the printer pins, the number of sub-grids or spots are known. But due to misalignments, deformations, artifacts or noise during producing the microarray images, these numbers may not be available. Thus, it is important that the gridding algorithm allows some flexibility in finding these parameters, as well as avoiding the use of other user-defined parameters. This is what the thresholding methods endeavor to do, by automatically finding the best number of thresholds (sub-grids or spots) – more details in the next section.

3.2.2 Using Multi-level Thresholding for Analyzing ChIP-Seq/RNA-Seq Data

In ChIP-seq and RNA-seq analysis, a protein is first cross-linked to DNA and the fragments subsequently pruned. Then, the fragments ends are sequenced, and the resulting reads are aligned to the genome. The result of read alignments produces a histogram in such a way that the x axis represents the genome coordinate and the y axis the frequency of the aligned reads in each genome coordinate. The aim is to find the significant peaks corresponding to enriched regions. For this reason, a non-overlapping moving window is used. By starting from the beginning, a dynamic window of minimum size t is being applied to the histogram and each window that could be analyzed separately. The size of the window could be different for each window to prevent truncating a peak before its end. Thus, for each window a minimum number of t bins is used and, by starting from the end of previous window, the size of window is increased until a zero value in the histogram is reached.

The aim is to obtain vectors $C_{w_i} = [c^1_{w_i}, ...c^n_{w_i}]^t$, where w_i is the i^{th} window and C_{w_i} is the vector that contains n threshold coordinates which correspond to the i^{th} window. Figure 3.1 depicts the process of finding the peaks corresponding to the regions of interest for the specified protein. The input to the algorithm includes the reads and the output of the whole process is the location of the detected significant peaks by using optimal multilevel thresholding combined with our recently proposed α index.

3.3 Automatic Detection of the Number of Clusters

Finding the correct number of clusters (number of sub-grids or spots or the number of regions in each window in ChIP-seq/RNA-seq analysis) is one of the most challenging issues.

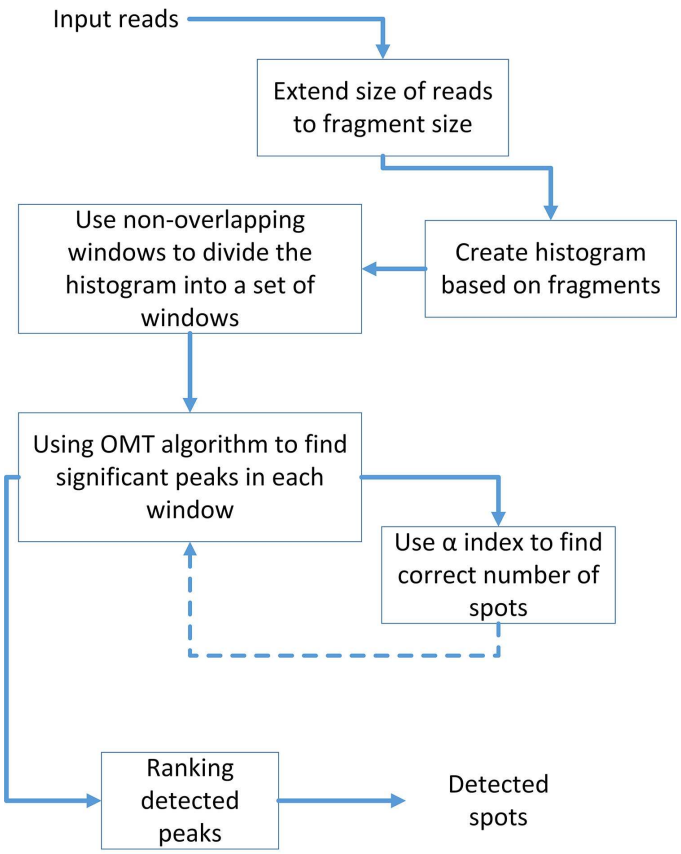

Figure 3.1: Schematic representation of the process for finding significant peaks.

This stage is crucial in order to fully automate the whole process. For this, we need to determine the correct number clusters or thresholds prior to applying multi-level thresholding methods. This is found by applying an index of validity (derived from clustering techniques) and testing over all possible number of clusters (or thresholds) from 2 to $\sqrt{n}$, where n is the number of bins in the histogram. We have recently proposed the $\alpha(x)$ index, which is the result of a combination of a simple index and the well-known I index [24] as follows:

$$\alpha(K) = \sqrt{K}\frac{I(K)}{A(K)} = \frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{3.5}$$

For maximizing $I(K)$ and minimizing $A(K)$, the value of $\alpha(K)$ must be maximized. Thus, the best number of thresholds K^* based on the α index is given by:

$$K^* = \underset{1 \leq K \leq \delta}{\operatorname{argmax}}\, \alpha(K) = \underset{1 \leq K \leq \delta}{\operatorname{argmax}} \frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{3.6}$$

3.4 Comparison of Transcriptomics Data Analysis Algorithms

3.4.1 DNA Microarray Image Gridding Algorithms Comparison

A conceptual comparison of microarray image griding methods based on their features is shown in Table 3.1. The methods included in the comparison are the following: (i) Radon transform sub-gridding (RTSG) [8], (ii) Bayesian simulated annealing gridding (BSAG) [4], (iii) genetic-algorithm-based gridding (GABG) [9], (iv) hill-climbing gridding (HCG) [6], (v) maximum margin microarray gridding (M^3G) [10], and the optimal multilevel thresholding algorithm for gridding (OMT) [11]. As shown in the table, OMT does not need any number-based parameter, and hence making it much more powerful than the other methods. Although the index or thresholding criterion can be considered as a "parameter", this can be fixed by using the between class criterion. In a previous work, we have "fixed" the index of validity to the α *index* and the *between class* as the thresholding criterion [11]. As can also be observed in the table, most algorithms and methods require the use of user-defined and subjectively fixed parameters. One example is the GABG, which needs to adjust the mutation and crossover rates, probability of maximum and minimum thresholds, among others. It is critical then to adjust these parameters for specific data, and variations may

occur across images of different characteristics.

3.4.2 Comparison of Algorithms for ChIP-Seq and RNA-Seq Analysis

A conceptual comparison between thresholding algorithms and other ChIP and RNA-Seq methods based on their features is shown in Table 3.2. The methods included in the comparison are the following: (i) GLobal Identifier of Target Regions (GLITR) [25], (ii) Model-based Analysis of ChIP-seq (MACS) [18], (iii) PeakSeq [19], (iv) quantitative enrichment of sequence tags (Quest) [26], (v) SICER [27], (vi) Site Identification from Short Sequence Reads (SiSSRs) [28], (vii) Tree shape Peak Identification for ChIP-seq (T-PIC) [20], and (viii) the optimal multilevel thresholding algorithm, OMT. As shown in the table, all algorithms require some parameters to be set by the user based on the particular data to be processed, including *p*-values, FDR, number os nearest neighbors, peak height, valley depth, window length, gap size, among others. OMT is the algorithm that requires almost no parameter at all. Only the average fragment length is needed, but this parameter can be easily estimated from the underlying data. In practice, if enough computational resources are available, the fragment length would not be needed, since the OMT algorithm could be run directly on the whole histogram.

3.5 Experimental Analysis

This section is necessarily brief and reviews some experimental results as presented in [11]. For the experiments, two different kinds of DNA microarray images have been used, which were obtained from the Stanford Microarray Database (SMD) the Gene Expression Omnibus (GEO). The images have different resolutions, number of sub-grids and spots.

Table 3.1: Conceptual comparison of recently proposed DNA microarray gridding methods.

Method	Parameters	Sub-grid Detection	Spot Detection	Automatic Detection No. of Spots	Rotation
RTSG	n: Number of sub-grids	√	×	×	√
BSAG	α ,β: Parameters for balancing prior and posterior probability rates	×	√	√	√
GABG	μ , c :Mutation and Crossover rates, p_{max}: probability of maximum threshold, p_{low}: probability of minimum threshold, f_{max} : percentage of line with low probability to be a part of grid, T_p: Refinement threshold	√	√	√	√
HCG	λ , σ: Distribution parameters	×	√	√	×
M^3G	c: Cost parameter	×	√	√	√
OMT	None[1]	√	√	√	√

[1] The only parameters that would be needed in the proposed method are the "thresholding criterion" and the "index of validity". These two "parameters" are methodological, not number-based, and hence making OMT less dependent on parameters.

Table 3.2: Conceptual comparison of recently proposed methods for ChIP-seq and RNA-seq data.

Method	Peak selection criteria	Peak ranking	Parameters
GLITR	*n*: Classification by height and relative enrichment	Peak height and fold enrichment	Target FDR, number nearest neighbors for clustering
MACS v1.3.5	Local region Poisson *p* value	*p* value	*p*-value threshold, tag length, *m*-fold for shift estimate
PeakSeq	Local region binomial *p* value	*q* value	Target FDR
Quest v2.3	height threshold, background ratio	q value	KDE bandwidth, peaks height, sub-peak valley depth, ratio to background
SICER v1.02	*p* value from random background model, enrichment relative to control	*q* value	Window length, gap size, FDR (with control) or *E*-Value (no control)
SiSSRs v1.4	$N^+ - N^-$ sign change, $N^+ + N^-$ threshold in region	*p* value	FDR, $N^+ + N^-$ threshold
T-PIC	Local height threshold	*p* value	Average fragment length, significance *p* value, minimum length of interval
OMT	number of ChIP reads minus control reads in window	volume	Average fragment length

We have used the between-class variance as the thresholding criteria, since it is the one that delivers the best results. All the sub-grids in each image are detected with a 100% accuracy, and also spot locations in each sub-grid can be detected efficiently with an average accuracy of 96.2% for SMD dataset and 96% for GEO dataset. Figure 3.2 shows the detected sub-

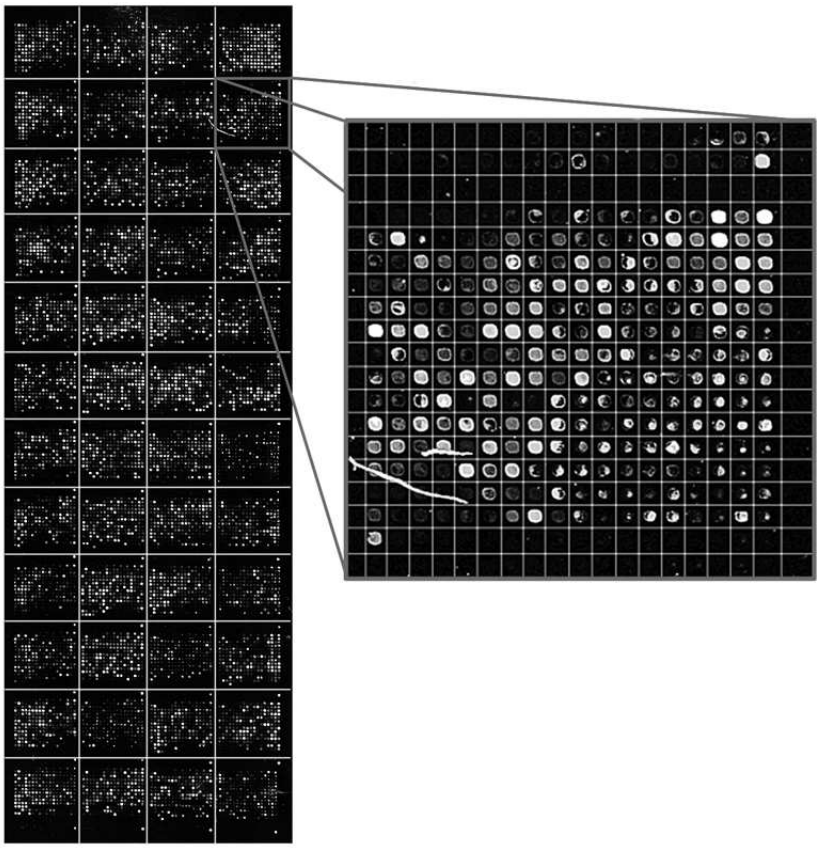

Figure 3.2: Detected sub-grids in AT-20387-ch2 microarray image (left) and detected spots in one of sub-grids (right).

grids from the AT-20387-ch2 image (left) and the detected spots in one of sub-grids (right). As shown in the figure, the proposed method precisely detects the sub-grids location at first, and in the next stage, each sub-grid is divided precisely into the corresponding spots with the same method.

In addition to this, some experimental, preliminary results for testing performance of the OMT algorithm on ChIP/RNA-seq data are shown here. We have used the FoxA1 dataset [18], which contains experiment and control samples of 24 chromosomes. The ex-

periment and control histogram were generated separately by extending each mapped position (read) into an appropriately oriented fragment, and then joining the fragments based on their genome coordinates. The final histogram was generated by subtracting the control from the experiment histogram. To find significant peaks, we used a non-overlapping window with the initial size of 3000bp. To avoid truncating peaks in boundaries, each window is extended until the value of the histogram at the end of the window becomes zero. Figure 3.3 shows three detected regions for chromosomes 9 and 17 and their corresponding base pair coordinates. It clear from the pictures that the peaks contain a very high number of reads, and then these regions are quite likely to represent binding sites, open reading frames or other bio-markers. A biological assessment of these bio-markers can corroborate this.

3.6 Discussion and Conclusion

Transcriptomics provide a rich source of data suitable for pattern analysis. We have shown how multilevel thresholding algorithms can be applied to an efficient analysis of transcriptomics and genomics data by finding sub-grids and spots in microarray images, as well as significant peaks in high-throughput next generation sequencing data. OMT can be applied to a wide range of data from different sources and with different characteristics, and allows data analysis such as sub-grid and spot detection in DNA microarray image gridding and also for detecting significant regions on ChIP and RNA-seq data. OMT has been shown to be sound and deal with noise in experiments and it is able to use on different approaches with a little change – this is one the most important features of this algorithm.

Thresholding algorithms, though shown to be quite useful for transcriptomics and genomics data analysis, are still emerging tools in these areas, and open the possibility for

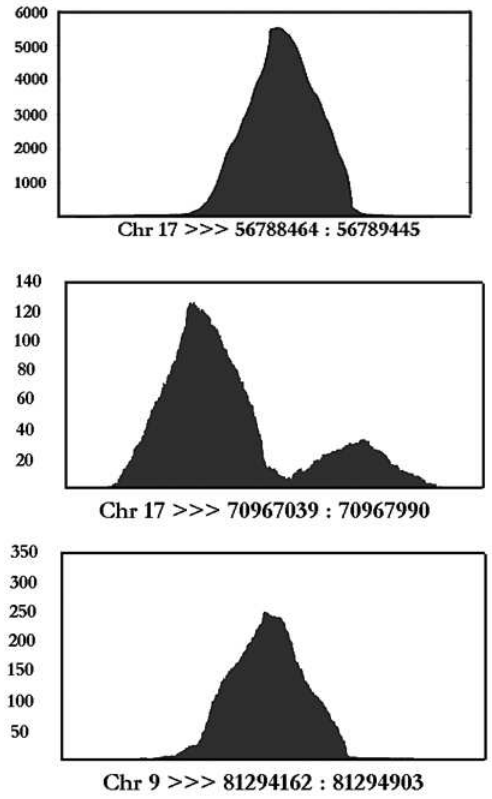

Figure 3.3: Three detected regions from FoxA1 data for chromosomes 9 and 17. The *x* axis corresponds to the genome position in bp and the *y* axis corresponds to the number of reads.

further advancement. One of the problems that deserves attention is the use of other thresholding criteria, including minimum error, entropy-based and others. For these two criteria the algorithm still runs in quadratic or *n*-logarithmic complexity, and which make the whole process sluggish. Processing a whole genome or even a chromosome for finding peaks in ChIP or RNA-seq is still a challenge, since it involves histograms with several million bins. This makes it virtually impossible to process a histogram at once, and so it has to be divided into several fragments. Processing the whole histograms at once is one of the open and challenging problems that deserve more investigation. Next generation sequence data analysis

is an emerging and promising area for pattern discovery and analysis, which deserve the attention of the research community in the field.

Bibliography

[1] J. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology, 9(1):34, 2011.

[2] M. Katzer, F. Kummer, and G. Sagerer. A Markov Random Field Model of Microarray Gridding. Proceeding of the 2003 ACM Symposium on Applied Computing, pages 72–77, 2003.

[3] J. Angulo and J. Serra. Automatic Analysis of DNA Microarray Images Using Mathematical Morphology. Bioinformatics, 19(5):553–562, 2003.

[4] B. Ceccarelli and G. Antoniol. A Deformable Grid-matching Approach for Microarray Images. IEEE Transactions on Image Processing, 15(10):3178–3188, 2006.

[5] G. Antoniol and M. Ceccarelli. A Markov Random Field Approach to Microarray Image Gridding. Proc. of the 17th International Conference on Pattern Recognition, pages 550–553, 2004.

[6] L. Rueda and V. Vidyadharan. A Hill-climbing Approach for Automatic Gridding of cDNA Microarray Images. IEEE Transactions on Computational Biology and Bioinformatics, 3(1):72–83, 2006.

[7] Y. Qi, A. Rolfe, K.D. MacIsaac, G. Gerber, D. Pokholok, J. Zeitlinger, T. Danford, R. Dowell, E. Fraenkel, T. S. Jaakkola, R. Young, and D. Gifford. High-resolution computational models of genome binding events. Nat Biotech, 24(8):963–970, 2006.

[8] L. Rueda. Sub-grid Detection in DNA Microarray Images. Proceedings of the IEEE Pacific-RIM Symposium on Image and Video Technology, pages 248–259, 2007.

[9] E. Zacharia and D. Maroulis. Micoarray image gridding via an evolutionary algorithm. IEEE International Conference on Image Processing, pages 1444–1447, 2008.

[10] D. Bariamis , D. Maroulis and D. Iakovidis. M^3G: Maximum Margin Microarray Gridding. BMC Bioinformatics, 11:49, 2010.

[11] L. Rueda and I. Rezaeian. A fully automatic gridding method for cdna microarray images. BMC Bioinformatics, 12:113, 2011.

[12] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10(1):57–63, 2009.

[13] A. Barski and K. Zhao. Genomic location analysis by chip-seq. Journal of Cellular Biochemistry, (107):11–18, 2009.

[14] P.J. Park. Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genetics, 10(10):669–680, 2009.

[15] M. Buck, A. Nobel, and J. Lieb. Chipotle: a user-friendly tool for the analysis of chip-chip data. Genome Biology, 6(11):R97, 2005.

[16] W. Johnson, W. Li, C. Meyer, R. Gottardo, J. Carroll, M. Brown, and X.S. Liu. Model-based analysis of tiling-arrays for chip-chip. Proceedings of the National Academy of Sciences, 103(33):12457–12462, 2006.

[17] D. Reiss, M. Facciotti, and N. Baliga. Model-based deconvolution of genome-wide dna binding. Bioinformatics, 24(3):396–403, 2008.

[18] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, , and X.S. Liu. Model-based analysis of chip-seq (macs). Genome Biology, 9(9):R137, 2008.

[19] J. Rozowsky, G. Euskirchen, R. Auerbach, Z. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. Nat Biotech, 27(1):66–75, 2009.

[20] V. Hower, S. Evans, and L. Pachter. Shape-based peak identification for chip-seq. BMC Bioinformatics, 11(81), 2010.

[21] C. Wang, J. Xu, D. Zhang, Z. Wilson, and D. Zhang. An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics, 41(1):117–129, 2008.

[22] L. Rueda. An Efficient Algorithm for Optimal Multilevel Thresholding of Irregularly Sampled Histograms. Proceedings of the 7th International Workshop on Statistical Pattern Recognition, pages 612–621, 2008.

[23] M. Luessi, M. Eichmann, G. Schuster, and A. Katsaggelos. Framework for efficient optimal multilevel image thresholding. Journal of Electronic Imaging, 18, 2009.

[24] U. Maulik and S. Bandyopadhyay. Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(12):1650–1655, 2002.

[25] G. Tuteja, P. White, J. Schug, and KH. Kaestner. Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Res, 37(17):e113, 2009.

[26] A. Valouev, D. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Meth, 5(9):829–834, 2008.

[27] C. Zang, DE. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics, 25(15):1952–1958, 2009.

[28] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo proteindna binding sites from chip-seq data. Nucleic Acids Research, 36(16):5221–5231, 2008.

Chapter 4

A New Algorithm for Finding Enriched Regions in ChIP-Seq Data

4.1 Introduction

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) is a technique that provides quantitative and genome-wide mapping of target protein binding events [1, 2]. In ChIP-Seq, a protein is first cross-linked to DNA and the fragments subsequently sheared. Following a size selection step that enriches for fragments of specified lengths, the fragments ends are sequenced, and the resulting reads are aligned to the reference genome. Detecting protein binding sites from massive sequence-based datasets with millions of short reads represents a truly bioinformatics challenge that requires considerable computational resources, in spite of the availability of programs for ChIP-chip analysis [3–6].

With the increasing popularity of ChIP-Seq technology, the demand for peak finding methods has increased the need to develop new algorithms. Although due to mapping

challenges and biases in various aspects of existing protocols, identifying peaks is not a straightforward task.

Different approaches have been proposed for detecting peaks on ChIP-Seq/RNA-Seq mapped reads. Zhang et al. presented a *model-based analysis of ChIP-Seq data* (MACS), which analyzes data generated by short read sequencers [7]. It models the length of the sequenced ChIP fragments and uses it to improve the spatial resolution of predicted binding sites. A two-pass strategy called *PeakSeq* has been presented in [8]. This strategy compensates for signals caused by open chromatin, as revealed by the inclusion of the controls. The first pass identifies putative binding sites and compensates for genomic variation in mapping the sequences. The second pass filters out sites not significantly enriched compared to the normalized control, computing precise enrichments and significance. *Tree shape Peak Identification for ChIP-Seq* (T-PIC) is a statistical approach for calling peaks that has been recently proposed in [9]. This approach is based on evaluating the significance of a robust statistical test that measures the extent of pile-up reads. Specifically, the shapes of putative peaks are defined and evaluated to differentiate between random and non-random fragment placements on the genome. Another algorithm for identification of binding sites is *site identification from paired-end sequencing* (SIPeS) [10], which can be used for identification of binding sites from short reads generated from paired-end Illumina ChIP-Seq technology.

One of the problems of the existing methods is that the locations of the detected peaks could be non-optimal. Moreover, for detecting these peaks all methods use a set of parameters that may cause variations of the results for different datasets. In the proposed method, both of these issues have been addressed by proposing a new peak finder algorithm based on *optimal multi-level thresholding* coupled with a model to find the best number of peaks based on clustering techniques for pattern recognition. The results of our experiments show

that our method can achieve a higher degree of accuracy than previously proposed peak finders while providing flexibility when applying it to different datasets.

4.2 The Peak Detection Method

4.2.1 Overview of the Method

In ChIP-Seq, a protein is first cross-linked to DNA and the fragments subsequently pruned. Then, the fragments ends are sequenced, and the resulting reads are aligned to the genome. The result of reading the alignments produces a histogram in such a way that the *x*-axis represents the genome coordinates (i.e., each bin corresponds to a single base in the genome), and the *y*-axis represents the frequency of the aligned reads in each genome coordinate. The aim is to find significant peaks corresponding to enriched regions. Each peak can be seen as homogeneous group (cluster) which is well separated from the others by means of "valleys". In that sense, the problem can be formulated as *one-dimensional clustering*. Figure 4.1 depicts the process of finding the peaks corresponding to the regions of interest for the specified protein. Each module is explained in detail in the next few sections.

4.2.2 Creating Histogram

The first step of the algorithm consists of converting the Input BED file containing the position and direction of each read to a histogram. Each read should be extended to a fragment length. The fragment length is the only parameter to be input by the user, even though the fragment length can be easily estimated from the underlying data. In practice, if enough computational resources are available, the fragment length would not be needed, since the OMT algorithm could be run directly on a whole chromosome.

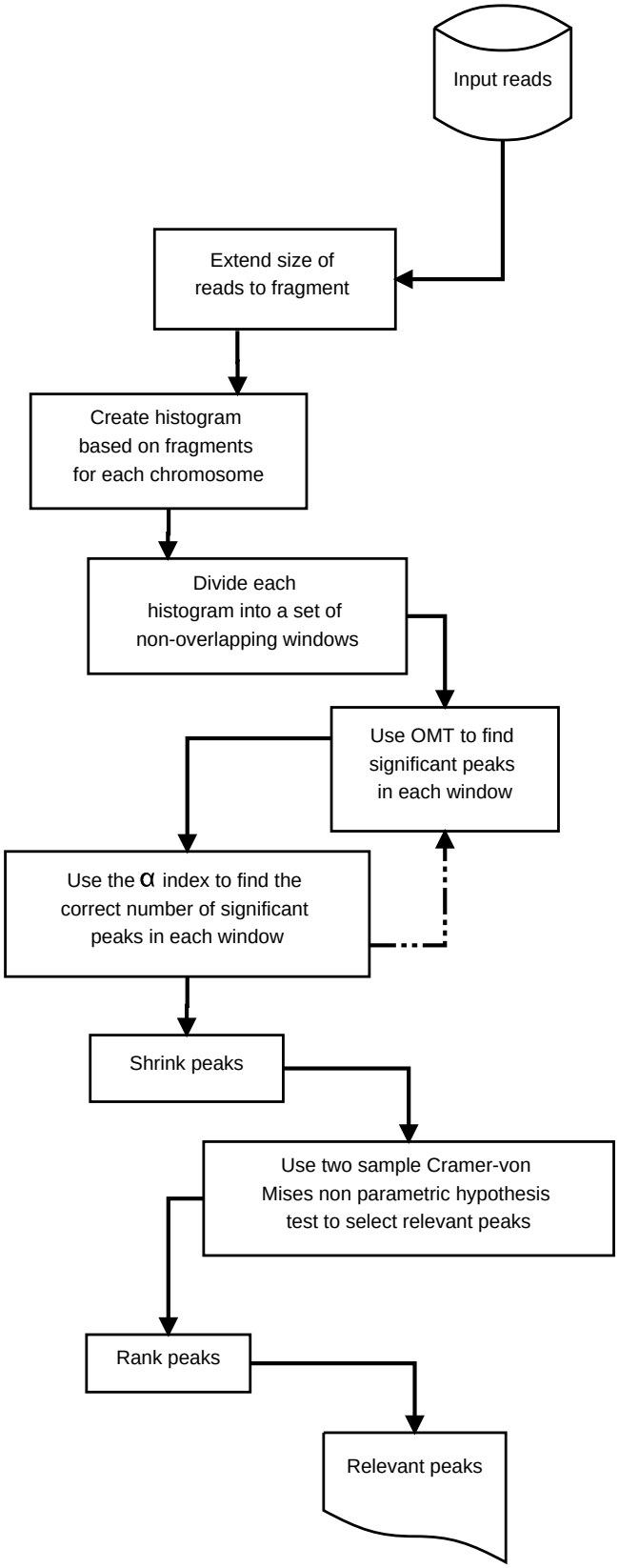

Figure 4.1: Schematic representation of the process for finding significant peaks by using OMT.

After extending each read to a fragment based on the direction of each read (forward or backward), each of them is aligned to the reference genome based on its coordinates. Afterwards, for each chromosome, separate histograms for experiment and control data are created for further processing. However, when dealing with a full chromosome, the number of bins is sufficiently large that it is rather difficult to process it all at once; this is also due to the fact that we need to find the optimal number of peaks. For this reason, a non-overlapping sliding window is used. By starting from the beginning of the chromosome, a sliding window of minimum size t is applied to the histogram and each window is analyzed separately. The sizes of the windows are not necessarily equal to prevent truncating a peak before its end. Thus, for each window, a minimum number of t bins is used and, by starting from the end of the previous window, the size of the window is increased until a zero value in the histogram is reached. We consider a minimum of $t = 3,000$ in order to ensure that a window covers at least one peak of typical size.

The aim is to obtain vectors $C_{w_i} = [c^1_{w_i}, ...c^{n_i}_{w_i}]^t$, where w_i is the i^{th} window and C_{w_i} is the vector that contains n_i thresholds which correspond to the i^{th} window.

4.2.3 Using OMT for Analyzing ChIP-Seq Data

Multi-level thresholding is one of the most widely-used techniques in different problems of signal and image processing, including segmentation, classification and object discrimination. This technique is an excellent approach for one-dimensional clustering, since it finds an optimal solution efficiently, e.g., in polynomial time. Given a histogram with frequencies or probabilities for each bin, the aim of multi-level thresholding is to divide the histogram into a number of groups (or classes) of contiguous bins in such a way that a criterion is optimized. In peak detection, we create a histogram based on fragments (reads). The his-

togram is then processed (see below) to obtain the optimal thresholding that will determine the locations of the peaks.

Consider a histogram H, an ordered set $\{1,2,\ldots,n-1,n\}$, where the ith value corresponds to the ith bin and has a probability, p_i. The histogram, H, can be obtained by counting the number of aligned reads. We also consider a threshold set T, defined as an ordered set $T = \{t_0, t_1, \ldots, t_k, t_{k+1}\}$, where $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} = n$ and $t_i \in \{0\} \cup H$. The problem of multi-level thresholding consists of finding a threshold set, T^*, in such a way that a function $f : H^k \times [0,1]^n \to \mathbb{R}^+$ is maximized/minimized. Using this threshold set, H is divided into $k+1$ classes: $\zeta_1 = \{1,2,\ldots,t_1\}$, $\zeta_2 = \{t_1+1, t_1+2, \ldots, t_2\}$, ..., $\zeta_k = \{t_{k-1}+1, t_{k-1}+2, \ldots, t_k\}$, $\zeta_{k+1} = \{t_k+1, t_k+2, \ldots, n\}$. A few criteria for multi-level thresholding have been proposed [11]. We consider the between-class variance criterion, which aims to maximize the inter-class separability of the classes, and which is proportional to:

$$\Psi_{\mathrm{BC}}(T) = \sum_{j=1}^{k+1} \omega_j \mu_j^2 \tag{4.1}$$

where $\omega_j = \sum_{i=t_{j-1}+1}^{t_j} p_i$, $\mu_j = \frac{1}{\omega_j} \sum_{i=t_{j-1}+1}^{t_j} i p_i$.

A dynamic programming algorithm for *optimal* multi-level thresholding was proposed in our previous work [11], which is an extension for irregularly sampled histograms. For this, the criterion has to be decomposed as a sum of terms as follows:

$$\Psi(T_{0,m}) = \Psi(\{t_0, t_1, \ldots, t_m\}) \triangleq \sum_{j=1}^{m} \psi_{t_{j-1}+1, t_j}\,, \tag{4.2}$$

where $1 \leq m \leq k+1$ and the function $\psi_{l,r}$, where $l \leq r$, is a real, positive function of $p_l, p_{l+1}, \ldots, p_r$, $\psi_{l,r} : H^2 \times [0,1]^{l-r+1} \to \mathbb{R}^+ \cup \{0\}$. If $m = 0$, then $\Psi(\{t_0\}) = \psi_{t_0,t_0} =$

$\psi_{0,0} = 0$. Full details of the thresholding algorithm can be found in [11]. The optimal thresholding is the one that maximizes the between-class variance (or, conversely, it minimizes the within-class variance). The algorithm runs in $O(kn^2)$ for a histogram of n bins, and has been further improved to achieve linear complexity for some criteria, i.e. $O(kn)$, by following the approach of [12].

4.2.4 Automatic Detection of the Best Number of Peaks

Finding the correct number of peaks (the number of regions in each window) is one of the most challenging issues. This stage is crucial in order to fully automate the whole process. For this, we need to determine the correct number peaks prior to applying the multi-level thresholding method. This is found by using an index of validity derived from clustering techniques. We have recently proposed the $\alpha(K)$ index [13], which is the result of a combination of a simple index, $A(K)$, and the well-known I index [14] as follows:

$$\alpha(K) = \sqrt{K}\frac{I(K)}{A(K)} = \frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{4.3}$$

where $E_K = \Sigma_{i=1}^{K}\Sigma_{k=1}^{n_i} p_k||k - z_i||$, $D_K = \underbrace{\max}_{i,j=1}^{K}||z_i - z_j||$, n is the total number of bins in the window, K is the number of clusters, z_k is the center of the kth cluster, t_i is the ith threshold found by optimal multilevel thresholding and $p(t_i)$ is the corresponding number of reads in the histogram.

For maximizing $I(K)$ and minimizing $A(K)$, the value of $\alpha(K)$ must be maximized. Thus, the best number of thresholds K^* based on the α index is given by:

$$K^* = \underset{1\leq K\leq\delta}{\operatorname{argmax}}\,\alpha(K) = \underset{1\leq K\leq\delta}{\operatorname{argmax}}\,\frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{4.4}$$

To find the optimal number of clusters (thresholds), we compute and compare values of $\alpha(K)$ over all possible numbers of clusters (or thresholds) from 2 to $\sqrt{n}/2$, where n is the size of window. The one with the maximum value of $\alpha(K)$ is the best number of clusters (thresholds).

4.2.5 Relevant Peaks Selection

After finding the locations of the detected peaks, in a two step process, significant peaks are selected. In the first step, the effective area of each peak is found by shrinking the peak. For this, by starting from the summit of the peak, we move to left and right separately until we reach a zero number of reads. In the second step, the two sample Cramer-von Mises non parametric hypothesis test [15], with $\alpha = 0.01$, is used to accept/reject peaks based on the comparison between experiment and control histograms corresponding to each peak. The reason for using the Cramer-von Mises test is that it can detect differences in distributions with higher statistical power than the commonly used two-sample Kolmogorov-Smirnov test [15]. Finally, those peaks which are accepted by the Cramer-von Mises test are ranked and returned as the final relevant peaks.

4.3 Experimental Results

To evaluate the proposed model, we have used various datasets, including the *FoxA1* dataset [7] which contains experiment and control samples of 24 chromosomes, and four transcription factors (with a total of 6 antibodies) for *Drosophila melanogaster* using published data from the Eisen lab [16] (available at the NCBI GEO database [17] , accession no. GSE20369). As in [9], the experiment and control histograms were generated separately

by extending each mapped position (read) into an appropriately oriented fragment, and then joining the fragments based on their genome coordinates. The final histogram was generated by subtracting the control from the experiment histogram. To find significant peaks, we used a non-overlapping window whose initial size is 3,000bp. To avoid truncating peaks in boundaries, each window is extended until the value of the histogram at the end of the window becomes zero. Figure 4.2 shows three detected regions for chromosomes 1,17 and 20 respectively, and their corresponding base pair coordinates in the FoxA1 dataset. It is clear from the plots that the peaks contain a very large number of reads, and then these regions are quite likely to represent binding sites, open reading frames or other biomarkers.

Computing the enrichment score for each method proceeds as follows. Random intervals from the genome are created by selecting the same number of intervals with the same lengths from each chromosome as in the called peaks but with random starting locations. Then, the number of occurrences of the binding motif in the called peaks and the random intervals are counted. The enrichment score is the ratio of the number of occurrences in the called peaks divided by the number of occurrences in the random intervals.

4.3.1 Comparison with Other Methods for ChIP-Seq Analysis

Table 4.1 shows a comparison between OMT and two recently proposed methods, MACS [7] and T-PIC [9]. As shown in the table, the number of significant peaks detected by OMT is higher than those of the other two methods. This implies that OMT is able to find significant peaks that are not detected by the other two methods. Also, the enrichment ratio for OMT is far higher than MACS and higher than T-PIC. Moreover, the average size of the peaks is smaller than the other two methods which implies that OMT is able to detect significant peaks more precisely.

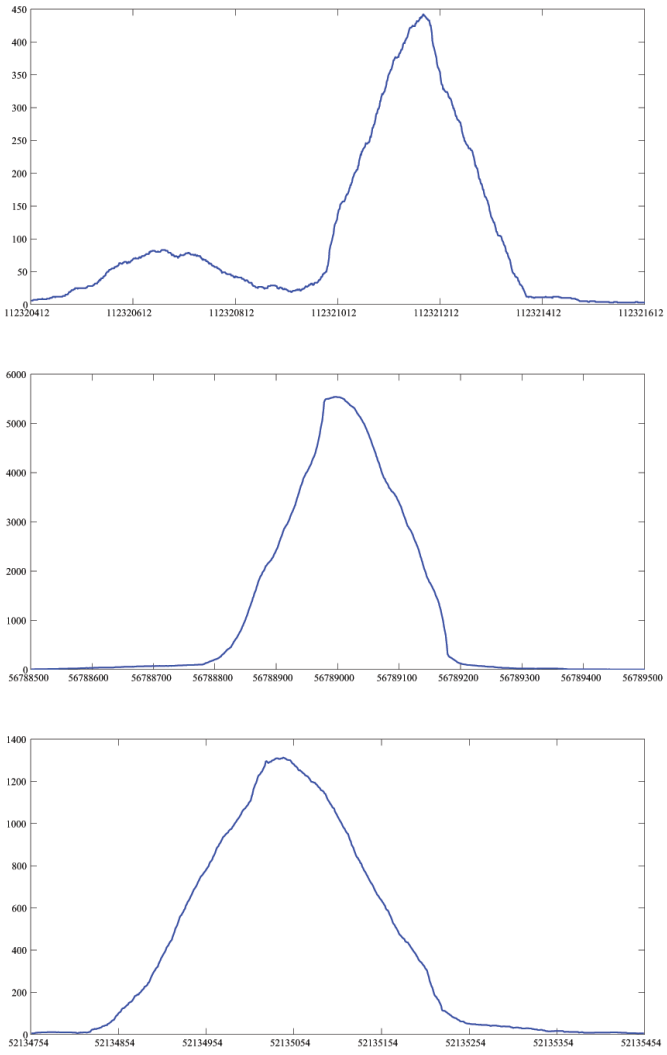

Figure 4.2: Three detected regions from the FoxA1 dataset for chromosomes 1 (top), 17 (middle) and 20 (bottom). The *x*-axis corresponds to the genome position in bp and the *y*-axis corresponds to the number of reads.

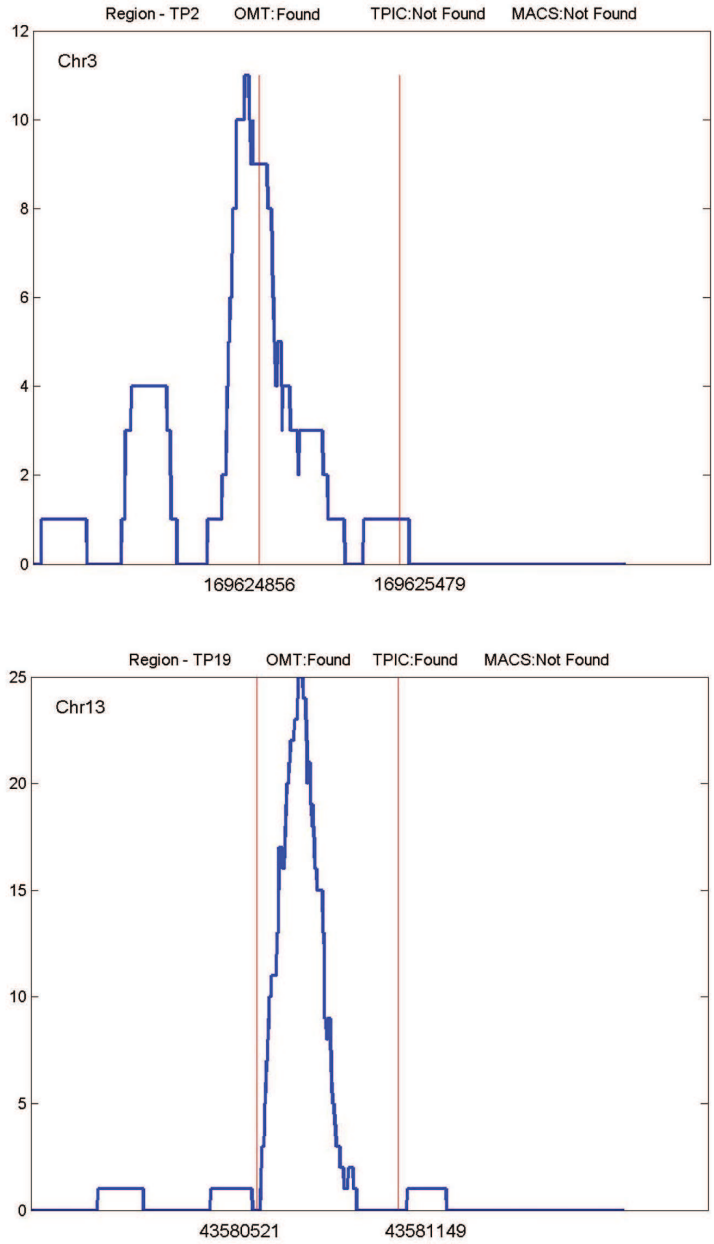

Figure 4.3: Two true positive regions in chromosomes 3 and 13 of FoxA1 dataset. The *x*-axis corresponds to the genome position in bp and the *y*-axis corresponds to the number of reads. Both peaks are detected by OMT but only the bottom one is detected by T-PIC, while none of them is detected by MACS.

Table 4.1: Comparison between OMT and two recently proposed methods, MACS and T-PIC, based on the number and mean length of detected peaks, and enrichment score.

Dataset	Method of Comparison	OMT	T-PIC	MACS
FoxA1	Detected peaks	20,032	17,619	13,639
	Mean length of peaks	306	510	394
	Enrichment ratio	2.62	2.54	1.68
CAD	Detected peaks	12,825	8,119	4,652
	Mean length of peaks	449	986	1,596
	Enrichment ratio	1.08	0.84	0.96
GT	Detected peaks	4,526	3,553	2,904
	Mean length of peaks	687	912	1,204
	Enrichment ratio	3.42	2.33	1.54
HB1	Detected peaks	8,356	5,481	6,857
	Mean length of peaks	253	991	1,124
	Enrichment ratio	1.93	1.69	1.62
HB2	Detected peaks	5,782	4,337	3,928
	Mean length of peaks	235	1,092	1,248
	Enrichment ratio	1.96	1.63	1.59
KR1	Detected peaks	15,324	11,891	9,804
	Mean length of peaks	350	872	1,635
	Enrichment ratio	2.14	1.75	1.54
KR2	Detected peaks	15,476	11,717	9,652
	Mean length of peaks	347	863	1,597
	Enrichment ratio	2.23	1.78	1.58

Also, Table 4.2 shows a summary of prediction for the proteins by each method. Each value shows the percentage of detected peaks by each method which are also detected by the other methods. For example, OMT detects 90.1% of the peaks detected by MACS while MACS only detects 59.7% of significant peaks detected by OMT in the FoxA1 dataset. This demonstrates the wide spectrum and specificity of the proposed OMT algorithm.

A conceptual comparison of OMT with other proposed algorithms based on their features is shown in Table 4.3. As shown in the table, the other algorithms require some

Table 4.2: Percentage of common peaks detected by each method in the comparison, related to each protein of interest.

		OMT	T-PIC	MACS
FoxA1	OMT	100	78.8	59.7
	T-PIC	99.4	100	64.4
	MACS	90.1	83.6	100
		OMT	T-PIC	MACS
CAD	OMT	100	50.8	28.8
	T-PIC	79.3	100	62.2
	MACS	98.1	95.5	100
		OMT	T-PIC	MACS
GT	OMT	100	50.5	21.2
	T-PIC	65.1	100	57.7
	MACS	78.9	85.1	100
		OMT	T-PIC	MACS
HB1	OMT	100	49.6	42.7
	T-PIC	79.6	100	69.2
	MACS	84.1	90.7	100
		OMT	T-PIC	MACS
HB2	OMT	100	63.1	43.2
	T-PIC	81.7	100	68.9
	MACS	88.4	91.3	100
		OMT	T-PIC	MACS
KR1	OMT	100	73.1	50.5
	T-PIC	84.6	100	66.6
	MACS	97.4	98.1	100
		OMT	T-PIC	MACS
KR2	OMT	100	73.6	60.2
	T-PIC	84.4	100	66.9
	MACS	97.1	97.7	100

parameters to be set by the user based on the particular data to be processed, including *p*-values, *m*-fold, window length, among others. OMT is the algorithm that requires the smallest number of parameters. Only the average fragment length is needed. However, the

Table 4.3: Conceptual comparison of recently proposed methods for *ChIP – Seq* data.

Method	**Peak selection criteria**	**Peak ranking**	**Parameters**
GLITR	*n*: Classification by height and relative enrichment	Peak height and fold enrichment	Target FDR, number nearest neighbors for clustering
MACS	local region Poisson *p*-value	*p*-value	*p*-value threshold, tag length, *m*-fold for shift estimate
PeakSeq	Local region binomial *p* value	*q* value	Target FDR
Quest v2.3	height threshold, background ratio	*q* value	KDE bandwidth, peaks height, sub-peak valley depth, ratio to background
SICER v1.02	*p* value from random background model, enrichment relative to control	*q* value	Window length, gap size, FDR (with control) or *E*-Value (no control)
SiSSRs v1.4	$N^+ - N^-$ sign change, $N^+ + N^-$ threshold in region	*p* value	FDR, $N^+ + N^-$ threshold
T-PIC	local height threshold	*p*-value	average fragment length, significance *p*-value, minimum length of interval
OMT	number of ChIP reads minus control reads in window	*p*-value	average fragment length

fragment length could be easily estimated from the underlying data, if enough computational resources were available, the fragment length would not be needed, since the OMT algorithm could be run directly on the whole chromosome.

4.3.2 Biological Validation

We have also biologically validated the peaks detected by OMT on the results of independent qPCR experiments for the FoxA1 protein. For this, we considered 25 true positives and 7 true negatives (regions) reported in [18]. The results of other two well-known meth-

ods, T-PIC and MACS, are included in the comparison. Table 4.4 shows the result of this biological validation on each method. As the other two methods, OMT has been able to reject all true negatives. Although OMT finds a larger number of regions, OMT shows very high sensitivity, finding more true positives than T-PIC and MACS. As an example, two true positive regions in chromosomes 3 and 13 of FoxA1 are shown in Figure 4.3. Both peaks are detected by OMT but only the bottom one is detected by T-PIC and none of them is detected by MACS.

An issue that deserves attention is the fact that some true positives found by qPCR show very low peaks in the CHIP-Seq experiments. We have visually inspected all true positive regions in the CHIP-Seq experiments, and found that 10 out of 25 of these regions have a maximum number of reads less than 5. This indicates that the CHIP-Seq experiment basically "disagrees" with qPCR on these genomic regions of interest. Then, it would not be up to the peak finding algorithm to detect these true positives. The proposed algorithm, OMT, however, finds all other true positives.

Table 4.4: Comparison of OMT, MACS and T-PIC, based on the number of true positive (TP) and true negative (TN) detected peaks.

	OMT	**T-PIC**	**MACS**
TP	15	13	12
TN	0	0	0

4.4 Discussion and Conclusion

We have presented a multi-level thresholding algorithm that can be applied to an efficient analysis of ChIP-Seq data to find significant peaks. OMT can be applied to high-throughput

next generation sequencing data with different characteristics, and allows us to detect significant regions on ChIP-Seq data. OMT has been shown to be sound and efficient in experiments and has the ability to be applied to various types of next generation sequencing data. When compared to other recently proposed methods, OMT shows to be more accurate, and use fewer parameters.

The proposed method offers new avenues for future research. One of these is to apply the OMT algorithm on the whole chromosome instead of using a set of windows as a way to reduce the number of parameters. Also, using other indices of validity and thresholding criteria could increase the accuracy of the method. Moreover, the proposed method could be applied on other datasets and proteins of interest. All these are issues that we are currently investigating.

Bibliography

[1] A. Barski and K. Zhao, Genomic location analysis by chip-seq. Journal of Cellular Biochemistry. no. 107, pp. 11–18, 2009.

[2] P. Park, Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genetics. vol. 10, no. 10, pp. 669–680, 2009.

[3] M. Buck, A. Nobel, and J. Lieb, Chipotle: a user-friendly tool for the analysis of chip-chip data. Genome Biology. vol. 6, no. 11, p. R97, 2005.

[4] W. Johnson, W. Li, C. Meyer, R. Gottardo, J. Carroll, M. Brown, and X. Liu, Model-based analysis of tiling-arrays for chip-chip. Proceedings of the National Academy of Sciences. vol. 103, no. 33, pp. 12 457–12 462, 2006.

[5] Y. Qi, A. Rolfe, K. MacIsaac, G. Gerber, D. Pokholok, J. Zeitlinger, T. Danford, R. Dowell, E. Fraenkel, T. S. Jaakkola, R. Young, and D. Gifford, High-resolution computational models of genome binding events. Nature Biotechnology. vol. 24, no. 8, pp. 963–970, 2006.

[6] D. Reiss, M. Facciotti, and N. Baliga, Model-based deconvolution of genome-wide dna binding. Bioinformatics. vol. 24, no. 3, pp. 396–403, 2008.

[7] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, , and X. Liu, Model-based analysis of chip-seq (macs). Genome Biology. vol. 9, no. 9, p. R137, 2008.

[8] J. Rozowsky, G. Euskirchen, R. Auerbach, Z. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. Gerstein, Peakseq enables systematic scoring of chip-seq experiments relative to controls. Nature Biotechnology. vol. 27, no. 1, pp. 66–75, 2009.

[9] V. Hower, S. Evans, and L. Pachter, Shape-based peak identification for chip-seq. BMC Bioinformatics. vol. 11, no. 81, 2010.

[10] C. Wang, J. Xu, D. Zhang, Z. Wilson, and D. Zhang, An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics. vol. 41, no. 1, pp. 117–129, 2008.

[11] L. Rueda. An Efficient Algorithm for Optimal Multilevel Thresholding of Irregularly Sampled Histograms. Proceedings of the 7th International Workshop on Statistical Pattern Recognition. pp. 612–621, 2008.

[12] M. Luessi, M. Eichmann, G. Schuster, and A. Katsaggelos, Framework for efficient optimal multilevel image thresholding. Journal of Electronic Imaging. vol. 18, 2009.

[13] L. Rueda and I. Rezaeian, A fully automatic gridding method for cdna microarray images. BMC Bioinformatics. vol. 12, p. 113, 2011.

[14] U. Maulik and S. Bandyopadhyay. Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Trans. on Pattern Analysis and Machine Intelligence. vol. 24, no. 12, pp. 1650–1655, 2002.

[15] T. W. Anderson. On the Distribution of the Two-Sample Cramer-von Mises Criterion. Ann. Math. Statist.. vol. 33, pp. 1148–1159, 1962.

[16] RK. Bradley, XY. Li, C. Trapnell, S. Davidson, L. Pachter, HC. Chu, LA. Tonkin, MD. Biggin, MB. Eisen. Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. PLoS Biol. vol. 8, no. 3, p. e1000343, 2010.

[17] T. Barrett, D. Troup, S. Wilhite, P. Ledoux, and D. Rudnev, Ncbi geo: archive for high-throughput functional genomic data. nucl acids res. Journal of Cellular Biochemistry. vol. 37, pp. 885–890, 2009.

[18] M. Lupien, J. Eeckhoute, C. A. Meyer, Q. Wang, Y. Zhang, W. Li, J. S. Carroll, X. S. Liu, and M. Brown, FoxA1 Translates Epigenetic Signatures into Enhancer-driven Lineage-specific Transcription. Cell. vol. 132, no. 6, pp. 958–970, 2008.

Chapter 5

CMT: A Constrained Multi-level Thresholding Approach for ChIP-Seq Data Analysis

5.1 Introduction

Determining the interaction between a protein and DNA to regulate gene expression is a very important step toward understanding of many biological processes and disease states. ChIP-Seq is one of the techniques used for finding regions of interest in a specific protein that interacts with DNA [1–7]. The main process consists of Chromatin-immunoprecipitation (ChIP) followed by sequencing of the immuno-precipitated DNA with respect to the reference genome. In the first step, chromatin is isolated from cells or tissues and then fragmented. After pruning, the fragments are sequenced and aligned to the reference genome. These aligned fragments produce a histogram in such a way that the x-axis represents the genome coordinates and the y-axis represents the frequency of the aligned fragments in

each genome coordinate.

Detecting protein binding sites from large sequence-based datasets with millions of short reads represents a challenging bioinformatics problem that requires considerable computational resources, despite the availability of a wide range of tools for ChIP-chip data analysis [8–11]. The growing popularity of ChIP-Seq technology has increased the need to develop new algorithms for peak finding. Due to mapping challenges and biases in various aspects of the existing protocols, identifying relevant peaks is not a straightforward task.

Different approaches have been proposed for detecting peaks on ChIP-Seq and RNA-Seq mapped reads. Zhang *et al.* presented a *model-based analysis of ChIP-Seq data* (MACS), which analyzes the data generated by short read sequencers [12]. MACS models the length of the sequenced ChIP fragments and uses it to improve the spatial resolution of predicted binding sites. A two-pass strategy called *PeakSeq* has been presented in [13]. This strategy compensates for signals caused by open chromatin, as revealed by the inclusion of the controls. The first pass identifies putative binding sites and compensates for genomic variation in mapping the fragment sequences. The second pass filters out sites not significantly enriched compared to the normalized control, computing precise enrichments and significance of each detected peak. *Tree shape Peak Identification for ChIP-Seq* (T-PIC) is a statistical approach for calling peaks in ChIP-Seq data [14]. This approach is based on evaluating the significance of a robust statistical test that measures the extent of pile-up reads. Specifically, the shapes of putative peaks are defined and evaluated to differentiate between random and non-random fragment placements on the genome. Another algorithm for detecting relevant peaks is *site identification from paired-end sequencing* (SIPeS) [15], which can be used for identification of binding sites from short reads generated from paired-end Illumina ChIP-Seq technology. Qeseq is another method for analyzing the aligned sequence

reads from ChIP-Seq data and identifying enriched regions [4]. The algorithm consists of three main modules: relative enrichment estimation, cluster detection and filtering possible artifacts. It cycles between its first two modules by removing detected clusters and evaluating enrichment in the rest of signal. In the last step, a filter module is used to remove artifacts from the results.

One of the downsides of the existing methods is that they try to find all the enriched regions regardless of their length. These regions can be grouped by their length. For example, histone modification sites normally have a length of 50 to 60 kbp, while some other regions of interest like exons have a much smaller length of around 100 bp. Using these methods, there is no way to focus on regions with a specific length and all of the relevant peaks should be detected first. This is a time consuming task that forces the model to process all possible regions. To deal with this issue, *constrained multi-level thresholding* (CMT) is proposed in this paper. Using CMT, we are able to search a specific region with a certain length which consequently increases the performance of the model. CMT is also able to target as many regions as the other methods simply by increasing the range for minimum and maximum lengths of the regions. The minimum and maximum lengths of the regions can be adjusted by the user based on their needs. The results of the experiments show that the proposed model is able to achieve a higher degree of accuracy than the previously proposed methods.

5.2 Results

To evaluate the proposed model, we have used various datasets. The first dataset is *FoxA1* [12] which contains experiment and control samples of 24 chromosomes. The FoxA1 protein is known to cooperatively interact with estrogen receptor in breast cancer cells [16, 17]. We consider another six datasets which belong to four transcription factors (with a total

of 6 antibodies) for *Drosophila melanogaster* using published data from the Eisen lab [18] (available at the NCBI GEO database [19], accession no. GSE20369). These four transcription factors, namely Hunchback (HB), Krppel (KR), Giant (GT) and Caudal (CAD), have been obtained by immunoprecipitating binding regions with affinity purified rabbit polyclonal antibodies raised against the *D. melanogaster* versions of the key A-P regulators. The other dataset is a genome-wide map of the *H3K4ac* antibody with ability to covalent acetylations in histone [20], which occur mainly at the N-terminal tails of the histone, and that can affect transcription of genes.

As in [14], the experiment and control histograms were generated separately by extending each mapped position (read) into an appropriately oriented fragment, and then joining the fragments based on their genome coordinates. We compare CMT, MACS [12] and T-PIC [14]. Figure 5.1 shows a typical region detected in chromosome 1 by CMT, MACS and T-PIC along with the corresponding base pair coordinates in the FoxA1 dataset. As shown in the plot, all three methods found the position of the peak accurately.

Computing the enrichment score for each method proceeds as follows. Random intervals from the genome are created by selecting the same number of intervals with the same lengths from each chromosome as in the called peaks but with random starting locations. Then, the number of occurrences of the binding motif in the called peaks and the random intervals are counted. Table 5.1 shows the binding motifs corresponding to each dataset. The motifs for CAD, GT, HB, and KR datasets have been obtained from [21], while the binding motif for the FoxA1 dataset has been obtained from [22]. The enrichment score is the ratio of the number of occurrences in the called peaks divided by the number of occurrences at random intervals.

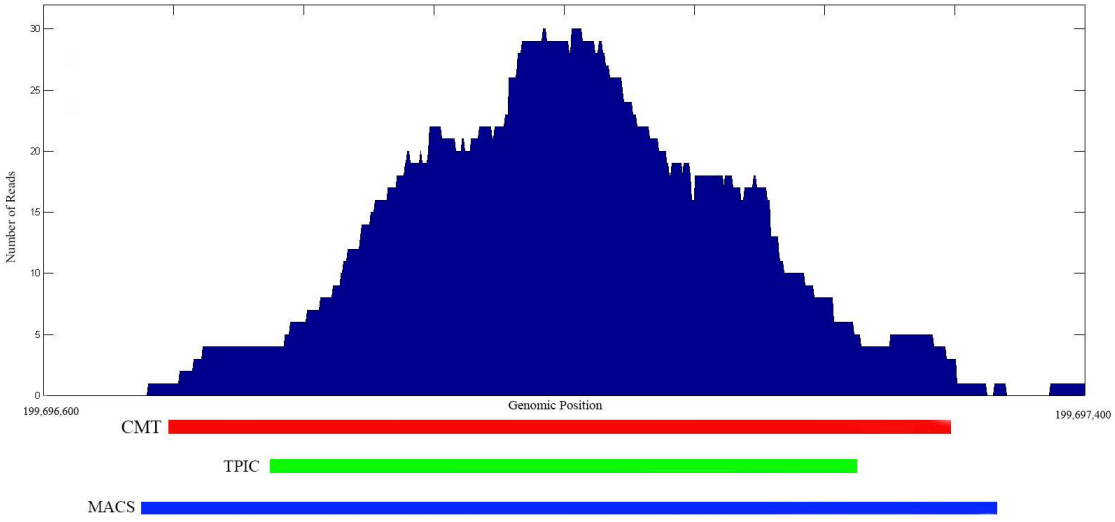

Figure 5.1: A detected region from the FoxA1 dataset for chromosome 1. The *x*-axis corresponds to the genome position in bp and the *y*-axis corresponds to the number of reads.

Table 5.1: Binding motifs corresponding to each dataset.

FoxA1	**CAD**	**GT**	**HB**	**KR**
TGCATG	TTTATTG , TTTATGA	TTACGTAA	TTTTTT	GANGGGT, AANGGGT

5.2.1 Comparison with Other Methods

Figure 5.2 shows the Venn diagram corresponding to each dataset for all three methods. We consider a peak detected by two methods to be overlapped, if the summit of the peak is located in the detected region by both of the methods. For example, Figure 5.1 shows an overlapping region detected by all three methods. In the FoxA1, KR1 and KR2 datasets, the number of regions selected by CMT is relatively higher than those of the other methods. These regions have mostly a small footprint which has not been detected by T-PIC or MACS. In the GT dataset, the numbers of regions detected by CMT and T-PIC are comparable. Interestingly, MACS detected only one fourth of the peaks detected by two other methods. In the HB1 and HB2 datasets, this case is inverted and MACS detects more regions than T-PIC and CMT. In the H3K4ac dataset, while the number of histone modification sites using CMT and T-PIC are comparable, we were not able to obtain any regions with minimum size of 2,000bp using MACS even after various parameter adjustments. Also, Table 5.2 shows a summary of prediction for the proteins found by each method. Each value represents the percentage of peaks detected by each method which are also detected by the other methods. For example, CMT detects 95.1% of the peaks detected by MACS, while MACS only detects 50.8% of significant peaks detected by CMT in the FoxA1 dataset. This demonstrates the wide spectrum and specificity of the proposed CMT algorithm. As mentioned earlier, since MACS was not able to detect wide peaks in H3K4ac dataset, the corresponding cells in Table 5.2 have been marked with *N/A* (not applicable).

Table 5.3 shows a comparison between the three peak finding algorithms considered in this paper. As shown in the table, in terms of enrichment ratio CMT is the best among these methods, overall. The difference between CMT, w.r.t. MACS and T-PIC is considerable in some datasets such as GT, HB1 and HB2. On the other hand, the average size of the

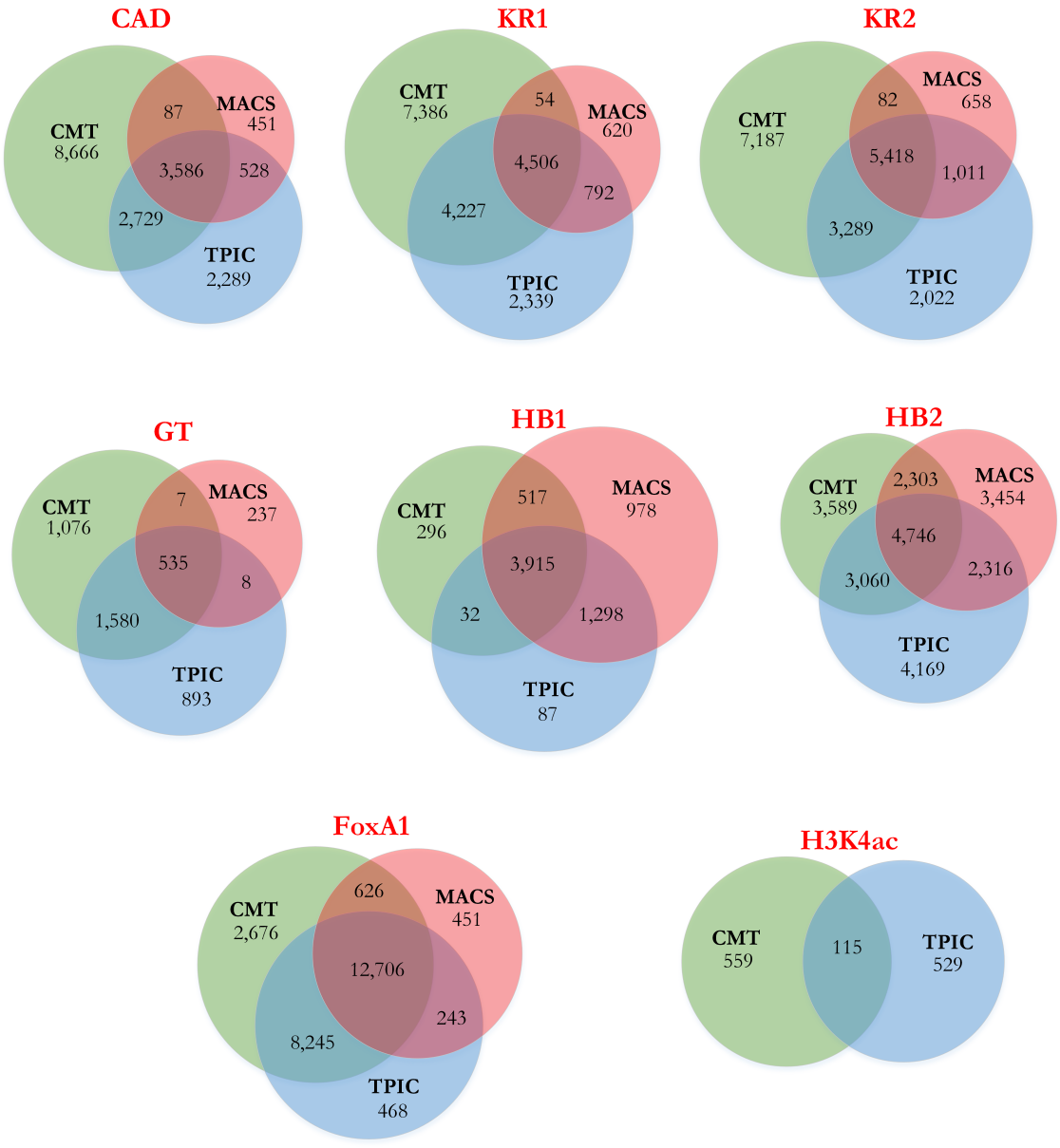

Figure 5.2: Venn diagrams corresponding to all datasets. Each Venn diagram shows the number of detected regions by CMT, MACS and T-PIC in each dataset along with the number of detected regions by each pair and all aformentioned methods.

Table 5.2: Percentage of common peaks detected by each method included in the comparison and related to each protein of interest.

		CMT	T-PIC	MACS
FoxA1	CMT	100	79.8	50.8
	T-PIC	96.7	100	59.8
	MACS	95.1	92.3	100
CAD	CMT	100	41.9	24.4
	T-PIC	72.7	100	47.3
	MACS	79.0	88.4	100
GT	CMT	100	66.1	16.9
	T-PIC	70.1	100	18.0
	MACS	68.9	69.0	100
HB1	CMT	100	82.9	93.1
	T-PIC	74.0	100	97.8
	MACS	66.1	77.7	100
HB2	CMT	100	85.3	64.2
	T-PIC	73.4	100	55.6
	MACS	66.7	67.1	100
KR1	CMT	100	54.0	28.2
	T-PIC	73.6	100	44.7
	MACS	76.4	88.7	100
KR2	CMT	100	54.5	34.4
	T-PIC	74.2	100	54.8
	MACS	76.7	89.6	100
H3K4ac	CMT	100	16.1	N/A
	T-PIC	16.7	100	N/A
	MACS	N/A	N/A	N/A

peaks is relatively smaller than those of the other two methods, which implies that CMT is able to detect significant peaks more precisely. This helps determine the actual footprint of a binding site accurately. We do not report the enrichment scores for the H3K4ac dataset, since the binding motifs for this dataset are not reported in [20]. In another comparison, using the FoxA1 dataset, we evaluate the enrichment score of those peaks that have been detected by one of the methods and missed by the other two. Table 5.4 shows the average size and enrichment score of CMT, MACS and T-PIC.

A conceptual comparison of CMT and other peak finding methods based on their features is shown in Table 5.5. As shown in the table, different algorithms require different sets of parameters for processing the data, including *p*-value, *m*-fold, window length, among others. CMT gives users the ability to fine tune the procedure based on their needs. Including the minimum and maximum range for regions of interest helps the procedure target regions within a specific range easily. It also boosts CMT to detect very small (or very large regions, depending on the parameters settings) more than T-PIC and MACS, as shown in Figure 5.2, where most of the peaks have a small footprint. This makes the peak detection process rather difficult for other methods. CMT overcomes this problem by using the specified ranges for minimum and maximum size of the target regions and scan the histogram with more emphasis on peaks within the specified range.

To compare the prediction specificity of these three methods, we swapped the ChIP and control samples, and calculated the false discovery rate (FDR) of each of these methods as follows:

$$FDR = \frac{No.\ control\ peaks}{No.\ of\ experiment\ peaks} \tag{5.1}$$

For example, if we have 100 peaks selected and by swapping the experiment and control

Table 5.3: Peak number, length and score comparison. Comparison between CMT, MACS and T-PIC based on the number and mean length of detected peaks and enrichment score.

Dataset	**Method of Comparison**	**CMT**	**T-PIC**	**MACS**
FoxA1	Mean length of peaks	277	303	373
	Enrichment ratio	2.39	2.42	1.83
CAD	Mean length of peaks	476	818	507
	Enrichment ratio	0.92	0.88	0.93
GT	Mean length of peaks	303	866	194
	Enrichment ratio	4.21	1.98	3.02
HB1	Mean length of peaks	365	920	429
	Enrichment ratio	2.03	1.57	1.80
HB2	Mean length of peaks	343	891	228
	Enrichment ratio	2.11	1.56	1.99
KR1	Mean length of peaks	517	728	492
	Enrichment ratio	1.91	1.83	1.95
KR2	Mean length of peaks	513	737	500
	Enrichment ratio	1.94	1.75	2.10

Table 5.4: Length and enrichement score comparison. Comparison between CMT, MACS and T-PIC the average length of detected peaks and enrichment score on FoxA1 dataset.

	CMT	**T-PIC**	**MACS**
Mean length of peaks	220	421	337
Enrichment ratio	2.74	2.92	1.67

samples and using the same parameters we obtain 30 peaks, then the FDR would be 30%. Figure 5.3 shows the comparison between CMT, MACS and T-PIC on the FoxA1 dataset based on the false discovery rate (FDR) and the number of selected peaks. As shown in the figure, while CMT and MACS act similarly, T-PIC falls behind with its higher FDR rate. There is a clear advantage for CMT in finding the top 1,000 regions, while from the 1,000 to 10,000 top regions, MACS yields a slightly lower FDR rate. Due to possible background noise in the data and also because the size of regions are relatively small, CMT is able to find peaks with lower FDR than T-PIC and MACS when we target a small subset of regions with high enrichment level.

From another perspective, we compared the true positive (TP) and false positive (FP) rates for each method. Figure 5.4 shows the ROC curve for CMT, T-PIC and MACS on the FoxA1 dataset. Also, Table 5.6 shows the corresponding area under curve (AUC) values. As shown in the plot and the table, CMT, again, performs better than the MACS and T-PIC.

5.2.2 Analysis of Genomic Features

We have also biologically validated the peaks detected by CMT on the results of independent qPCR experiments for the FoxA1 protein. We consider 25 true positives and 7 true negatives (regions) reported in [23]. The results of the other two well-known methods, T-PIC and MACS, are included in the comparison. Table 5.7 shows the results of this bio-

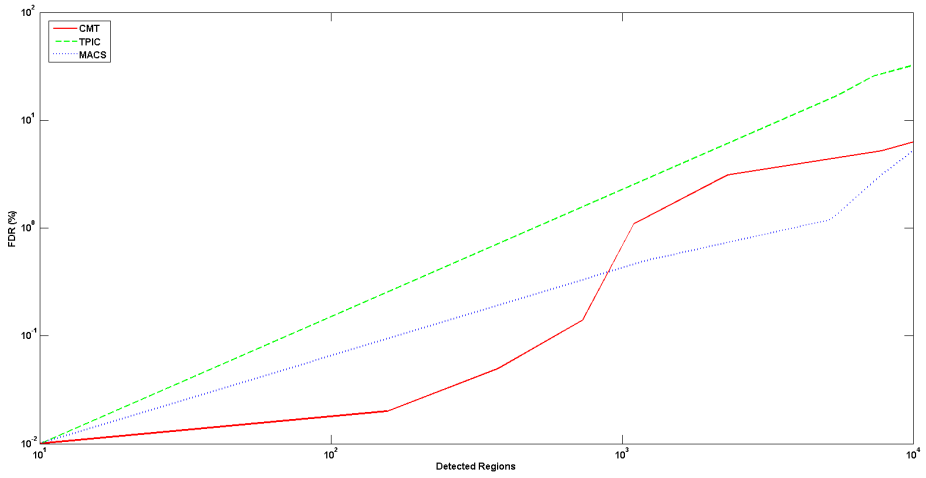

Figure 5.3: Comparison between CMT, MACS and T-PIC based on the FDR rate and number of peaks.

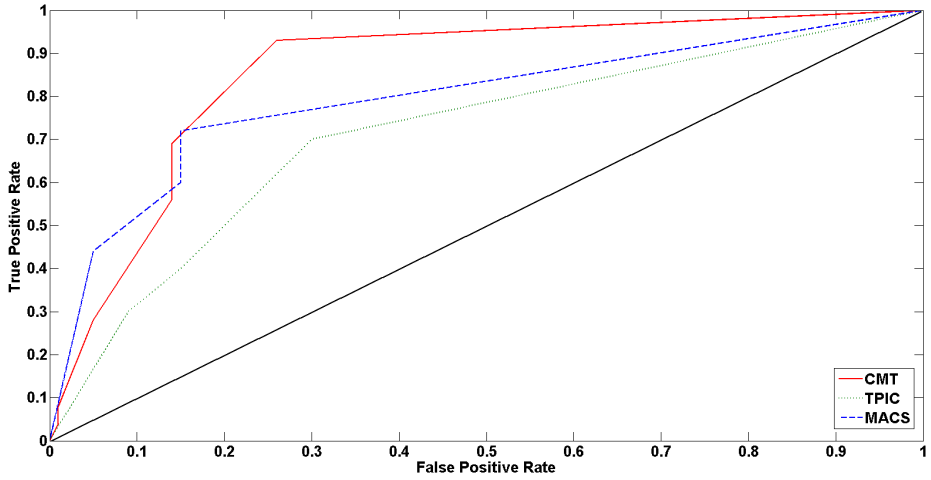

Figure 5.4: ROC curve corresponding to CMT, T-PIC and MACS.

Table 5.5: Conceptual comparison of recently proposed methods for finding peaks in ChIP-Seq data.

Method	**Peak selection criteria**	**Peak ranking**	**Parameters**
GLITR	*n*: Classification by height and relative enrichment	Peak height and fold enrichment	Target FDR, number of nearest neighbors for clustering
MACS	Local region Poisson *p*-value	*p*-value	*p*-value threshold, tag length, *m*-fold for shift estimate
PeakSeq	Local region binomial *p* value	*q* value	Target FDR
Quest v2.3	Height threshold, background ratio	*q* value	KDE bandwidth, peaks height, sub-peak valley depth, ratio to background
SICER v1.02	*p* value from random background model, enrichment relative to control	*q* value	Window length, gap size, FDR (with control) or *E*-Value (no control)
SiSSRs v1.4	$N^+ - N^-$ sign change, $N^+ + N^-$ threshold in region	*p* value	FDR, $N^+ + N^-$ threshold
T-PIC	Local height threshold	*p*-value	average fragment length, significance *p*-value, minimum length of interval
Qeseq	Local enrichment significance	*p*-value	no parameter
CMT	Height threshold and volume difference	fold enrichment	average fragment length, minimum and maximum region size, cut-off, minimum supported reads

Table 5.6: Area under curve (AUC) comparison between CMT, MACS and T-PIC, based on the number of false positive (FP) and true positive (TP) detected peaks.

	CMT	**T-PIC**	**MACS**
AUC	0.856	0.794	0.712

logical validation of each method. As the other two methods, CMT has been able to reject all true negatives. Although CMT finds a larger number of regions, it shows a high sensitivity, finding more true positives than T-PIC and MACS. As an example, one of the true positive regions in chromosome 3 is shown in Figure 5.5. The region is detected by CMT but not by T-PIC or MACS.

Table 5.7: True positive and true negative peak comparison. the comparison of CMT, MACS and T-PIC is based on the number of true positive (TP) and true negative (TN) detected peaks.

	CMT	**T-PIC**	**MACS**
TP	14	13	12
TN	0	0	0

In another experiment, using the information gathered from the UCSC Genome Browser on the *NCBI36/hg19* assembly, the genomic features of each detected peak have been investigated. We assigned a genomic feature to a peak if that peak overlaps with the region containing that genomic feature. A detected peak can be aligned to more than one genomic feature. For example, if a specific peak overlaps with a gene and exon simultaneously, we count that peak as both gene *and* exon. Table 5.8 shows the percentage of regions that are located in gene, promoter, intron and exon areas as well as inter-genetic regions. CMT was able to detect more regions corresponding to genes, promoters and exons, while the percentage of detected regions within introns and inter-genetic areas by CMT is less than the

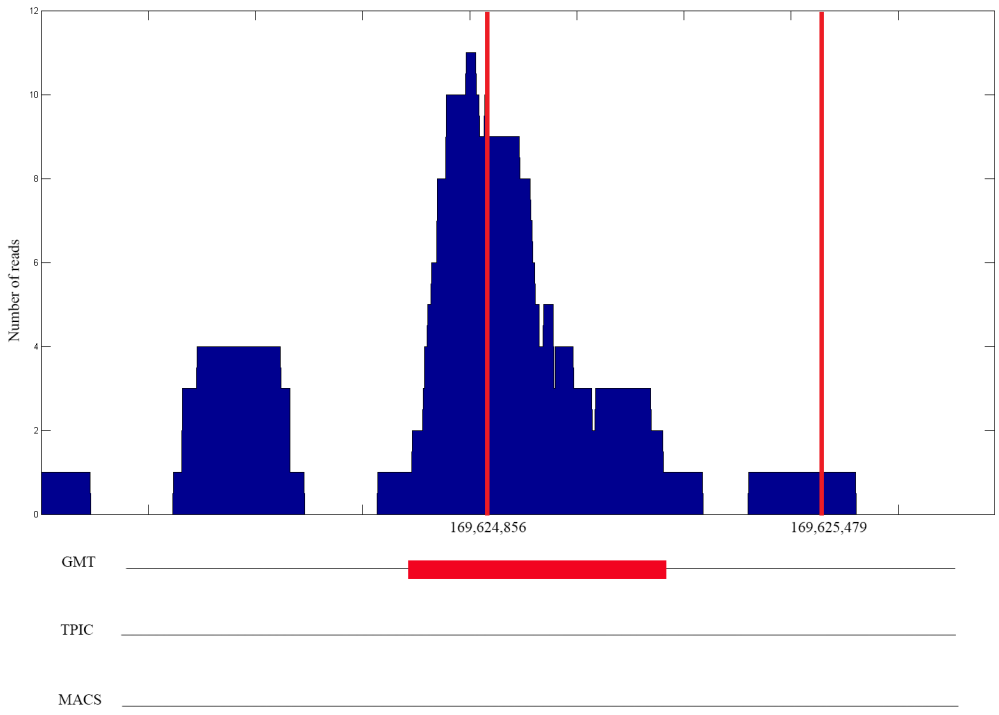

Figure 5.5: One of the true positive regions located in chromosome 3 of the FoxA1 dataset. The red lines show the actual location of the previously verified true positive region. The *x*-axis corresponds to the genome position in bp and the *y*-axis corresponds to the number of reads. The peak is detected by CMT but not by T-PIC or MACS.

percentage of detected regions by MACS and T-PIC. We have also analyzed the genomic features of the peaks detected by each method and not by the others. Table 5.9 shows the result of this analysis. As shown in the table, again, CMT found more genes, exons and promoters than T-PIC and MACS, while it found less peaks corresponding to the non-coding regions.

Table 5.8: Comparison of CMT, MACS and T-PIC, based on the percentage of detected regions that are associated with different genomic features.

Method	# of Regions	Genes		Exons		Introns		Promoters		Inter-genetic Regions	
		Regions	%	Regions	%	Regions	%	Regions	%	Regions	%
MACS	14,026	12,249	87.3	967	6.9	12,438	88.7	676	4.8	7,338	52.3
T-PIC	21,662	19,041	87.9	1,721	7.9	18,731	86.5	934	4.3	10,989	50.7
CMT	26,253	23,311	88.8	2,231	8.5	22,143	84.3	1,226	4.7	13,053	49.7

Table 5.9: Comparison of CMT, MACS and T-PIC, based on the percentage of detected regions detected by one method and not by the others.

Method	**Genes**	**Exons**	**Introns**	**Promoters**	**Inter-genetic Regions**
MACS	70.5 %	7.5 %	71.4 %	3.8 %	57.4%
T-PIC	67.7 %	9.8 %	68.4 %	2.8 %	57.5%
CMT	89.1 %	10.2 %	68.5 %	4.3 %	47.2 %

5.2.3 Targeting a Specific Range of Regions Using Constraints

There are different types of regions of interest within the genome with various lengths. Some of the regions are long-range in the sense that have a length of up to 60 kbp such as histone modifications sites. Some other regions are mid-range such as DNA polymerase binding sites, or genes in which the length of the corresponding regions can vary from 1 to 20 kbp. There are also some regions of interest with a very small footprint such as exons of length approximately 100 bp and transcription factor binding sites of length around 10 bp.

To find a specific type of biomarker, it is better to search for regions within a certain range in the genome. Finding all regions of interest corresponding to a target protein and selecting only those regions that are wide enough to be a histone modification site or a gene increase the computational complexity of the method without adding any benefit to the analysis. Using a constraint-based model helps us target only those regions that are in a specified range. Moreover, the sensitivity of the algorithm can be adapted dynamically to target the regions of interest based on the specified range with higher accuracy.

5.3 Methods

The aim is to find significant peaks corresponding to regions that interact with the protein of interest. Roughly speaking, each peak can be seen as a cluster which is separated from its neighbours by “valleys”. In that sense, the problem can be formulated as a *one-dimensional clustering* problem. Figure 5.6 depicts the process of finding the peaks corresponding to the regions of interest for the specified protein. After extending each read to a fragment, a histogram is created for each chromosome using those fragments. In the next step, relevant peaks are selected by CMT after fine tuning the exact position of the regions. Finally, by

comparing each region with the corresponding region in the control histogram, the relevant peaks are selected.

5.3.1 Creating the Histogram

The first step of the method consists of creating a histogram using the input BED file containing the position and direction of the reads. Each read should be extended to a fragment length, which is related to the settings used to shearing the DNA. This parameter can be input by the user, even though the fragment length can be easily estimated from the underlying data if enough computational resources are available.

After extending each read to a fragment length based on the direction of each read, each fragment is aligned to the reference genome based on its coordinates. Afterwards, for each chromosome, two separate histograms for experiment and control datasets are created for further processing. Each bin in the histogram corresponds to a nucleotide.

5.3.2 The Constrained Thresholding Algorithm

For each chromosome, the corresponding experiment histogram, which is obtained from the previous step, is analyzed separately using the constraint-based algorithm. In this algorithm, each region is treated as an independent cluster. By starting from the beginning of the chromosome and based on the minimum and maximum ranges of the target regions (determined by user), the best point to divide the histogram is found.

Although various parametric and non-parametric thresholding methods and criteria have been proposed, the three most important streams are Otsu's method, which aims to maximize the separability of the classes measured by means of the sum of between-class variances [24], the criterion that uses information theoretic measures in order to maximize the

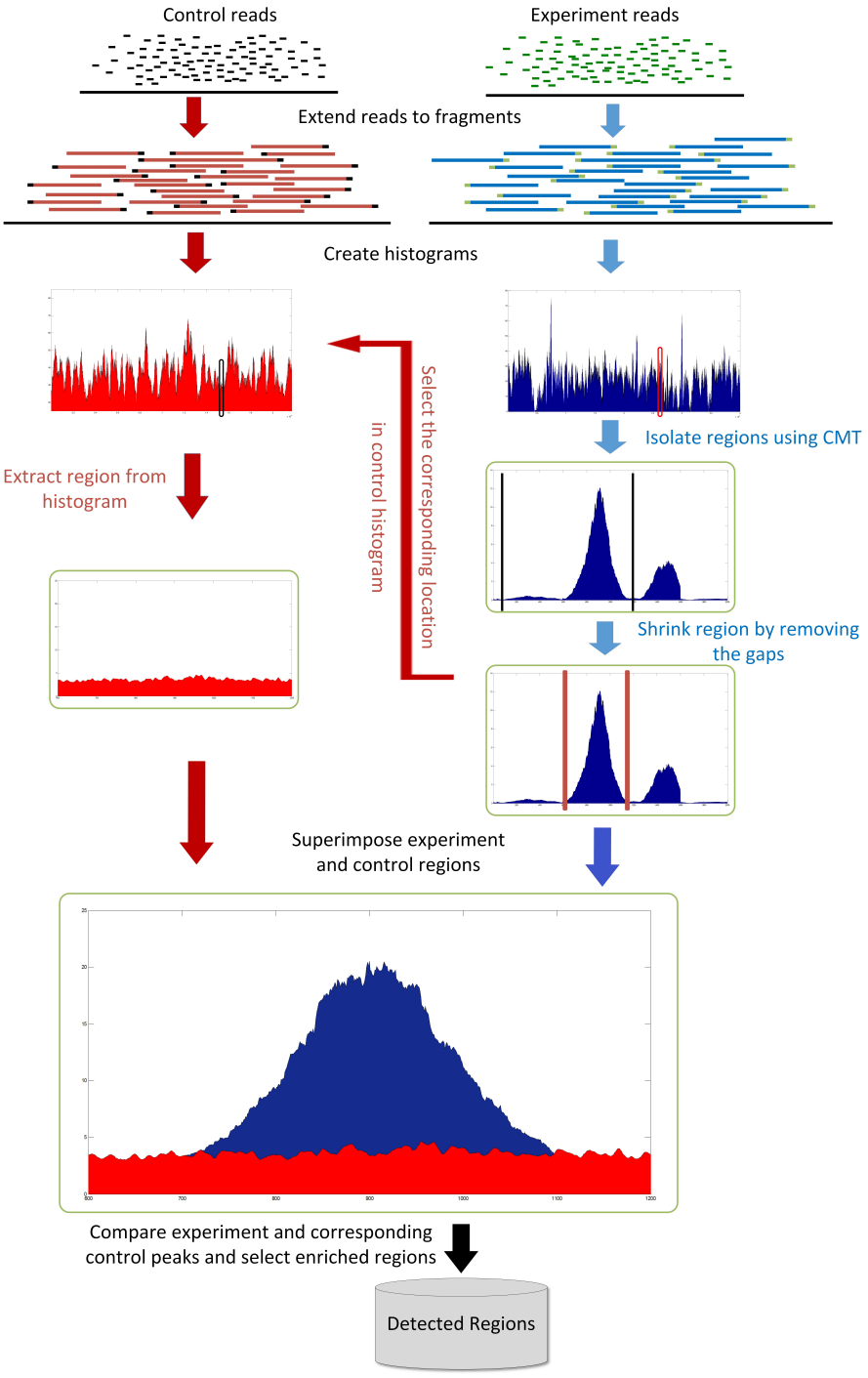

Figure 5.6: Schematic diagram of the pipeline for finding significant peaks.

separability of the classes [25], and the minimum error criterion [26]. In this work, we use the between-class variance criterion [24] because it provides higher accuracy.

Consider a histogram H, an ordered set $\{1,2,\ldots,n-1,n\}$, where the ith value corresponds to the ith bin and has a probability, p_i. Also, consider a threshold set T, defined as an ordered set $T = \{t_0, t_1, \ldots, t_k, t_{k+1}\}$, where $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} = n$ and $t_i \in \{0\} \cup H$. The aim of CMT is to find the values of t_i within a window starting from the current position and based on the given minimum and maximum length defined by the user.

The between class variance criterion is given by:

$$\Psi_{BC} = \omega_1 \mu_1^2 + \omega_2 \mu_2^2, \tag{5.2}$$

where $\omega_1 = \sum_{i=1}^{t^*} p_i$, $\mu_1 = \frac{1}{\omega_1} \sum_{i=1}^{t^*} i \times p_i$, $\omega_2 = \sum_{i=t^*}^{n} p_i$ and $\mu_2 = \frac{1}{\omega_2} \sum_{i=t^*}^{n} i \times p_i$.

The aim is to obtain t^* for each potential region in such a way that Ψ_{BC} is maximized for that window. Figure 5.7 depicts the procedure for finding threshold t^*. The sub-optimal threshold t^* can be found by sliding the blue line between min and max and compute Ψ_{BC} respectively. The best point to separate two neighbour peaks is the one that maximizes Ψ_{BC}.

The final output of the model consists of two vectors, $S_i = [s_1, ..., s_n]^t$ and $E_i = [e_1, ..., e_n]^t$, where s_i and e_i are the start and end position of the i^{th} detected region respectively and n is the number of detected peaks. Although this method is not optimal, its worst-case time complexity is $O(n)$, where n is the number of genomic positions (nucleotides) in a chromosome.

5.3.3 Gap Skipping

After aligning the reads to the reference genome, and depending on the number of reads obtained from the experiment, the fragments may cover a small fraction of the genome and

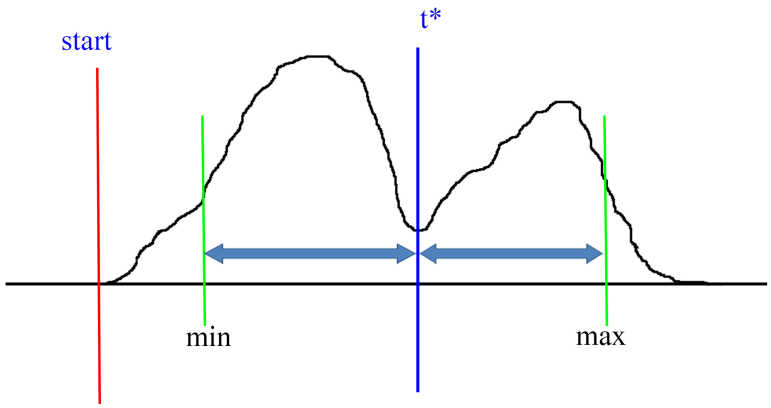

Figure 5.7: An example of finding the threshold t^* using the CMT algorithm.

leave very large gaps between neighbour regions. To speed up the peak finding process, gaps are skipped by computing the maximum height of each window. If that height does not surpass the minimum acceptable height for the region, that window is skipped and no further analysis is done on the regions within that window. The minimum acceptable height is a user-adjustable value that specifies how many reads a region should support to make it acceptable as a possible region of interest.

5.3.4 Selecting Enriched Regions

After finding the potential regions, they have to be shrunk from the borders for removing possible empty gaps on the left and right sides of the region. Starting from the highest point of the region, the start and end borders are moved to left and right respectively until the height of the region in both of those points reaches a value below a cut-off level. The cut-off level is adjustable by the user. The default value is 1, which means that the algorithm will isolate the continuous part of the region that contains at least one fragment aligned to

those positions.

In the next step, the isolated experiment regions detected in the previous step are compared to their corresponding regions in the control histogram. A region in the experiment histogram is considered as an enriched region if it satisfies the following properties:

- the size of the region should be within the acceptable ranges defined by the user, and
- there should be a k-fold difference between the squared density of the experiment region and the control region as follows:

$$V_e \geq K \times V_c \tag{5.3}$$

where $V_e = \Sigma_{i=start}^{end} e_i^2$, $V_c = \Sigma_{i=start}^{end} c_i^2$; e_i and c_i are the heights of the experiment and control regions at position i respectively. Also, K is a user-defined parameter (whose default value is 2), and corresponds to the minimum acceptable fold change between experiment and control.

The regions that satisfy the aforementioned criteria are considered enriched and are used for further processing and biological validation.

Implementation

CMT has been implemented in `C++`. It runs on x86 systems using the Windows operating system. The executable version of the code is available at http://luisrueda.cs.uwindsor.ca/software/CMT-ChIP-Seq.rar. The source code is available upon request. A readme file is included in the downloadable package.

Bibliography

[1] Barski A, Zhao K (2009) Genomic location analysis by ChIP-Seq. Journal of Cellular Biochemistry 107: 11–18.

[2] Park P (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genetics 10: 669–680.

[3] Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics 13: 840–852.

[4] Micsinai M, Parisi F, Strino F, Asp P, Dynlacht BD, et al. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. Nucleic acids research 40: e70–e70.

[5] Jackman RW, Wu CL, Kandarian SC (2012) The ChIP-seq-Defined Networks of Bcl-3 Gene Binding Support Its Required Role in Skeletal Muscle Atrophy. PloS one 7: e51478.

[6] Auerbach RK, Chen B, Butte AJ (2013) Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. Bioinformatics : 1–2.

[7] Stower H (2013) DNA Replication: ChIP-seq for human replication origins. Nature Reviews Genetics 14: 78–78.

[8] Buck M, Nobel A, Lieb J (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biology 6: R97.

[9] Johnson W, Li W, Meyer C, Gottardo R, Carroll J, et al. (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proceedings of the National Academy of Sciences 103: 12457-12462.

[10] Qi Y, Rolfe A, MacIsaac K, Gerber G, Pokholok D, et al. (2006) High-resolution computational models of genome binding events. Nature Biotechnology 24: 963–970.

[11] Reiss D, Facciotti M, Baliga N (2008) Model-based deconvolution of genome-wide DNA binding. Bioinformatics 24: 396-403.

[12] Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, et al. (2008) Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9: R137.

[13] Rozowsky J, Euskirchen G, Auerbach R, Zhang Z, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology 27: 66–75.

[14] Hower V, Evans S, Pachter L (2010) Shape-based peak identification for ChIP-Seq. BMC Bioinformatics 11.

[15] Wang C, Xu J, Zhang D, Wilson Z, Zhang D (2008) An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics 41: 117–129.

[16] Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nature genetics 38: 1289–1297.

[17] Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, et al. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958–970.

[18] RK Bradley, XY Li, C Trapnell, S Davidson, L Pachter, HC Chu, LA Tonkin, MD Biggin, MB Eisen (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol 8: e1000343.

[19] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic acids research 41: D991–D995.

[20] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics 40: 897–903.

[21] Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, et al. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic acids research 36: 2547–2560.

[22] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.

[23] Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, et al. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958–970.

[24] N Otsu (1979) A threshold selection method from gray-level histograms. IEEE Trans on Systems, Man and Cybernetics SMC-9: 62–66.

[25] J Kapur , P Sahoo and A Wong (1985) A new method for gray-level picture thresholding using the entropy of the histogram. Computer Vision Graphics and Image Processing 29: 273–285.

[26] J Kittler and J Illingworth (1986) Minimum error thresholding. Pattern Recognition 19: 41–47.

Chapter 6

Identifying Informative Genes for Prediction of Breast Cancer Subtypes

6.1 Introduction

Despite advances in treatment, breast cancer remains the second leading cause of cancer related deaths among females in Canada and the United States. Previous studies have revealed that breast cancer can be categorized into at least five subtypes, including basal-like (Basal), luminal A, (LumA), luminal B (LumB), HER2-enriched (HER2), and normal-like (Normal) types [1, 2]. These subtypes have their own genetic signatures, and response to therapy varies dramatically from one subtype to another. The variability among subtypes holds the answer to how to better design and implement new therapeutic approaches that work effectively for all patients. It is clinically essential to move toward effectively stratifying patients into their relevant disease subtype prior to treatment.

Techniques such as breast MRI, mammography, and CT scan, can examine the phenotypical mammary change, but provide little effective information to direct therapy. Genomic

techniques provide high-throughput tools in breast cancer diagnosis and treatment, allowing clinicians to investigate breast tumors at a molecular level. The advance of microarray approaches have enabled genome-wide sampling of gene expression values and/or copy number variations. The huge amount of data that has been generated has allowed researchers to use unsupervised machine learning approaches to discover characteristic "signatures" that have since established distinct tumor subtypes [1]. Tumor subtyping has explained a great deal about some of the mysteries of tumor pathology [3], and has begun to enable more accurate predictions with regard to response to treatment [4]. While offering enormous opportunity for directing therapy, there are some challenges arising in the analysis of microarray data. First, the number of available samples (e.g. patients) is relatively small compared to the number of genes measured. The sample size typically ranges from tens to hundreds because of costs of clinical tests or ethical constraints. Second, microarray data is noisy. Although the level of technical noise is debatable [5], it must be carefully considered during any analysis. Third, due to technical reasons, the data set may contain missing values or have a large amount of redundant information. These challenges affect the design and results of microarray data analysis.

This current study focuses on identifying a minimal number of genes that will reliably predict each of the breast cancer subtypes. Being a field of machine learning, pattern recognition can be formulated as a feature selection and classification problem for multi-class, high-dimensional data using two traditional schemes. The first applies a multi-class "feature selection" method directly followed by a classifier to measure the dependency between a particular feature and the multi-class information. A well-known example of the feature selection method is the minimum redundancy maximum relevance (mRMR) method proposed in [6] and [7]. The second traditional scheme is the most common of the two and

treats the multi-class feature selection as multiple binary-class selections. Methods using multiple binary class selections differ in how to bisect the multiple classes. The two most popular ways to solve this problem are one-versus-one and one-versus-all [8]. In this paper, we propose a novel and flexible hierarchial framework to select discriminative genes and predict breast tumor subtypes simultaneously. The main contributions of this paper can be summarized as follows:

1. We implement our framework using *Chi2* feature selection [9] and a *support vector machine (SVM) classifier* [10] to obtain biologically meaningful genes, and to increase the accuracy for predicting breast tumor subtypes.

2. We use a novel feature selection scheme with a hierarchial structure, which learns in a cross-validation framework from the training data.

3. We establish a flexible model where any feature selection and classifier can be embedded for use.

4. We discover a new, compact set of biomarkers or genes useful for distinguishing among breast cancer types

6.2 Related Work

Using microarray techniques, scientists are able to measure the expression levels for thousands of genes simultaneously. Finding relevant genes corresponding to each type of cancer is not a trivial task. Using hierarchical clustering, Perou and colleagues developed the original 5 subtypes of breast cancer based on the relative expression of 500 differentially expressed genes [1]. It has since been demonstrated that combining platforms to include

DNA copy number arrays, DNA methylation, exome sequencing, microRNA sequencing and reverse-phase protein arrays may define these subtypes even further [2]. It is postulated that there are, indeed, upward of over 10 different forms of breast cancer with differing prognosis [11]. Other groups have tailored analysis toward refining the patient groups based on relative prognosis, reducing the profile for one subtype to a 14-gene signature [12]. Given any patient subtype, obtained through one or several platforms, we hypothesize that machine learning approaches can be used to more accurately determine the number of genes required to reliably predict a subtype for a given patients.

On the other hand, modeling today's complex biological systems requires efficient computational techniques designed in articulated model, and used to extract valuable information from existing data. In this regard, pattern recognition techniques in machine learning provide a wealth of algorithms for feature extraction and selection, classification and clustering. A few relevant approaches are briefly discussed then.

An entropy-based method for classifying cancer types was proposed in [13]. In entropy-classed signatures, the genes related to the different cancer subtypes are selected, while the redundancy between genes is reduced simultaneously. Recursive feature addition (RFA) has been proposed in [14], which combines supervised learning and statistical similarity measures to select relevant genes to the cancer type. A mixture classification model containing a two-layer structure named as mixture of rough set (MRS) and support vector machine (SVM) was proposed in [15]. This model is constructed by combining rough sets and SVM methods, in such a way that the rough set classifier acts as the first layer to determine some singular samples in the data, while the SVM classifier acts as the second layer to classify the remaining samples. In [16], a binary particle swarm optimization (BPSO) was proposed. BPSO involves a simulation of the social behavior in organisms such as bird

flocking and fish schooling. In BPSO, a small subset of informative genes is selected where the genes in the subset are relevant for cancer classification. In [17], a method for selecting relevant genes in comparative gene expression studies was proposed, referred to as *recursive cluster elimination* (RCE). RCE combines *k*-Means and SVM to identify and score (or rank) those gene clusters for the purpose of classification. *k*-Means is used initially to group the genes into clusters. RCE is then applied to iteratively remove those clusters of genes that contribute the least to classification accuracy. In the work described in this paper we used the original five breast cancer subtypes to determine whether our proposed hierarchial tree-based scheme could reduce the gene signature to a reliable subset of relevant genes.

6.3 Methods

First, we describe the training phase for gene selection and breast cancer subtyping, and then we describe how the model can be used in predicting subytpes in a clinical setting. The complete gene profile of each breast cancer subtype is compared against the others. Each subtype varies in the genes that are associated with it, and in the accuracy with which those genes predict that specific subtype. The subtypes are then organized by two main criteria. The first criterion is the level of accuracy with which the selected genes identify the given subtype. The second criterion is the number of genes identified. Clearly applying two or more gene selection criteria is a multi-objective problem in optimization [18]. In this study, we use the rule that select the smallest subset of genes that yields the highest accuracy. Therefore, a subtype that is predicted with 95% accuracy by five genes is ranked higher than a subtype for which 20 genes are required to acquire the same accuracy. The subtype that is ranked highest is removed and the procedure is repeated for the remaining subtypes comparing each gene profile against the others. The highest ranked subtype is

again removed and becomes a leaf on the hierarchical tree (see Fig. 6.1). Therefore, each leaf on the tree becomes a distinct subtype outcome.

6.3.1 Training Phase

We give an example of such a tree to illustrate our method in Fig. 6.1. Suppose there are five subtypes, namely $\{C_1, \cdots, C_5\}$. The training data is a $m \times n$ matrix $D = \{D_1, \cdots, D_5\}$ corresponding to the five subtypes. D_i, of size $m \times n_i$, is the training data for class C_i. m is the number genes and n_i is the number of samples in subtype C_i. $n = \sum_{i=1}^{5} n_i$ is the total number of training samples from all five classes. First of all, feature selection and classification are conducted, in a cross-validation fashion, for each class against the other classes. For example, suppose subtype C_3 obtains the highest rank based on accuracy and the number of genes contributing to that accuracy. We thus record the list of the particular genes selected and create a leaf for that subtype. We then remove the samples of the subtype, which results in $D = \{D_1, D_2, D_4, D_5\}$ and continue the process in the same fashion. Thus, at the second level, subtype C_5 yields the highest rank, and hence its gene list is retained and a leaf is created. Afterward the training data set becomes $D = \{D_1, D_2, D_4\}$ for the third level. We repeat the training procedure in the same fashion until there is no subtype to classify. At the last level, two leaves are created, for C_4 and C_2, respectively.

6.3.2 Prediction Phase

Once the training is complete, we can apply the scheme to predict breast cancer subtypes. Given the gene expression profile of a new patient, a sequence of classification steps are performed by tracing a path from the root of the tree toward a leaf. At each node in the path, only the genes selected in the training phase are tested. The process starts at the first

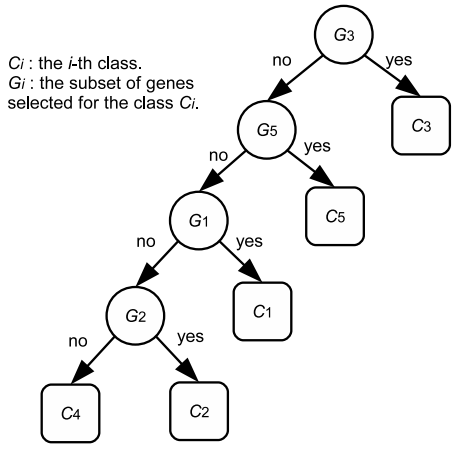

Figure 6.1: Determining breast cancer type using selected genes.

level (root of the tree), in which case only the genes selected for C_3, namely G_3 are tested. If the patient's gene profile is classified as a positive sample, then the prediction outcome is subtype C_3, and the prediction phase terminates. Otherwise, the sequence of classification tests is performed in the same fashion, until a leaf is reached, in which case the prediction outcome is the subtype associated with the leaf that has been reached.

6.3.3 Characteristics of The Method

Our structured model has the following characteristics. First, it involves a greedy scheme that tries the subtype which obtains the most reliable prediction and the smallest number of genes first. Second, it conducts feature selection and classification simultaneously. Essentially, it is a specific type of decision tree for classification. The differences between the proposed model and the traditional decision tree includes: i) each leaf is unique, while one class usually has multiple leaves in the later; ii) classifiers are learned at each node, while the traditional scheme learns decision rules; and iii) multiple features can be selected, while in the traditional scheme each node corresponds to only one feature. Third, the proposed

model is flexible as any feature selection method and classifier can be embedded. Obviously, a classifier that can select features simultaneously also applies, (e.g. the l_1-norm SVM [19]).

6.3.4 Implementation

In this study, we implement our model by using Chi2 feature selection [9] and the state-of-the-art SVM classifier [10]. These two techniques are briefly described briefly next. Chi2 is an efficient feature selection method for numeric data. Unlike some traditional methods which discretize numeric data before conducting feature selection, Chi2 *automatically* and *adaptively* discretizes numeric features and selects features as well. It keeps merging adjacent discrete statuses with the lowest χ^2 value until all χ^2 values exceed their confidence intervals determined by a decreasing significant level, while keeping consistency with the original data. If, finally, a feature has only one discrete status, it is removed. The χ^2 value of a pair of adjacent discrete statuses or intervals is computed by the χ^2 statistic, with 1 degree of freedom, as follows:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{6.1}$$

where n_{ij} is the number of samples in the i-th interval and j-th class, and e_{ij} is the expected value of n_{ij}. e_{ij} is defined as $r_i \frac{c_j}{n}$ where $r_i = \sum_{j=1}^{k} n_{ij}$, $c_j = \sum_{i=1}^{2} n_{ij}$, and n is the total number training samples.

Based on these selected genes, the samples are classified using SVM [10]. Soft-margin SVM is applied in our current study. SVM is a linear maximum-margin model with decision function $d(x) = \text{sign}[f(x)] = \text{sign}[w^{\mathrm{T}}x + b]$ where w is the normal vector of the separating

hyperplane and b is the bias. Soft-margin SVM solves the following problem in order to obtain the optimal w and b:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C^{\mathrm{T}}\xi \\ &\text{s.t. } Z^{\mathrm{T}}w + by \geq 1 - \xi \\ &\qquad \xi \geq 0, \end{aligned} \tag{6.2}$$

where ξ is a vector of slack variables, C is a vector of constant that controls the trade-off between the maximum margin and the empirical error, y is a vector that contains the class information (either -1 or +1), and Z contains the normalized training samples with its i-th column defined as $z_i = y_i x_i$ [20]. Since optimization of the SVM involves inner products of training samples, by replacing the inner products by a kernel function, we can obtain a kernelized SVM.

For the implementation, the Weka machine learning suite was used [21]. A gene selection method based on the χ^2 feature evaluation algorithm was first used to find a subset of genes with the best ratio of accuracy/gene number [9]. For classification, LIBSVM [22] in Weka is employed. The *Radial basis function* (RBF) kernel is used with the LIBSVM classifier without normalizing samples and with default parameter settings.

6.4 Computational Experiments and Discussions

6.4.1 Experiments

In our computational experiment, we analyzed Hu's data [23]. Hu's data (CEO accession number GSE1992) were generated by three different platforms including Agilent-011521

Human 1A Microarray G4110A (feature number version) (GPL885), Agilent-012097 Human 1A Microarray (V2) G4110B (feature number version) (GPL887), and Agilent Human 1A Oligo UNC custom Microarrays (GPL1390). Each platform contains 22,575 probesets, and there are 14,460 common probesets among these three platforms. We used SOURCE [24] to obtain 13,582 genes with unique unigene IDs in order to merge data from different platforms. The dataset contains 158 samples from five subtypes of breast cancer(13 Normal, 39 Basal, 22 Her2, 53 LumA and 31 LumB). The sixth subtype Claudin is excluded from our current analysis as the number of samples of this class is too few (only five). However, we will investigate this subtype in our future work.

To evaluate the accuracy of the model, 10-fold cross-validation is used. As shown in Table 6.2, using all genes decreases the overall accuracy of the model, since many of the genes are irrelevant or redundant. For example, using all 13,582 genes, the overall accuracy is just 77.84%; while using a ranking algorithm and taking the top 20 genes for prediction brings the accuracy up to 86.70%. Table 6.1 shows the top 20 genes ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the five subtypes. Using the proposed hierarchical decision-tree-based model, makes the prediction procedure more accurate. While the accuracy of prediction between LumA and LumB is relatively low compared to the other classes. This is because of the very high similarity and overlap between samples of these two classes. The overall accuracy of the model, as shown in Table 6.2, is 95.11%. This is very interesting since only 18 genes are used to predict the subtypes that the patient belongs to. these 18 genes have been obtained by selecting 6 genes per node and decreasing them one by one as long as the accuracy of the model keeps consistent. As a matter of fact, our method is able to increase its accuracy from around 86% to 95% by using a new subset of genes based on the proposed method containing only 18

genes.

Table 6.1: Top 20 genes ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the five subtypes.

Rank	Gene Name		Rank	Gene Name		Rank	Gene Name		Rank	Gene Name
1	FOXA1		6	THSD4		11	DACH1		16	ACOT4
2	AGR3		7	NDC80		12	GATA3		17	B3GNT5
3	CENPF		8	TFF3		13	INPP4B		18	IL6ST
4	CIRBP		9	ASPM		14	TTLL4		19	FAM171A1
5	TBC1D9		10	FAM174A		15	VAV3		20	CYB5D2

Fig. 6.2 shows the tree learned in the training phase and the set of genes selected at each step. The selected genes are contained in each node, a patient's gene expression profile is used to feed the tree for prediction, each leaf represents a subtype, and the accuracy at each classification step is under the corresponding node.

From this figure, we can see that the Basal subtype is chosen first as it obtains the highest accuracy, 99.36% to classify patients from the other subtypes including Normal, Her2, LumA and LumB. Then the samples of Basal are removed for the second level. The Normal subtype is chosen then, since it achieves the highest accuracy (95.79%) to separate samples from the other subtypes, including Her2, LumA and LumB. From previous studies, it is well-known that the subtypes LumA and LumB are very difficult to be identified among all subtypes. This is the reason for why LumA and LumB appear at the bottom of the tree. After removing other subtypes, LumA and LumB can avoid misclassification on the other subtypes. In spite of this drawback, the accuracy for separating LumA and LumB is as high as 88.1%.

As shown in Figure 6.2, there is no overlap between the genes selected among the different clusters. This result provides interesting new biomarkers for each breast cancer subtype. Some of the selected genes have been previously indicated in cancer (highlighted

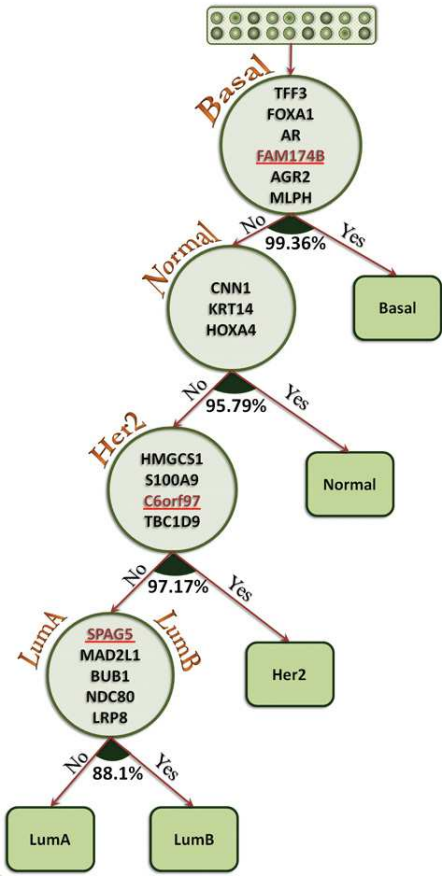

Figure 6.2: Determining breast cancer type using selected genes.

in black in Figure 6.2), while others have emerged as interesting genes to be investigated. For example, TFF3 and FoxA1 genes are predictably indicated in Basal subtype. Another feature of the proposed hierarchical model is that the number of genes in each node has been optimized to give the best ratio of accuracy and number of selected genes. For this, at first, 10 genes with highest rank have been selected for each node. Then, out of those selected genes, those with lower rank are removed step by step as long as the accuracy of classification using the remaining genes don't get decreased.

Table 6.2: Accuracy of classification using LibSVM Classifier

Classification Method	Gene Selection Method	# of Genes	Accuracy	Precision	Recall	F-measure
LibSVM	—	all genes	77.84%	0.802	0.778	0.749
LibSVM	Chi-Squared	20	86.70%	0.866	0.867	0.864
Proposed Method	Proposed Method	18	95.11%	0.951	0.951	0.951

6.4.2 Biological Insight

We used FABLE to determine if the genes selected by our approach are biologically meaningful. Fast Automated Biomedical Literature Extraction (FABLE) is a web-based tool to search through MEDLINE and PubMed databases. The genes that are related to tumors reported in the literature are highlighted in black in Figure 6.2. Those not yet reported are underlined and colored in red. We can see that 15 out of 18 genes have been found in the literature. This implies that our approach is quite effective in discovering new biomarkers.

We also explored the reasons for the high performance of our method. First, the subtypes that are easily classified are on the top of the tree, while the harder subtypes are considered only after removing the easier ones. Such a hierarchical structure can remove the disturbance of other subtypes, thereby allowing us to focus on the most difficult subtypes, LumA/B. Second, combining gene selection when building the classifier allows us to select genes that contribute to prediction accuracy. Third, our tree-based methodology is quite flexible; any existing gene selection measure and classification technique can be embedded in our model. This will allow us to apply this model to subtypes as they become more rigorously defined using other platforms such as copy number variation. Furthermore, our method could be applied to groups of patients stratified based on responses to specific treatments. Collectively, having a small, yet reliable number of genes to screen is more cost

effective and would allow for subtype information to be more readily applied in a clinical setting.

6.5 Conclusion and Future Work

In this study, we proposed a novel gene selection method for breast cancer subtype prediction based on a hierarchical, tree-based model. The results demonstrate an impressive accuracy to predict breast cancer types using only 18 genes. Herein, we propose a novel gene selection method for breast cancer subtype prediction based on a hierarchical, tree-based model. The results demonstrate an impressive accuracy to predict breast cancer subtypes using only 18 genes in total. Moreover, Most of the selected genes are shown to be related to breast cancer based on previous studies, while a few are yet to be investigated. As future work, we will validate these results using cell lines that fall within a known subtype. We will determine whether our predicted 18 gene array can accurately denote which subtype each of these cell lines falls under. This hierarchical, tree-based model can narrow down analysis to a relatively small subset of genes. Importantly, the method can be applied to more refined stratification of patients in the future, such as subtypes derived using a combination of platforms, or for groups of patients that have been subdivided based on response to therapy. Using this computational tool we can determine the smallest possible number of genes that need to be screened for accurately placing large populations of patients into specific subtypes of cancer or specified treatment groups. This could contribute to the development of improved screening tools, providing increased accuracy for a larger patient population than that achieved by Oncotype DX, but allowing for a cost effective approach that could be widely applied to the patient population.

Bibliography

[1] Perou, C.M., et al.: Golecular Portraits of Human Breast Tumours. Nature. 406, 747–752 (2000)

[2] Perou, C.M., et al.: Comprehensive Molecular Portraits of Human Breast Tumours. Nature. 490, 61–70 (2012)

[3] Chandriani,S., Frengen, E., Cowling, V.H., Pendergrass, S.A., Perou, C.M., Whitfield, M.L., Cole, M.D.: A Core MYC Gene Expression Signatures is Prominent in Basal-Like Breast Cancer but only Partially Overlaps the Core Serum Response. PLOS One. 4(8), e6693 (2009)

[4] van't Veer, L.J., et al.: Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. Nature. 415(6871), 530–536 (2002)

[5] Klebanov, L., Yakovlev, A.: How High is The Level of Technical Noise in Microarray Data?. Biology Direct. 2, 9 (2007)

[6] Ding, C., Peng, H.: Munimun Redundancy Feature Selection from Microarray Gene Expression Data. Journal of Bioinformatics and Computational Biology. 3(2), 185–205 (2005)

[7] Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(8), 1226–1238 (2005)

[8] Li, T., Zhang, C., Ogihata, M.: A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Vased on Gene Expression. Bioinformatics. 20(15), 2429–2437 (2004)

[9] Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. In: IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE Press, New York (1995)

[10] Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)

[11] Curtis, C., et al.: The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. Nature. 486(7403), 346–352 (2012)

[12] Hallett, R.M., Dvorkin-Gheva, A., Bane, A., Hassell, J.A.: A Gene Signature for Predicting Outcome in Patients with Basal-Like Breast Cancer. Scientific Reports. 2, 227 (2012)

[13] Liu, X., Krishnan, A., Mondry, A.: An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data. BMC Bioinformatics. 6, 76 (2005)

[14] Liu, Q., Sung, A.H., Chen, Z., Liu, J., Huang, X., Deng, Y.: Feature Selection and Classification of MAQC-II Breast Cancer and Multiple Myeloma Microarray Gene Expression Data. PLoS One. 4(12), e8250 (2009)

[15] Zeng, T., Liu, J.: Mixture Classification Model Based on Clinical Markers for Breast Cancer Prognosis. Artificial Intelligence in Medicine. 48, 129–137 (2010)

[16] Mohamad, M.S., Omatu, S., Deris, S., Yoshioka, M.: Particle Swarm Optimization for Gene Selection in Classifying Cancer Classes. Artificial Life and Robotics. 14(1), 16–19 (2009)

[17] Yousef, M., Jung, S., Showe, L., Showe, M.: Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. BMC Bioinformatics. 8, 144 (2007)

[18] Li, Y., Ngom, A., Rueda, L.: A Framework of Gene Subset Selection Using Multiobjective Evolutionary Algorithm. LNBI/LNCS. 7632, 38–48 (2012)

[19] Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-Norm Support Vector Machines. In: NIPS, MIT Press, Cambridge, MA (2004)

[20] R. O. Duda and P. E. Hart and D. G. Stork: Pattern Classification. Wiley-Interscience, New York (2006)

[21] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter. 11(1), 10–18 (2009)

[22] Chang, C.-C, Lin, C.-J.: LIBSVM: a Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology. 12, 27:1–27:27 (2011)

[23] Hu, Z., et al.: The Molecular Portraits of Breast Tumors are Conserved Across Microarray Platforms. BMC Genomics. 7, 96 (2006)

[24] Diehn, M., et al.: SOURCE: a Unified Genomic Resource of Functional Annotations, Ontologies, and Gene Expression Data. Nucleic Acids Research. 31(1), 219–223, (2003) Available at http://smd.stanford.edu/cgi-bin/source/sourceSearch

Chapter 7

Conclusion and Future Works

Transcriptomics provide a rich source of data suitable for pattern analysis. We have shown how multilevel thresholding algorithms can improve transcriptome data analysis in different ways. We proposed OMTG, an efficient parameterless framework for DNA microarray image analysis. By adapting the method to analyze next generation sequencing data, we proposed OMT, a robust and versatile peak finder for finding significant peaks in high-throughput next generation sequencing (ChIP-Seq) data. Using different datasets, and various computational and biological analysis steps, it has been shown that both OMT and OMTG are sound and robust to noise in experiments. It is also able to be used on different approaches with a little change – this is one the most important features of this algorithm. We also proposed a constraint-based multi-level thresholding algorithm to find enrichment regions with a specific range using ChIP-Seq data. Moreover, we proposed a novel multi-class breast cancer subtype prediction framework with the ability of obtaining biologically meaningful genes that can accurately predict breast tumor subtypes.

7.1 Conclusion

This thesis introduced new pattern recognition and image processing and analysis methods for transcriptomics data analysis. The methods proposed in this thesis have been shown to work mostly free of parameters and perform efficiently on real-life datasets from different sources. The main contributions of the thesis can be summarized as follows:

1. Chapters 2 and 3:
    (a) Proposing OMTG, a new method for separating sub-grids and spot centers in cDNA microarray images.
    (b) OMTG uses no parameter which makes it a desirable method for gridding microarray images with different structure and resolution without any need for adjustment and tuning.
    (c) Proposing a new validity index for detecting the correct number of sub-grids and spots in microarray image.
    (d) Proposing a refinement procedure used to increase the accuracy of spot detection.
2. Chapters 4 and 5:
    (a) Proposing OMT, a multi-level thresholding based method for finding significant peaks in ChIP-Seq data.
    (b) OMT can be applied to high-throughput next generation sequencing data with different characteristics.
    (c) It has been shown that OMT is statistically sound and robust in experiments and has the ability to be applied to various types of next generation sequencing data.

(d) Comparing to other recently proposed methods, OMT shows to be more accurate and use fewer parameters.

(e) Proposing CMT, a constraint based multi-level thresholding method to find significant peaks within a specific range in ChIP-Seq data.

(f) Unlike other methods, which find all types of regions at once and then select peaks with desired length, targeting specific regions with a certain range is one of the main advantages of CMT that increase the performance of the algorithm in comparison with the other methods.

3. Chapter 6:

(a) Proposing a hierarchical, tree-based gene selection method for breast cancer subtype prediction.

(b) Obtaining an impressive accuracy of more than 95% for predicting breast cancer types using only 18 genes in total.

(c) Most of the selected genes are shown to be related to breast cancer based on previous studies

(d) Providing a computational tool for determining the smallest possible number of genes that need to be screened for accurately placing large populations of patients into specific subtypes of cancer or specified treatment groups.

7.2 Future Work

Considering the huge amount of data generated by different biological platforms, manual analysis of these data is simply impossible. Using supervised and unsupervised machine

learning techniques can provide a variety of efficient and robust models to analyze data. Some of the possible future works are the following:

- Thresholding algorithms are still emerging tools in these areas, and open the possibility for further advancement.
- One of the problems that deserves attention is the use of other thresholding criteria, including minimum error, entropy-based and others in finding the optimal number of spots in a sub-grid and the optimal number of sub-grids in a DNA microarray image.
- Processing a whole genome or even a chromosome for finding peaks in ChIP or RNA-seq is still a challenge, since it involves processing histogram with millions of bins. Processing different part of the histogram in parallel could improve the performance of the peak finding algorithm.
- Next generation sequence data analysis is an emerging and promising area for pattern discovery and analysis, which deserves the attention of the research community in the field.
- One of the future works can be applying the OMT algorithm on the whole chromosome instead of using a set of windows as a way to reduce the number of parameters.
- Using other indices of validity such as minimum error and entropy-based; and other thresholding criteria could increase the accuracy of the method.
- Pathway and biological analysis of selected genes in terms of their real-life performance in identifying breast cancer subtypes and accurately denote which subtype each of these cell lines falls under.

- The proposed hierarchical model can be applied to more refined stratification of patients in the future, such as subtypes derived using a combination of platforms, or for groups of patients that have been subdivided based on response to therapy.

Vita Auctoris

Iman Rezaeian was born in 1979 in Kazeroon, Iran. He received his Bachelors degree in Computer Engineering from Islamic Azad University, Shiraz, Iran, in 2002, and his Master in Computer Engineering and Information Technology from Amirkabir University of Technology, Tehran, Iran, in 2008. His research interests include pattern recognition, machine learning, bioinformatics, data mining, cancer research and next generation sequencing data analysis.