2010

Three essays on analytical models to improve early detection of cancer

Chaitra Gopalappa
University of South Florida

Three Essays on Analytical Models to Improve Early Detection of Cancer

by

Chaitra Gopalappa

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Major Professor: Tapas K. Das, Ph.D.
Alex Savachkin, Ph.D.
Susana Lai-Yuen, Ph.D.
Rebecca Sutphen, M.D.
Selen Cremaschi, Ph.D.

Date of Approval:
May 4, 2010

Keywords: disease progression and intervention, bioinformatics, health care systems, applied stochastic, computational probability, applied optimization, agent-based simulation, wavelet based signal processing

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

THREE ESSAYS ON ANALYTICAL MODELS TO IMPROVE EARLY DETECTION OF CANCER

Chaitra Gopalappa

ABSTRACT

Development of approaches for early detection of cancer requires a comprehensive understanding of the cellular functions that lead to cancer, as well as implementing strategies for population-wide early detection. Cell functions are supported by proteins that are produced by active or *expressed* genes. Identifying cancer *biomarkers*, i.e., the genes that are expressed and the corresponding proteins present only in a cancer state of the cell, can lead to its use for early detection of cancer and for developing drugs. There are approximately 30,000 genes in the human genome producing over 500,000 proteins, thereby posing significant analytical challenges in linking specific genes to proteins and subsequently to cancer. Along with developing diagnostic strategies, effective population-wide implementation of these strategies is dependent on the behavior and interaction between entities that comprise the cancer care system, like patients, physicians, and insurance policies. Hence, obtaining effective early cancer detection requires developing models for a systemic study of cancer care.

In this research, we develop models to address some of the analytical challenges in three distinct areas of early cancer detection, namely proteomics, genomics, and disease progression. The specific research topics (and models) are: 1) identification and quantification of proteins for obtaining biomarkers for early cancer detection (mixed integer-nonlinear programming (MINLP) and wavelet-based model), 2) denoising of

gene values for use in identification of biomarkers (wavelet-based multiresolution denoising algorithm), and 3) estimation of disease progression time of colorectal cancer for developing early cancer intervention strategies (computational probability model and an agent-based simulation).

PREFACE

My time at USF has been an exciting adventure and I am in deep gratitude to all who made it a memorable experience. My research as an engineer in such an interdisciplinary area would not have been successful had it not been for the collaborative effort and guidance of several dedicated researchers who have been my mentors over these years.

I would like to immensely thank my advisor Dr. Tapas Das for encouraging and guiding me to work in an area which I was always intrigued about, cancer research, but had never thought could be involved in as an engineer. His dedication to research and hard work has been truly inspirational. This roller coaster ride would not have been exciting or successful had it not been for the committed support and guidance of so many people, who were always willing to share their knowledge and provide generous advice and guidance: Drs. Rebecca Sutphen, Eric Thomas, John Koomen, Steven Enkemann, Sean Yoder, and Steven Eschrich, and Ms. Tricia Holtje from Moffitt Cancer Center and Research Institute, to whom I am truly thankful for their sincere guidance in this area in which I had no prior knowledge; Drs. Selen Cremaschi from University of Tulsa and Seza Orcun from Purdue University for their committed support, and also for making me feel at home during my three month visit to Purdue to work on my research; Drs. Susana Lai-Yuen and Alex Savachkin whose excellent feedback at various phases of my dissertation helped shape my future tasks; Dr. Jose Zayas-Castro, Chair of Industrial Engineering, whose mentorship and willingness to help is truly inspiring; Drs. Brad Doebbeling and David Haggstrom from VA, In-

diana University, for their guidance; Dr. Karen Liller from Graduate School, USF, for her support and encouragement towards interdisciplinary research while working on the Graduate Student Challenge Grant; my teammates on the Challenge grant Barbara Davila and Jael Rodriguez from Medical School and Dayna Martinez from Engineering, USF, without whose collaborative effort the proposal would not have been successful; Dr. Pekny from Purdue University and Lori Losee from Regenstrief foundation for their support; Jackie Stephens and Gloria Latter from Industrial Engineering for their warm support; and several people at USF with who I have briefly met but each discussion has helped me build my research knowledge. My time at USF would not have been so pleasantly memorable if not for my friends at USF with whom I have had several research discussions, student chapter activities, and social times, Diana Prieto, Wilkistar Otieno, Patricio Rocha, Qingwei Li, and Andres Uribe.

I had never thought I would venture this far into education, and I completely owe it my mom, sister, and grandparents, for their love, support, sacrifices, and confidence in me. Amma, Thaata, Mummy, and Reshma, thank you for standing by me and guiding me, at the same time, helping me build my own character and personality. I am extremely proud to have you as my family. I cannot express enough my gratitude for the dedicated love, support, motivation, honest expression of difference of opinion, and sincere critiques of my best friend, study companion, hiking partner, and soulmate Vishnu Nanduri, that has kept me motivated and encouraged as I begin my research career.

CHAPTER 1

INTRODUCTION

Cancer is a result of abnormal behavior of cells with characteristics including uncontrolled growth of cells, and over time, invasion to adjacent tissues and sometimes metastasizing to different locations of the body via the lymph or blood. Due to this timeline of cancer, early detection and intervention is crucial. Achieving early detection involves developing diagnostic tools which requires a comprehensive understanding of the cellular functions that lead to cancer (cellular level), and also strategies for effective population-wide implementation of the developed tools (strategic level). Before looking into the cause of cancer and the need for analytical models to improve early detection, it is interesting to look into the evolution of our knowledge of cancer over the centuries.

1.1 Evolution of Cancer Research

The American Cancer Society in its recent article ([4]) has compiled several interesting facts of the history of cancer over the past centuries, like the earliest scientific cancer research, evolution of theories for the cause of cancer, and availability of treatment and survival. It has been noted that, the earliest known description of cancer (the word cancer was not used and was coined only around 300 B.C by Hippocrates) was in 1600 B.C in an Egyptian textbook on trauma surgery, where it was described as a disease with no treatment. Since then, there has been tremendous amount of research and treatment advances, and as of now, the article reports that more than

2 out of 3 people diagnosed with cancer survive at least 5 years. However, early detection is crucial for improved chances of survival as can be seen by the survival rates per stage of diagnosis in Figure 1.1, which is a summary of the 1996-2006 cancer statistics reported by the National Cancer Institute's SEER (Surveillance, Epidemiology and End Results) Program ([5]). The Figure plots the 5-year relative survival for different primary sites of cancer, and as can be seen for all these sites, survival when diagnosed at late stage of cancer (distant) is much lower compared to early stage (local). Therefore, developing tools for early detection of cancer is essential.

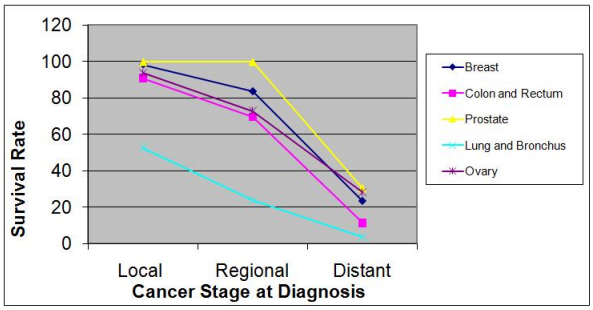

Figure 1.1. Five-year relative survival (in %) for different primary sites of cancer

As in any disease, developing diagnosis tools for early detection and further developing effective treatments require an in-depth understanding of the cause of cancer. It is interesting to present here some of the theories for the cause of cancer with the evolution of knowledge over the centuries [4]. In 300 B.C. we had the humoral theory, where it was believed that cancer was caused by excess black bile, one of the four fluids or humor (blood, phlegm, yellow bile, and black bile) that were required to be in balance for a person to be healthy. In the late 1700's we had the lymph theory, where it was believed that cancer was composed of fermenting lymph fluid. In 1838 it was realized that cancer was made of cells and not fluid, however, it was not know at the time that it was the normal cells that become cancerous (blastema theory). Until 1920s some believed that cancer was caused by trauma. It is interesting to point out

here that the earliest mention of cancer in 1600 B.C was in a trauma surgery book. It was only after the mid 20th century with the discovery of the chemical structure of DNA do we know cancer as of today, as being caused by mutated genes. Without overlooking the foundation of knowledge laid by past research, though there has been thousands of years of work in this area, so much of what we know of cancer today is based on recent advancements, and as stated by the American Cancer Society in [4] "*Scientists have learned more about cancer in the last 2 decades than has been learned in all the centuries preceding*". It is interesting to point the evolution of our knowledge from: a balance of four fluids would keep us healthy, to: it is a coordinated effort of several cell components, like 30,000 genes, and 500,000 proteins; thus, the vastness of data creating the need for analytical models.

1.2 Brief Description of Cell and Cause of Cancer

The cell is a functional unit of all organisms, and the nucleus of the cell contains genetic information in the form of DNA. The DNA is considered to be a blue print that provides instructions for all cellular functions. Cell cycle is a sequence consisting of DNA replication, growth, and division of cells to form two new cells. Thus the genetic information is carried from cell to cell.

1.2.1 Chromosomes, DNA, Genes, and Proteins

Humans cells have 23 pairs of chromosomes where each parent contributes one to each pair. Chromosome is a very long DNA molecule that carry portions of hereditary information. Genes are sections of DNA that carry on the chromosomes that determine specific human characteristics. The human genome consists of around 30,000 genes, with different genes active or *expressed* in different parts of the body. Active genes produce or *translate* proteins, and more than 500,000 proteins are translated by

the human genome. These proteins are the main components that drive the metabolic activities in a cell which lead to functioning of the different parts of the body. Cancer is a result of the dysfunction in cellular level metabolic activities

1.2.2 Cancer Biomarkers

Identification of the gene or protein *biomarkers* of cancer, i.e., the specific genes that are expressed or the proteins that are present only in cancerous cells, could lead to its use as cancer diagnostic tools and further drug discovery. Identifying biomarkers of the earliest stage of cancer can lead to its use as a screening tool for early detection of cancer. Since response to treatment varies across individuals even for the same type of cancer, identifying biomarkers could also allow for a more individualized treatment approach. For example, it is possible that different sets of proteins could cause the same type of cancer, thereby, identifying behavior of drugs on these protein sets could lead to a more targeted treatment. It may be noted that engineering personalized rather than standardized medicines has been identified under one of the Grand Challenges for the 21^{st} century - *Engineering Better Medicine*, that was laid out by the National Academy of Engineering, thus signifying the identification of cancer biomarkers.

1.3 Need for Analytical Models in Achieving Early Cancer Detection

As is clear from the data in Figure 1.1, early detection of cancer is essential for improved chances of survival. Advancements in technology, biochemistry, and medicine has increased our knowledge of the cause of cancer. However, utilizing this knowledge to develop tools and achieve effective early detection is faced with several challenges both biochemical and analytical, two of which can be broadly classified as follows.

1. Challenges in identification of gene and protein biomarkers (cellular level): The vastness of data and complexity of the cell creates analytical hurdles in biomarker discovery. As an example, we consider the challenges faced in the general procedure in extracting biomakers. The procedure includes translating chemical presence of cell components (e.g., genes and proteins in cell samples) to numerical signals, processing signals to extract meaningful data (e.g., gene expression, or amount of protein), and processing data to extract cancer biomarkers. Some of the challenges include irreproducibility of sample, which arises due to factors like the same genes producing different proteins under different conditions of the cell, or proteins changing after production thus resulting in proteins with different metabolic activities. Each time the sample is converted to numerical signal, the vastness of data makes it infeasible to manually identify the values of the cell components, like all genes that are expressed and its level of expression. Therefore, analytical models need to be developed to automatically identify the type and amount of the components present in a sample.

2. Effective population-wide implementation of early detection strategies (strategic level): Achieving effective population-wide early cancer detection extends beyond discovery of cancer screening tools. Utilizing the advances in screening tools requires development of effective screening strategies and moreover its implementation at a systemic level, i.e, in the general population. The system is comprised of independent decision making entities like the population, physicians, and insurance policies. Each entity has its own goals that could sometime conflict with that of the others, and hence, achieving early cancer detection is subject to the behavior and interaction between these system entities. Therefore, developing effective population-wide early cancer detection strate-

gies requires building a model of the entire cancer care system, using which, implementation strategies can be developed and analyzed.

1.3.1 Research Topics Addressed in this Dissertation

As part of this dissertation, we developed mathematical models to address analytical challenges in the areas of disease progression (strategic level), genomics, and proteomics (cellular level). The remaining part of this chapter provides a brief description of the specific research topics of this dissertation, and Chapters 2, 3, and 4, will cover each of these topics in depth.

1. Probability model for estimating disease progression of colorectal cancer: Per American Cancer Society, colorectal cancer (CRC) is the third most common cause of cancer related deaths in the United States. Experts estimate that about 85% of CRCs begin as precancerous polyps, early detection and treatment of which can significantly reduce the risk of CRC. Hence, it is imperative to develop population-wide intervention strategies for early detection of polyps. Development of such strategies requires precise values of the population-specific rates of incidence of polyp and its progression to cancerous stage. There has been a considerable amount of research in recent years on developing screening based CRC intervention strategies. However, these are not supported by population-specific estimates of progression rates. This research addresses this need by developing a probability model that estimates polyp progression rates considering race and family history of CRC; note that, it is ethically infeasible to obtain polyp progression rates through case studies. We use the estimated rates to simulate the progression of polyps in the population of the State of Indiana. The simulation also includes the screening procedure constructed as

per the current screening guidelines for colorectal cancer, and the screening compliance by the population of Indiana.

2. Mathematical model to remove noise from gene data generated by microarrays: Microarray technology for measuring gene expression values has created significant opportunities for advances in disease diagnosis and treatment planning. However, random noise introduced by the sample preparation, hybridization, and scanning stages of microarray processing creates inaccuracy in the estimates of gene expression levels. Literature presents several methodologies for noise reduction, which can be broadly categorized as: 1) model based approaches for estimation and removal of hybridization noise, 2) approaches using commonly available image denoising tools, and 3) approaches involving control samples. In this research we present a novel methodology for identifying and removing hybridization and scanning noise from microarray images, using a dual tree complex wavelet transform based multiresolution analysis coupled with bivariate shrinkage thresholding. The key features of our methodology include consideration of specific characteristics of microarray images and the noise distribution, and the ability to work with a single microarray without needing a control. Our methodology is first benchmarked on a fabricated data set that mimics a real microarray probe data set. Thereafter, our methodology is tested on data sets obtained from a number of Affymetrix GeneChip human genome HG-U133 Plus 2.0 arrays processed on HCT-116 cell line. The results indicate an appreciable improvement in the quality of the microarray data.

3. Analytical processing of data for estimating protein values: Identification of protein biomarkers in blood and urine samples can provide a non-invasive screening tool for early detection of cancer. Developing such screening tools requires iden-

tifying all the proteins that are produced in the body and further identifying those that distinguish cancer cases from controls. Due to the large number of proteins produced by the human body (500,000), most of which are present in small amounts in blood and urine, the biochemical processing of the blood or urine samples need to be accompanied by automatic mathematical algorithms that identify and quantify the proteins. Developing such mathematical algorithms is a challenging task due to vastness of data which is further complicated by the complex biochemical nature of the problem. In this research, we develop wavelet and optimization algorithms for analytical processing of the data to obtain protein information across several samples, using which, biomarkers can be identified.

CHAPTER 2

MODEL FOR COLORECTAL POLYP PROGRESSION

2.1 Introduction

Colorectal cancer (CRC) is the third most common cause of cancer related deaths in the U.S. Most CRCs begin as precancerous polyp ([6, 7]), referred to as adenoma-carcinoma sequence ([8]). Employing effective population-wide strategies for early detection and treatment at precancerous stages can lead to significant reduction in CRC mortalities. Literature presents a considerable amount of research on developing CRC screening strategies with varying tests and time lines, and examining their influence on mortality rates. However, developing feasible intervention strategies requires a system-based model of cancer care that must consider, in addition to screening alternatives, various other interacting elements of the system, including the social-behavioral traits of the people and the physicians, and the parameters of the insurance policies. Two important processes of the system-based cancer care model are: *polyp incidence* and *polyp progression*. Polyps follow a natural incidence and progression, and upon diagnosis, drive the behavior and interaction of the system elements. Thus, precise models portraying the incidence and the progression processes are fundamental to developing effective intervention strategies. In this research, our attention is focused on developing a probability model to estimate progression rates of colorectal polyps.

The literature contains a considerable number of simulation and mathematical models for CRC screening strategies ([9, 10, 11, 12, 13, 14, 15, 1, 16], and CISNET models [17]). All of these cited works have a natural history component for the incidence and progression of polyps, most of which are modeled using variants of Markovian techniques. The main inputs required for these Markovian models are the *incidence rates* of polyps, and *progression rates* between stages, e.g., the inverse of the time that polyps take to progress from adenoma (pre-malignant) to carcinoma. The incidence rates have been estimated based on case study results involving randomized screening and follow-up. However, for progression rates, a case study approach is not feasible, since it is unethical to keep a diagnosed polyp under observation without treatment. As a result, most of the above models use progression rates that are derived based on expert opinion, obtained either by convening a panel or by utilizing the data presented in [18] and [19].

Mathematical models, in contrast to the models based on expert opinion, can incorporate characteristics like race and family history of CRC, and hence estimate population-specific progression rates. The National Academy of Engineering, under one of the Grand Challenges for the 21^{st} century - *Engineering Better Medicine*, noted the need for engineering personalized rather than standardized medicines since "people differ in susceptibility to disease and response to medicine." Also, progression rates in the pre-diagnosis phase may vary between populations, thus underscoring the need for population-specific progression rates. Although literature presents numerous models and cost based analysis of CRC screening strategies, and numerous models and case studies on incidence rates, mathematical models to estimate polyp progression rates have been limited ([20], [21]). The study presented in [20] uses the data from the national colonoscopy screening database of Germany to develop a statistical approach to obtain annual transition rate and the 10 year cumulative risk of CRC specific to sex

and age groups. The transition rates are only for after the onset of advanced adenoma (polyp≥1cm). A Markov model to estimate progression rates for stages after the onset of cancer, and the incidence of cancer is presented in [21]. The model was built based on the results of a case study conducted on a population with high-risk of CRC.

In this research, we present a probability model that was developed for estimating polyp progression rates, specific to race and family history status, from the incidence of polyp to carcinoma and between stages of carcinoma. Note that, estimating progression rates, and thus time to progress, from incidence of polyp to carcinoma are of vital importance for developing early pre-cancer intervention strategies. The polyp progression pathways considered in our model are depicted in Figure 2.1 and are described as follows.

CRC Pathways- Polyp incidence and stages of progression: Most CRCs originate as visible precancerous polyps and only a small percentage arise as flat carcinoma. Not all polyps are pre-malignant and hence only some progress to carcinoma. While CRCs generally develop from polyps greater than 1cm, carcinoma has also been diagnosed in polyps between 6mm and 9mm ([19]). In this research, we consider three possible pathways for polyp progression before the onset of cancer. Figure 2.1 depicts the three pathways and the four stages of colorectal cancer. Polyp-pathways 1 and 2 refer to the progression types that begin with a visible adenoma polyp before progressing to cancer; this was adopted from the pathways presented by [19]. Non-polyp pathway refers to the cancers arising from flat polyps ([22]). After the in-situ stage, the polyp progresses through the three stages of invasive cancer: local, regional, and distant, which when related to Dukes classification of cancer ([23]), correspond to stages A+B, C, and D, respectively. Note that, while most cancers begin as pre-cancerous polyp, not all polyps progress to cancer, i.e, remain in the benign stages. Hence, polyps can be categorized into progressive and non-progressive ([19]). In what follows, we

present our probability model, results estimated from the probability model, the model validation, and concluding remarks.

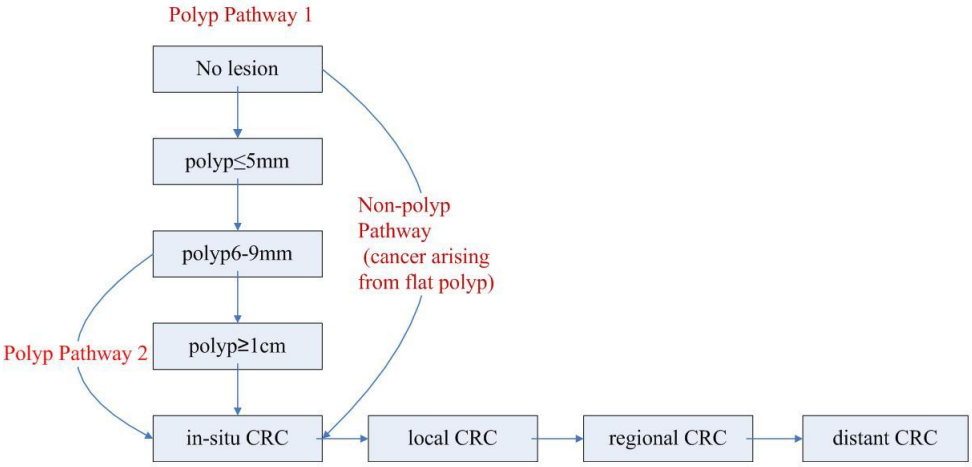

Figure 2.1. CRC pathways: Polyp incidence and stages of progression

2.2 Probability Model to Estimate Progression Rates

One of the main inputs required to develop the progression rate model is the *incidence rate* of polyps specific to patient's age, race, and family history status of CRC. In this section, we first present our method for estimating incidence rates, followed by the development of a probability model for estimating progression rates. The probability model was developed based on the assumption that non-polyp pathway contributes towards 15% of cancers, and 70% of cancers from the polyp pathways arise from pathway 1 (based on expert opinions cited in [19] and [11]). All notation used in the model presented in this section are summarized in Table 2.1.

2.2.1 Polyp Incidence Rates

We estimated the probability of incidence of $polyp \leq 5mm$, which is the first visible stage on the polyp pathways, by utilizing the case studies presented in [24] and

Table 2.1. Notation used in the probability model

Symbol	Description
p_5, $\mathcal{S}$, $\mathcal{L}$, $\mathcal{R}$, $\mathcal{D}$	events of incidence of polyp$\leq$5mm, in-situ CRC, local CRC, regional CRC, distant CRC
p_{5_t}, $\mathcal{S}_t$, $\mathcal{L}_t$, $\mathcal{R}_t$, $\mathcal{D}_t$	p_5, $\mathcal{S}$, $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{D}$ at time t
$\tilde{\mathcal{S}}_t$, $\tilde{\mathcal{L}}_t$, $\tilde{\mathcal{R}}_t$, $\tilde{\mathcal{D}}_t$	events of prevalence of in-situ CRC, local CRC, regional CRC, and distant CRC, at time t
$\tilde{\mathcal{C}}_t$	event of prevalence of CRC (in-situ, local, regional, or distant), at time t
$\tilde{\mathcal{I}}\mathcal{C}_t$	event of prevalence of *invasive* CRC (local, regional, or distant), at time t
F	random variable denoting family history of CRC of a randomly selected individual
R	random variable denoting race of a randomly selected individual
A_t	random variable denoting age (in years) of an individual at time t
L	random variable denoting length of life in years
Z^+	set of positive integers
N_S	number of stages from p_5 to $\mathcal{S}$
N_L	number of stages from $\mathcal{S}$ to $\mathcal{L}$
N_D	number of stages from initial event (p_5 or $\mathcal{S}$) to distant CRC ($\mathcal{D}$)
T_i^j	time to progress from event i to j
$\lambda_{i\|rf}^j$	progression rate from event i to event j given $R = r$ and $F = f$

[25]. The study in [24] presents data on the number of positive sigmoidoscopy results (indicating polyps) in a population that had tested negative (indicating normal) three years back. Using this repeated test data, we estimated the probability of incidence of polyp$\leq$5mm per year. Studies have shown that the rate of polyp incidence varies with age, however, [24] did not contain enough data to estimate the incidence probabilities based on age groups. Therefore, we used the statistics in [25], which presents age-based data of patients undergoing polyp detection and removal at the Rochester Methodist Hospital, Rochester, Minnesota. It has also been observed in the literature that a family history of CRC increases the risk of developing CRC ([26] and [25]). Therefore, we considered population specific incidence rates, which we estimated as follows.

Let $p_{5_{t_1}}$ denote an event of incidence of polyp$\leq$5mm at time t_1, and $A_{t_1} \in Z^+$ be the random variable denoting the age at t_1, where Z^+ denotes the set of positive integers. Let R and F be the random variables denoting the race and the status of family history of CRC, respectively, of a randomly selected individual. We consider $R = \{Caucasian,\ African\ American,\ Other\}$, and $F = \{0, > 0\}$, i.e., F has 2 outcomes, either no family history of CRC (event $F = 0$) or atleast one case of CRC

in the family (event $F > 0$). We computed the joint probabilities of $p_{5_{t_1}}$ and A_{t_1} in interval $[a, b]$, for given events of $F = f$ and $R = r$ as follows. Let the probability of $p_{5_{t_1}}$ given $R = r$ and $F = f$ (i.e., $P\{p_{5_{t_1}}|R = r \cap F = f\}$) be denoted by $P_{rf}\{p_{5_{t_1}}\}$. Applying the definition of conditional probability, and since $p_{5_{t_1}}$ and $R = r \cap F = f$ are dependent events we can write that

$$P_{rf}\{p_{5_{t_1}}\} = \frac{P\{p_{5_{t_1}} \cap R = r \cap F = f\}}{P\{R = r \cap F = f\}}, \tag{2.1}$$

and also since $P_{rf}\{p_{5_{t_1}}\}$ and $a \leq A_{t_1} \leq b$ are dependent events we can write that

$$P_{rf}\left\{(a \leq A_{t_1} \leq b) \cap p_{5_{t_1}}\right\} = P_{rf}\left\{(a \leq A_{t_1} \leq b)|p_{5_{t_1}}\right\} P_{rf}\left\{p_{5_{t_1}}\right\}. \tag{2.2}$$

The probability values for the elements on the right hand side of equations (2.1) and (2.2) are estimated using data from: [24] for probabilities of $p_{5_{t_1}}$; [25] for probability distributions based on A_{t_1} and F; and [27] for probability distributions based on R. A detailed description of the estimation is presented below.

1. Estimating $P\{F = f \cap R = r\}$: Using the definition of conditional probability, since $F = f$ and $R = r$ are dependent events, we can write $P\{F = f \cap R = r\} = P\{F = f|R = r\}P\{R = r\}$. To compute $P\{F = f|R = r\}$, we consider that the number of CRCs per family (i.e., the family history status F) is Poisson distributed with mean μ_r for a given race. We estimate μ_r for each race as $\{\frac{Number of CRC cases in the population (i.e., CRC prevalence count)}{Total population} * Average family size\}$. The CRC prevalence count for year 2006 for each race was obtained from SEER (Surveillance Epidemiology and End Results) database ([28]), which presents CRC statistics of the U.S. population. The total population count in year 2006 for each race was obtained from the U.S. census data. The average family size

of the U.S. population is 3.20, as reported by census ([29]). With the inclusion of second degree relatives, we assume that the average family size for all race is 7. Using the Poisson distribution probability density function, we compute $P\{F = f|R = r\} = \frac{(\mu_r)^f e^{-\mu_r}}{f!}$, and $P\{R = r\}$ can be easily computing using the U.S. census data. For equation (2.1), we compute $P\{F = 0 \cap R = r\}$ as above, and $P\{F > 0 \cap R = r\} = (1 - P\{F = 0|R = r\})P\{R = r\}$. We present below the estimates of μ_r and $P\{F > 0|R = r\}100$ in Tables 2.2 and 2.3

Table 2.2. Estimates of the Poisson parameter μ_r

R = All Race	R = Caucasian	R = African American
0.026	0.028	0.018

Table 2.3. Percentage of population with family history of CRC given race ($P\{F > 0|R = r\}$*100)

R = All Race	R = Caucasian	R = African American
2.55	2.77	1.77

2. Estimating $P\{p_{5_{t_1}} \cap R = r \cap F = f\}$ and $P_{rf}\left\{(a \leq A_{t_1} \leq b)|p_{5_{t_1}}\right\}$: The article in [24] presents part of the results of a large-scale randomized Prostate, Lung, Colorectal, and Ovarian Screening Trial (PLCO) [30], that was conducted to test the effect of various screening tests on mortalities from the cancers. As part of the colorectal cancer study, initially, a population in the age group 55-74 was screened for colorectal polyps. On the population that tested negative (indicating normal), a repeated screening test was conducted three years after the initial screen test. The number of positive (indicating presence of polyp) screen results from the repeated test has been presented in [24]. The diagnosed polyps have been categorized into sizes <0.5, 0.5-0.9, and ≥1.0 cm. Since all polyps should have started as <0.5cm with the event of incidence occurring during

one of the three years between tests, we estimate the probability of incidence of polyp≤5mm at an arbitrary year t_1 ($P\{p_{5_{t_1}}\}$) as $\frac{1}{3} * \frac{Numberofpeopletestedpositive}{Totaltested}$, where, we multiply by $\frac{1}{3}$ assuming that there were equal number of incidences in each of the three years. However, note that, age groups 40-54 and >74 were not part of the study population in [24], and hence, the above estimate of $P\{p_{5_{t_1}}\}$ will only apply to population in age 55-74. Therefore, to estimate $P\{p_{5_{t_1}}\}$ for the required population (i.e., age>40) we perform simple mathematical calculations using: i) the percentage distribution of polyps across age groups from [25], and ii) percentage distribution of U.S. population across age groups, taken from the U.S. census data.

Applying the definition of conditional probability, we compute $P\{p_{5_{t_1}} \cap R = r \cap F = f\} = P\{(R = r \cap F = f)|p_{5_{t_1}}\}P\{p_{5_{t_1}}\}$. Since, when given event $p_{5_{t_1}}$, we do not have data to determine dependence of events $F = f$ and $R = r$, we assume independence and compute $P\{(R = r \cap F = f)|p_{5_{t_1}}\}P\{p_{5_{t_1}}\} = P\{R = r|p_{5_{t_1}}\}P\{F = f|p_{5_{t_1}}\}P\{p_{5_{t_1}}\}$. We can estimate $P\{R = r|p_{5_{t_1}}\} = \frac{Numberofpolypcasesinracer}{Totalpolypcases}$, however, since we did not have suitable data to determine this proportion, we approximated $P\{R = r|p_{5_{t_1}}\} = \frac{NumberofCRCcasesinracer}{TotalCRCcases}$, data for which was obtained from the Indiana database ([27]). We estimate $P\{F = 0|p_{5_{t_1}}\} = \frac{Numberofpolypcaseswithfamilyhistoryof CRC}{Totalpolypcases} = 0.14$ using data presented in [25], and $P\{F > 0|p_{5_{t_1}}\} = 0.86$. Also, $P_{rf}\{(a \leq A_{t_1} \leq b)|p_{5_{t_1}}\}$ is equated to the proportion of polyps in respective age groups as presented in [25].

Note that, we consider the minimum age for developing a polyp as 40 years, since risk of cancer below 40 is low based on discussions presented in [31] and [32]. The report by the American Cancer Society in [31] notes that 90% of CRCs are diagnosed in individuals above the age of 50. Also, the expert panel from the U.S. Multisociety Task Force on Colorectal Cancer suggests a starting screening age of

50 years and 40 years for individuals without and with a family history of colorectal polyps, respectively ([32]).

2.2.2 Probability Model to Estimate Progression Rates

Based on expert opinion ([20], [9]), we consider that the progression times for the following events of incidences: polyp≤5mm to in-situ CRC, in-situ to local CRC, local to regional CRC, and regional to distant CRC are exponentially distributed with event dependent parameters (progression rates). It may be noted that accurate estimation of the progression rate from the incidence of pre-cancerous polyp (polyp≤5mm) to carcinoma (in-situ), is crucial for developing effective pre-cancer intervention strategies.

Let $\lambda^{j}_{i|rf}$ denote the progression rate from event i to event j given $R = r$ and $F = f$. In what follows, we present models for estimating $\lambda^{j}_{i|rf}$ for pre-cancer event (from incidence of polyp≤5mm to incidence of in-situ) considering polyp pathways 1 and 2 (see Figure 2.1), and post-cancer events (between incidence of different CRC stages) considering all pathways.

2.2.2.1 Estimating Pre-Cancer Progression Rate

Pre-cancer progression rate, which we will denote as $\lambda^{\mathcal{S}}_{p_5|rf}$, refers to the inverse of the expected time to progress from incidence of the first stage of visible polyp ≤5mm (p_5) to incidence of in-situ CRC ($\mathcal{S}$), given $R = r$ and $F = f$.

Note: Not all p_5 progress to S. Those that do progress are called progressive polyps and the rest non-progressive. In this research, the estimation of $\lambda^{\mathcal{S}}_{p_5|rf}$ considers cases of both progressive and non-progressive polyps. In other words, if the random value for the time to progress to in-situ, selected from the distribution exponential($\lambda^{\mathcal{S}}_{p_5|rf}$), is such that it exceeds the natural life, then the polyp is considered non-progressive.

We now present the model for estimating $\lambda^{\mathcal{S}}_{p_5|rf}$. Let $\mathcal{S}_{t_2}$ denote the event of incidence of in-situ CRC at time t_2. Recollect that $p_{5_{t_1}}$ denotes p_5 at time t_1, and A_t denotes age at time t, then following polyp pathways in Figure 2.1, $t_2 > t_1$ (as represented in Figure 2.2). For a population with $\mathcal{S}_{t_2}$, since events of $p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta$ are mutually exclusive (i.e., $\sum_\alpha \sum_{\beta>\alpha}(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta) = 1$) and exhaustive, we can apply the total probability rule and write that

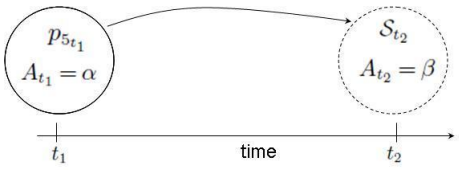

Figure 2.2. Event of incidence of polyp≤5mm at time t_1 ($p_{5_{t_1}}$) and its progression to an event of incidence in-situ CRC at time t_2 ($\mathcal{S}_{t_2}$), with age at t_1 and t_2 as α and β, respectively

$$P(\mathcal{S}_{t_2}) = \sum_\alpha \sum_{\beta>\alpha} P\left\{\mathcal{S}_{t_2}|(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} P\left\{p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta\right\}. \tag{2.3}$$

Let $T^{\mathcal{S}}_{p_5}$ be a random variable denoting the time to progress from p_5 to $\mathcal{S}$. Referring to Figure 2.2, $P\left\{\mathcal{S}_{t_2}|(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\}$ is equivalent to $P\left\{T^{\mathcal{S}}_{p_5} = \beta - \alpha|(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\}$, therefore, we can rewrite equation (2.3) as,

$$P(\mathcal{S}_{t_2}) = \sum_\alpha \sum_{\beta>\alpha} P\left\{T^{\mathcal{S}}_{p_5} = \beta - \alpha|(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} P\left\{p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta\right\}. \tag{2.4}$$

Applying conditional probability, equation (2.4) can be written as,

$$\begin{aligned} P(\mathcal{S}_{t_2}) = \sum_\alpha \sum_{\beta>\alpha} & P\left\{T^{\mathcal{S}}_{p_5} = \beta - \alpha|(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} \\ & P\left\{p_{5_{t_1}}|(A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} P\left\{A_{t_1} = \alpha \cap A_{t_2} = \beta\right\}, \end{aligned} \tag{2.5}$$

where, $P\left\{p_{5_{t_1}}|(A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\}$ can simply be written as $P\left\{p_{5_{t_1}}|A_{t_1}=\alpha\right\}$, since incidence of polyp$\leq$5mm at time t_1 is only dependent on age at t_1 and not on age at any future time t_2. Using the estimate from equation (2.2), $P\left\{p_{5_{t_1}}|A_{t_1}=\alpha\right\}$ can be computed as $\frac{P\left\{p_{5_{t_1}} \cap (a\leq A_{t_1}\leq b)\right\}}{b-a+1}\frac{1}{P(A=\alpha)}$, $a \leq \alpha \leq b$, i.e., by applying conditional probability and assuming constant rate of incidence within each age interval. Note that, the assumption is in accordance with that in the microsimulation model MISCAN-colon, that evaluates CRC screening policies ([9]), and whose input parameter values were based on expert estimates presented in meetings at the National Cancer Institute. Further, in equation (2.5), $P\left\{(T^{\mathcal{S}}_{p_5}=\beta-\alpha)|(p_{5_{t_1}} \cap A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\}$ can be substituted as $\lambda^{\mathcal{S}}_{p_5}e^{-\lambda^{\mathcal{S}}_{p_5}(\beta-\alpha)}$, where $\lambda^{\mathcal{S}}_{p_5}$ denotes progression rate from p_5 to $\mathcal{S}$. The remaining term on the right hand side of (2.5) can be estimated using population demographics from U.S. census data. Hence, $\lambda^{\mathcal{S}}_{p_5}$ can be estimated using (2.5) if the probability on the left hand side is available. However, $P\{\mathcal{S}_{t_2}\}$ is unknown, and is infeasible to estimate with the currently available data as explained below.

Let us divide the in-situ CRC stage as a series of sequential events $\{s_0, s_1, s_2, \cdot\cdot\cdot, s_n\}$, where s_i is an event indicating i time units of cancer progression in the in-situ stage. To relate $\mathcal{S}$ to s_i, we need to consider smaller time units (e.g., day), in which case $\mathcal{S}$ is equivalent to s_0, denoting the event of *epoch* of *incidence* of in-situ. Note that, for a diagnosed case of in-situ CRC, it is not possible to determine the value of i, and hence, we cannot obtain data related to the occurrence of each event s_i. Therefore, it is not feasible to estimate $P\{\mathcal{S}\}$. However, it is possible to estimate the probability of event of *prevalence* of in-situ CRC, i.e., $P\{\cup_{i=1}^{n} s_i\}$, as equal to the proportion of people in stage in-situ CRC in a randomized screening trial (we will denote $P\{\cup_{i=1}^{n} s_i\}$ at an arbitrary time t_2 as $P\{\tilde{\mathcal{S}}_{t_2}\}$). Due to the unavailability of a suitable randomized study that can be used to estimate $P\{\tilde{\mathcal{S}}_{t_2}\}$, we estimated $P\{\tilde{\mathcal{C}}_{t_2}\}$, which is the probability of prevalence of CRC at t_2. That is,

$P\{\tilde{\mathcal{C}}_{t_2}\} = P\{\tilde{\mathcal{S}}_{t_2} \cup \tilde{\mathcal{L}}_{t_2} \cup \tilde{\mathcal{R}}_{t_2} \cup \tilde{\mathcal{D}}_{t_2}\}$, where, the events in the probability term on the right hand side of the equation denote the prevalences of in-situ CRC, local CRC, regional CRC, and distant CRC, respectively, at time t_2. Therefore, Figure 2.2 is modified to include the above changes and is presented as Figure 2.3. As illustrated by Scenarios 1 through 4 in the Figure, for a randomly chosen individual at t_2, $\tilde{\mathcal{C}}_{t_2}$ corresponds to an event of prevalence of one of the CRC stages.

To reflect the above changes we modify equation (2.5) as follows,

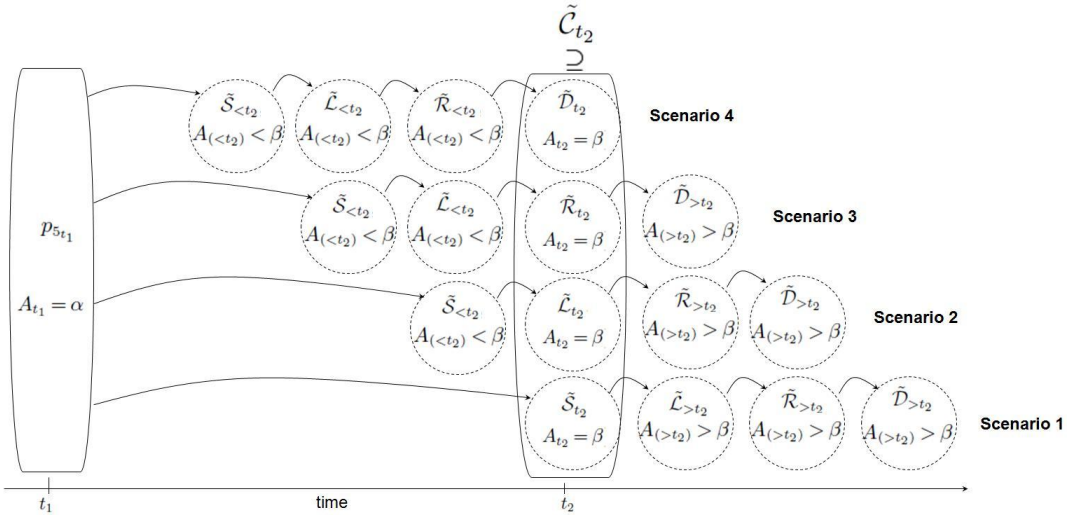

Figure 2.3. Event of incidence of polyp≤5mm at time t_1 ($p_{5_{t_1}}$) and its progression to an event of prevalence of CRC at time t_2 ($\tilde{\mathcal{C}}_{t_2}$) (i.e., either $\tilde{\mathcal{S}}_{t_2}$, $\tilde{\mathcal{L}}_{t_2}$, $\tilde{\mathcal{R}}_{t_2}$, or $\tilde{\mathcal{D}}_{t_2}$), with age at t_1 (A_{t_1}) and t_2 (A_{t_2}) as α and β, respectively

$$P\{\tilde{\mathcal{C}}_{t_2}\} = \sum_{\alpha} \sum_{\beta > \alpha} P\left\{T_{p_5}^{\mathcal{S}} \leq \beta - \alpha | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} P\left\{p_{5_{t_1}} | A_{t_1} = \alpha\right\} P\left\{A_{t_1} = \alpha \cap A_{t_2} = \beta\right\} \quad (2.6)$$

where, left hand side has been replaced with $P\{\tilde{\mathcal{C}}_{t_2}\}$. Also, we have $T_{p_5}^{\mathcal{S}} \leq \beta - \alpha$ instead of $= \beta - \alpha$, since, an occurrence of $\tilde{\mathcal{C}}$ at t_2 does not necessarily mean occurrence of s_0 (i.e., $s_0\tilde{\mathcal{C}}$). Also since, $\mathcal{S}$ (or s_0) is the first of the set of chronological events that make $\tilde{\mathcal{C}}$, age at $\mathcal{S}$ is $\leq \beta$ implying that $T_{p_5}^{\mathcal{S}} \leq \beta - \alpha$. $P\{\tilde{\mathcal{C}}_{t_2}\}$ was estimated using data from

[33], which presents screen results from CRC counseling and screening conducted by 10 health departments in 15 diverse counties in the state of North Carolina, as part of a pilot study on cancer coordination and control.

Referring to Figure 2.3, the upper bound on $T^{\mathcal{S}}_{p_5}$, i.e., $T^{\mathcal{S}}_{p_5} = \beta - \alpha$, is represented by Scenario 1, while a lower bound would be represented by Scenario 4. To quantify the lower bound, we can consider a value of one year, however, this may not be realistic and can be explained with the following examples. For a combination of values for $\{A_{t_1} = \alpha, A_{t_2} = \beta\}$ consider an example of $\{A_{t_1} = 40, A_{t_2} = 43\}$. Referring to Scenario 4, if $T^{\mathcal{S}}_{p_5} = 1$, it will imply that age at $\tilde{\mathcal{S}}_{t_2}$ is 41, and age at $\tilde{\mathcal{D}}_{t_2}$ is 43, i.e., it takes 2 years to progress from in-situ to distant CRC. Similarly, if instead, we consider an example of $\{A_{t_1} = 40, A_{t_2} = 70\}$, if $T^{\mathcal{S}}_{p_5} = 1$, it will imply that age at $\tilde{\mathcal{S}}_{t_2}$ is 41, and age at $\tilde{\mathcal{D}}_{t_2}$ is 70, i.e., it takes 29 years to progress from in-situ to distant CRC, while it took only 1 year to progress from polyp$\leq$5mm to in-situ CRC. This is highly unlikely since the progression between cancer stages is faster compared to precancer stages. Therefore, in order to place a more realistic lower bound, we consider equal time to progress between stages, and referring to Scenario 4, we write $T^{\mathcal{S}}_{p_5} \geq \frac{(\beta-\alpha)N_S}{N_D}$, where N_S and N_D are the number of stages from polyp$\leq$5mm to in-situ CRC and polyp$\leq$5mm to distant CRC, respectively. As an example, for polyp pathway 1 in Figure 2.1, $N_S = 3$ and $N_D = 7$. Note that, the equal time between stages was an assumption made only for obtaining a more realistic lower bound, than using an arbitrary value of one year, and was not an assumption on the progression rate estimation.

The modified equation can be written as,

$$P(\tilde{\mathcal{C}}_{t_2}) = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T^{\mathcal{S}}_{p_5} \leq (\beta-\alpha) | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} P\left\{p_{5_{t_1}} | A_{t_1} = \alpha\right\} P\{A_{t_1} = \alpha \cap A_{t_2} = \beta\} \quad (2.7)$$

where,

$$P\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T_{p_5}^{\mathcal{S}} \leq (\beta-\alpha)|(p_{5_{t1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} = \left[1 - e^{-\lambda_{p_5}^{\mathcal{S}}(\beta-\alpha)}\right] - \left[1 - e^{-\lambda_{p_5}^{\mathcal{S}} N_S\left(\frac{\beta-\alpha}{N_D}\right)}\right]$$

since the bounds on the progression time is equivalent to the event $T_{p_5}^{\mathcal{S}} \leq (\beta - \alpha) \bigcap T_{p_5}^{\mathcal{S}} \geq \frac{N_S(\beta-\alpha)}{N_D}$. Note that, the formulation in equation (2.7) will imply that every individual with p_5 at t_1 will live through to time t_2, however in reality this is not the case. Therefore, denoting $L \in Z^+$ as a random variable indicating length of life, we write (2.7) as,

$$P(\tilde{\mathcal{C}}_{t_2}) = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T_{p_5}^{\mathcal{S}} \leq (\beta-\alpha)|(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} P\left\{p_{5_{t_1}}|A_{t_1} = \alpha\right\} P\left\{A_{t_1} = \alpha \cap A_{t_2} = \beta\right\} P\left\{L > \beta\right\}$$

where,

$$\mathrm{P}\left\{\tfrac{N_S(\beta-\alpha)}{N_D} \leq T_{p_5}^{\mathcal{S}} \leq (\beta-\alpha)|(p_{5_{t1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} = \left[1 - e^{-\lambda_{p_5}^{\mathcal{S}}(\beta-\alpha)}\right] - \left[1 - e^{-\lambda_{p_5}^{\mathcal{S}} N_S\left(\frac{\beta-\alpha}{N_D}\right)}\right] \tag{2.8}$$

In equation (2.8), the only unknown value is the progression rate $\lambda_{p_5}^{\mathcal{S}}$ which can be computed by iteratively incrementing its value to where it best fits (2.8). Before doing so however, in order to obtain population-specific progression rates, we estimate $\lambda_{p_5|rf}^{\mathcal{S}}$, which denotes $\lambda_{p_5}^{\mathcal{S}}$ given race $(R = r)$ and family history status $(F = f)$, as follows.

Since events of $R = r \cap F = f$ are mutually exclusive (i.e., $\sum_r \sum_f P\{R = r \cap F = f\} = 1$) and exhaustive, applying the total probability rule we can write,

$$P(\tilde{\mathcal{C}}_{t_2}) = \sum_r \sum_f P\left\{\tilde{\mathcal{C}}_{t_2} | (R = r \cap F = f)\right\} P\left\{R = r \cap F = f\right\}$$
$$= \sum_r \sum_f P\left\{\tilde{\mathcal{C}}_{t_2} \cap R = r \cap F = f\right\}. \tag{2.9}$$

Note that, equation (2.8) for a given $R = r \cap F = f$ is equivalent to $P\{\tilde{\mathcal{C}}_{t_2} | (R = r \cap F = f)\}$ of equation (2.9). Therefore, we can write

$$P\left\{\tilde{\mathcal{C}}_{t_2} \cap R = r \cap F = f\right\} =$$
$$\left[\sum_\alpha \sum_{\beta > \alpha} P_{rf}\left\{\frac{N_S(\beta - \alpha)}{N_D} \leq T_{p_5}^{\mathcal{S}} \leq (\beta - \alpha) | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\}\right.$$
$$\left. P_{rf}\left\{p_{5_{t_1}} | A_{t_1} = \alpha\right\} P_r\left\{A_{t_1} = \alpha \cap A_{t_2} = \beta\right\} P_r\left\{L > \beta\right\}\right]$$
$$P\left\{R = r \cap F = f\right\} \quad \forall r \forall f, \tag{2.10}$$

where,

$$P_{rf}\left\{\tfrac{N_S(\beta-\alpha)}{N_D} \leq T_{p_5}^{\mathcal{S}} \leq (\beta - \alpha) | (p_{5_{t1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\}$$
$$= \left[1 - e^{-\left(\lambda_{p_5|rf}^{\mathcal{S}}\right)(\beta - \alpha)}\right] - \left[1 - e^{-\left(\lambda_{p_5|rf}^{\mathcal{S}}\right) N_S \left(\frac{\beta - \alpha}{N_D}\right)}\right].$$

Note that, occurrences of $P_{rf}\{(.)\}$ in (2.10) (and henceforth) represent $P\{(.)|(R = r \cap F = f)\}$, and have been written as such for notational convenience. Applying conditional probability, we can write $P\{\tilde{\mathcal{C}}_{t_2} \cap R = r \cap F = f\} = P\{(R = r \cap F = f)|\tilde{\mathcal{C}}_{t_2}\} P\left\{\tilde{\mathcal{C}}_{t_2}\right\}$. Since not enough data is available to determine the dependence of events $F = f$ and $R = r$ when given $\tilde{\mathcal{C}}_{t_2}$, we assume independence and write $P\left\{\tilde{\mathcal{C}}_{t_2} \cap R = r \cap F = f\right\} = P\left\{R = r | \tilde{\mathcal{C}}_{t_2}\right\} P\left\{F = f | \tilde{\mathcal{C}}_{t_2}\right\} P\left\{\tilde{\mathcal{C}}_{t_2}\right\}$. As mentioned earlier, $P\left\{\tilde{\mathcal{C}}_{t_2}\right\}$ is estimated using data presented in [33]. We compute $P\{(R =$

$r)|\tilde{\mathcal{C}}_{t_2}\} = \frac{Number\ of\ CRC\ cases\ in\ race\ r}{Total\ number\ of\ CRC\ cases}$, where the required numbers are obtained from the Indiana State Department of Health database ([34]). We consider $P\{(F = f)|\tilde{\mathcal{C}}_{t_2}\} = 0.2$ based on the observations reported by the American Cancer Society in [35]. Note that, estimates of $P\{R = r \cap F = f\}$ were earlier obtained for equation (2.1), the details of which are described in Section2.2.1. Therefore, the only unknown element in (2.10) is $\lambda^{\mathcal{S}}_{p_5|rf}$, and hence, can be easily estimated for all values of r and f.

2.2.2.2 Estimating Post-Cancer Progression Rates

This section discusses the estimation of progression rates between CRC stages, i.e, between stages in-situ, local, regional, and distant, with events of incidences denoted as $\mathcal{S}$, $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{D}$, respectively. A similar model as that developed for estimating $\lambda^{\mathcal{S}}_{p_5|rf}$ in equation (2.10) can be used in estimation of the progression rates between the CRC events. For example, consider event of incidence of in-situ at time t_2 ($\mathcal{S}_{t_2}$) and consider $\tilde{\mathcal{IC}}_{t_3}$, the event of prevalence of *invasive* CRC at time t_3 (i.e., $P\{\tilde{\mathcal{IC}}_{t_3}\} = P\{\tilde{\mathcal{L}}_{t_3} \cup \tilde{\mathcal{R}}_{t_3} \cup \tilde{\mathcal{D}}_{t_3}\}$), as represented by Figure 2.4. We can estimate the progression rate from $\mathcal{S}$ to $\mathcal{L}$ given $R = r$ and $F = f$, denoted as $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$, by the following equation

$$P_{rf}\{\tilde{\mathcal{IC}}_{t_3}\} = \sum_{\alpha}\sum_{\beta>\alpha} P_{rf}\left\{\frac{N_L(\beta-\alpha)}{N_D} \leq T^{\mathcal{L}}_{\mathcal{S}} \leq (\beta-\alpha)|(\mathcal{S}_{t_2} \cap A_{t_2} = \alpha \cap A_{t_3} = \beta)\right\} P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\} P_r\{A_{t_2} = \alpha \cap A_{t_3} = \beta\} P_r\{L > \beta\} \quad \forall r, \forall f \quad (2.11)$$

where,

$$\mathrm{P}_{rf}\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T^{\mathcal{L}}_{\mathcal{S}} \leq (\beta-\alpha)|(\mathcal{S}_{t_2} \cap A_{t_2} = \alpha \cap A_{t_3} = \beta)\right\} =$$

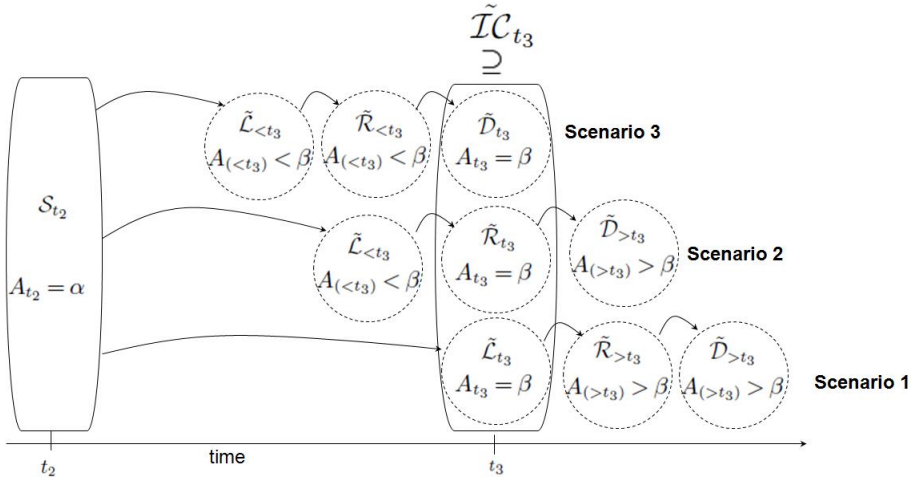

Figure 2.4. Event of incidence of in-situ CRC at time t_2 ($\mathcal{S}_{t_2}$) and its progression to an event of prevalence of invasive CRC at time t_3 ($\tilde{\mathcal{I}}\tilde{\mathcal{C}}_{t_3}$) (i.e., either $\tilde{\mathcal{L}}_{t_3}$, $\tilde{\mathcal{R}}_{t_3}$, or $\tilde{\mathcal{D}}_{t_3}$), with age at t_2 (A_{t_2}) and t_3 (A_{t_3}) as α and β, respectively

$$\left[1 - e^{-\left(\lambda^{\mathcal{L}}_{\mathcal{S}|rf}\right)(\beta-\alpha)}\right] - \left[1 - e^{-\left(\lambda^{\mathcal{L}}_{\mathcal{S}|rf}\right)N_L\left(\frac{\beta-\alpha}{N_D}\right)}\right],$$

$T^{\mathcal{L}}_{\mathcal{S}}$ denotes time to progress from $\mathcal{S}$ to $\mathcal{L}$, and N_L and N_D in the lower bound of $T^{\mathcal{L}}_{\mathcal{S}}$ now represent the number of stages from $\mathcal{S}$ to $\mathcal{L}$ and $\mathcal{S}$ to $\mathcal{D}$, respectively. $P_{rf}\{\tilde{\mathcal{I}}\tilde{\mathcal{C}}\}$ at an arbitrary time t_3 can be estimated as $P_{rf}\{\tilde{\mathcal{I}}\tilde{\mathcal{C}}\} = P_{rf}\{\tilde{\mathcal{I}}\tilde{\mathcal{C}} \cap \tilde{\mathcal{C}}\} = P_{rf}\{\tilde{\mathcal{C}}\}P_{rf}\{\tilde{\mathcal{I}}\tilde{\mathcal{C}}|\tilde{\mathcal{C}}\}$. $P_{rf}\{\tilde{\mathcal{C}}\}$ is obtained as earlier from [33], and $P_{rf}\{\tilde{\mathcal{I}}\tilde{\mathcal{C}}|\tilde{\mathcal{C}}\}$ is obtained using CRC diagnosed data from the Indiana database [36]. Therefore, from equation (2.11), we can estimate $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$ if the only unknown $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$ can be computed, since the rest of the terms can be obtained similar to the equivalent terms in equation (2.10). However, as explained in Section 2.2.2.1, it is infeasible to estimate $P\{\mathcal{S}\}$ using results from randomized screening trials, hence it is also not feasible to estimate $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$ using screening trials. Note however that, at this point in the model, we have estimated $\lambda^{\mathcal{S}}_{p_5|rf}$, the progression rate from p_5 to $\mathcal{S}$. Therefore, using the estimated values of $P_{rf}\left\{p_{5_{t_1}} \cap (a \leq A_{t_1} \leq b)\right\}$ from Section 2.2.1 and $\lambda^{\mathcal{S}}_{p_5|rf}$

from Section 2.2.2.1, we developed a model to estimate $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$, which is explained below.

The schematic of the probability model developed for estimating $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$ is presented in Figure 2.5, and can be interpreted as follows. Consider p_5 at time t_1 and its progression to $\mathcal{S}$ at t_2, $t_1 < t_2$. Now considering age at t_1 and t_2, if $c \leq A_{t_2} \leq d$ and $a \leq A_{t_1} \leq b$, then $\{\forall a, b, c, d : [a, b] \leq [c, d]\}$. For example, considering age intervals [40,49], [50,64], [65,74], and [75,], if $A_{t_2} = [40, 49]$ then $A_{t_1} = [40, 49]$, and if $A_{t_2} = [50, 64]$ then $A_{t_1} = [40, 49]$ or $A_{t_1} = [50, 64]$. Accordingly, $P_{rf}\{\mathcal{S}_{t_2} \cap (c \leq A_{t_2} \leq d)\}$ can be estimated by using $P_{rf}\left\{p_{5_{t_1}} \cap (a \leq A_{t_1} \leq b)\right\}$ and $\lambda^{\mathcal{S}}_{p_5|rf}$, $\forall a, b, c$, and d, as follows,

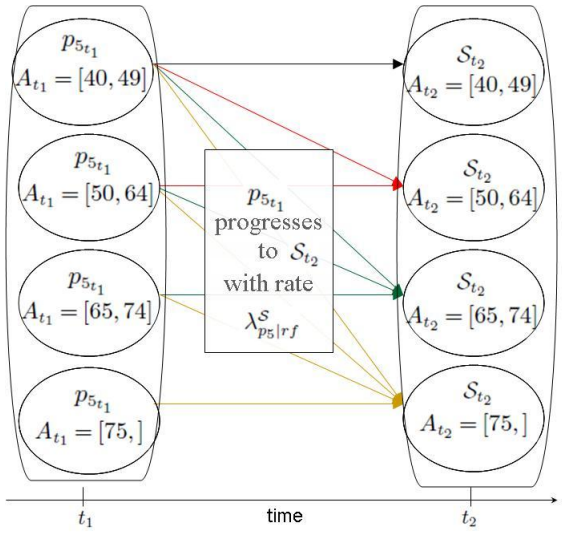

Figure 2.5. Event of incidence of polyp≤5mm at t_1 ($p_{5_{t_1}}$) and its progression to an event of incidence of in-situ CRC at t_2 ($\mathcal{S}_{t_2}$), with age $a \leq A_{t_1} \leq b$ and $c \leq A_{t_2} \leq d$ such that $[a, b] \leq [c, d]$

$$P_{rf}\{\mathcal{S}_{t_2} \cap (c \leq A_{t_2} \leq d)\} = \sum_{[a,b] \leq [c,d]} P_{rf}\left\{p_{5_{t_1}} \cap (a \leq A_{t_1} \leq b)\right\} \left[\sum_{k=a}^{b-1} \sum_{m=\max(k+1,c)-k}^{d-k} P_{rf}\{T^{\mathcal{S}}_{p_5} = m|(p_{5_{t_1}} \cap A_{t_1} = k)\} P\{L > m + k\}\right] \tag{2.12}$$

where, $P_{rf}\{T^{\mathcal{S}}_{p_5} = m\} = \left(\lambda^{\mathcal{S}}_{p_5|rf}\right) e^{-m\left(\lambda^{\mathcal{S}}_{p_5|rf}\right)}$. Note that, equation (2.12) was derived by a simple application of the total probability rule. For example, considering one specific age at A_{t_2} as 50, we can write $P_{rf}\{\mathcal{S}_{t_2} \cap A_{t_2} = 50\} = \sum_{\alpha<50} P_{rf}\{\mathcal{S}_{t_2} \cap A_{t_2} = 50|(p_{5_{t_1}} \cap A_{t_1} = \alpha)\}P_{rf}\{p_{5_{t_1}} \cap A_{t_1} = \alpha)\} = P_{rf}\{T^{\mathcal{S}}_{p_5} = 50-\alpha|(p_{5_{t_1}} \cap A_{t_1} = \alpha)\}P_{rf}\{p_{5_{t_1}} \cap A_{t_1} = \alpha)\}$. Equation (2.12) however has been written for an age interval, and has the variable L which denotes the length of life of an individual.

$P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$, required for equation (2.11), can now be estimated from $P_{rf}\{\mathcal{S}_{t_2} \cap (c \leq A_{t_2} \leq d)\}$ by applying conditional probability, and by considering a constant rate of incidence within each age interval, which is in accordance to literature as explained earlier. Equation (2.11) can now be solved to obtain $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$. The progression rates from local to regional CRC and regional to distant CRC can similarly be estimated by cyclically computing the probability of event of incidence (similar to computation of $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$ using equation (2.12)) followed by estimation of the progression rates (similar to estimation of $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$ using equation (2.11)). Note that, for stages past in-situ, L also includes survival based on stage of cancer in addition to the natural life of an individual.

2.3 Results: Estimated Incidence and Progression Rates

In Tables 2.4 and 2.5, we present rates of p_5 for polyp pathways and rates of $\mathcal{S}$ for non-polyp pathway, respectively, for different combinations of age, race, and family history status. For example, in Table 2.4, the percentage of incidence of polyp$\leq$5mm at age $50 \leq A \leq 64$ and $R = Caucasian$ and $F > 0$ is 4.25. The mean times to progress from event i to event j given $R = r$ and $F = f$, i.e., $\frac{1}{\lambda^{j}_{i|rf}}$, are presented in Tables 2.6 and 2.7, for polyp pathways and non-polyp pathway, respectively. The

values in Tables 2.4 through 2.7 will serve as input for developing a model of polyp progression. Such a model is essential for developing CRC intervention strategies.

Table 2.4. $P_{rf}(p_5 \cap a \leq A \leq b)100$: Percentage incidence of polyp≤5mm at age $[a, b]$, given $R = r$ and $F = f$, for polyp pathways

	All Race		R = Caucasian		R = African American	
Age Group $[a, b]$	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
$[40, 49]$	0.74	0.12	0.73	0.13	0.90	0.10
$[50, 64]$	4.33	0.70	4.25	0.74	5.28	0.58
$[65, 74]$	4.54	0.73	4.46	0.78	5.53	0.61
$[75,]$	2.13	0.34	2.09	0.36	2.59	0.29

Table 2.5. $P_{rf}(\mathcal{S} \cap a \leq A \leq b)100$: Percentage incidence of in-situ CRC at age $[a, b]$, given $R = r$ and $F = f$, for non-polyp pathway

	All Race		R = Caucasian		R = African American	
Age Group $[a, b]$	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
[40,49]	0.025	0.002	0.026	0.003	0.024	0.002
[50,64]	0.257	0.026	0.265	0.029	0.252	0.018
[65,74]	0.330	0.034	0.339	0.038	0.318	0.023
[75,]	0.344	0.040	0.346	0.044	0.332	0.026

Table 2.6. Mean times to progress from event i to event j, given $R = r$ and $F = f$ ($\frac{1}{\lambda^j_{i|rf}}$) on polyp pathways

	All Races		R = Caucasian		R = African American	
event $i \rightarrow$ event j	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
$p_5 \rightarrow$ in-situ [a]	23	41.6	21.5	39	29	50
in-situ $\rightarrow$ local	3.4	3.4	3.5	3.5	3.1	3.1
local $\rightarrow$ regional	5	5	4.5	4.5	3.5	4
regional $\rightarrow$ distant	0.95	0.95	0.95	0.95	0.88	0.9

[a] : $\lambda^{\mathcal{S}}_{p_5|rf}$ was estimated considering progressive and non-progressive polyp

2.3.1 Comparison of Results

The expected progression time estimates (i.e., Tables 2.6 and 2.7) can be used to compute one-step transition probabilities needed to build Markov models (as in [1]) for polyp progression. For example, probabilities for $R = Caucasian$ and $F = 0$ are depicted in Figure 2.6. We compared the transition probabilities derived from our model with those compiled by [1], who analyze the cost-effectiveness of screening for a population without a family history of cancer. Though the study in [1] is based

Table 2.7. Mean times to progress from event i to event j, given $R = r$ and $F = f$ ($\frac{1}{\lambda^j_{i|rf}}$) on non-polyp pathway

	All Races		R = Caucasian		R = African American	
event $i \rightarrow$ event j	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
in-situ $\rightarrow$ local	3.4	3.3	3.4	3.4	3.1	3.1
local $\rightarrow$ regional	3.5	4.8	4	4.5	3.5	3.5
regional $\rightarrow$ distant	0.9	0.95	0.9	0.95	0.9	0.9

Table 2.8. One-step transition probabilities between stages: Comparing results presented in this research to literature presented in [1]

Literature (Leshno et. al. (2003))		This research	
From $\rightarrow$ To Stages	Transition Probability	From $\rightarrow$ To Stages	Transition Probability
low-risk polyp $\rightarrow$ high risk polyp	0.02	$p_5 \rightarrow$ polyp$\geq$1cm	0.035
high-risk polyp $\rightarrow$ local CRC	0.05 (0.02-0.10)	polyp$\geq$1cm $\rightarrow$ local CRC	0.06
local CRC $\rightarrow$ regional CRC	0.28 (0.20-0.35)	local CRC $\rightarrow$ regional CRC	0.20
regional CRC $\rightarrow$ distant CRC	0.63 (0.50-0.70)	regional CRC $\rightarrow$ distant CRC	0.65

on the population of Israel, the reason for our comparison is to only check if our estimates are within commonly observed ranges, and is not meant as a validation. The polyp stages considered in [1] are low risk polyps (<1cm), high risk ($\geq$1cm) polyps, local CRC, regional CRC, and distant CRC, which, as seen in Figure 2.6, are slightly different from that in our model. Therefore, to obtain a rough comparison, we assumed equal progression time between stages p_5 and $\mathcal{S}$ (see pathway 1 in Figure 2.1), and computed the transition probability from p_5 to polyp$\geq$1cm. As shown in Table 2.8, for similar stages (i.e., rows 2, 3, and 4), the transition probabilities obtained from our model are comparable to that assumed in [1]. Using our mathematical modeling approach of progression rate estimation, we can further compute population-specific transition probabilities to build Markov models for developing effective CRC intervention strategies.

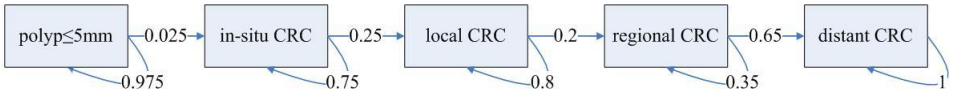

Figure 2.6. One-step transition probabilities for polyp pathway 1 (R= Caucasian, F=0)

2.4 Validation of Progression Rates Estimated from Probability Model

In order to validate the progression rates estimated in Section 2.2, we used a simulation based approach as follows. A simulation model was constructed such that it initially generates a population based on a user-input demographics data of specific populations. For validation purpose, we considered two different populations: population of the State of Indiana and population of the clinical trial described in [37, 38, 39], that was conducted in the State of Minnesota. Further, the simulation model was built such that it executes the following three events every year for each person in the population: *event 1*) updating age of each person, and creating new births and generating mortalities in the population; *event 2*) the natural incidence and progression of polyps using values presented in Tables 2.4 through 2.7; and *event 3*) screening based on the actual compliance rates of the corresponding population. Note that, based on change in age (through event 1), event 2 generates polyps and handles its natural progression until a successful screen (through event 3) leads to the polyp's diagnosis. The number and stage of the new cases of polyps that are diagnosed each year are recorded. For validation, the simulated statistics on diagnosed cases of CRCs are compared with the actual statistics of the corresponding population. The simulation model was constructed in *Repast* ([40]), a java agent-based modeling framework. The reason for using an agent-based approach is for ease of including the behavior and interaction between the system entities (including physician and insurance policies), which is a part of our future research for obtaining cancer intervention strategies. Simulation events 2 and 3, that were mentioned above, are described using flowcharts in Figures 2.7 and 2.8. We present below the details of the validation on the two populations.

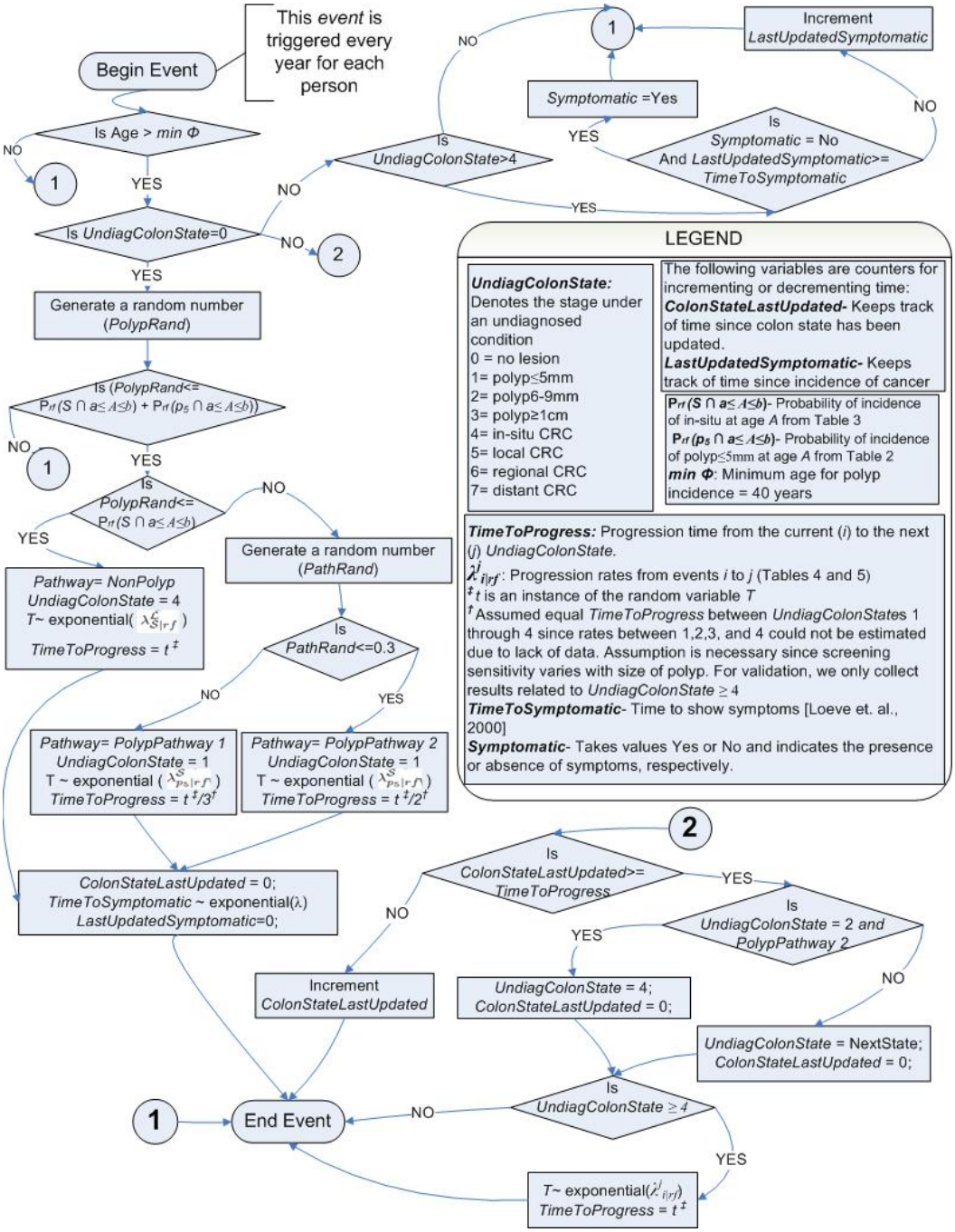

Figure 2.7. Flowchart of simulation Event 2: Incidence and progression of polyps

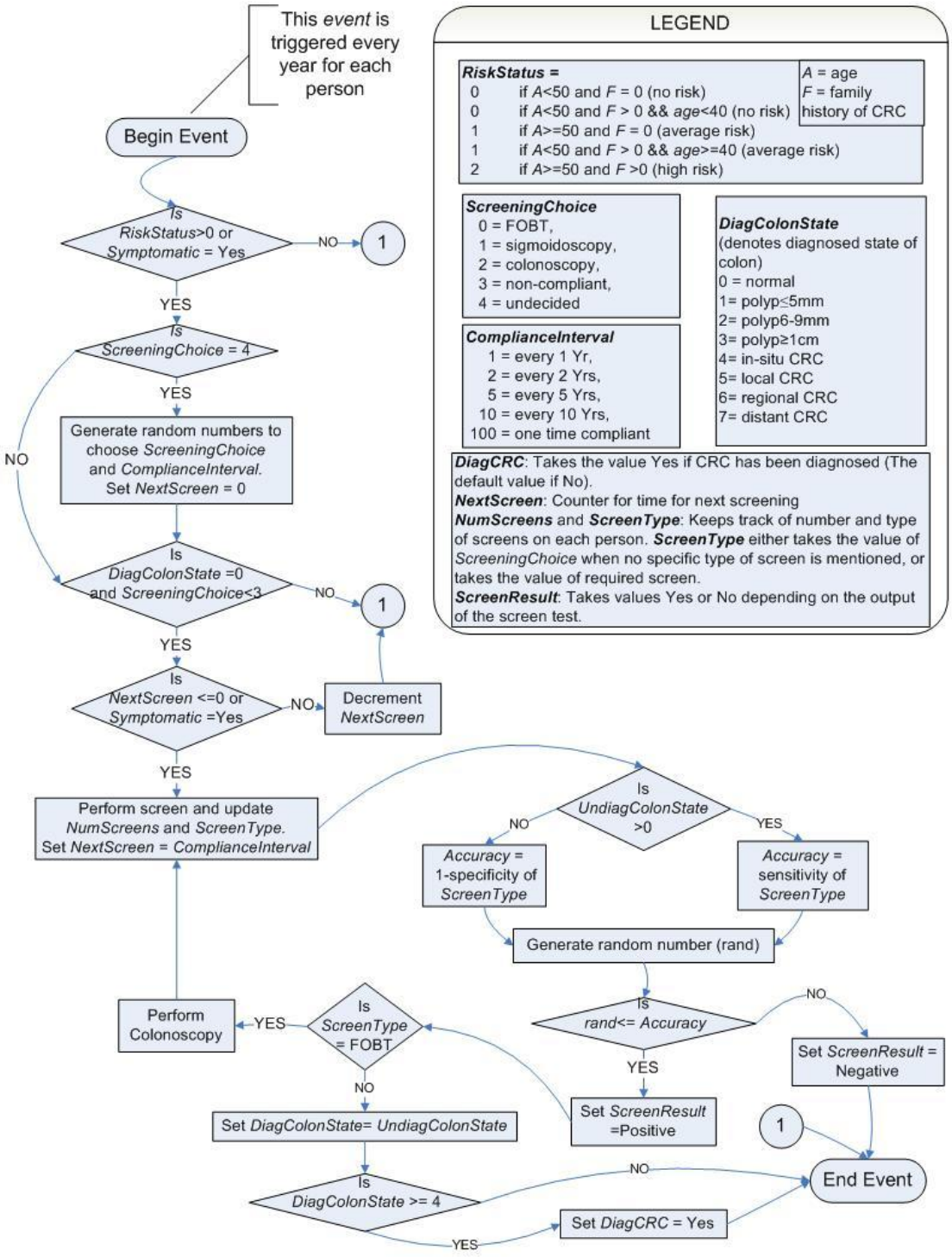

Figure 2.8. Flowchart of simulation Event 3: Screening

2.4.1 Simulation of the Indiana Population

The actual proportion of population in different race, sex, and age groups, as estimated from census data, was used to generate an initial sample population for the State of Indiana. The mortality and birth rates required for event 1 were also obtained from the census data. The incidence and progression rates required for event 2 were obtained from Tables 2.4 through 2.7. The screening rates required for event 3 were computed using the data obtained through a survey, which was conducted in year 2001 by the Indiana State Department of Health as part of the Behavioral Risk Factor Surveillance System ([2]). The survey considered three screening options: FOBT (fecal occult blood test), colonoscopy, and sigmoidoscopy. The sensitivity and specificity of the screening tests were taken from the values used in an intervention study conducted in year 2006 ([3]), and were also used in the MISCAN-Colon microsimulation model ([9]). The values of the screening parameters used in the simulation are summarized below.

2.4.1.1 Screening Parameters

The parameters used for percentage of population compliant to screening, screening sensitivity and specificity are summarized below in Tables 2.9, 2.10, and 2.11.

Table 2.9. Percentage of population compliant to screening ([2])

Screening Type	Percentage Compliant
FOBT	43
Sigmoidoscopy	19
Colonoscopy	19
NeverCompliant	19

Table 2.10. Screening sensitivity ([3])

Stage	FOBT	Sigmoidoscopy	Colonoscopy
poly<5mm	0.02	0.75	0.80
poly6-9mm	0.02	0.85	0.85
poly>1cm	0.05	0.95	0.95
in-situ	0.05	0.95	0.95
invasive CRC	0.60	0.95	0.95

Table 2.11. Screening specificity ([3])

FOBT	Sigmoidoscopy	Colonoscopy
0.98	0.95	0.90

2.4.1.2 Assigning Family History Status

For each person in the simulation, we need to determine the family history status. We assume that $F \sim Poisson(\mu_r)$ as described earlier while estimating $P\{F = f \cap R = r\}$ in Section 2.2.1. Note that, the CRC proportion (i.e., $\frac{Number of CRC cases}{Total population}$ in each race) for the State of Indiana is equivalent to the National proportion. Also, the average family size for the State of Indiana is 3.05 which is approximately equal to the National average (see estimation of $P\{F = f \cap R = r\}$ in Section 2.2.1). Therefore, to generate random numbers in the simulation, we use μ_r as presented in Section 2.2.1.

The simulation was run with a sample population of 10,000 for 30 trials. As presented in Tables 2.12 and 2.13, the confidence interval (CI) of the simulated results, related to new cases of CRCs diagnosed over 5 years, were compared with the actual 2000-2004 diagnosed cases of CRC available on the Indiana State Department of Health website ([27, 36]). Table 2.12 presents CRC counts per 100,000 of population. The large range in CI for the African American population can be attributed to its small percentage (8%) of the total population of Indiana. Table 2.13 presents the percentage distribution of CRCs among various stages at the time of diagnosis. As

can be seen in both tables, the actual values lie in between the simulated CI. Note that, some of the actual values in Table 2.13 fall on the boundary of the simulated CI, which can be attributed to the fact that about 6.5% of actual CRC cases did not have a stage identifier (un-staged). It may be noted that Table 2.13 serves as a verification because, the percentage distribution of diagnosed CRC in different stages was initially used in the probability model. However, Table 2.12 serves as a validation, as the simulated CRC results presented in the Table are not equivalent to the cancer prevalence probabilities ($P\{\tilde{\mathcal{C}}_{t_2}\}$ estimated using [33]) that were initially used in the probability model. The difference between the two is that:

1. The cancer prevalence probabilities used in the probability model were estimated using data from *clinical trial* studies (not from Indiana database). However, we compare the simulated diagnosed cancer counts with the *actual diagnosed counts* in Indiana.

2. The second difference is inherent in *clinical trial rates* versus *actual diagnosed counts* itself. The former statistic includes all cases of cancer in the population, since, in a clinical trial, all participants get screened (ignoring screen sensitivity which is the same in both cases). However, the latter statistic does not include all cases, since, in an actual population, not everyone is compliant to screening.

Therefore, the accuracy of the simulated diagnosed cancer counts are dependent on the population's screening compliance rates as well as the polyp natural incidence rates and expected progression times estimated from the probability model. Since the compliance rates were computed from the Indiana population database, the results in Table 2.12 serve as a validation of the probability model.

Table 2.12. Simulated vs. actual Indiana CRC counts per 100,000 of population

Stages	Race	Simulated 95% CI		Actual Indiana Counts
		Lower CI	Upper CI	
	All Race	48.83	57.55	56.02
local + regional + distant	Caucasian	52.50	61.97	57.68
	African American	31.97	56.53	47.93
in-situ + local + regional + distant	All Race	52.98	61.90	60.70

Table 2.13. Simulated vs. actual Indiana values for stage at time of diagnosis as percentage of total CRC counts

CRC Stage	Simulated 95% CI		Actual Indiana Values
	Lower CI	Upper CI	
in-situ	6.02	9.04	7.70
local	34.99	41.88	34.94
regional	29.58	36.75	33.75
distant	17.99	23.76	17.12
un-staged	NA	NA	6.50

NA- Not Applicable

2.4.1.3 Results from the Simulation Model

The above simulation model was developed to validate the probability model. However, the combination of the probability model followed by the simulation, as constructed in this research, serves as a model in itself for obtaining certain polyp related estimates of interest. One such set of estimates is related to the progressive polyps. The simulated CIs on the maximum likelihood estimate of the exponential distribution parameter, i.e., on the mean *time to progress* (in years) from p_5 to $\mathcal{S}$ given $R = Caucasian$, are presented in Table 2.14. It may be seen that, our estimated value for the progression from polyp$\leq$5mm to in-situ CRC (row 1 of the Table) compares well with expert opinion in ([18]), where, an average time of approximately 10 years to progress from adenomatous polyp (mainly <1cm) to invasive CRC (local CRC and beyond) is suggested. Use of a mathematical modeling approach for estimating population-specific values, as in this research, allows us to quantify any variations across populations. See, for example, Table 2.14, which shows shorter progression time to cancer for a population with $F > 0$. We also obtained results for the

proportion of polyp≤5mm progressing to in-situ CRC (i.e., proportion of progressive polyps), which are presented in Table 2.15 for $R = Caucasian$. Note that, the proportion of progressive polyps in a population with $F > 0$ is approximately 1.8 times as much as that in a population with $F = 0$. Though it known that a family history of CRC increases the life-time chances of cancer, such mathematical quantifications (Tables 2.14 and 2.15) of polyp progression could not be found in the literature.

Table 2.14. Estimated confidence interval on mean time to progress from polyp≤5mm to in-situ CRC (in years) according to family history

Family History(F)	Upper 95% CI	Lower 95% CI
All[a]	10.7	8.3
$F = 0$	12.1	9.4
$F > 0$	9.9	7.7

[a] includes $F = 0$ and $F > 0$

Table 2.15. Estimated proportion of p_5's progressing to $\mathcal{S}$

Family History(F)	Proportion (in %)
All[a]	20.9
$F = 0$	19.5
$F > 0$	34.2

[a] includes $F = 0$ and $F > 0$

2.4.2 Simulation of Minnesota Study

The authors in [37, 38, 39] present a clinical trial conducted in Minnesota, where a population in age group 50-80 years with no history of cancer was recruited and randomly divided into 3 groups. Groups 1 and 2 were subject to annual and biennial FOBT screening, respectively, and group 3 was a control group. The objective of the study was to identify the difference in CRC related mortality rates among the three groups, and hence analyze the effect of annual and biennial FOBT screening on mortalities. Phase I of the study was conducted from 1978 to 1982, and continued to Phase II from Feb 1986 to Feb 1992. Study groups 1 and 2 were simulated, separately, by utilizing the values for proportions of people in age ranges 50-59, 60-69, and 70-80

that were given by [37], as follows. The simulation first generated people between age 0-30 years, with proportions of people in age ranges 0-9, 10-19, and 20-30 equal to the proportions of people in age ranges 50-59, 60-69, and 70-80, respectively, of the actual study. The simulation was first run for 50 years so that the population is now between age group 50-80 years, and then later run for a period of 14 years representing the timeline of the actual clinical trial. The mortality and birth rates (event 1) were kept at zero during the first 50 years. Event 2, i.e., polyp incidence and progression was run during the entire period, and for which the rates were obtained from Tables 2.4 through 2.7. During the first 50 years, any symptomatic cases of CRC were removed from the simulation in order to remove existing diagnosed cases of cancer, and the proportion of population in the three age groups were adjusted. The exponential mean time to symptomatic was taken as per the times (preclinical to clinical) considered in the MISCAN-colon model ([10]), whose parameters, as mentioned earlier, were based on expert estimates presented in meetings at the National Cancer Institute. During the 14 year run that represented the actual study, study groups 1 and 2 were subject to annual and biennial screening, respectively, as per screening details and test sensitivity provided by [37] (event 3).

The simulated cases of CRC during 13 years of the study were compared with the actual cases given by [37] and the results (represented as CRC cases per 1000 population) are presented in Figures 2.9 and 2.10 for annual and biennial screening groups, respectively. Since the simulated screening intervals matched that in the clinical trial, the accuracy of the simulated CRC cases is dependent on the natural polyp progression, whose rates were estimated from the probability model. Moreover, tracking a population and comparing results over a 13 year period is a stronger analysis of the polyp progression, and therefore, Figures 2.9 and 2.10 can be used for validating the probability model.

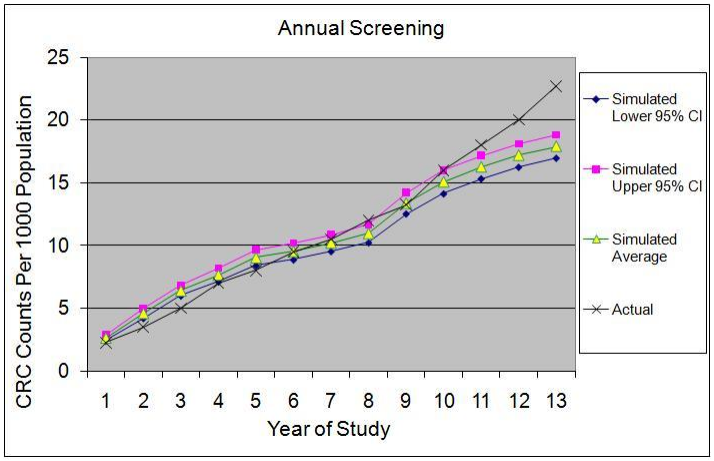

Figure 2.9. Comparing simulated versus actual CRC cases per 1000 population for annual group of the Minnesota study

As seen in the Figures, there is a good agreement between the actual and simulated values, with only slight deviations in the annual case. While the actual values for biennial group falls within the simulated 95% confidence interval in most years, the actual values for annual group are lower during Phase 1 of the study and higher during Phase II of study. The aforementioned deviations in values can be explained as follows.

1. Lack of data on the number of screenings: It is noted by [37] that not everyone participated in the scheduled number of screens, which was 11 for annual group and 6 for biennial group ([38]) over the entire duration of study. For each person, let X= number of screens obtained during study duration. Using information provided for each group in the study, data could be extracted for the following features: for annual group - percentage of people with $X \geq 1$, $X \geq 6$, $X \geq 9$, and $X = 11$; and for biennial group - percentage of people with $X \geq 1$, $X \geq 3$, $X \geq 5$, and $X = 6$. Note that, when we compare the information available for X between the two groups, biennial group has more information compared to that of annual group. Under the biennial group, the extracted information

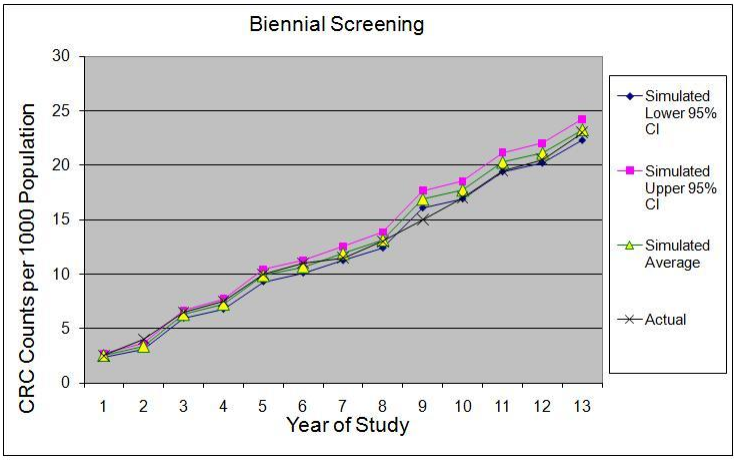

Figure 2.10. Comparing simulated versus actual CRC cases per 1000 population for biennial group of the Minnesota study

was used to obtain probability that $X = 1$ *or* 2, i.e., $P(1 \leq X \leq 2)$, and was split uniformly between $P(X = 1)$ and $P(X = 2)$, and similarly $P(3 \leq X \leq 4)$ was split between $P(X = 3)$ and $P(X = 4)$. However, under the annual group, $P(1 \leq X \leq 5)$, $P(6 \leq X \leq 8)$, and $P(9 \leq X \leq 10)$ had to be split between 5, 3, and 2 values of X, respectively. This lack of information for annual group can be considered to partly account for the difference in simulated to actual values. Also, under both groups, if an individual had participated in n number of screens, it was assumed that it was the first n of the scheduled screens, while in reality the screenings could have been spread over the period of study. We can expect that this assumption will cause greater and noticeable difference, between simulated and actual values, for the annual group than the biennial group because, e.g., $X = 5$ under biennial screening is spread over only 6 possible screening schedules while $X = 5$ under annual screening is spread over 11 possible screening schedules. From the assumption stated above and lack of data for the annual screen, we can expect that there will be more than actual number of screens during the initial period of study and hence more CRCs

get diagnosed in the simulation. Further, we can also expect that, along with causing lesser than actual number of screens during the latter period of study, increased screening in the beginning reduces the number of polyps that progress to CRC, and hence further reducing CRC incidence during the latter period. This is evident in Figure 2.9.

2. Lack of information on screening outside of the study: As mentioned in the study, participants could have obtained screening from a source outside the study during and in-between the two phases, and yearly updates were obtained on any diagnosed cases of CRC. Since information on outside screening was not available, it was assumed that a person undergoes screening outside of study in symptomatic cases only. The time to symptomatic was extracted from the MISCAN-colon parameters presented in [10]. The above assumption together with the fact that advanced stages of CRC tend to be more symptomatic than earlier stages, we can expect that, while comparing simulated minus actual CRC cases per stage at diagnosis, the value will be higher for distant stage compared to regional compared to local. This trend is evident in Tables 2.16 and 2.17, for annual and biennial groups, respectively, which presents the CRC cases per 1000 population according to stage at diagnosis over the 13 year period. Note that, in the clinical trial, staging of CRC was done by using Dukes classification. We considered Dukes stages A and B as equivalent to local, and Dukes stages C and D as equivalent to regional and distant, respectively.

Based on our assumptions 1) that an individual undergoes the first n of the scheduled screens, and 2) that outside screening was done only in symptomatic cases, we can hypothesize that the percentages of diagnosed CRCs in stages regional and distant are more during the latter period of study compared to those in the initial period.

Table 2.16. CRC counts per 1000 population and stage at diagnosis - For annually screened group of Minnesota study

Simulated 95% CI			Actual Counts	
CRC Stage	Lower CI	Upper CI	Dukes CRC Staging	Actual Values
local	8.83	9.76	A+B	13.3
regional	4.64	5.36	C	5.6
distant	2.87	3.56	D	2.3
un-staged	NA	NA	un-staged	2.1
Total CRC	16.34	18.68	Total CRC	23.3

Table 2.17. CRC counts per 1000 population and stage at diagnosis - For biennially screened group of Minnesota study

Simulated 95% CI			Actual Counts	
CRC Stage	Lower CI	Upper CI	Dukes CRC Staging	Actual Values
local	9.51	10.76	A+B	11.6
regional	6.77	7.73	C	6.1
distant	4.66	5.49	D	3
un-staged	NA	NA	un-staged	2.1
Total CRC	20.94	23.98	Total CRC	22.8

The reasoning leading to the hypothesis can be explained as follows. Assumption 1 generates more than actual screenings during the initial few years, thus causing less cases to reach a symptomatic stage. Consequently, the latter duration has lesser than actual number of screens causing more cases to reach symptomatic stage. Since advanced stages of CRC are more symptomatic than earlier stages, we can thus hypothesize that, the consequence of assumption 1 combined with assumption 2 causes diagnosis of greater percentage of regional and distant cases during latter period of study compared to initial period. This hypothesis can be verified by the numbers in Tables 2.18 and 2.19. As speculated, the percentage of CRCs in regional and distant stage is higher at the end of year 13 compared to that at the end of year 5.

2.5 Concluding Remarks

Precise values of polyp incidence and progression rates are crucial for developing population-wide CRC intervention strategies. Polyp incidence rates for population groups characterized by age, race, and family history of CRC, were estimated by using

Table 2.18. Stage at diagnosis as percentage of total CRC counts - For annual group of Minnesota study

	Simulated 95% CI				Actual Values	
	End of 5 Years		End of 13 Years		End of 13 Years	
CRC Stage	Lower CI	Upper CI	Lower CI	Upper CI	Dukes CRC Staging	Actual Values
local	58.57	64.63	51.20	55.24	A+B	57.08
regional	20.99	26.60	27.07	29.86	C	24.03
distant	12.43	16.78	16.62	20.00	D	9.87
un-staged	NA	NA	NA	NA	un-staged	9.01

Table 2.19. Stage at diagnosis as percentage of total CRC counts - For biennial group of Minnesota study

	Simulated 95% CI				Actual Values	
	End of 5 Years		End of 13 Years		End of 13 Years	
CRC Stage	Lower CI	Upper CI	Lower CI	Upper CI	Dukes CRC Staging	Actual Values
local	48.34	53.56	43.28	46.70	A+B	50.88
regional	28.93	34.35	30.48	34.29	C	26.75
distant	15.13	19.70	20.88	24.36	D	13.16
un-staged	NA	NA	NA	NA	un-staged	9.21

data from the literature. The data sources included clinical CRC screening trials, population databases, and evidence-based reports from state health departments and national institutes. The natural progression timeline of polyps, on the other hand, could not be directly estimated using observed data, since it is infeasible to allow a diagnosed polyp to progress naturally without treatment. Hence, we developed a probability model to estimate population-specific rates of polyp progression. The probability model was constructed based on known concepts of the natural progression of polyps. Thereafter, using the model, data from the above mentioned sources were synthesized to estimate progression rates. These rates are characterized by race and family history of CRC and correspond to both progressive and non-progressive polyps. The estimated incidence and progression rates (presented in Tables 2.4 through 2.7) were used to simulate the natural history of colorectal polyps for the population in the State of Indiana and a subset of the population in the State of Minnesota. The simulation results were used to validate the probability model. The simulation model also yielded 1) the expected time for progressive polyps to reach in-situ CRC from

the polyp≤5mm stage, and 2) the proportion of polyps reaching in-situ CRC (i.e., progressive polyps). Tables 2.14 and 2.15 present results for the progression time and proportion of the progressive polyps for population with and without family history of CRC.

2.5.1 Discussion of Results

The polyp progression related values available in the literature are mainly experience based approximations and are not population-specific. Though the literature contains several data sources related to the incidence of either precancerous polyp or carcinoma, these data had not been synthesized to mathematically estimate population-specific polyp progression times (as in Tables 2.6 and 2.7). Mathematical estimation will help identify and quantify any variation across populations which are critical for developing early intervention strategies.

The probability model was developed to estimate rates considering both progressive and non-progressive polyps, and was subsequently used in the simulation model to obtain statistics of progressive polyps. Such an approach is significant, since, while it is known that a family history of CRC increases the risk of cancer, quantification of the increased risk based on proportion of progressive polyps (as in Table 2.15) has not been presented in the literature. Also, the risk based on progression time from polyp≤5mm to in-situ CRC had not been mathematically quantified, as in Table 2.14. Consideration of both progressive and non-progressive polyps also supports development of more comprehensive intervention strategies comprising resource needs and allocation.

2.5.2 Accuracy and Estimation Errors

Though our model estimates polyp progression rates specific to race and family history of CRC, for better accuracy of the estimates, it is essential that the model be expanded to consider other dependent factors. Examples of dependent factors include, number and histology of polyps, classification of first and second degree relatives with CRC, and personal history of other medical conditions. However, inclusion of these factors in the model would require significant additional data, which is currently unavailable. Also, estimating progression rates between each of the pre-in-situ stages would help to simulate a more comprehensive natural progression of polyps. As in any estimated value, the progression rates presented in this manuscript could contain some error. While synthesizing data from various sources is beneficial for estimating the rates, the variation in the data acquisition processes across these sources could induce some estimation error. For example, the value of $P\{F > 0|p_{5_{t_1}}\}$ was estimated based on diagnosed data at the Rochester Methodist hospital ([25]). However, the value of $P\{F > 0|\tilde{\mathcal{C}}_{t_2}\}$ was estimated based on expert observation reported by the American Cancer Society. Since these values were based on either large amount of data or long-term observations, we expect the error to be relatively small. Due to the unavailability of required data, some of the values were based on assumptions, hence creating room for estimation error. For example, based on expert opinion in the literature, the time to progress was assumed to follow exponential distribution. Though we cannot empirically ascertain this assumption, based on the validation results, we believe that the exponential distribution is a good alternative. It may be noted that, this research presents a model framework that has the potential to estimate population-specific progression rates the accuracy of which can be improved

as more data becomes available. Such a model is significant for obtaining population-specific intervention strategies.

2.6 Future Research

Obtaining effective cancer intervention strategies encompasses not only development of screening strategies, but also analyzing factors pertaining to the availability of resources such as the patient's access to physician and hospital, and effective dissemination of evidence based information to the population. For example, it would be useful to assess the population's compliance to screening guidelines based on features like the patient's knowledge of cancer screening tests and cancer risk factors. This knowledge can be related to the patient's access to information through interaction with their physician and/or through other sources. The model can then be used for a cost-effectiveness analysis of programs to increase risk awareness and its impact on reduction in cancer cases. It would also be interesting to model the impact of insurance policies under different system settings. Therefore, in addition to simulating the population entity, we need to include entities like physicians and insurance policies, and their interactions. Note that, the current simulation model has been constructed as an agent-based model for the convenience of developing such a system-based simulation, which is part of our future research. Such a systems approach will allow for a more realistic analysis of feasible intervention strategies.

In summary, the probability model in the current state can be considered a base model that presents potential for use in developing population-specific intervention strategies. The work presented here can be used to support the need for collection of specific data required for analyzing and identifying more population-specific factors of interest. While the model developed in this manuscript has been specifically applied to colorectal cancer, most diseases follow a similar pattern, i.e., incidence

and progression. Following a similar procedure, but with disease specific modeling details, the current framework could be utilized for estimating progression and developing intervention strategies for other cancers as well. In addition, by inclusion of a transmission model in the current framework, cost-effective analysis of prevention programs for infectious diseases like HIV/AIDS could be developed.

CHAPTER 3

A DENOISING METHODOLOGY FOR MICROARRAY

3.1 Introduction

Discovery of the human genome generated significant anticipation for better understanding of the roles played by the genes on cell behavior and the resulting impact on human health [41, 42]. Most diseases result from cell dysfunction, which can be traced to alteration in the structure of one or more genes (biomarkers). Alterations could be in the form of abnormal increase or decrease in the expression (activity) levels of genes and their patterns. Identifying different patterns in biomarker genes could hold the key to disease diagnosis and individualized treatment planning, which is of vital interest as identified by the National Academy of Engineering under one of the Grand Challenges- Engineering Better Medicine. Hence, our ability to first accurately measure the expression levels of the genes is crucial towards identifying gene biomarkers. The microarray technology has revolutionized the field of genomics by offering the capability to measure expression levels of tens of thousands of genes simultaneously. However, during the process of gene expression estimation, noise from various sources gets added to the expression value. The noise generally originates during the phases of sample preparation, hybridization, and scanning [43]. The sample preparation noise initiates from the process of RNA amplification [44], and the hybridization noise refers to the randomness in the process of RNA binding to

the probes. The sources of scanning noise include leak of external light, variations in laser intensity, and presence of dirt [45, 43].

A microarray is a tiny chip (1.28cm X 1.28cm) which is divided into approximately a million squares which we will refer to as probe squares. A number of such probe squares across the array are randomly allocated to each gene [44]. The complexity of the arrangement of microarray probes and the types of inherent noise render a significant challenge for denoising. Some of the existing methods for denoising microarray data can be found in [46, 47, 48, 49, 50, 51]. In [46] the authors develop statistical models for analyzing hybridization and cross-hybridization, and use measures of cross-hybridization to improve the quality of gene expression estimates. A probability model characterizing the nature of molecular-binding (hybridization) in affinity based biosensors is presented in [47]. The methodology developed in [48] employs a multiresolution approach, in which a 2-D stationary wavelet transform is applied across a microarray image. Various other methodologies [49, 50, 51], focus on denoising of microarrays involved in identifying differentially expressed genes. These methods generally require multiple arrays or two-color microarrays. Noise boundary models are developed in [49] using two replicate chips of normal tissues, and the resulting threshold boundaries are applied on fold change obtained between cancer tissue and normal tissue. Two-color microarrays are denoised in [51] by considering the control and experimental sets as two component vector arrays and use multi-channel image processing techniques.

The disadvantages in the use of control sample based approaches of denoising are as follows.

1. Use of such methods, that require processing of up to two additional microarray chips, adds significant cost (over a thousand dollars) in disease diagnosis and treatment planning applications.

2. During microarray chip processing, different amounts of hybridization and scanning noise get added each time the process is performed, even under a controlled environment. Therefore, when two replicate chips of control samples are processed, the difference in their gene expressions is a result of noise that are specific to the images of the two chips. Hence, the difference in gene expressions of the control images cannot be used to derive a noise threshold for denoising other microarray images, since they are likely to contain different sets of noise.

3. In methods that use the difference in final gene expression values, between control and case samples, it is difficult to differentiate noise from actual signal for the case of low expressed genes. That is, while significant difference in high expressed genes can be easily identified, significant difference in low expressed genes could be falsely classified as noise.

We present a novel and comprehensive methodology for removing hybridization and scanning noises from Affymetrix microarray images before the image data is processed by Affymetrix software to obtain final gene expression values. The method uses data from within the image that needs to be denoised and does not require control samples. Since noise arises from a variety of sources, the methodology uses a multiresolution analysis approach on the image to effectively isolate the noise at different frequencies. The image is decomposed using a dual tree complex wavelet transform, which is shown to have better properties with regard to shift invariance, directional selectivity, and perfect reconstruction, compared to transforms using real wavelet functions [52]. Natural images are generally known to contain a small number of edges, and hence when it is decomposed at different frequencies it results in a small number of large coefficients [53, 54]. This facilitates the process of separation of noise from significant data using thresholding [53]. It was noticed that a

similar direct use of the existing multiresolution denoising technique was ineffective when applied to microarray images. We identified two major features of microarray images that were the cause of the ineffectiveness: 1) presence of numerous edges in microarray images results in a vast number of large coefficients during decomposition, hence hindering the noise separation, and 2) non-Gaussian (Poisson) characteristic of the hybridization noise added to the denoising inefficiency, since most thresholding methods require error to be Gaussian. A more detailed discussion of the topics of *presence of numerous edges* and *Poisson noise* in microarray images is presented in Section 3.2.2. To alleviate the above difficulties, we developed two forms of data transformations: 1) extraction of the probe squares that are assigned to a gene on a microarray, and construction of a separate dyadic *subimage* for each gene, and 2) for each subimage, Gaussian transformation of the Poisson noise. For each modified subimage data, we apply a dual tree complex wavelet transform followed by bivariate shrinkage thresholding. The thresholding technique underlying the bivariate shrinkage method considers the interdependencies of the detail coefficients in adjacent levels of decomposition producing better performance [55]. Thereafter, the denoised version of the microarray image is created using the probe squares from the denoised subimages. Final expression values of the genes are obtained from the probes using Affymetrix software.

Since it is not possible to know the true values (ground truth) of the expression levels of the genes for a tissue sample, it is difficult to assess the performance of denoising on a microarray. We address this problem by constructing a sample data set mimicking a *subimage* for a gene, and use it to benchmark our methodology. We then implement our methodology on Affymetrix GeneChip human genome HG-U133 Plus 2.0 array [56] data sets, obtained by processing a sample from HCT-116 cell line at the Microarray Core Facility at Moffitt Cancer Center and Research Institute. Each

HG-U133 Plus 2.0 array contains about 1.3 million probes and is used to measure expressions of the entire human genome (about 38,500 established genes). A chip is processed on the cell line and multiple scans of the chip are obtained. Using these multiple datasets we conduct statistical comparisons to establish the denoising performance.

In summary, the contributions of the methodology include 1) identification of distinguishing features between microarray images and natural images, i.e., features that were the cause of ineffective denoising, 2) obtaining a strategy for reconfiguring microarray images to resemble natural images, and 3) developing a strategy for estimating noise parameters from within the microarray image being denoised (which obviates the high cost of using images from multiple chips and hence, eliminates the influence of different noise unique to the respective images). Also, estimating noise by the use of multiple instances (probesquares) of the same gene, from within the raw image, will avoid removal of valid data.

3.2 A Novel Denoising Methodology for Microarrays

Both hybridization and scanning noise in microarrays are inherently nonuniform across a microarray, and hence are localized in space. Also, due to the variety of their sources, they appear at different frequencies. These frequency and space localizations make microarray noise an ideal candidate for wavelet based multiresolution analysis. Prior to the presentation of our methodology, we provide, for unfamiliar readers, a brief outline of the 2-D wavelet decomposition technique and the properties of a dual tree complex wavelet transform.

3.2.1 Overview of Wavelet Based Multiresolution Analysis

Wavelet's multiresolution approach to data representation has been a major breakthrough in the field of signal processing. Its applications range from data compression, data denoising, to real-time process monitoring [57, 58]. Wavelets consist of basis functions, where, a basis is made up of a scaling function (Φ) and a mother wavelet (Ψ). The mother wavelet is a short wave and therefore has its energy localized in time, unlike sinusoids in Fourier transforms. A one-dimensional signal g(t) can be represented by translations of a scaling function and translations and dilations of the mother wavelet as follows.

$$g(t) = \sum_{k \in Z} c_{j_0;k} \Phi_{j_0;k}(t) + \sum_{j \geq j_0} \sum_{k \in Z} d_{j;k} \Psi_{j;k}(t)$$

where, $\Phi_{j_0;k}(t) = 2^{j_0/2}\Phi(2^{j_0}t-k)$, $\Psi_{j;k}(t) = 2^{j/2}\Psi(2^{j}t-k)$, and $j = j_0, j_0+1, ...; k \in Z$, where Z is a set of integers. Translation is a shift in the location of the function along the axis and is represented by change in the value of k. Dilations or change in frequency, represented by j, are attained by change in the width of the wavelet function. Wavelet transforms use scaling (low pass filter) and wavelet (high pass filter) functions to decompose a signal at different resolutions or *scales* and obtain their scaling coefficients, $c_{j_0,k}$ (approximations), and wavelet coefficients, $d_{j,k}$ (details), respectively. Two-dimensional signals (images) are decomposed by passing the signal through low and high pass filters on the rows of the signal, the output of which is again passed through low and high pass filters on the columns. This leads to the creation of horizontal, vertical, and diagonal detail coefficients as shown below.

$$g(x,y) = \sum_{k,l \in Z} c_{j_0;k,l} \Phi_{j_0;k,l}(x,y) + \sum_{i} \sum_{j \geq j_0} \sum_{k,l \in Z} d^{(i)}_{j;k,l} \Psi^{(i)}_{j;k,l}(x,y)$$

where, $k \in Z$ and $l \in Z$ are translation indices along x and y axes respectively. The scaling function $\Phi_{j_0;k,l}(x,y)$ is obtained by the tensor product of the scaling functions applied on the rows and columns respectively, which is written as $\Phi(x) \bigotimes \Phi(y)$. The detail functions are obtained as follows.

$$\Psi^{(1)}_{j;k,l}(x,y) = \Phi(x) \bigotimes \Psi(y) \qquad \text{Horizontal Detail}$$

$$\Psi^{(2)}_{j;k,l}(x,y) = \Psi(x) \bigotimes \Phi(y) \qquad \text{Vertical Detail}$$

$$\Psi^{(3)}_{j;k,l}(x,y) = \Psi(x) \bigotimes \Psi(y) \qquad \text{Diagonal Detail}$$

Since wavelets are localized in space and frequency, the wavelet coefficients obtained from decomposition contain a few large values and a large number of small values [53, 54]. This is an important property of wavelet decomposition and holds the key to the application of wavelets in signal denoising. Signals are decomposed using wavelets, and denoising is achieved by employing thresholding methods for identification of significant *large* coefficients, and for removal of noisy parts. The coefficients are then reconstructed to obtain noise free signal.

Discrete wavelet transforms use wavelets that are real. After decomposition of an image, the approximation and detail coefficients obtained are each equal to the original length of the signal. Hence, the signal is downsampled to remove odd-numbered coefficients. This alters the shift invariance property. That is, if there is a shift in the input signal, wavelet coefficients at different scales undergo a major change in energy distribution. Use of real wavelets cause another problem called directional selectivity. It is known that detail coefficients contain energy distributed in both positive and negative gradients. However, it is not possible to differentiate between the two orientations as all gradients are obtained as output from a single filter [52]. Com-

plex wavelets can be used to overcome these drawbacks. In complex wavelets, the phases vary approximately linearly with input shift and can be designed such that the magnitudes vary very slowly, thus making them approximately shift invariant [52]. Output from complex wavelet contain complex coefficients since the complex filters either emphasize on positive frequencies and reject negative frequencies, or vice-versa, thus achieving directional selectivity. However, while using complex wavelets, if we decompose an image to more than one level, we cannot achieve perfect reconstruction [52]. Use of dual tree complex wavelet transform (DT CWT) alleviates the above problem and at the same time achieves shift invariance and directional selectivity. In what follows, brief descriptions of DT CWT and a bivariate shrinkage thresholding technique that is used in our denoising methodology are presented.

3.2.1.1 Dual-Tree Complex Wavelet Transform

DT CWT uses the concept that shift invariance can be achieved for real DWT by eliminating down-sampling ([52]). However, instead of eliminating down-sampling, DT CWT achieves shift invariance by using two parallel fully decimated trees where the delays of the filters in the second tree are one sample offset from the first at level one and half a sample different at further levels [52, 59]. This construction also offers the perfect reconstruction property. The dual tree transform is interpreted as a complex transform by considering the outputs from the two trees as real and imaginary parts of complex wavelet coefficients. This also offers DT CWT the directional selectivity property [52].

3.2.1.2 Bivariate Shrinkage Thresholding

For any thresholding technique, the amount of noise removed is dependent on its ability to identify the coefficients that relates to noise. A feature that distinguishes

Bivariate Shrinkage (BiShrink) method ([55]) from other potential thresholding techniques that are available in the literature is the consideration of interscale dependency property of detail coefficients. It is well known that large/small value of the wavelet coefficients usually propagate through the scales [55, 53]. BiShrink method generates models to identify this interdependency between adjacent scales. Using maximum likelihood estimators for the variance of noise and that of data at two adjacent scales, and assuming noise to be Gaussian$(0, \sigma)$, the BiShrink method estimates noise-free values of the detail coefficients using *maximum a posterior* (MAP) estimators ([55](Model 3)). A study performed on DT CWT decomposed image coefficients ([55]) demonstrates that bivariate shrinkage exhibit better denoising performance compared to other thresholding methods.

3.2.2 Microarray Denoising Methodology

In multiresolution based denoising, the image is decomposed using the selected wavelet basis followed by application of the thresholding strategy. However, microarray image denoising cannot follow a similar procedure due to the difficulties related to large number of edges and non-Gaussian noise, and hence, two different strategies were developed. Before giving a detailed description of the two features of microarray images and the two developed strategies, we give a brief overview of a microarray chip, cRNA extraction, and hybridization of cRNAs to DNA strands.

3.2.2.1 Microarray Chip

A DNA Microarray, for example, a GeneChip human genome HG-U133 Plus 2.0 array, is divided into over 1,300,000 minute squares called *probes* [56]. Each probe square is embedded with millions of copies of a short DNA strand that correspond to

a single gene. The 1.3 million probe squares are allocated to over 54,676 genes with each gene represented in eleven randomly selected probe squares across the chip.

3.2.2.2 cRNA Extraction

RNAs are extracted from the cell sample and synthesized to obtain cDNA (complimentary DNA) and further cRNA. Hence, all genes that are expressed in the cell will contain its corresponding cRNAs in the extract, with number of cRNAs being proportional to gene expression. The number of cRNAs in the extract are amplified to ensure proper hybridization. Molecules of a chemical called biotin are attached to the cRNA strands.

3.2.2.3 Hybridization of cRNAs to DNA Strands

The cRNA extract is poured over the microarray. If the cRNA finds a matching DNA strand (hence representing same gene) the cRNA will bind i.e., *hybridize* to the strand. Therefore, the extent of hybridization is proportional to the expression of the gene. A fluorescent stain is then run over the array which sticks to the biotin that acts as a molecular glue. The microarray is scanned to extract the intensity of glow which is proportional to extent of hybridization. The millions of strands of DNA in each probe square are represented by only a small matrix of pixel values, e.g., 7x7 in HG-U133 Plus 2.0. Each probe square is thus represented by a matrix of pixels whose intensities represent the extent of hybridization in the region, and hence the corresponding gene expression value. Thus, a microarray image is comprised of matrices of intensities of probe squares across the entire chip.

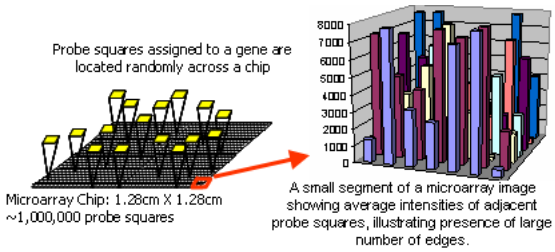

Figure 3.1. Randomly varying intensities of adjacent probe squares depicts presence of large number of edges in a microarray image

3.2.2.4 Strategy 1: A Separate Subimage for each Gene

As mentioned above, eleven randomly selected probe squares are allocated to each of the 54,676 genes. Thus, it is highly likely that adjacent probe squares are allocated to different genes with varying levels of expression. The pixel intensities of a probe square are proportional to the extent of cRNA binding, and hence adjacent probe squares have varying intensities. Also note that, these intensities vary among pixels even within a probe square[44]. Moreover, a mismatch square is placed below every probe square assigned to a gene (perfect match) [60]. These mismatch squares are designed to measure the extent of undesired bindings, and thus have very differing intensities from that of the perfect match probes. As evident from above, the microarrays, by design, have inherent variations between adjacent probe squares in addition to the variations within. Thus, the microarray images have a very large number of edges contrary to natural images. Figure 3.1 shows a schematic representation of the wide variations in intensity among adjacent probe squares in a microarray. Such images, with high intensity variations, when decomposed, generate a large number of high value detail coefficients. Consequently, during thresholding, it becomes difficult to identify the large coefficients that correspond to noise. This results in poor denoising.

Our strategy involves separating the microarray image into multiple subimages, one for each gene, where each subimage is created as a collection of the probesquares assigned to a gene. The idea behind this strategy is that, intensity variations across probe squares of the same gene are much less compared to that across probe squares of different genes. Hence, creating a subimage for each gene will have far less number of edges compared to that of the original microarray image, thus leading to better denoising. The subimages are obtained as follows. The library file (CDF file) of the microarray chip contains a list of all genes along with their i) x and y coordinates of all corresponding probe squares, and ii) the number of rows and columns of pixels for each probe square. The DAT file contains the pixel values of the microarray chip. Using the CDF file, the location of the first probe square of gene 1 is identified and the corresponding matrix of pixel values is extracted from the DAT file. Similarly the matrix of pixel values of the remaining probe squares of gene 1 are extracted and the matrices are placed sequentially row-wise to obtain the subimage in Figure 3.2. Some of the matrices are repeated to ensure that the subimages are dyadic, which is a requirement for wavelet based multiresolution analysis. Similar subimages are obtained for all genes. As depicted in Figure 3.2, the subimages are individually considered for denoising. After which, the original probe square intensities on the microarray image are replaced with those extracted from the denoised subimages. Further analysis of the image, that involve determining the final gene values [60] is carried out using existing procedures of Affymetrix, namely GCOS and MAS. As shown in Figure 3.2, the subimage data, prior to decomposition, undergo another transformation, which is described next.

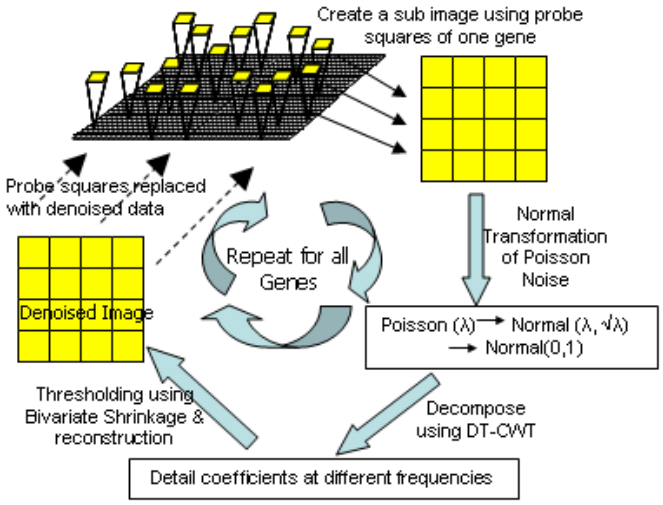

Figure 3.2. Construction of a dyadic subimage, transformation of poisson to normal, denoising and replacing denoised data to reconstruct microarray image

3.2.2.5 Strategy 2: Noise Characterization and Transformation

The noise in gene expression values is a combination of sample preparation noise, hybridization noise, and scanning noise. Sample preparation noise arises from: inherent inability of the experimental procedure to achieve a target cRNA amplification rate during cRNA extraction, and variation in the amplification rate among the cRNAs (a Poisson noise [43]). However, sample preparation noise is small in proportion to the other two noise types. Hybridization noise is attributed to the probabilistic nature of hybridization and is proportional to the gene expression value (pixel intensity). Such an intensity dependent noise is called *shot noise* and is known to have a Poisson distribution. Scanning noise, on the other hand, is independent of expression values and may be induced by the presence of dirt, reflection of light, and other random causes. Readers are referred to [43] for a detailed characterization of the noise types. In what follows, we explain our strategy for estimating the noise parameters using pixel intensities from a single chip that needs to be denoised.

Isolation of *sample preparation* noise is not feasible using a single chip, since identifying the noise parameter requires the use of multiple replicates of the same sample. Moreover, since the magnitude of this error is small, we did not attempt to remove this noise. The probabilistic variation in hybridization induces Poisson noise in gene expression values [43]. Hybridization noise arise from two types of binding errors, i) where a cRNA strand might not bind even when a matching DNA strand is present (missed binding), and ii) when a cRNA might bind to a non-matching DNA strand (false binding). Therefore, the number of binding errors per pixel is Poisson distributed. Note that, Poisson distributed random variables take non-negative values, and therefore we consider the number of binding errors as equal to the absolute value of the difference between the number of false bindings and the number of missed bindings. Therefore, for each pixel, the Poisson noise has either been added (when false binding is the dominating type of binding error) or subtracted (when missed binding is the dominating type of binding error) from the actual intensity. As is clear from above, the variation among the pixel intensities within a probe square is also induced by variation in hybridization. Therefore, we estimate the hybridization noise Poisson parameter λ from the absolute value of intensity variations among the pixels of a probe square. Also, since the denoising method using BiShrink thresholding requires that all error is $\sim \mathcal{N}(\mu, \sigma^2)$, with $\mu = 0$, we use a normal approximation of the Poisson distribution. This approximation can be justified as follows. Since each probe square is represented by a small number of pixels, each pixel reflects a collection of millions of DNA strands. If we divide these DNA strands into several groups and consider the number of binding errors for each group, then, the sum of the binding errors over all groups of the pixel would be normally distributed based on the central limit theorem. Therefore, we can write that hybridization noise $\sim \mathcal{N}(\mu, \sigma^2)$ with,

$\mu = \lambda$ and $\sigma^2 = \lambda$. The *scanning* noise, due to the nature of its sources, is assumed to be Gaussian with mean zero.

In what follows, we describe the estimation of the mean of the hybridization noise, and then apply a standard normal transformation since, as mentioned earlier, BiShrink thresholding requires a noise with mean zero. Since hybridization noise is proportional to intensity, noise estimation was carried out separately for each probe square within a subimage. Let p^n represent the array of pixel intensities of the n^{th} probe square of a subimage. Each pixel intensity value $p^n(x, y)$, where, x and y are array coordinates, is modeled as follows:

$$p^n(x, y) = E^n + \epsilon_H^n(x, y) + \epsilon_S^n(x, y), \tag{3.1}$$

where E^n denote the true expression value of the gene as measured by the n^{th} probe square, $\epsilon_H^n(x, y)$ denote the normal transformation of hybridization noise with $\mu = \lambda^n$ and $\sigma^2 = \lambda^n$, and $\epsilon_S^n(x, y)$ denote the normal distributed scanning noise with parameters $(0, \sigma^n)$. Note that, the hybridization and scanning noises are independent of each other, where hybridization noise is based on incorrect hybridization, while scanning noise is caused by external features like presence of dirt. We take the *median* of the pixel values of the n^{th} probe square, denoted by $\hat{E}^n$ as an estimate of E^n. Define $d^n(x, y)$ as

$$d^n(x, y) = p^n(x, y) - \hat{E}^n. \tag{3.2}$$

Then we can write that $d^n(x, y)$ contains two normally distributed noise components, sum of which also has a normal distribution with parameters $\mu = \lambda^n$ and $\sigma = \sqrt{\lambda^n} + \sigma^n$. Note that, as explained earlier, the hybridization noise is a positive value which is either added or subtracted from E^n. Positive values of $d^n(x, y)$ denote pixels where noise has been added, while negative values denote pixels where noise has

been subtracted. Therefore, to estimate the value of λ using a maximum likelihood estimate principle, we use the absolute value of $d^n(x, y)$, and write that

$$\lambda^n = \frac{1}{N} \sum_x \sum_y |d^n(x, y)| \tag{3.3}$$

where, N is the size of the array. Note that $d^n(x, y)$ contains a combination of hybridization noise and scanning noise. While scanning noise is considered to have zero mean, the hybridization noise has a mean of λ^n. Therefore, in order to obtain a total noise mean of zero, we apply a standard normal transformation of $d^n(x, y)$ (denoted by $d_T^n(x, y)$) as follows

$$d_T^n(x, y) = \frac{|d^n(x, y)| - \lambda^n}{\sqrt{\lambda^n}}. \tag{3.4}$$

Therefore, $d_T^n(x, y)$ now contains hybridization noise and scanning noise with a total sum of mean of zero. Finally, we obtain the modified individual pixel values, denoted by $p_M^n(x, y)$, as follows. From $p^n(x, y)$, we remove $d^n(x, y)$ which contained a non-zero mean from hybridization error. By retaining the sign of the $d^n(x, y)$, hence representing the direction in which hybridization noise was included, we replace $d^n(x, y)$ by $d_T^n(x, y)$, and write that

$$p_M^n(x, y) = p^n(x, y) - d^n(x, y) + sign(d^n(x, y))|d_T^n(x, y))|. \tag{3.5}$$

Therefore, $p_M^n(x, y)$ now contains the true expression value along with the normally distributed hybridization and scanning noises, with a total mean noise of zero.

3.2.3 Steps of the Denoising Procedure

1. Begin with the DAT file of the microarray image and the CDF file of the corresponding chip. The DAT file contains the pixel values, i.e., the raw image data. The CDF file is the library file that contains a list of genes, the coordinates of all of its corresponding perfect match and mismatch probe squares, and the number of rows and columns of each probe square.

2. Using the information on the coordinates and the number of rows and columns from the CDF file, extract the intensity values of all perfect match (PM) probe squares of a gene from the DAT file. (In this research, some of the C codes available at [61], a public software repository, were used for the purpose.)

3. Copy onto a text file the data arrays that represent the probe squares, to form the rows and columns of a subimage of maximum possible dyadic size. For example, typically a gene on the GeneChip human genome HG-U133 Plus 2.0 array is represented by 11 PM probe squares where each probe square is (suppose) a 7×7 array of intensity values. Therefore, forming a subimage of 3×3 probe squares will give us an array of intensity values of size 21×21. In order to make this dyadic i.e., 32×32, the first 11 rows and columns are repeated to form the last 11 rows and columns respectively. Since the subimage thus created involves only the first 9 out of the 11 PM probe squares of the gene, the last 9 PM probe squares are used to create another dyadic subimage in a similar manner (note that 7 probes are included in both subimages). This allows us to denoise all the probe squares.

4. Since Poisson noise is proportional to the pixel data intensity, conduct on each probe square the Gaussian transformation of Poisson noise as given in Equation (3.5) to create two modified subimages.

5. Denoise each of the two modified subimages by applying a sequence of DT CWT decomposition, Bivariate Shrinkage, and reconstruction, using the software available at [62].

6. Replace the data in the DAT file that corresponds to the PM probe squares of the gene by using the denoised intensity values from the subimages.

7. Repeat steps 2 through 6 for the mismatch (MM) probe squares of the gene.

8. Repeat steps 2 through 7 with PM and MM squares of all remaining genes.

9. Create the CEL file by taking the 75th percentile value of each probe square, and process the CEL file using Affymetrix software to arrive at the final expression values of the genes.

The denoising methodology was coded in C and a C-MATLAB interface was used to access MATLAB programs from [62]. The processing of all programs was carried out using the high performance computing resources provided by Research Computing Department at USF [63].

3.3 Numerical Validation using Simulated and Affymetrix Microarray Data

In this section we present the results of the tests used to measure the performance of our denoising methodology. Peak Signal-to-Noise Ratio (PSNR) is one of the commonly adopted measure of performance for techniques used in denoising of natural

images [64, 53]. PSNR measurement generally involves denoising an image that was created by adding a known quantity of Gaussian noise to a clean image. The value of PSNR, calculated in decibels, is inversely proportional to the noise that remains after denoising, and hence a higher value represents better denoising. The PSNR is calculated as follows.

$PSNR = 20log_{10}\left(\frac{256}{\epsilon}\right)$, where, $\epsilon = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(s_k - d_k)^2}$,

s_k denotes the k^{th} pixel intensity of the clean image, d_k is the corresponding value of the denoised image, and N denotes the total number of pixels. Clearly, a clean image is essential in establishing the performance of a denoising methodology using PSNR. Due to the unavailability of a clean microarray image, we first tested our methodology on a simulated image as described below.

3.3.1 Testing Denoising Performance on a Simulated Image

Since our methodology separately denoises each subimage created for a gene, we fabricated a *clean* dyadic subimage of size 32×32 pixels. As described earlier in the example in Step 3 of Section 3.2.3, the simulated subimage consisted of 9 probe squares. We also simulated another subimage with 16 probe squares (obtaining 28 X 28 pixels and repeating four rows and columns to get 32 X 32). Similar denoising performance was noted in both subimages, and the one presented here is for the 16 probe square subimage. The pixel intensities within a probe square were kept constant. Mimicking the actual values of a subimage from a real microarray data set, different intensity values for the probe squares across the simulated subimage were chosen. Two separate sets of noise having Poisson and Gaussian distributions were added to the simulated subimage, and subsequently denoised using our methodology. Poisson noise, proportional to the pixel intensity of the probe squares, and Gaussian

noise, with different values of the variance parameter for different trials, were added to the clean image using functions available within the MATLAB software.

The PSNR values for different error combinations are presented in Table 3.1 and 3.2. PSNR in Table 3.1 was obtained by considering k in s_k and d_k as individual pixel index, while in Table 3.2 k is a probe square index and s_k and d_k represent the 75^{th} percentile of the pixel intensities of the k^{th} probe square. Different image types that are considered in the tables (see column 1) are as follows: (1) noisy image, where PSNR calculation is based on pixel values of the clean and noise added (without denoising) images, (2) denoised image without Poisson transformation, where PSNR is calculated using clean and denoised images and the denoising is carried out without including Gaussian transformation of the Poisson noise, and (3) denoised image with Poisson transformation, where PSNR is calculated using clean and denoised images and denoising is done using the complete methodology, i.e., including Gaussian transformation of the Poisson noise. The following observations are made

Table 3.1. PSNR value estimated by taking error, epsilon, as difference between individual values

Image Type		Noise Added		
		Poisson*+ Normal(0,10)	Poisson*+ Normal(0,20)	Poisson*+ Normal(0,30)
Noisy Image		18.57	16.57	15.8
Denoised Image	Without Poisson Transformation	22.45	18.49	18.07
	With Poisson Transformation	33.32	32.11	31.14

*Parameter of the Poisson noise added is proportional to intensity of corresponding probe square

from Tables 3.1 and 3.2. In both tables, the higher PSNR values for the denoised images with/without using Poisson transformation indicate a significant improvement compared to the PSNR values of the noisy images. According to the literature on

Table 3.2. PSNR value estimated by taking error, ϵ, as difference between 75th percentile of probesquare

Image Type		Noise Added		
		Poisson+ Normal(0,10)	Poisson+ Normal(0,20)	Poisson+ Normal(0,30)
Noisy Image		27.72	27.85	27.45
Denoised Image	Without Poisson Transformation	26.08	27.75	26.39
	With Poisson Transformation	35.2	33.83	32.91

denoising studies ([64, 53]), the PSNR values obtained here can be considered high indicating a desirable denoising performance. A comparison of the PSNR values obtained without and with Gaussian transformation of the Poisson noise (2nd and 3rd rows of the numbers respectively), shows a significant difference, thus clearly establishing the importance of including the transformation strategy in our methodology. In what follows, we present results from application of our methodology on a set of Affymetrix microarray data.

3.3.2 Testing Denoising Performance on Affymetrix Microarray Data

Having established the performance of our methodology on a simulated data set, we extended our testing on data sets obtained from Affymetrix GeneChip human genome HG-U133 Plus 2.0 arrays, processed on HCT-116 colorectal cancer cell line at the Microarray Core Facility of Moffitt Cancer Center and Research Institute. Due to the unavailability of clean images we could not compute PSNR to measure the extent of noise removal. Hence, we adopted a strategy of testing noise reduction through coefficient of variation (CV) of the original and denoised data sets. The premise of the strategy is that, multiple instances of a data set with random noise are likely to have a higher CV compared to that from corresponding data sets from which some of the noise has been removed. However, it may be noted that though a

reduction of CV is indicative of noise removal, it cannot be directly translated into the extent of noise removed.

An assessment of denoising performance should ideally be based on *gene expression values* obtained from before and after denoising. However, to obtain the gene expression values, the data sets (pixel intensities) are processed using Affymetrix GCOS or MAS5 software. These softwares perform data transformations, like background correction, before converting pixel intensities into gene expression values. As a result, the gene expression values will reflect the impact of our denoising method as well as that of the transformations performed in GCOS or MAS5. Thus, it would be difficult to estimate the effect of our denoising methodology alone when gene expression values are used. Hence, instead, we used the pixel intensities. We conducted two different denoising performance measurement tests as described below.

3.3.2.1 Analyzing CV of Probe Squares Across Multiple Scans of a Microarray

In this test we collected three different scans of a single chip resulting in data sets containing same hybridization noise but different scanning noise. The data sets were denoised individually using our methodology, and subsequently CEL files were created by representing each probe square with a single value equal to the 75th percentile of its pixel intensities. For each probe square, CV of its values across the three scans was computed from the CEL file data. The above procedure for CV calculation for each probe square was also repeated using the data sets prior to denoising (i.e., using the original data). In order to assess denoising performance, we divided the range of CV (0 to 1) into multiple subdivisions. We chose finer subdivisions near the lower range of CV values and wider divisions towards the higher end, since a majority of the probe squares had a relatively smaller CV. The number of probe squares under each

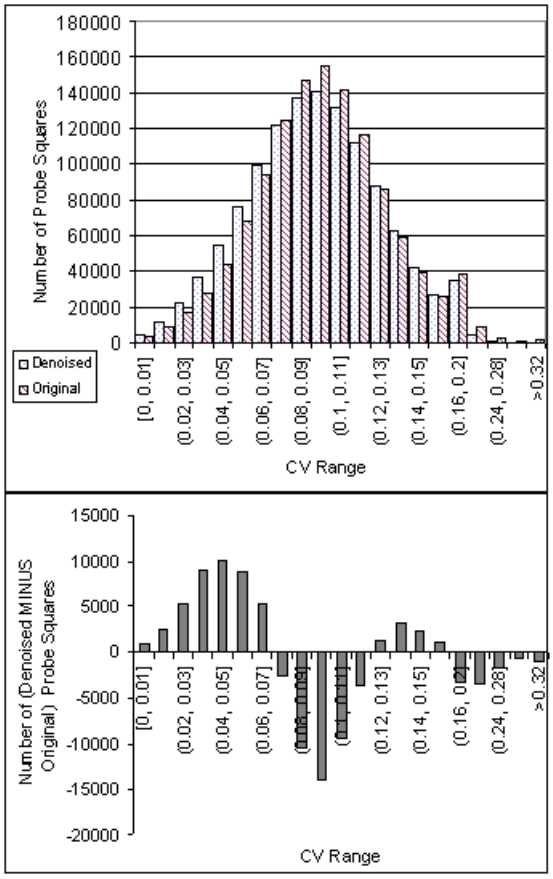

Figure 3.3. Distribution of probe squares showing impact of denoising

CV range was counted for both original and denoised data sets and plotted as shown in the top plot of Figure 3.3. The shift in the histogram of denoised probes (indicated by dotted bars) towards the lower CV range (leftwards) indicates the reduction in noise after denoising. The bottom plot of Figure 3.3 shows the *difference* between the denoised and original number of probe squares under different CV ranges. A positive rectangle for a CV range implies that the number of probe squares in the denoised image is greater than the number of probe squares in the original image. The figure shows four distinct zones of CV values: [0-0.07]; (0.07-0.12]; (0.12-0.16]; (0.16-1.0];

we henceforth refer to these as zones 1, 2, 3, and 4 respectively. The following are interpretations of the data observed in these zones.

1. Zone 1 shows a total increase of 42,289 probe squares in the denoised set. Clearly, these probe squares had higher CV values (zones 2 through 4) in the original set. This is indicative of noise reduction.

2. Zone 4, with highest CV values, shows that the denoised set has much less number of probe squares (9,998) than in the original set, which indicates that after denoising many probe squares in this zone have reduced their CV values and thus migrated to the left (zones 3 through 1).

3. Probe distributions in zones 2 and 3 are resultant of the migrations described for zones 1 and 4.

4. Zones 1 and 4 indicate a CV reduction for 52,287 probes. However, the total number of probes that experienced a CV reduction is likely to be higher, since it is difficult to assess the exact number. In perspective of the total number of probes squares in a microarray (approximately one million), this number ($\geq 52,287$) might seem small. However, the bottom plot of Figure 3.3 shows that a vast majority of the probe squares have CV lower than 0.24. Thus it can be concluded that the original data had a small amount of noise, some of which has been removed.

From the above, we can conclude that our methodology is capable of reducing noise introduced during the scanning process. Tests performed on two other data sets yielded similar results. However, as mentioned earlier, the extent of noise removal could not be ascertained.

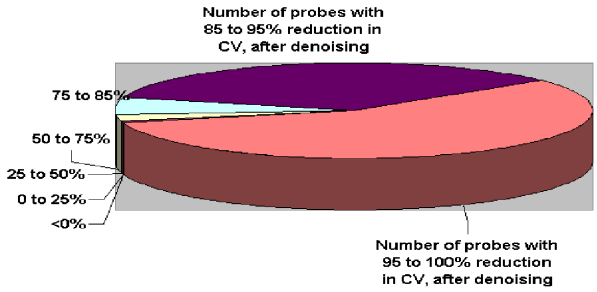

Figure 3.4. Proportion of probe squares in various ranges of R^s_{CV} for Scan 1 data

3.3.2.2 Analyzing CV within Probe Squares of a Microarray

In order to assess performance of our methodology in removing noise introduced during hybridization, we compared, between original and denoised data sets, the variation across pixel values of each probe square in a data set obtained from a single scan. The CV across each probe square s, was calculated from original and denoised DAT files, which were used to obtain the percentage reduction in CV (R^s_{CV})due to denoising as,

$$R^s_{CV} = \frac{CV^{(s)}_{orig} - CV^{(s)}_{denoi}}{CV^{(s)}_{orig}} \times 100$$

Figure 3.4 presents a pie-chart that shows the proportions of the probe squares in various ranges of R^s_{CV} values for Scan 1 data set. Data sets from scans 2 and 3 yielded very similar results. It is apparent from the figure that almost all probe squares have had a high reduction in CV. However, due to the nature of the wavelet based denoising approach, the reduction in CV values cannot be attributed completely to the reduction in hybridization error. A better testing strategy would be to hybridize multiple chips from a single sample, scan them to obtain the DAT files, and denoise them separately. Comparing the CV of each probe square using data from the CEL files of these multiple chips, in a manner similar to that presented above for the test

using multiple scans, will give us a better estimate of hybridization noise reduction. Due to the immediate unavailability of such a dataset, we were unable to perform the stated testing strategy.

3.4 Conclusions

A novel multiresolution analysis based methodology, using a dual-tree complex wavelet transform and bivariate shrinkage thresholding, for removing hybridization and scanning noise from microarray data is presented. Two specific microarray data features that are not conducive to wavelet based denoising are identified, for which specific data transformation strategies are developed. The aforementioned features are: presence of excessive number of edges in microarray images, and non-Gaussian (Poisson) distribution of the hybridization noise. The strategies involve creation and denoising of separate subimages for each gene (instead of the complete microarray image), and Gaussian transformation of the Poisson noise prior to denoising. A comprehensive approach to denoising microarray data, as presented in here, is not available in the open literature. We believe that our approach to removing noise from microarrays will increase the reliability of gene expression values for use in disease diagnosis and treatment planning.

Testing the performance of any image denoising methodology requires noisy image along with its clean base image. Since it is not possible to have a clean base image for a microarray, testing of our methodology was first carried out using data simulated to mimic microarray probes. When noise was introduced to the simulated data, our methodology was able to remove them to a significant extent as evidenced by PSNR (peak signal to noise ratio) values. The testing was then continued with image data sets obtained from Affymetrix microarrays processed on colorectal cancer cell samples. Due to unavailability of clean base image, coefficient of variation (CV) based

comparison strategies were developed to obtain measures of denoising performance. A considerable extent of quality improvement was noticed for the tested Affymetrix data sets.

Though our methodology is capable of identifying hybridization and scanning noise with the use of a single chip, identification of sample preparation noise still remains an open challenge. The reduction of the portion of sample preparation noise which is due to randomness in amplification rate, is dependent on the techniques used to control or maintain a constant rate. This is considered beyond the scope of our current research. However, it is possible to extend our methodology to identify the portion of sample preparation noise which is due to the probabilistic nature of amplification. This could involve creating subimages with probe squares of the same gene from multiple chips. Since the distribution of this noise is also Poisson, like that of hybridization, probe squares from the same location on the multiple chips will have Poisson noise of magnitude equal to the sum of the two individual Poisson noise types. We believe, therefore, that the application of our current methodology will offer desirable performance. However, since the sample preparation noise is only a small portion of the total noise, the requirement of multiple chips may not be economically justified, except in cases where multiple chips are naturally involved in the experiment.

CHAPTER 4

ANALYTICAL PROCESSING OF CHROMATOGRAMS

4.1 Introduction

Proteome, the set of proteins translated by the genome, is the main component that drives the metabolic activities in a cell. Hence, the proteome is anticipated to hold the key to the pathway of diseases like cancer and therefore play an important role in cancer diagnosis and treatment. However, application of this theory into practice requires identification and quantification of the proteins produced in the cell, which is a challenging task because of the complexity of the proteome. While an organism has a constant genome, its proteome varies from cell to cell. Proteins are produced or *translated* by genes that are active or *expressed*, and different genes are expressed in different parts of the body. The 35,000 genes of the human genome are estimated to translate more that ten times as many proteins, where the type and quantity of protein produced is based on the coded information contained in the gene. In extreme cases, an individual gene could have a coding capacity of about 1000 proteins [65]. Due to variations in the translation process, the same gene could produce different forms of the protein (protein isoforms) that differ in their function of metabolic activities [66]. Further, the structure of the protein could change after production (post-translational modifications (PTM)) [67], hence creating proteins with functions that are different from the original. This complex structure of the proteome creates challenges in identifying and quantifying proteins produced in a

cell, and thus several proteins and its function are still unknown. Identifying protein *biomarkers*, i.e., the sets and quantity of proteins present in the cell only during a disease state, will help in diagnosis and further in drug discovery for the respective disease. Since early diagnosis of cancer lead to better chances of survival, identifying protein biomarkers related to early stages of cancer is essential.

Advancements in the area of protein profiling, a technique for detecting proteins in tissue, blood, or urine samples [68, 69], has led to the identification of a number of protein biomarkers. However, thus far, very few have been identified, e.g., in ovarian cancer, CA125 is the only clinically used biomarker for diagnosis with a sensitivity of only 50% for early stages of cancer [70, 71]. Other biomarkers that have been identified lack the required specificity and hence cannot be applied to population-wide screening. Along with the complex structure of the proteome, the current lack of efficient biomarkers can be attributed to challenges such as chemical separation of the proteins [72]. The separation, generated using inherent difference in property of proteins, is required for identifying and measuring the quantify of each protein. Incomplete separation causes low quantity (i.e., low abundance) proteins to be masked by those in high abundance and hence several significant proteins might yet remain unidentified [73]. Another significant challenge is during the analytical quantification of protein values from the plots, called chromatograms, generated during protein profiling. Chromatogram is a visual representation of the protein separation, where, the measure of certain chemical components in a sample that are reflective of the separated proteins are plotted. In this research, we develop a methodology for addressing challenges in protein identification and quantification from chromatogram data. Before giving a detailed description of a chromatogram and discussing the research problem, we briefly explain some of the protein profiling techniques.

Most protein profiling techniques, e.g., 2-dimensional electrophoresis (2DE), matrix assisted laser desorption and ionization (MALDI), and surface-enhanced laser desorption and ionization (SELDI), use protein characteristics like hydrophobicity, ionization mass, and electrophoresis [67, 74, 75] to separate proteins. However, small differences in hydrophobicity or large ionization mass between different proteins hinders protein separation [67]. In this research, we develop a methodology to process chromatograms generated by Beckman Coulter ProteomeLab Protein Fractionation-2 Dimension (PF2D). PF2D follows a two dimensional approach, where, separation in the first dimension is based on the varying isoelectric point (pI) of proteins and the sample is separated into fractions based on pH. In the second dimension the proteins in the fractions are further separated using the property of hydrophobicity ([76]). This approach is found to account, to some extent, for post-translational modifications of proteins and to improve reproducibility of protein detection, which are weaknesses of the other profiling techniques [67, 76]. Also, the increased capacity of PF2D to separate proteins provides an additional leverage in analyzing low abundance proteins, which otherwise would have been masked by high abundance proteins. The use of PF2D in proteomics has been well-established technically [77, 78, 79, 76, 80], and has been shown to be promising for biomarker research [81, 82]. However, the task of quantitative processing of the chromatograms is complicated because, while high *abundance* proteins can be easily quantified from these plots, identifying and quantifying low abundance proteins is a challenge. The rest of this chapter is structured as follows: description of the PF2D chromatogram, quantitative challenges in its processing, and current literature are discussed in Section 4.2; the methodology for quantitative processing is presented in Section 4.3; and the results of the methodology are presented in Section 4.4.

4.2 Description of PF2D Chromatograms and Quantitative Challenges

A sample chromatogram output generated from the PF2D second dimensional separation, that was applied on a fraction from first dimension separation, is presented in Figure 4.1. Approximately 17 such fractions, separated based on pH, are obtained for one blood or urine sample, however, the number can be varied based on requirements. During the second dimensional separation, proteins in the fractions are detected by the amount of UV absorbance and as in Figure 4.1 the absorbance unit (AU) is plotted against retention time ([76]). Interpreting the plot, the spikes or peaks are indicative of the presence of proteins, and based on the effectiveness of separation, each peak represents one or more proteins. The area underneath the peak is equivalent to the amount of the protein present in the sample. Analytical processing of the chromatogram involves identifying the presence and location of peaks, and estimating the area of each peak. Further, by comparing equivalent peaks across samples, peaks that distinguish cancer cases from controls (potential biomarkers) can be obtained. To identify the protein content of the required peaks, the corresponding portion of the fraction can be chemically analyzed. This feature of PF2D, which allows future analysis of fractions of interest, is an advantage over other protein profiling techniques ([76, 67]). One of the significant reasons for the lack of effective protein biomarkers can be attributed to the challenges in the analytical processing of chromatograms. We explain below the steps and challenges in PF2D chromatogram processing.

1. Baseline Correction: During the generation of the chromatogram, ideally, the presence of proteins should reflect as a positive value and absence a value zero. However, due to difficulties in setting the correct baseline, the signal drifts to either side of zero as can be seen in Figure 4.1. Hence, the first task is to identify the baseline for the generated signal.

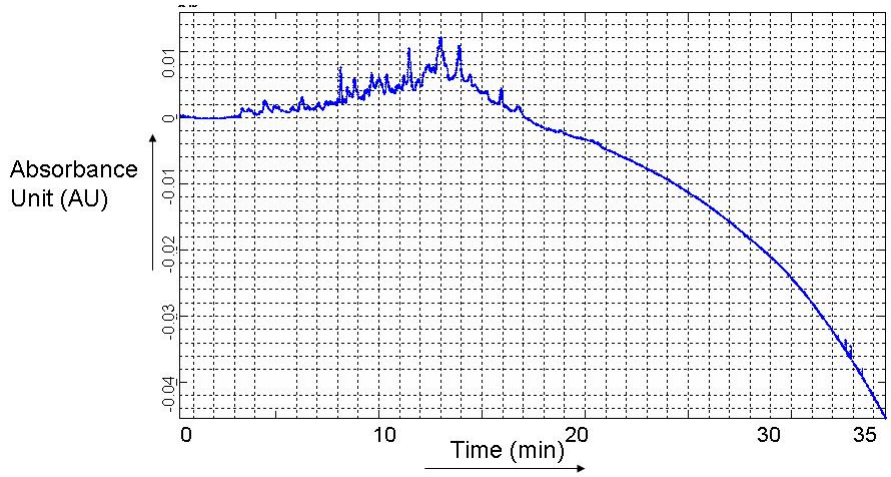

Figure 4.1. Sample PF2D signal

2. Peak Identification: Figure 4.2 presents an enlarged image of a section of Figure 4.1. Note that, though we can visually identify peaks in the signal, it is a tedious and manually infeasible task as each sample contains thousands of peaks and obtaining biomarkers requires analyzing hundreds of samples. Therefore, developing a methodology for mathematical identification of peaks is essential. While large peaks, like B and D in Figure 4.2, can be easily identified using mathematical tools, it is a challenging task to differentiate smaller valid peaks, like C and E, from noise. Since most proteins are present in small quantities in blood and urine, successful identification of small peaks would be essential for detection of potential low abundance biomarker proteins [83].

3. Intensity Quantification: Obtaining biomarkers involves identifying proteins that are produced in distinguishably different quantities across cases and controls. Hence, the third step involves quantifying the area of each peak which is equivalent to the quantity of the proteins represented by the peak. Estimating the area is a challenging task due to the presence of overlapping peaks, e.g., A and B in Figure 4.2, which are caused by incomplete separation of proteins.

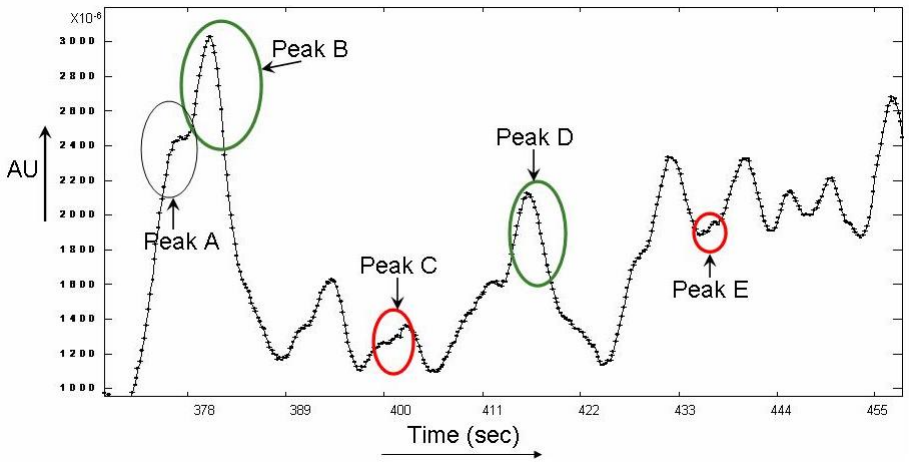

Figure 4.2. Peak identification and intensity quantification

Since each peak is most likely representative of a set of proteins rather than an individual protein, it might seem fair to group peaks A and B and estimate the area under the combined peaks, which is a relatively easier task. However, this will be undesirable for identifying distinguishable proteins across cases and controls, the reason for which can be explained as follows. As seen in Figure 4.2, most of the small peaks overlap with the larger peaks. Hence, if these peaks are grouped, when comparing peaks across cases and controls, the high value of the large peaks will mask any significant difference that exists across the smaller peaks. Hence, developing an algorithm that identifies and quantifies all peaks is essential.

4. Peak Alignment: Ideally, identical peaks, i.e., those that represent the same set of proteins across samples, should be aligned in time since they are detected based on protein properties. However, as highlighted in Figure 4.3, due to chemical causes ([79]), there is a horizontal shift in the occurrence of peaks in sample 1 when compared to identical peaks in sample 2. Hence, for comparison

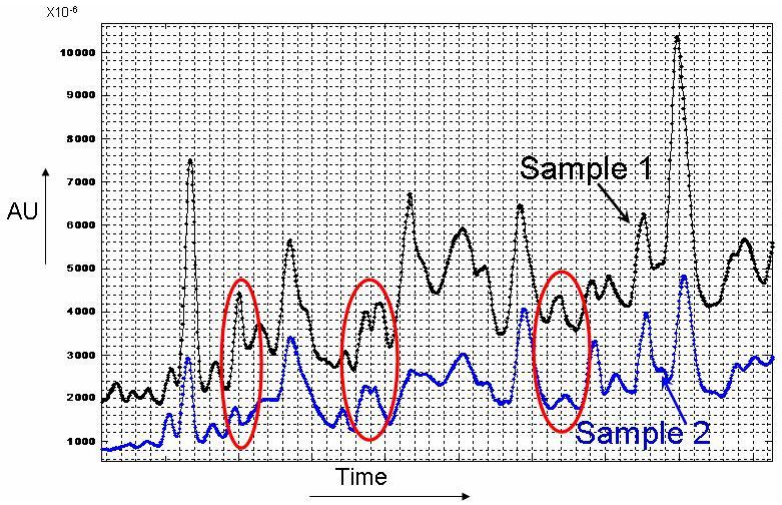

Figure 4.3. Horizontal shift of peaks

across cases and controls, identical peaks across all samples need to be aligned along x-axis (time).

Currently, 32 Karat is the software that is mainly used for quantitative processing of PF2D chromatograms ([76]). The drawback of the software is that it is not completely automatic but manual, like setting targets for aligning peaks and peak identification, and multiple sample alignments is also not possible. While mathematical methodologies for processing of PF2D chromatograms are limited, literature presents methodologies for other protein profiling techniques, where, the quantitative steps involved in processing are similar to that explained above. Most commonly used protein profiling techniques are mass spectrometry (MS) based, where the proteins are detected based on mass-to-charge ratio (m/z) unlike pH or hydrophobicity in PF2D. The spectral signals generated by MS techniques contain the number of ions, indicating intensity of proteins, versus the corresponding m/z values. Current methodologies for quantitative processing of SELDI and MALDI based MS signals have been reviewed in [84] and [85], respectively, which can be summarized as follows. Methodologies for baseline correction include dividing the signal into windows

and scaling to local minima in each window ([86, 87]), using a global moving average ([88]), or linear interpolation ([89, 90, 91]). Some methodologies smooth the signal, to remove noise before peak identification technique, by applying a moving average filter ([92, 93]) or a Gaussian filter ([90, 94]). The crest of the peak is identified as the local maxima and considered as a valid peak if signal-to-noise ratio (SNR) is greater than a user-defined threshold ([90, 93, 94, 95, 96]), or in addition, also use a baseline threshold and ignore all peaks below this value ([88]). SNR of a peak is usually obtained as the local average divided by the local standard deviation. Other techniques of peak identification, along with SNR, use a baseline value to identify beginning and end of peaks and apply a threshold for peak width ([97]), or estimate slopes and use a threshold for both sides of the crest to determine a valid peak ([89]). While the above techniques are successful in identifying high abundance peaks, identification of low abundance peaks is subject to surrounding signal region thus creating inconsistencies in results. Since noise varies along the signal, small peaks in the vicinity of little or no noise or in the presence of large peaks might have a low SNR value, even equal or lower than that of a noise peak in other parts of the signal. Hence, it is not possible to set an optimal threshold. Methodologies in [98, 99, 95, 96, 86] use wavelet based signal processing tools (extracting signal at multiple frequencies/scales) to identify peaks.

The methods presented in literature have been successful in automatically identifying certain peaks, and as presented by [84] and [85], based on identification of known proteins, the method in [98] performs better than other techniques. However, the current methods do not include techniques for identification of small peaks or overlapping peaks. Most methods use thresholds in differentiating noise from valid peaks, and it has been seen that low thresholds increase false discovery rate, while high thresholds miss several valid peaks. In this research, our main focus is in identifi-

cation and quantification of small peaks and overlapping peaks. Since blood circulates throughout the body, it contains proteins from all parts of the body. There are about 500,000 proteins produced in the human body most of which are present in small amounts in blood and urine, and several of which are yet to be discovered. Therefore, identification of the small peaks that represent low abundance proteins is essential for obtaining potential biomarker proteins [83]. Also, due to incomplete protein separation, most of the small peaks overlap with and occur on shoulder of large peaks. If these overlaps are not separated, during the process of identifying peaks that distinguish cancer cases from controls, any significant difference in the small proteins will be masked by the larger proteins. Hence, in this research, we develop continuous wavelet transforms based methods to address challenges with peak identification and quantification especially those that have low abundance or overlap.

Peak alignment is a task encountered in several areas in biochemistry and the literature presents considerable number of methods based on time warp or sequence alignment, and several of which perform peak alignment prior to peak detection. The authors in [100] present a method for peak alignment of PF2D chromatogram. The method has been derived from dynamic time warp based on dynamic programming that was originally developed for speech recognition, and has been used for alignment of chromatograms in other biochemistry techniques [101, 102, 48]. In such methods, the signal is usually divided into number of sections and each section is stretched or shortened by shifting its ends within a slack parameter, and interpolating the signal points in between. The shift that provides the best correlation coefficient with the target signal is retained. Such a method will change the profile of the peak and therefore its area, which is not suitable in our application as we are interested in identifying proteins that have distinguishing quantities (areas) across cases and controls (biomarkers). Other similar methods based on techniques of time warp are presented

in [103, 104]. A scale-space representation, similar in concept to a wavelet function, but using a Gaussian function and for peak alignment of multiple spectra is presented in [105]. By obtaining Dirac components as sum of weighted distance between peaks from all spectra (samples) and convolving with the Gaussian function, the common peak location across spectra is obtained as the local maxima of the resulting function. However, its performance with overlapping peaks and low abundant peaks is dependent on setting appropriate coefficient weights derived from prior knowledge of biology, which is not feasible to obtain in our application. In this research, we first perform the peak identification and quantification followed by alignment of the identified peaks by using a simple model based on mixed integer-nonlinear programming. We follow such a sequence as individual peak quantification is essential for obtaining biomarkers and also to avoid any changes in profile, especially in low abundant peaks, that might result if alignment is conducted first.

4.3 Methodology for Quantitative Processing of PF2D Chromatograms

In this research, we developed a mathematical model using continuous wavelet transforms for baseline correction, peak identification, and quantification, and an optimization model for peak alignment. In what follows, we briefly discuss wavelet transforms and explain in detail the methodologies for the quantitative processing.

4.3.1 Continuous Wavelet Transforms for Baseline Correction, and Peak Identification and Quantification

Wavelet is a short wave, and note that, in Chapter 3, we briefly discussed discrete wavelet transforms for removing noise from microarray images. For chromatogram processing, we use continuous wavelet transforms (CWT), defined as the convolution of the normalized form of a function $\psi(t)$, called the mother wavelet, with the data

signal $x(t)$ to obtain wavelet transforms $T(a, b)$ at *scale* a and *translation* b ([106]). We will discuss later in this section the advantage of using continuous over discrete wavelet transforms for chromatogram processing. The mathematical representation of CWT is written as

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt \tag{4.1}$$

Scale can be referred to as the stretched/squeezed form of a wavelet and translation as the position of the center of the wavelet on the x-axis. By varying the values of a, the signal can be decomposed or represented at different frequencies and the value of the transform at varying b can be obtained. In this research, we use the Mexican hat wavelet function (Figure 4.4) which is written as ([106])

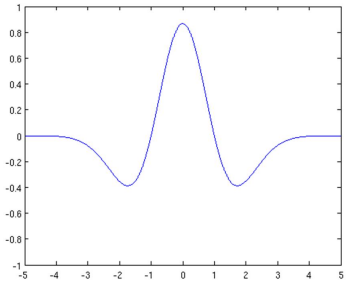

Figure 4.4. Mexican hat wavelet

$$\psi\left(\frac{t-b}{a}\right) = \left[1 - \left(\frac{t-b}{a}\right)\right] e^{-0.5\left[1-\left(\frac{t-b}{a}\right)\right]} \tag{4.2}$$

By using such a peak shaped wavelet function, identification of peaks in a chromatogram can be achieved by utilizing values of $T(a, b)$ which is explained as follows. As an illustration of scaling and translation, we depict in Figure 4.5 a wavelet at fixed translation b and at three scales a_1, a_2, and a_3, $(a_1 < a_2 < a_3)$, and in Figure 4.6 a wavelet at a fixed scale and three translations b_1, b_2, and b_3. Visually, it can be

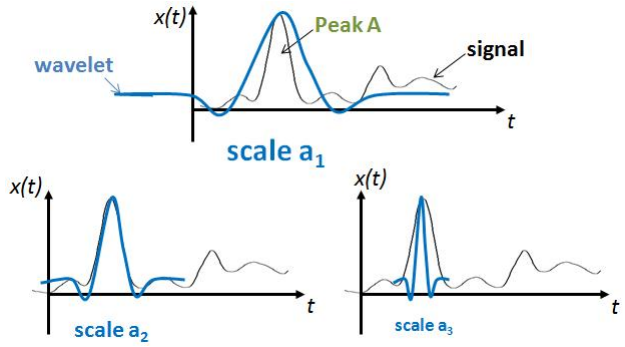

Figure 4.5. A wavelet at a fixed translation and at three different scales

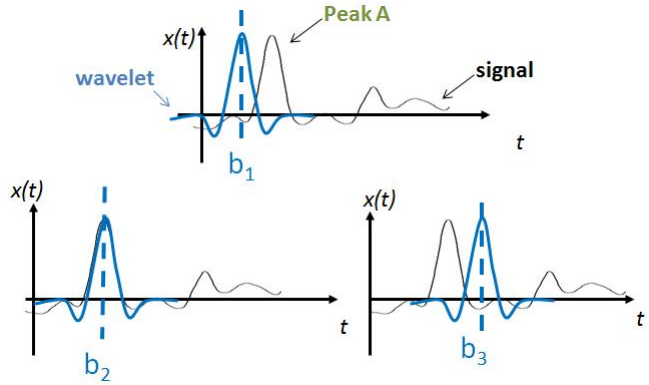

Figure 4.6. A wavelet at a fixed scale and three different translations

seen that the wavelet at scale a_2 and translation b_2 is the best match for representing peak A, and mathematically it will be seen that $T(a_2, b_2)$ will attain the maximum value among all a and all b in the neighborhood of the peak. We can explain this mathematical behavior by considering the positive or negative yield of the convolution $(x(t)\psi\left(\frac{t-b}{a}\right))$ in Equation 4.1 using Figure 4.7, as follows. For a wavelet with fixed a and b, Figure 4.7 indicates with a '+' or '-' sign the positive or negative value resulting from the convolution at different time segments (t) marked by the vertical lines along the x-axis. The time segments where the wavelet and signal are both positive or both negative result in a positive value for the convolution, and yields a negative value when they are of opposite signs ([106]). Hence, in our example in Figures 4.5 and 4.6, considering the different values of a and b it can be seen that $T(a_2, b_2)$ will provide

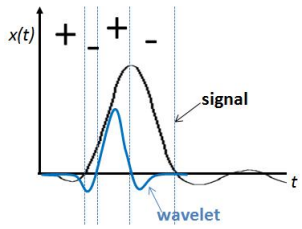

Figure 4.7. Positive and negative contributions of convolution of the wavelet with a signal

the maximum value. Sample values of $T(a, b)$, also commonly referred to as *wavelet coefficients*, at five scales and all translations for the length of the signal are plotted in Figure 4.8. As seen in the Figure, the use of CWT provides continuous translations thus generating continuous coefficients, which also appear as peaks with the use of suitable wavelet functions like the Mexican hat used here. Hence, the local maxima of coefficients at each scale represents the presence of peaks, and for each peak k, the best match scale ($\hat{a_k}$) and translation ($\hat{b_k}$) are obtained as the scale and translation of the local maxima ($T_k(\hat{a_k}, \hat{b_k})$) where b_k is in the neighborhood of k. From the set of identified local maxima, we further develop techniques to differentiate between noise and actual peak, which is explained later in this section. Further, since the baseline of the signal changes slowly thus making it monotonic around a peak, the convolution with a compact symmetric wavelet (e.g., Mexican hat) provides an automatic baseline correction ([98]). Notice that the baseline correction is visible in Figure 4.8, where, the coefficients take a zero baseline while the original signal had a baseline around 1500. Note that, in case of clear peaks, i.e., no overlaps, the value of $T_k(\hat{a_k}, \hat{b_k})$ (see Equation 4.1) is proportional to the area underneath the corresponding signal peak k, and the location of the crest of k is $\hat{b_k}$. However, in case of overlap peaks, neither $T_k(\hat{a_k}, \hat{b_k})$ or $\hat{b_k}$ will give an estimate of the peak quantity or location, respectively. Therefore, in this research, we also develop a methodology for quantifying overlapping peaks.

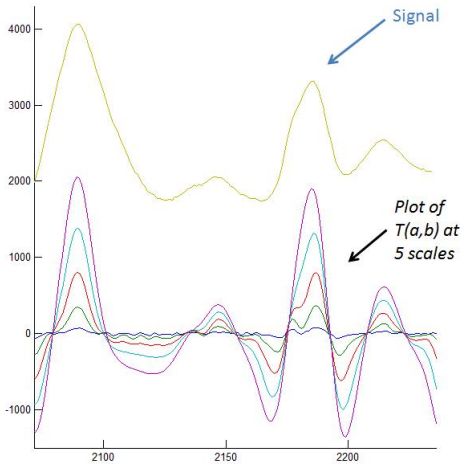

Figure 4.8. Sample plot of wavelet coefficients

Using the above knowledge of CWT, we describe below the steps for identification and quantification of peaks, including techniques for small peaks and overlapping peaks.

1. Obtain CWT coefficients $T(a, b), (\forall a, \forall b)$. The resulting T is a two dimensional matrix of size MxN, where, M is the number of scales and N is the length of the signal.

2. Using slopes, identify local maxima at each scale to obtain a matrix P of size MxN. If a local maxima is found at $T(a, b)$, $P(a, b) = T(a, b)$ else $P(a, b) = 0$.

3. Link local maxima (non-zero $P(a, b)$'s) across scales: As discussed earlier, the optimal scale $\hat{a}$ for a peak corresponds to the maximum value across scales, i.e, $P(\hat{a}, b) = max_a P(a, b)$. However, for any peak in the signal, its local maxima across scales does not occur at the same translation, i.e., if $P_k(1_k, b1_k)$ and $P_k(2_k, b2_k)$ are the local maxima corresponding to a peak k at scales 1 and 2, respectively, $b1_k \neq b2_k$ for most peaks. Therefore, to identify $\hat{a}_k$, the non-zero values of $P(a, b)$ across all scales that correspond to peak k need to be first linked. To achieve this, beginning at the lowest scale (we start with 2 since

scale 1 mostly represents noise), perform the following steps at each scale. Note that, the concept of linking local maxima is similar to obtaining ridges in [98], however, there are changes in the procedure that we adopt.

(a) For each $(\bar{a}, \bar{b})$ where $P(\bar{a}, \bar{b}) \neq 0$, using a window size w find a matrix index (a, b), such that, $(a \in 1 : \bar{a} - 1, b \in \bar{b} - w : \bar{b} + w)$ and $P(a, b) \neq 0$, i.e, find a local maxima in any of the previous scales in the neighborhood $\bar{b} \pm w$. The search is done in more than one previous scale, since, based on features of the signal, the local maxima might not occur at all scales. Note that, instead of setting a fixed value for w like in the literature, we estimate peak dependent value which is explained at the end of this section.

(b) If found, suppose at only one index $(\bar{a} - 1, \bar{b} - 4)$, i.e., $P(\bar{a} - 1, \bar{b} - 4) \neq 0$, then set $P(\bar{a}, \bar{b} - 4) = P(\bar{a}, \bar{b})$ and $P(\bar{a}, \bar{b}) = 0$, i.e, shift the location of the local maxima at scale $\bar{a}$ to align with that of the previous scales indicating that they belong to the same peak. If found in more than one index in the neighborhood $\bar{b} \pm w$, move it to the nearest location.

(c) If not found, i.e., $P(a, b) = 0 \; \forall (a \in 1 : \bar{a} - 1, b \in \bar{b} - w : \bar{b} + w)$, then retain the value of $P(\bar{a}, \bar{b})$, thus indicating that it is a new peak.

4. Location of peaks: Each non-zero column $\bar{b}$ in the modified P indicates the presence of a signal peak. For peak k, set the optimal scale as $\hat{a_k}$ where, $P_k(\hat{a_k}, \bar{b_k}) = \max_{a_k} P_k(a_k, \bar{b_k})$, and for now, consider the quantity of k as $Q(k) = P_k(\hat{a_k}, \bar{b_k})$. Set the location of the crest of k as $L(k) = \bar{b_k}$, thus corresponding to the location of the local maxima at the lowest scale, the reason for which can be explained as follows. For clear non-overlapping peaks, the location will be equal to the original best fit $\hat{b_k}$ in T, i.e., corresponding to $T_k(\hat{a_k}, \hat{b_k}) = P_k(\hat{a_k}, \bar{b_k})$. However, in case of overlapping peaks, column $\hat{b_k}$ will be skewed away from the location of

the peak crest. Now, consider the translation of the wavelet at a point $b1$ that corresponds to the crest of a peak in the signal $x(t)$. At lower scales where the width of the wavelet is very small, it is most likely that the wavelet is completely contained within the region of the signal peak for translations $b1 \pm n$ for a small n. Hence, the signs of the convolution along different segments of the wavelet (recollect Figure 4.7) will be the same at all $b_1 \pm n$, thus resulting in a maximum value of T at a translation where $x(t)$ would have the maximum profile, i.e., at $b1$. Therefore, the location of the local maxima at the lower scales is most likely to correspond to the crest or very close to the crest of the signal peak. See Figure 4.9, that contains a sample plot of $T(a, b)$, $x(t)$ (signal), and $P(a, b)$, for a better understanding of the relationship between the variables. Note that, the figure shows only 10 scales for ease of illustration. Due to the overlapped peaks, the location of $T_k(\hat{a_k}, \hat{b_k})$ is skewed away from the peak crest, however, the location corresponding to the translation at the lowest scale ($\bar{b_k}$) provides a better estimate.

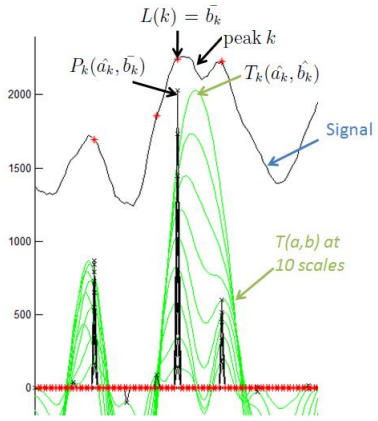

Figure 4.9. Location of peaks

5. Identification of valid peaks: For each peak k perform the following steps.

   (a) Estimate the width ($W(k)$) of the peak formed by the wavelet coefficients at scale $\hat{a_k}$ by searching around $\hat{b_k}$. Identify the highest scale at which the peak is recorded, $H(k)$, as the highest non-zero scale in $P_k(a_k, \bar{b_k})$.

   (b) Retain k as a valid peak by setting thresholds for $W(k)$, $\frac{Q(k)}{W(k)}$ and $H(k)$.

6. Quantification of overlapping peaks: Peak quantity is obtained by utilizing the knowledge of the positive and negative contributions to T arising from the convolution (refer to Figure 4.7 for the convolution contributions), and is explained as follows. Notice that, for any clear peak k, while $\hat{b_k}$ (i.e., the translation that provides the maximum transform $T_k(\hat{a_k}, \hat{b_k})$) is equivalent to the location of the peak crest, $T_k(\hat{a_k}, b_k) \leq 0$ when b_k corresponds to the peak troughs. Let $b_k = s_k(a)$ and $b_k = e_k(a)$ denote two translations (referring to (s)tart and (e)nd) at scale a in the neighborhood of peak k such that, $s_k(a) < e_k(a)$ and $T_k(a_k, s_k(a)) \leq 0$, $T_k(a_k, s_k(a) + 1) > 0$, $T_k(a_k, e_k(a)) \leq 0$, and $T_k(a_k, e_k(a) - 1) > 0$. Now consider two peaks i and j that overlap, and since they are almost always not of the same frequency, $\hat{a_i} \neq \hat{a_j}$, and let $\hat{a_i} < \hat{a_j}$. Due to the mathematical form of the convolution, the functions $T_i(\hat{a_i}, s_i(\hat{a}) : \hat{b_i} : e_i(\hat{a}))$ and $T_j(\hat{a_j}, s_j(\hat{a}) : \hat{b_j} : e_j(\hat{a}))$ will be such that, there will be some amount of overlap of the regions $s_i(\hat{a}) : e_i(\hat{a})$ and $s_j(\hat{a}) : e_j(\hat{a})$. The overlaps could either be complete, i.e, $s_j(\hat{a}) \leq s_i(\hat{a}) \leq e_i(\hat{a}) \leq e_j(\hat{a})$ (scenario 1), or partial, i.e, either $s_j(\hat{a}) \leq s_i(\hat{a}) \leq e_j(\hat{a}) \leq e_i(\hat{a})$ or $s_i(\hat{a}) \leq s_j(\hat{a}) \leq e_i(\hat{a}) \leq e_j(\hat{a})$ (scenario 2). Scenario 1 will most likely occur in cases when the smaller peak i is a low abundant peak occurring on the shoulder of a large peak, e.g., as in Figure 4.10, which presents the signal and its wavelet coefficients $T(a, b)$ at 25 scales. Utilizing the above knowledge and identifying overlaps of scenario 1, estimate

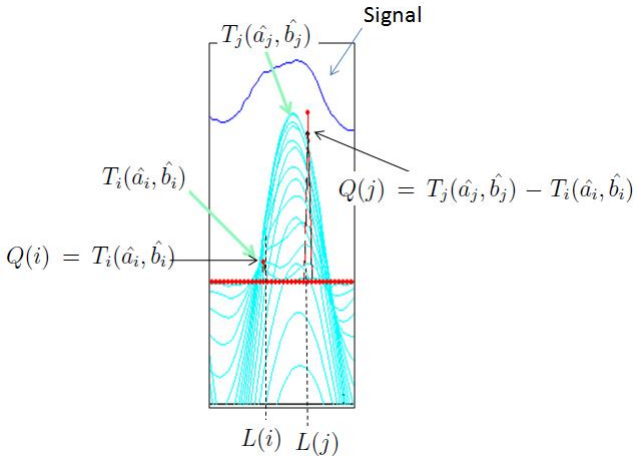

Figure 4.10. Estimating area of overlapping peaks under Scenario 1

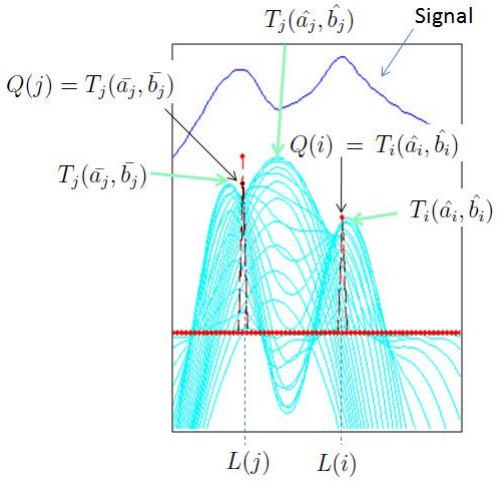

Figure 4.11. Estimating area of overlapping peaks under Scenario 2

$Q(i) = T_i(\hat{a_i}, \hat{b_i})$ and $Q(j) = T_j(\hat{a_j}, \hat{b_j}) - T_i(\hat{a_i}, \hat{b_i})$. Scenario 2 represents incomplete separation of two protein sets, where, the overlap will cause the coefficient to be maximum at a scale larger than the actual scale of the larger peak j (e.g., see Figure 4.11). Therefore, by identifying overlaps of scenario 2, and by searching through scales below $\hat{a_j}$, obtain a scale $\bar{a}$ where $s_j(\bar{a}) < e_j(\bar{a}) < s_i(\bar{a}) < e_i(\bar{a})$ or $s_i(\bar{a}) < e_i(\bar{a}) < s_j(\bar{a}) < e_j(\bar{a})$. Identify the translation of the local maxima at $\bar{a_j}$ as $\bar{b_j}$. Now, quantify the peaks as $Q(i) = T_i(\hat{a_i}, \hat{b_i})$ and $Q(j) = T_j(\bar{a_j}, \bar{b_j})$ (see Figure 4.11).

Note that, the algorithm presented in the above steps identifies low abundant peaks as well. For example, unlike use of signal-to-noise ratio (SNR) as in most methods in the literature, the thresholds for peak identification used in this research ($W(k)$ and $\frac{Q(k)}{W(k)}$ and $H(k)$) are only dependent on the peak under review and not on the neighborhood signal. Since noise varies along sections of the signal, using the characteristics of the peak alone helps obtain a better identification. Also, for each peak k at each scale a, the search window w for linking local maxima is set as $w = s_k(a) : e_k(a)$, where the translations $b_k = s_k(a)$ and $b_k = e_k(a)$ are as explained in Step 6 of the algorithm. Notice that, as a increases, $e_k(a) - s_k(a)$ increases or remains the same and $s_k(a) \leq s_k(a-1) \leq e_k(a-1) \leq e_k(a))$. Therefore, instead of setting a constant window size, determining the value based on the feature of the peak, as discussed above, provides a better search criteria. While incomplete chemical separation of proteins created overlapping peaks, using wavelet features to identify and quantify the value of each of these overlapping peaks provides as much protein separation as possible. Such separation allows for better identification of small significant biomarker peaks, where otherwise, the non-separation would have caused the large peaks to mask significant differences across small peaks.

4.3.2 Optimization Model for Peak Alignment

Alignment of peaks is achieved by developing an optimization algorithm. The objective of the algorithm is to minimize the magnitude distance between two signals, since misaligned peaks are likely to have more magnitude. Note that, in the context of peak alignment, a *signal* is an array whose length equals the length of the original chromatogram. The array elements take value equal to peak intensity if its array index corresponds to location of identified peak (obtained using Q and L from Section 4.3.1), and takes a zero value at all other index. Distance is estimated as the sum of

magnitude difference between elements of the two signals. The optimization model developed for *peak alignment* is as follows. Let A and B be two signals, where the objective is to align B against A, i.e., for every peak in A we need to find a matching peak from B. Let $B_m = AlignedB$. The optimization model for the alignment is formulated as follows.

$$
\begin{aligned}
Objective: Minimize \sum_x |A(x) - B_m(x)| + \sum_y \theta(y) & \\
B_m(x) = \sum_y B(y) * \beta(y,x) \qquad & \forall(x) \qquad (4.3)\\
\sum_x \beta(y,x) \leq 1 \qquad & \forall(y) \qquad (4.4)\\
\sum_y \beta(y,x) \leq 1 \qquad & \forall(x) \qquad (4.5)\\
\sum_y \beta(y,x)y - \sum_y \beta(y,j)y \sum_y \beta(y,x) \geq 0 \qquad & \forall(x, \forall(j<x)) \qquad (4.6)\\
\sum_x B(y) * \beta(y,x) + \theta(y) = B(y) \qquad & \forall(y) \qquad (4.7)\\
\beta(y,x) = 0 \quad \forall|y-x| > \alpha, & \qquad (4.8)\\
where \;\; \alpha = maximum\ possible\ shift\ on\ x-axis(time-axis) & \\
\beta(y,x)\ \epsilon\ \{0,1\} \qquad & \forall(y, \forall(x)) \\
B_m(x)\ \epsilon\ \mathcal{R} \qquad & \forall(x) \\
\theta(y)\ \epsilon\ \mathcal{R} \qquad & \forall(y) \\
& \qquad (4.9)
\end{aligned}
$$

The objective function minimizes 'distance + unmatched', where, 'distance' equals the sum of difference between aligned peak intensities and 'unmatched' equals sum of peak intensities in $B(y)$ that did not find a matching peak in $A(x)$. Constraint 4.3 tracks peak shifts by using binary variables, Constraints 4.4 and 4.5 avoid duplication of peaks, Constraint 4.6 ensures that the sequential order of the peaks is maintained,

Constraint 4.7 keeps track of peaks in $B(x)$ that did not find a matching peak in $A(x)$, and Constraint 4.8 places a bound on the peak shift. Since peaks were expressed over pH scale, the shift of peaks is limited to a range of neighboring pH and hence search beyond the range is not required.

4.3.2.1 Transformation from Nonlinear to Linear

Note that, constraint 4.6 contains a nonlinear term. Since solving a nonlinear model is much more difficult than a linear model, for computational efficiency, we linearize constraint 4.6 as follows.

$$\sum_{y} \beta(y,x) = \gamma(x) \qquad \forall(x) \qquad (4.10)$$

$$\delta(x) + \gamma(x) = 1 \qquad \forall(x) \qquad (4.11)$$

$$\sum_{y} \beta(y,x) * y - \sum_{y} \beta(y,j) * y + \delta(x) * 100000 \geq 0 \quad \forall(x, \forall(j < x)) \qquad (4.12)$$

$$\delta(x) \; \epsilon \; \{0,1\} \qquad \forall(x)$$

$$\gamma(x) \; \epsilon \; \{0,1\} \qquad \forall(x)$$

Hence the above constraints will replace constraint 4.6.

4.3.2.2 Heuristics to Minimize Execution Time of Optimization Algorithm

The optimization algorithm described above for peak alignment was solved using CPLEX optimization solver. For a pair of data signals, each with length 150 peaks, the number of constraints will equal 30,225, and will increase to 97,950 for a length of 300 peaks. The average signal length for a sample can be approximately 2250 peaks, making the problem very complex which can take forever to solve. Therefore, we developed a heuristic approach to minimize the execution time which can be

described as follows. Note that, for any given peak, the distance of shift is limited due to its occurrence based on chemical property. Therefore we adopt the following two step procedure of breaking down the signal into multiple smaller fragments:

1. Every sample has a small set of common high abundance proteins. Therefore, in the first step, we consider a signal as an array of high abundance proteins only. The optimization algorithm is applied to align these high abundance proteins across two samples.

2. The point of alignments in step 1 is used as a breakpoint, to divide the entire data set into multiple smaller signal fragments. Each signal fragment will consist of an array of a small number of peaks. Therefore, in step 2, the algorithm is applied on each set of signal fragment separately, and the peaks within each fragment are aligned.

4.4 Results and Discussion

Our algorithm was tested on PF2D chromatograms obtained on urine samples of ovarian cancer patients. The samples were obtained as part of the Tampa Bay Ovarian Cancer Coalition (TBOCC)- a population-based study conducted by a team from Moffitt Cancer Center and Research Institute, Tampa. The study is being performed in the Tampa Bay metropolitan region of the State of Florida, where about 400 blood and urine samples will be prospectively obtained from ovarian cancer patients and also from a matching control. The cancer cases and controls are matched based on demographics and risk factors like age, menopausal status, and race. This study, to our knowledge, is the largest prospective collection of preoperative population-based samples in the US.

4.4.1 Peak Detection

For comparison of peak detection capacity, peaks in PF2D chromatograms were identified using our algorithm and MassSpecWavelet ([98]), which is a peak detection method also based on complex wavelet transforms. Note that, MassSpecWavelet was originally developed for SELDI-TOF mass spectrometry based spectrums, whose protein separation procedure differs from that of PF2D as was explained earlier in Section 4.2, however, with similar challenges in protein identification. We use MassSpecWavelet for comparison as it has been shown to perform comparatively better that the other algorithms ([85]) for peak identification. MassSpecWavelet differentiates between noise and valid peak based on two main thresholds, signal-to-noise ratio (SNR) and amplitude of the local maxima of the peak. The SNR is estimated as the ratio of local maxima to the local noise level, where, the noise level within a local window size around the peak is estimated as the 95th percentile of the coefficients at the first scale. Since MassSpecWavelet was developed for MS spectrums, we also compare the peak detection capacity of the two algorithms on MS spectrums. We present below, the comparison of results on PF2D chromatograms followed by MS spectrums. Note that, all visual verifications mentioned below were based on expert opinion, biochemists and oncologists.

4.4.1.1 Comparison of Peak Identification on PF2D Chromatograms

The thresholds in MassSpecWavelet and our algorithm were varied to obtain different number of peak selections, and the ensuing results with approximately same number of selections in both algorithms were compared. While large peaks are identified in both algorithms, it was found that our algorithm performs better in identification of small peaks and overlapping peaks. A sample result is presented by considering

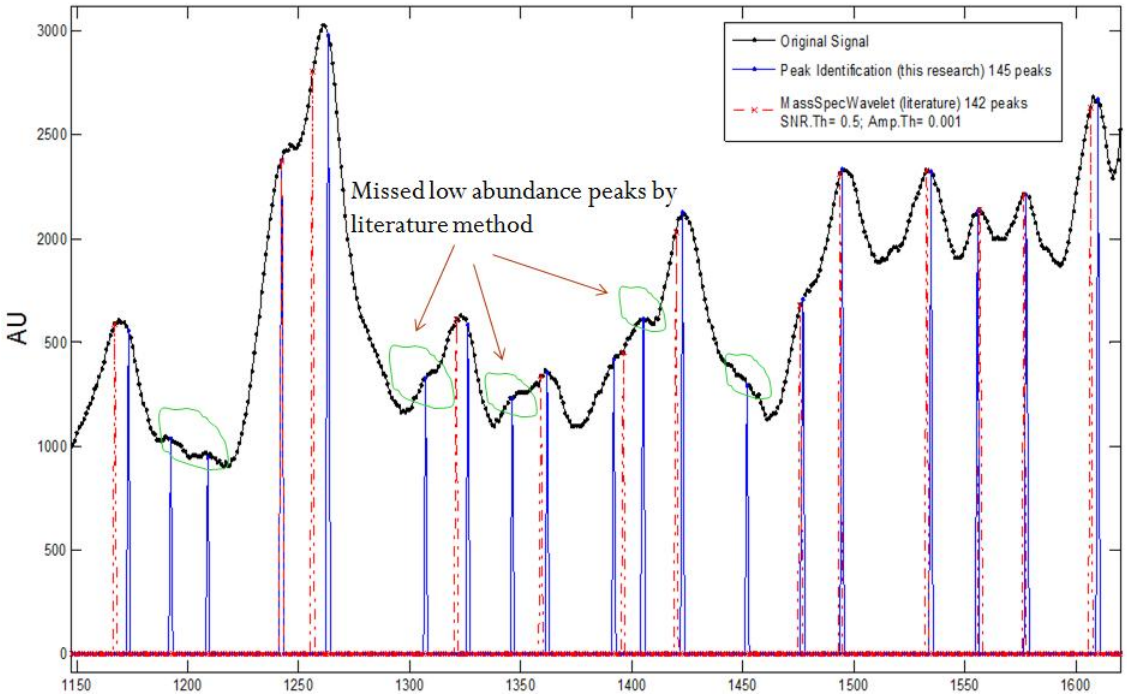

Figure 4.12. Peak identification on section of chromatogram with several peaks

two sections of the same chromatogram, one with several valid peaks (Figure 4.12) and the other with lot of noise (Figure 4.13). The blue lines indicate location of peaks identified by our algorithm and red lines by MassSpecWavelet, where the number of peaks identified are 145 and 142, respectively. Notice that, the section of the chromatogram shown in Figure 4.12 contains several small peaks and overlapping peaks, and as circled, several of which are not identified by MassSpecWavelet. Moreover, MassSpecWavelet identifies several false peaks in the tail end of the chromatogram that usually has just a few valid peaks (Figure 4.13) and also misses some valid peaks, which is more clearly seen in the enlarged portion between data points 6500 to 7100. Note that, the false detection of peaks by our algorithm is much lesser as can be seen in the Figure. Though the chemical content of the small circled peaks, that were missed by MassSpecWavelet in Figure 4.12, is not confirmed, it may be noted that, such automatic identification of all peak like appearances is essential, the reason

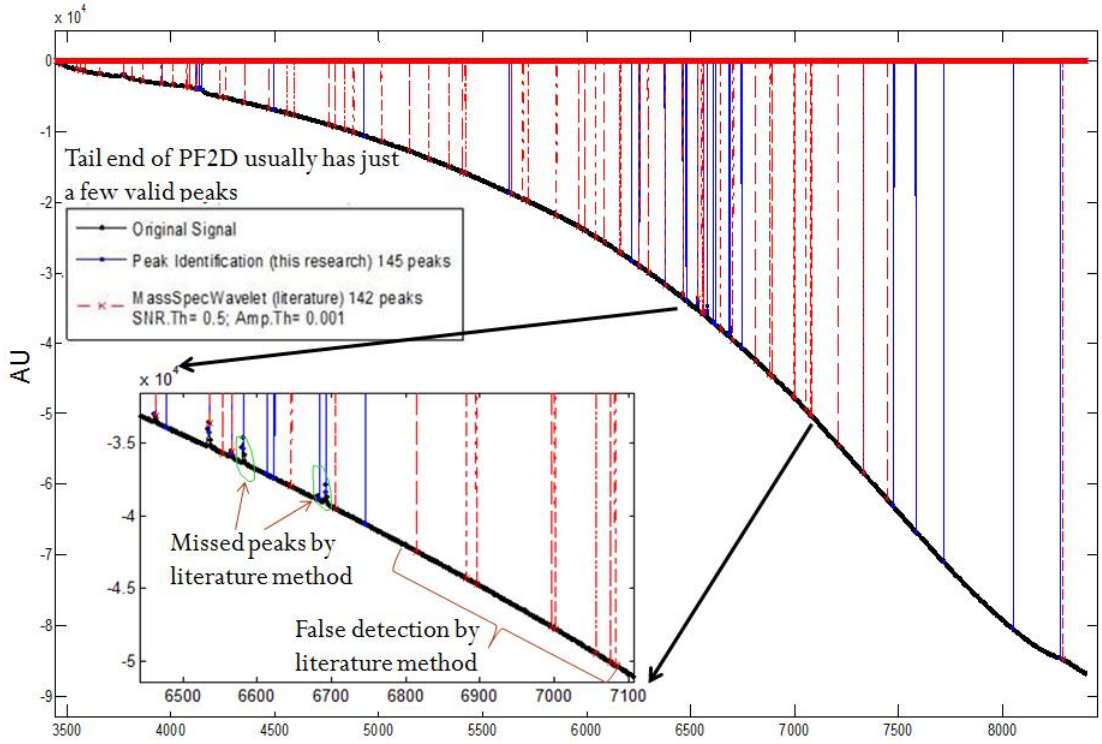

Figure 4.13. Peak identification on section of chromatogram with high noise

for which can be explained as follows. Since there are about 500,000 proteins most of which are present in small amounts in urine and blood, obtaining the undiscovered small biomarker proteins requires identification of peaks that distinguish cancer cases from controls. The vast number of small peaks makes it impossible to manually identify distinguishing peaks, and moreover the distinguishing characteristics might be present in the form of a pattern or set, rather than individual peaks, hidden from the naked eye. Hence, the identification of distinguishing peaks requires the use of a mathematical model, and therefore, an automatic algorithm that identifies the location and quantifies the area of all peak like appearances.

In order to test the capacity of identification of small and overlapping peaks by MassSpecWavelet, the thresholds of peak detection were further lowered. As can be seen in Figures 4.14 and 4.15, where the thresholds in MassSpecWavelet have been re-

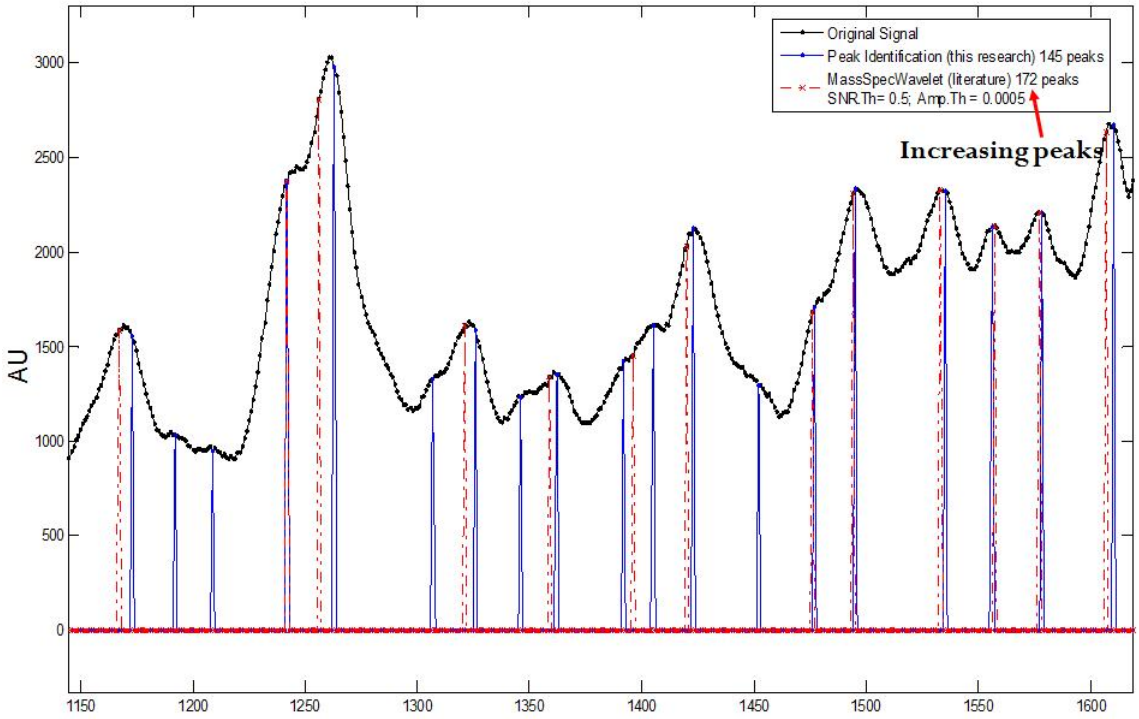

Figure 4.14. Sample results of peak identification: Increasing peaks selected by MassSpecWavelet to 172

duced to increase the number of peaks to 172 and 202, respectively, MassSpecWavelet again misses the small and overlapping peaks.

4.4.1.2 Comparison of Peak Identification on MS Spectrum

Both algorithm (ours and MassSpecWavelet) were applied on MALDI TOF mass spectrometry (MS) based Aurum data ([107] publicly available at https: // proteomecommons. org/). The advantage of using Aurum data is that it provides the list of know valid peaks. Hence, we estimated the peak detection sensitivity of both algorithms, and compared sensitivities corresponding to thresholds that provide equivalent number of identified peaks in both algorithms. The number of peaks and sensitivity results are presented in Table 4.1. For our algorithm, all thresholds were kept constant except for $\frac{Q(k)}{W(k)}$, which was varied to identify different number of peaks.

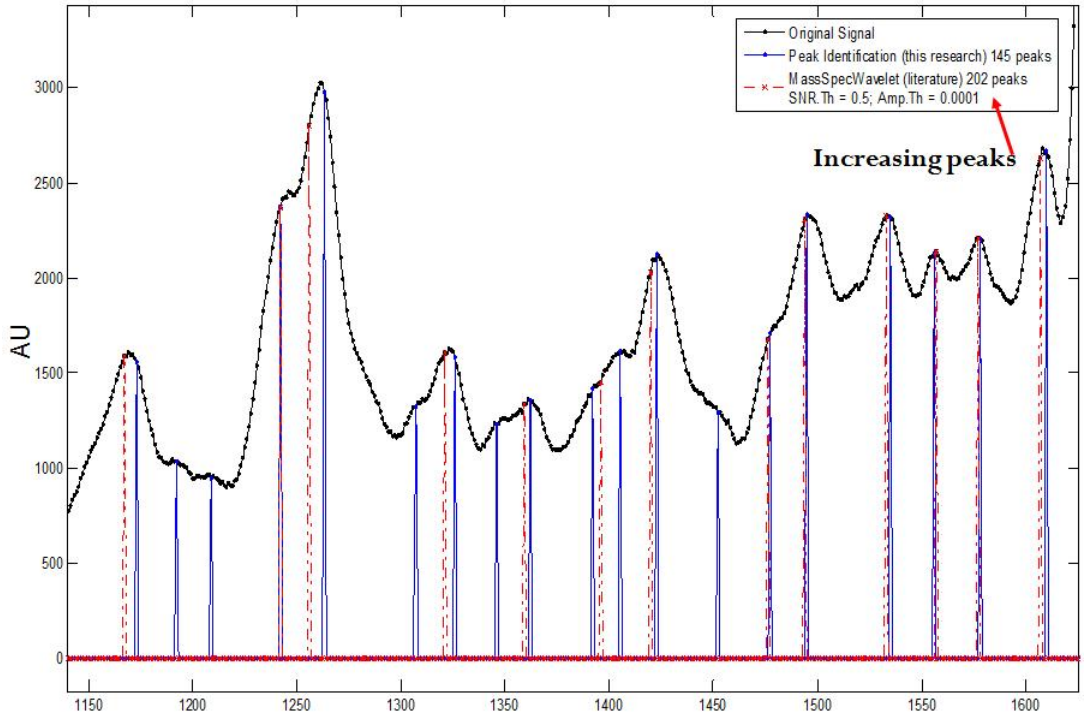

Figure 4.15. Sample results of peak identification: Increasing peaks selected by MassSpecWavelet to 202

In the Table, rows 1 to 3 present results where MassSpecWavelet's threshold for SNR was kept constant while amplitude was varied to identify different number of peaks. Rows 4 to 6 present results where SNR was varied and amplitude was slightly tuned so as to obtain an equivalent of 195 peaks to match that obtained by our algorithm. As can be seen, our algorithm provides better sensitivity for all cases.

4.4.2 Peak Alignment

Results of peak alignment across samples were visually validated on the PF2D chromatograms. Sample alignment is explained in Figure 4.16 and a larger section of the alignment is presented in Figure 4.17. As explained in Figure 4.16, the objective of the optimization model was to align peaks in Sample 2 against peaks in Sample 1. The dark blue lines indicate the original location of the identified peaks (using

Table 4.1. Sensitivity of detection of known peaks in the Aurum data

Our Algorithm		MassSpecWavelet	
Number of Identified Peaks	Sensitivity (%)	Number of Identified Peaks (Thresholds: SNR, Amplitude)	Sensitivity (%)
155	82.7	151 (1, 0.01)	65.4
209	84.6	210 (1, 0.0075)	69.2
305	90.4	305 (1, 0.0049)	75.0
195	81.8	197 (3, 0.007)	67.3
195	81.8	201 (2, 0.007)	67.3
195	81.8	195 (1, 0.0075)	67.3

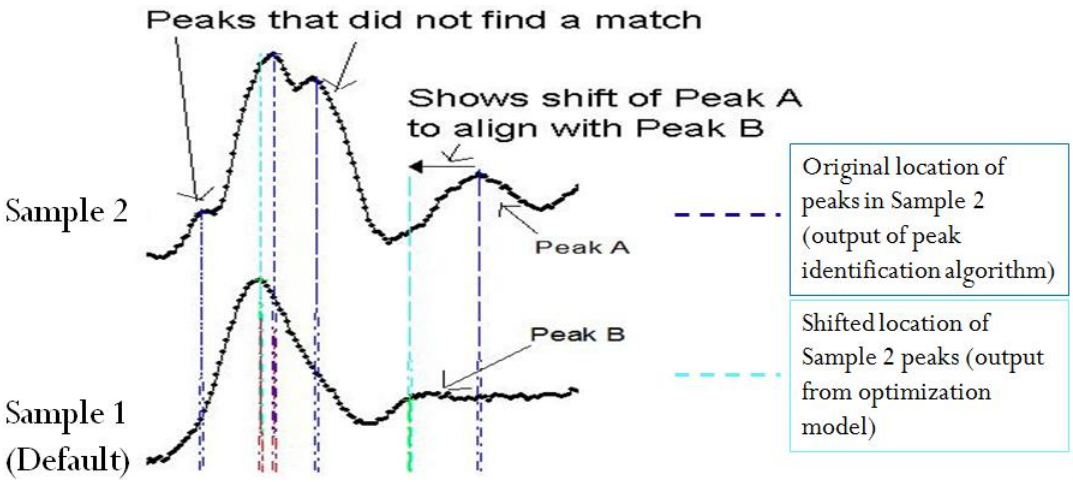

Figure 4.16. Peak alignment on PF2D data was visually validated

L from Section 4.3.1), and the light blue lines indicate the shifted location (output from the optimization model in Section 4.3.2). Notice that, from the example shown in the Figure, visually speaking we can say that peaks A and B represent the same protein, hence confirming the results from the optimization model (where the light blue line corresponds to crest of peak B indicating that peak A has been aligned with peak B). Similar visual confirmation of alignments was conducted on several sections of multiple datasets.

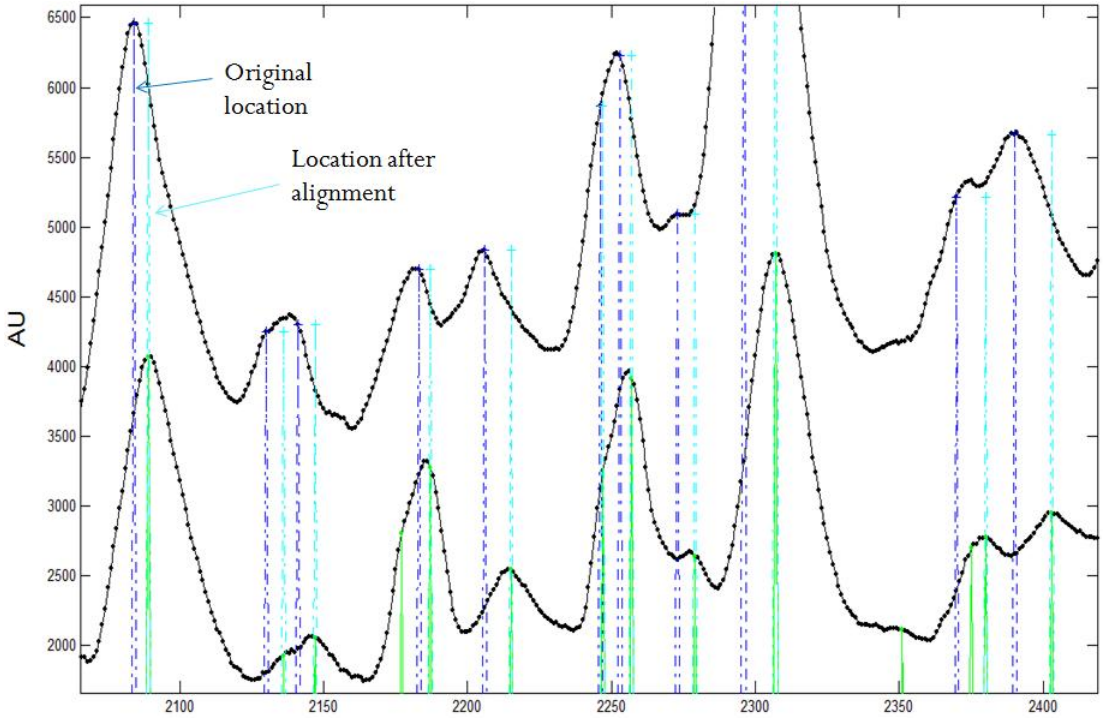

Figure 4.17. A portion of PF2D chromatograms illustrating peak alignments

REFERENCES

[1] M. Leshno, Z. Halpern, and N. Arber. Cost-effectiveness of colorectal cancer screening in the average risk population. *Health Care Management Science*, 6:165 174, 2003.

[2] Epidemiology Resource Center/Data Analysis Team Indiana State Department of Health. Indiana health behavior risk factors 2001 state survey data. url{http://www.in.gov/isdh/reports/brfss/2001/index.htm, August 2002. Last accessed - Feb 03, 2009.

[3] I. Vogelaar, M. van Ballegooijen, D. Schrag, R. Boer, S. J. Winawer, J. D. F. Habbema, and A. G. Zauber. How much can current interventions reduce colorectal cancer mortality in the U.S. *Cancer*, 107(7):1624– 1633, October 2006.

[4] American Cancer Society. The history of cancer. url{http: //www.cancer.org /docroot /CRI /content /CRI _2 _6x _the _history _of _cancer _72.asp, March 2010. Last accessed - May 14, 2010.

[5] Kosary C.L. Krapcho M. Neyman N. Aminou R. Waldron W. Ruhl J. Howlader N. Tatalovich Z. Cho H. Mariotto A. Eisner M.P. Lewis D.R. Cronin K. Chen H.S. Feuer E.J. Stinchcomb D.G. Edwards B.K. (eds) Altekruse, S.F. Seer cancer statistics review, 1975-2007, National Cancer Institute, Bethesda, MD.

[6] B. Morson. The polyp-cancer sequence in the large bowel. *Proceedings of Royal Society of Medicine*, 67:451– 457, 1974.

[7] A. I. Neugut, S. Jacobson, J, and I. DeVivo. Epidemiology of colorectal adenomatous polyps. *Cancer Epidemiology, Biomarkers and Prevention*, 2:159– 176, March/April 1993.

[8] A. Leslie, F. A. Carey, N. R. Pratt, and R. J. C. Steele. The colorectal adenoma-carcinoma sequence. *British Journal of Surgery*, 89:845– 860, 2002.

[9] F. Loeve, R. Boer, G. J. Oortmarssen, M. V. Ballegooijen, and J. D. F Habbema. The miscan-colon simulation model for the evaluation of colorectal cancer screening. *Computers and Biomedical Research*, 32:13– 33, 1999.

[10] F. Loeve, M. L. Brown, R. Boer, M. van Ballegooijen, G. J. van Oortmarssen, and J. D. F Habbema. Endoscopic colorectal cancer screening: a cost-saving analysis. *J Natl Cancer Inst*, 92(7):557– 563, April 2000.

[11] S. Roberts, L. Wang, R. Klein, R. Ness, and R. Dittus. Development of a simulation model of colorectal cancer. *ACM Transactions on Modeling and Computer Simulation*, 18(1), December 2007.

[12] P. R. Harper and S. K. Jones. Mathematical models for the early detection and treatment of colorectal cancer. *Health Care Management Science*, 8:101– 109, 2005.

[13] R. K. Khandker, J. D. Dulski, J. B. Kilpatrick, R. P. Ellis, J. B. Mitchell, and W. B. Baine. A decision model and cost-effectiveness analysis of colorectal cancer screening and surveillance guidelines for average-risk adults. *International Journal of Technology Assessment in Health Care*, 16(3):799– 810, 2000.

[14] J.L. Wagner, S. Tunis, M. Brown, A. Ching, and R. Almeida. The cost effectiveness of colorectal cancer screening in average-risk adults. *In: Young GP, Rozen P, Levin B, eds. Prevention and early detection of colorectal cancer. Philadelphia: WB Saunders*, pages 321– 356, 1996.

[15] R. T. Clemen and C. J. Lacke. Analysis of colorectal cancer screening regimens. *Health Care Management Science*, 4:257– 267, 2001.

[16] S. Vijan, E. W. Hwang, T. P. Hofer, and R. A. Hayward. Which colon cancer screening test a comparison of costs, effectiveness, and compliance. *The American Journal of Medicine*, 111:593– 601, December 2001.

[17] National Cancer Institute NCI. Cisnet-cancer intervention and surveillance modeling network. http://cisnet.cancer.gov/colorectal/profiles.html http:// cisnet.cancer.gov/ publications/ Colorectal, 2007. Last accessed: Feb 03- 2009.

[18] S. J. Winawer, R. H. Fletcher, L. Miller, F. Godlee, M. H. Stolar, C. D. Mulrow, S. H. Woolf, S. N. Glick, T. G. Ganiats, J. H. Bond, L. Rosen, J. G. Zapka, S. J. Olsen, F. M. Giargiello, J. E. Sisk, R. Van Antwerp, C. Brown-Davis, D. A. Marciniak, and R. J. Mayer. Colorectal cancer screening: Clinical guidelines and rationale. *Gastroenterology*, 112:594– 642, 1997.

[19] F. Loeve, R. Boer, A. G. Zauber, M. van Ballegooijen, G. J. van Oortmarssen, S. J. Winawer, and J. D. F. Habbema. National polyp study data: Evidence for regression of adenomas. *Int. Journal of Cancer*, 111:633– 639, 2004.

[20] H. Brenner, M. Hoffmeister, C. Stegmaier, G. Brenner, L. Altenhofen, and U. Haug. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. *Gut*, 56:1585– 1589, 2007.

[21] J-M. Wong, M-F. Yen, M-S. Lai, W. Duffy, S., R. A. Smith, and T. H-H. Chen. Progression rates of colorectal cancer by duke's stage in a high-risk group: Analysis of selective colorectal cancer screening. *The Cancer Journal*, 10(3):160– 169, May/June 2004.

[22] R. M. Soetikno, T. Kaltenbach, R. V. Rouse, W. Park, A. Maheshwari, T. Sato, S. Matsui, and S. Friedland. Prevalence of nonpolypoid (flat and depressed) colorectal neoplasms in asymptomatic and symptomatic adults. *Journal of Americal Medical Association*, 299(9):1027– 1035, March 2008.

[23] C. E. Dukes. The classification of cancer of the rectum. *Journal of Pathological Bacteriology*, 35(3):323– 332, 1932.

[24] R.L. Schoen, Pinsky P.F., J. L. Weissfeld, R. S. Bresalier, T. Church, P. Prorok, and J. K. Gohagan. Results of repeat sigmoidoscopy 3 years after a negative examination. *Journal of the American Medical Association*, 290(1):41– 48, July 2003.

[25] G. C. Harewood and G. O. Lawlor. Incident rates of colonic neoplasia according to age and gender. *J Clinical Gastroenterology*, 39(10):894– 899, December 2005.

[26] M. Noe, P. Schroy, Babayan R. Demierre, M-F., and A. C. Geller. Increased cancer risk for individuals with a family history of prostate cancer, colorectal cancer, and melanoma and their associated recommendations and practices. *Cancer Causes Control*, 19:1– 12, 2008.

[27] Epidemiology Resource Center Indiana State Department of Health and Division of Chronic/Communicable Disease. Cancer incidence and mortality in indiana. http://www.in.gov/isdh/reports/cancerinc/2004/section2c.htm, 2004. Last accessed - Feb 03 2009.

[28] National Cancer Institute. Surveillance epidemiology and end results. http://seer.cancer.gov/faststats/selections.phpOutput, 2006. Last accessed - Feb 03 2009.

[29] U.S Census Bureau. 2006-2008 american community survey 3-year estimates. url{http://factfinder.census.gov/servlet/ACSSAFFFacts. Last accessed - Feb 03, 2009.

[30] National Cancer Institute. Prostate, lung, colorectal and ovarian cancer screening trial (plco). url{http://prevention.cancer.gov/programs-resources/groups/ed/programs/plco. Last accessed - Feb 03, 2009.

[31] American Cancer Society. Cancer facts and figures 2008. *Atlanta: American Cancer Society*, 2008.

[32] S. J. Winawer, R. H. Fletcher, D. Rex, J. Bond, R. Burt, Ferrucci J., T. Ganiats, T. Levin, S. Woolf, D. Johnson, L. Kirk, S. Litin, and C. Simmang. Colorectal cancer screening and surveillance: Clinical guidelines and rationaleupdate based on new evidence. *Gastroenterology*, 124:544– 560, 2003.

[33] L. Jenkins, D. Bradshaw, P. Cannon, J. Gierisch, and W. Freas. Colorectal cancer screening in local health departments - a pilot project of the north carolina advisory committee on cancer coordination and control and the north carolina division of public health. 2003.

[34] Epidemiology Resource Center Indiana State Department of Health and Division of Chronic/Communicable Disease. Cancer incidence and mortality in indiana. http://www.in.gov/isdh/reports/cancerinc/2004/section2r.htm, 2004. Last accessed - Feb 03 2009.

[35] American Cancer Society. Cancer facts and figures 2008-2010. *Atlanta: American Cancer Society*, 2008.

[36] Epidemiology Resource Center Indiana State Department of Health and Division of Chronic/Communicable Disease. Cancer incidence and mortality in indiana. http://www.in.gov/isdh/reports/cancerinc/2004/section2s.htm, 2004. Last accessed - Feb 03 2009.

[37] J. S. Mandel, J. H. Bond, T. R. Church, D. C. Snover, B. G. Mary, Schuman L. M., and F. Ederer. Reducing mortality from colorectal cancer by screening for fecal occult blood. *The New England Journal of Medicine*, 328(19):1365– 1371, May 1993.

[38] J. S. Mandel, T. R. Church, F. Ederer, and J. H. Bond. Colorectal cancer mortality: Effectiveness of biennial screening for fecal occult blood. *J. of National Cancer Institute*, 91(5):434– 437, March 1999.

[39] V. A. Gilbertsen, R. McHugh, L. Schuman, and S. E. Williams. The earlier detection of colorectal cancers. *Cancer*, 45:2899– 2901, June 1980.

[40] Argonne National Lab ANL. Repast-recursive porous agent simulation toolkit. http://repast.sourceforge.net/, 2007. Last accessed: Feb 03- 2009.

[41] G. Piatetsky-Shapiro and P. Tamayo. Microarray data mining: Facing the challenges. *SIGKKD Explorations*, 5(2).

[42] B.H. Mecham, D.Z. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T.J. Mariani. Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics*, 18:308–315, 2004.

[43] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray. *Proceedings of the National Academy of Sciences*, 99(22):14031–14036, 2002.

[44] Affymetrix. http:// www.affymetrix.com /corporate /media /genechip_essentials /index.affx. Last accessed - May, 2007.

[45] R. Dror. Noise models in gene array analysis. *Report in fulfillment of the area exam requirement in the MIT Department of of Electrical Engineering and Computer Science*, June 2001.

[46] H. Vikalo, B. Hassibi, and A. Hassibi. A statistical model for microarrays, optimal estimation algorithms, and limits of performance. *IEEE Transactions on Signal Processing*, 54(6):2444–2455, 2006.

[47] A. Hassibi, S. Zahedi, R. Navid, R.W. Dutton, and T.H. Lee. Biological shot-noise and quantum-limited signal-to-noise ratio in affinity-based biosensors. *Journal of Applied Physics*, 97, 2005.

[48] X.H. Wang, R.S.H. Istepanian, and Y.H. Song. Microarray image enhancement by denoising using stationary wavelet transform. *IEEE Transactions on Nanobioscience*, 2(4):184–189, 2003.

[49] V.M. Aris, M.J. Cody, J. Cheng, J.J. Dermody, P. Soteropoulos, M. Recce, and P.P. Tolias. Noise filtering and non-parametric analysis of microarray data underscores disriminating markers of oral, prostate, lung, ovarian and breast cancer. *BMC Bioinformatics*, 5, 2004.

[50] X. Cai and G.B. Giannakis. Identifying differentially expressed genes in microarray experiments with model-based variance estimation. *IEEE Transactions on Signal Processing*, 54(6):2418–2426, 2006.

[51] R. Lukac, K.N. Plataniotis, B. Smolka, and A.N. Venetsanopoulos. A multichannel order-statistic techniques for cdna microarray image processing. *IEEE Transactions on Nanobioscience*, 3(4), 2004.

[52] N. Kingsbury. Image processing with complex wavelets. *Phil. Trans. R. Soc. Lond. A*, 1999.

[53] J. K. Romberg, H. Choi, and R. G. Baraniuk. Bayesian tree-structured image modeling using wavelet-domain hidden markov models. *IEEE Transactions on Image Processing*, 10(7):1056–1068, 2001.

[54] F. Abramovich, T.C. Bailey, and T. Sapatinas. Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society*, 49, 2000.

[55] L. Sendur and I. W. Selesnick. Bivariate shrinkage functions for wavelet-based denoisisng exploiting interscale dependency. *IEEE Transactions on Signal Processing*, 50(11), 2002.

[56] Affymetrix. Data sheet: Genechip human genome arrays. Technical report, Affymetrix, 2003-2004. Last accessed - January, 2008.

[57] S.G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 1989.

[58] R. Ganesan, T.K. Das, and V. Venkataraman. Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Transactions*, 36, 2004.

[59] N. G. Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Journal of Applied and Computational Harmonic Analysis*, 10(3):234–253, 2001.

[60] Affymetrix. Statistical algorithms description document. *Affymetrix Inc., CA, USA*, 2002.

[61] Moffitt Cancer Center & Research Institute and the University of South Florida. SRC sofware library. http://morden.csee.usf.edu/software/. Last accessed - January, 2008.

[62] I. Selesnick, S. Cai, K. Li, L. Sendur, and A. F. Abdelnour. Matlab implementation of wavelet transforms. http: // taco.poly.edu/ WaveletSoftware/ denoise.html. Last accessed - January, 2008.

[63] University of South Florida. Research computing. http://rc.usf.edu/. Last accessed - January, 2008.

[64] L. Sendur and I. W. Selesnick. Bivariate shrinkage with local variance estimation. *IEEE Signal Processing Letters*, 9(12):438–441, 2002.

[65] The Proteomics Center at Children's Hospital Boston. Introduction to proteomics. http:// www.childrenshospital.org/ cfapps/ research/ data_admin/ Site602/ mainpageS602P0.html. Last accessed - Nov 16th, 2008.

[66] G. Black. Mechanisms of alternative pre-messenger rna splicing. *Annual Review of Biochemistry*, 72:291–336, 2003.

[67] Schlautman J.D., W. Rozek, R. Stetler, R. L. Mosley, H. E. Gendelman, and P. Ciborowski. Multidimensional protein fractionation using proteomelab pf 2d for profiling amyotrophic lateral sclerosis immunity: A preliminary report. *Proteome Sci*, 6(26), 2008.

[68] J. Jessani and B. F. Cravatt. The development and application of methods for activity-based protein profiling. *Current Opinion in Chemical Biology*, 8:54–59, 2004.

[69] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow. The seldi-tof ms approach to proteomics: Protein profiling and biomarker identification. *Biochemical and Biophysical Research Communications*, 292:587592, 2002.

[70] 2007.http://www.emedicine.com/med/fulltopic/topic1698 Garcia, AA. Ovarian Cancer. eMedicine from WebMD. Accessed online, 11/01/2008.

[71] R. Sutphen, Y. Xu, G. D. Wilbanks, J. Fiorica, E. C. Grendys Jr., J. P. LaPolla, H. Arango, M. S. Hoffman, M. Martino, K. Wakeley, D. Griffin, R. B Blanco, A. B. Cantor, Y-J. Xiao, and J. P. Krischer. Lysophospholipids are potential biomarkers of ovarian cancer. *Cancer Epidemiol Biomarkers Prev*, 13(7):1185–1191, 2004.

[72] S. Bodovitz and T. Joos. The proteomics bottleneck: strategies for preliminary validation of potential biomarkers and drug targets. *Trends in Biotechnology*, 22(1):4–7, 2004.

[73] J.L. Luque-Garcia and T.A. Neubert. Sample preparation for serum/plasma profiling and biomarker identification by mass spectrometry. *Journal of Chromatography A*, 1153:259–276, 2007.

[74] H.J. An, S. Miyamoto, K. S. Lancaster, C. Kirmiz, B. Li, K. S. Lam, G. S. Leiserowitz, and C. B. Lebrilla. Profiling of glycans in serum for the discovery of potential biomarkers for ovarian cancer. *Journal of Proteome Research*, 5:1626–1635, 2006.

[75] N. Dossat, A. Mang, J. Solassol, W. Jacot, Maudelonde T. Daurs J-P. Lhermitte, L., and N. Molinari. Comparison of supervised classification methods for protein profiling in cancer diagnosis. *Cancer Informatics*, 3:295–305, 2007.

[76] M.H. Simoniam and E. Betgovargez. Proteome analysis of human plasma with the proteomelab PF2D system A1936A. 2003.

[77] A. Schramm, O. Apostolov, B. Sitek, K. Pfeiffer, K. Sthler, H.E. Meyer, W. Havers, and A. Eggert. Proteomics: techniques and applications in cancer research. *Klin Padiatr*, 215(6), 2003.

[78] J. Reinders, U. Lewandrowski, J. Moebius, Y. Wagner, and A. Sickmann. Challenges in mass-spectrometry based proteomics. *Proteomics*, 4(12):3686–3703, 2004.

[79] Y.K. Shin, H-J. Lee, J. S. Lee, and Y-K. Paik. Proteomic analysis of mammalian basic proteins by liquid-based two-dimensional column chromatography. *Proteomics*, 6(4):1143–1150, 2006.

[80] Beckman coulter: From tissues to targets: Proteomelab pf 2d protein fractionation system, br9436a. Technical report, Beckman Coulter, Inc., 2003.

[81] H.J. Lee, M-J. Kang, E-Y. Lee, S. Y. Cho, H. Kim, and Y. K. Paik. Application of a peptide-based pf2d platform for quantitative proteomics in disease biomarker discovery. *Proteomics*, 8(16):3371–3381, 2008.

[82] D.G. Ward, N. Suggett, Y. Cheng, W. Wei, H. Johnson, L. J. Billingham, T. Ismail, M. J. O. Wakelam, P. J. Johnson, and A. Martin. Identification of serum biomarkers for colon cancer by proteomic analysis. *Br J Cancer*, 94(12):1898–1905, 2006.

[83] H. B. Burke. Proteomics: Analysis of spectral data. *Cancer Informatics*, 1:15–24, 2005.

[84] A. Cruz-Marcelo, R. Guerra, amd Li Y. Vannucci, M., C. C. Lau, and T .K. Man. Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data. *Bioinformatics*, 24(19):2129–2136, 2008.

[85] C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC Bioinformatics*, 10(4), 2009.

[86] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[87] P. Du, R. Sudha, M. B. Prystowsky, and R. H. Angeletti. Data reduction of isotope-resolved lc-ms spectra. *Bioinformatics*, 23:1394–1400, 2007.

[88] J. W. H. Wong, G. Cagney, and H. M. Cartwright. Specalignprocessing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.

[89] D. Mantini, F. Petrucci, D. Pieragostino, P. DelBoccio, M. D. Nicola, C. D. Ilio, G. Federici, P. Sacchetta, S. Comani, and A. Urbani. Limpic: a computational method for the separation of protein malditof-ms signals from noise. *Bioinformatics*, 8(101), 2007.

[90] Y. Yasui, M. Pepe, M. L. Thompson, B. L. Adam, G. L. Wright, Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng. A data-analytic strategy for protein biomarker discovery:profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4:449–463, 2003.

[91] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. W. Lin, J. Z. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution lc-ms. *Bioinformatics*, 22:1902–1909, 2006.

[92] Proteinchip software 3.1 operation manual. Technical report, Ciphergen Biosystems Inc., 2002.

[93] X. Li, R. Gentleman, X. Lu, Q. Shi, J.D. Iglehart, Harris L., and A. Miron. *SELDI-TOF mass spectrometry protein data*, volume 1. Springer New York, 2005.

[94] C. A. Smith, E. J. Want, G. O. Maille, R. Abagyan, and G. Siuzdak. Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78:779–787, 2006.

[95] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M. C. Hung, and H. M. Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5:4107–4117, 2005.

[96] Y. V. Karpievitch, E. G. Hill, A. J. Smolka, J. S. Morris, K. R. Coombes, K. A. Baggerly, and J. S. Almeida. Prepms: Tof ms data graphical preprocessing tool. *Bioinformatics*, 23:264–265, 2007.

[97] M. Katajamaa, J. Miettinen, and M. Oresic. Mzmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22:634–636, 2006.

[98] P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):20592065, 2006.

[99] E. Lange, C. Gropl, K. Reinert, O. Kohlbacher, and A. Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. *Pacific Symposium on Biocomputing*, 11:243–254, 2006.

[100] S. Toppo, A. Roveri, M. P. Vitale, M. Zaccarin, E. Serain, E. Apostolidis, M. Gion, M. Maiorino, and F. Ursini. Mpa: A multiple peak alignment algorithm to perform multiple comparisons of liquid-phase proteomic profiles. *Proteomics*, 8:250–253, 2008.

[101] N.V. Nielsen, J. M. Cartensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805:17–35, 1998.

[102] D. Bylund, R. Danielsson, G. Malmquist, and K. Markides. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for parafac modelling of liquid chromatographymass spectrometry data. *Journal of Chromatography A*, 961:237–244, 2002.

[103] R. J. O. Torgrip, M. Aberg, B. Karlberg, and S. P. Jacobsson. Peak alignment using reduced set mapping. *Journal of Chemometrics*, 17:573–582, 2003.

[104] A. M. van Nederkassel, M. Daszykowski, P.H.C Eilers, and Y.V. Heyden. A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118:199–210, 2006.

[105] W. Yu, X. Li, J. Liu, B. Wu, K.R. Williams, and H. Zhao. Multiple peak alignment in sequential data analysis: A scale-space-based approach. *IEEE/ACM Transactions On Computational Biology and Bioinformatics*, 3(3), 2006.

[106] P. S. Addison. *The Illustrated Wavelet Transform Handbook*. Institute of Physics, Bristol and Philadelphia, 2002.

[107] J.A. Falkner, D.M. Veine, M. Kachman, A. Walker, J.R. Strahler, and P.C. Andrews. Validated maldi-tof/tof mass spectra for protein standards. *Journal of the American Society for Mass Spectrometry*, 18(5):850–855, 2007.

ABOUT THE AUTHOR

Chaitra Gopalappa received her Ph.D. in Industrial Engineering in 2010 and a Masters in Industrial Engineering in 2006 from University of South Florida. She received a B.E. in Industrial Engineering and Management from Visveswaraiah Technological University, India, in 2002. Her research areas of interest include disease progression and intervention, bioinformatics, and health care systems engineering. Her methodological areas of interest include applied stochastic modeling, computational probability, applied optimization, and wavelet based signal processing.

Chaitra was a lead investigator in an interdisciplinary challenge grant awarded as part of 'The 2008-2009 Graduate Student Challenge Grants: Building Research Partnerships Across Disciplines, University of South Florida'. She was a visiting research assistant at Purdue University (March 08-May 08), as part of Cancer Care Engineering project, Discovery Park, that was funded by Regenstrief Foundation, Indiana. Chaitra will join as a post-doctoral fellow in the Division of HIV/AIDS Prevention of Centers for Disease Control and Prevention at Atlanta, Georgia, starting August 2010.