11-15-2013

Towards the Prediction of Mutations in Genomic Sequences

Juan Carlos Martinez
jmart054@cs.fiu.edu

DOI: 10.25148/etd.FI13120420

Recommended Citation

Martinez, Juan Carlos, "Towards the Prediction of Mutations in Genomic Sequences" (2013). *FIU Electronic Theses and Dissertations*. 987.
https://digitalcommons.fiu.edu/etd/987

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

TOWARDS THE PREDICTION OF MUTATIONS IN GENOMIC SEQUENCES

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Juan Carlos Martinez

2013

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Juan Carlos Martinez, and entitled Towards the Prediction of Mutations in gnomic Sequences, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Jinpeng Wei

Malek Adjouadi

Naphtali Rishe

Shu-Ching Chen

Nelson Lopez-Jimenez

Sundaraja Sitharama Iyengar, Major Professor

Date of Defense: November 15, 2013

The proposal of Juan Carlos Martinez is approved.

Dean Amir Mirmiran.
College of Engineering and Computing

Dean Lakshmi N. Reddi.
University Graduate School

Florida International University, 2013

DEDICATION

Dedicated to mom.

ACKNOWLEDGMENTS

I would like to dedicate my work to my mother Jene Soto and my mentor Dr. Sitharama Iyengar for their strong and continuing support throughout my academic journey. I would also love to show my huge gratitude to all my FIU committee members: Dr. Chen, Dr. Adjouadi, Dr. Rishe and Dr. Wei for their valuable contribution and advice to my work. My immense gratitude also goes to my non-FIU committee member Dr. Lopez-Jimenez for assisting me with all the insights in the amazing and interesting areas of Medicine and Biology that I needed for my Bioinformatics work. Very special thanks are also given to Dr. Ning Xie who has been my unofficial co-advisor all this time and has provided me with extremely valuable advice.

ABSTRACT OF DISSERTATION

TOWARDS THE PREDICTION OF MUTATIONS IN GENOMIC SEQUENCES

by

Juan Carlos Martinez

Florida International University, 2013

Professor Sundaraja Sitharama Iyengar, Major Professor

Bio-systems are inherently complex information processing systems. Furthermore, physiological complexities of biological systems limit the formation of a hypothesis in terms of behavior and the ability to test hypothesis. More importantly the identification and classification of mutation in patients are centric topics in today's cancer research.

Many cancers have been traced to somatic mutations in different genes in the genome. Classification of cancer based on gene expression has provided insights into its complex landscape of multiple interactions between gene networks, as well as into possible treatment strategies. Next generation sequencing (NGS) technologies can provide genome-wide coverage at a single nucleotide resolution and at reasonable speed and cost. The unprecedented molecular characterization provided by NGS offers the potential for an individualized approach to treatment. These advances in cancer genomics have enabled scientists to interrogate cancer-specific genomic variants and compare them with the normal variants in the same patient. Analysis of this data provides a catalog of somatic variants, present in tumor genome but not in the normal tissue DNA. Determining the molecular signatures of genes mutated in cancer may help to predict the clinical outcome and carry out therapeutic modifications in treating the patients. However, predicting such signatures at the time of the tumor discovery is a major challenge. Several groups have reported lists of predictive genes and reported good predictive performance in terms of prognosis and potential for malignancy based on them. However, the gene lists differed widely and had only very few genes in common. The search for reliable molecular signatures has provided a fertile field for computational approaches. In this dissertation, we present a new computational framework to the problem of predicting the number of mutations on a chromosome for a certain patient, which is a fundamental problem in clinical and research fields. We begin this dissertation with the development of a framework system that is capable of utilizing published data from a longitudinal study of patients with acute myeloid leukemia (AML), who’s DNA from both normal as well as malignant tissues was subjected to NGS analysis at various points in time. By processing the sequencing data at the time of cancer diagnosis using the components of our framework which includes training mutations data, mutation extractions, normalization, subspace-based instance filtering, etc., we tested our framework by predicting the regions of the genome to be mutated at the time of relapse and, later, by comparing our results with the actual regions that showed mutations (discovered by sequencing their genomes at the time of relapse). We demonstrate that this coupling of the algorithm pipeline can drastically improve the predictive abilities of searching a reliable molecular signature. Arguably, the most important result of our research is its superior performance to other methods like Radial Basis Function Network, Sequential Minimal Optimization, and Gaussian Process. In the final part of this dissertation, we present a detailed significance, stability and statistical analysis of our model. A comparison performance of the results are presented. This work clearly lays a good foundation for future research for other types of cancer.

TABLE OF CONTENTS

CHAPTER PAGE

1. INTRODUCTION 1

2. LITERATURE REVIEW 4
2.1 Background 4
2.1.1 Bioinformatics..5
2.1.2 The DeoxyriboNucleic Acid (DNA)...5
2.1.3 Genes...7
2.1.4 Gene expression 7
2.1.5 DNA Mutations...10
2.1.6 Cancer Genomics Study..17
2.2 Related Work. 20
2.2.1 Fasta 20
2.2.2 Blast 21
2.2.3 RepeatMasker 21
2.2.4 Masker Aid..21
2.2.5 Tandem repeat finder 22
2.2.6 Reputer..23
2.2.7 The Vmatch large scale sequence analysis software 23
2.2.8 Amadeus 23
2.2.9 Allegro 24
2.2.10 Our work 24
2.2.11 Comparative chart...24
2.2.12 Other studies on cancer genomics...25
2.2.13 Clinical perspective...35

3. RESEARCH PROBLEM AND SPECIFIC AIMS 39
3.1 Motivation..39
3.2 Problem definition 40
3.3 Specific Aims..41
3.3.1 Aim # 1: To build a tool for statistical analysis on genomic data sets.....................41
3.3.2 Aim # 2: To utilize this data set and formulate a prediction model.........................41
3.3.3 Aim # 3: To facilitate the discovery of new functionalities in areas of the genome that had not been previously implicated in cancer. 41
3.4 Significance and impact of our work 43

4. DATA MINING AND STATISTICAL ANALYSIS 45
4.1 Background 45
4.2 Weka 45
4.2.1 Dataset..46
4.2.2 Classifier 47

4.2.3 Weka from Java – API 47
4.3 Statistical Analysis 47

5. MODELING CANCER RELATED GENOME SEQUENCES OF PATIENT DATA50
5.1 Introduction 50
5.2 Statistical analysis for our model 50
5.3 Mechanisms and Models 52

6. MATERIALS AND METHODS 56
6.1 Framework 57
6.2 Patient data 57
6.3 Variables to consider 62
6.4 Prediction model 63

7. RESULTS 67
7.1 Prediction of number of mutations per chromosome. 67
7.1.1 Significance of our experimental data 68
7.1.2 Analysis of performance measures to evaluate the success of our prediction 77
7.1.3 Significance, stability and statistical analysis of model 79
7.1.4 Comparison of performance of our model versus other algorithms 94
7.1.5 Concluding Remarks 98
7.2 Prediction of location of mutations. 100
7.3 Accuracy discussion 102

8. CONCLUSION 105
8.1 Research summary 105
8.2 Final statistical remarks 107
8.3 Foundation for further discoveries. 109
8.4 Application of our framework in the treatment of patients by clinical oncologists...110
8.5 Some potential enhancements. 112

SPECIAL RECOGNITIONS 113

APPENDICES 114

BIBLIOGRAPHY 143

VITA 153

LIST OF TABLES

TABLE PAGE

Table 2.1. Examples of genetic diseases caused by expanding tri-nucleotide repeats. 14

Table 2.2. Pathologic intronic insertions in humans. 16

Table 2.3. Comparative chart of our work versus other approaches. 25

Table 2.4. Comparison of Wavelet Analysis and Fourier transform (FT). 30

Table 6.1. Patient designation, gender and age. 59

Table 7.1. Number of mutations per chromosome in patient UPN 933124. 68

Table 7.2. Table for performance measures for our prediction model. 79

Table 7.3. Statistical analysis of errors for patient A. 81

Table 7.4. Statistical analysis of errors for patient B. 81

Table 7.5. Statistical analysis of errors for patient C. 82

Table 7.6. Statistical analysis of errors for patient D. 82

Table 7.7. Statistical analysis of errors for patient E. 83

Table 7.8. Statistical analysis of errors for patient F.

83

Table 7.9. Statistical analysis of errors for patient G. 84

Table 7.10. Statistical analysis of errors for patient H. 84

Table 7.11. Statistical analysis of all Patients. 85

Table 7.12. Coefficients for chromosome models for patient A. 86

Table 7.13. Coefficients for chromosome models for patient B. 87

Table 7.14. Coefficients for chromosome models for patient C. 88

Table 7.15. Coefficients for chromosome models for patient D. 89

Table 7.16. Coefficients for chromosome models for patient E. 90

Table 7.17. Coefficients for chromosome models for patient F. 91

Table 7.18. Coefficients for chromosome models for patient G. 92

Table 7.19. Coefficients for chromosome models for patient H. 93

Table 7.20. Comparison table for correlation coefficients for our model versus others. 97

Table 7.21. Location of mutations, both observed and predicted for patient UPN 933124. 102

Table 7.22. Summary of results for patient UPN 933124. 103

LIST OF FIGURES

FIGURE PAGE

Figure 2.1. Graphical representation of a DNA double helix. 6

Figure 2.2. Mechanism for gene expression. 8

Figure 2.3. Transcription process carried out by DNA polymerase. Blue bases: RNA product. 10

Figure 2.4. Transcription from DNA to RNA example and translation to protein. 10

Figure 2.5. DNA Mutation. 11

Figure 2.6. Types of DNA mutations. 12

Figure 2.7. Tandem repeat example. 13

Figure 2.8. Transposable element example. 15

Figure 2.9. Sample procedure of applying wavelet transform in cancer genome analysis. 28

Figure 2.10. The four types of mutations. 31

Figure 6.1. Dependency flow graph for our prediction framework. 58

Figure 6.2. Sample data files for patient UPN 933124. 62

Figure 6.3. Overview of geneMap.txt. 62

Figure 6.4. Weka input file. 64

Figure 6.5 Algorithm for finding mutations locations. 66

Figure 7.1 Histogram of absolute error for patient A. 69

Figure 7.2 Histogram of absolute error for patient B. 69

Figure 7.3 Histogram of absolute error for patient C. 70

Figure 7.4 Histogram of absolute error for patient D. 70

Figure 7.5 Histogram of absolute error for patient E. 71

Figure 7.6 Histogram of absolute error for patient F. 71

Figure 7.7 Histogram of absolute error for patient G. 72

Figure 7.8 Histogram of absolute error for patient H. 72

Figure 7.9 Histogram of relative error for patient A. 73

Figure 7.10 Histogram of relative error for patient B. 73

Figure 7.11 Histogram of relative error for patient C. 74

Figure 7.12 Histogram of relative error for patient D. 74

Figure 7.13 Histogram of relative error for patient E. 75

Figure 7.14 Histogram of relative error for patient F. 75

Figure 7.15 Histogram of relative error for patient G. 76

Figure 7.16 Histogram of relative error for patient H. 76

Figure 7.17 Comparison of performance of our model versus others. 98

Figure 7.18 Potential estimation error graph as more data is obtained. 100

ACRONYMS AND ABBREVIATIONS

DNA	Deoxyribonucleic Acid
A,C,T,G	Adenine (A), cytosine (C), guanine (G), and thymine (T)
RNA	Ribonucleic Acid
mRNA	Messenger RNA
TR	Tandem Repeat
TE	Transposable Element
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
OO	Object Oriented
ML	Machine Learning
IDE	Integrated Development Environment
JDK	Java Development Kit

CHAPTER 1

INTRODUCTION

In academia, as well as in industry, there have been a number of efforts in developing computer algorithms to aid the research in molecular biology. Due to the fact that both DNA and proteins can be directly represented as a sequence of characters (either bases or aminoacids, respectively), the bioinformatics field has come up with a large number of computer solutions with a strong emphasis on string algorithms. A significant number of studies has shown that cancer development brings about changes in DNA molecule at a nucleotide level (mutations) within the genome of an individual.

Many cancers have been traced to somatic mutations in different genes in the genome. [1]. The classification of cancer based on gene expression has provided much insight into the complex landscape of multiple interactions between gene networks, as well as into possible treatment strategies. Advances in cancer genomics have been accelerated with the emergence of high-throughput sequencing strategies that enable scientists to interrogate cancer-specific genomic variants and compare them with the normal variants in the same patient [2][3]. Next generation sequencing (NGS) technologies can provide genome-wide coverage at a single nucleotide resolution and at reasonable speed and cost. [4]. The unprecedented molecular characterization provided by NGS offers the potential for an individualized approach to treatment.

One of the contributions of such strategies includes the definition of relatively characteristic gene expression profiles, or molecular signatures that may have prognostic implications for targeted therapies. In cancer patients, the objective of NGS is to obtain and compare information about cancer and normal tissue DNAs. Analyses of the data provide a catalog of somatic variants present in tumor genome but not in the normal tissue DNA. The goal of the analysis is to reveal a putative drug target in the examined cancer which facilitates the selection of therapy, and improves the personalized risk assessment. Moreover, determining the molecular signatures of genes mutated in cancer may help to predict the clinical outcomes [5].

Predicting molecular signatures based on the initial events in cancer development may help researchers, as well as clinicians, to carry out therapeutic modifications in treating the patients. However, predicting such signatures at the time of the tumor diagnosis is a major challenge. Several groups have reported lists of predictive genes and reported good predictive performance in terms of prognosis and potential for malignancy based on them. However, the gene lists differed widely and had only very few genes in common. Furthermore, those types of predictions were based on computational approaches not involving NGS, such as microarray analysis, qPCR, and others [6][7][8] in various types of cancer such as colorectal [9], lung [10], prostate[11], breast cancer[12] and others. Additionally, the disadvantages of current predictive models in cancer are that they are focused on evaluating mutations that anticipate the risk of progression and its clinical impact on the length of patient’s survival. They are not intended to predict mutational events at molecular level, rather to detect and classify existing mutations, and utilize that information to make predictions in the behavior of the cancer disease.

In our work, by having the full sequence of the genome from both normal and cancer tissue at various points in time, we are able to extract valuable information to validate the predictions generated from the genome. Thus, we propose to search for longitudinal studies of cancer patients with such identified mutated genes by the use of NGS at different points in time. The mutation profile identified earlier in time will serve as the basis for the prediction later in time. Furthermore, we will compare the observed mutated locations at the time of relapse with the results of our own predictions.

Our goal in this study was to test a computational framework utilizing published data from a longitudinal study [13] of patients with acute myeloid leukemia (AML) whose DNA from both normal as well as cancer tissues were subjected to next generation sequencing analysis at various points in time. First of all, we processed the longitudinal sequencing data from cancer tissue. Secondly, we tested our framework by predicting the regions of the genome to mutate at the time of relapse. Finally, we compared our results with the observed mutated regions, identified by sequencing their genomes at the time of relapse. The accuracy of the predicted number of mutations ranged from 75 to 84%; the accuracy of the locations of mutations ranged from 69% to 88% in the patients analyzed. Our predictions agree well with the reported data. Although our approach is purely computational and does not provide insights into the putative molecular mechanisms by which the mutations occur, we are confident that it may assist researchers in determining the importance of a particular mutation in cancer progression, thus providing another tool to select candidates to target in drug development for cancer treatment.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

Over time our knowledge of genomes has developed at the same rate as their sequencing. Furthermore, sequencing capacity has grown significantly recently. The characterization of the genome starts from its sequencing or reading step and assembling. However, it does not end by the time its sequence is obtained. Thus, at this point, we are just at the beginning of a much more complex study. We are constantly looking for better coverage of the sequence and better resolution and, as a result, the assembly of the human genome is a multistep always-evolving process. The Genome Reference Consortium frequently produces a new version of the human reference genome for scientists to have a unique set of gene coordinates when comparing their results. The latest release from GRC is the 37th [14]. With every new release, a large set of analysis is required to understand the works of the genome, its functionalities, and how we can detect any anomalies in it as well as how we can predict future ones.

Nevertheless, before going into further details of how to understand the scope of our research of the genome, it is necessary to have a minimum background on how it works [111]. In the next pages, we will provide some basic concepts in the Biology field that will ease the understanding of the rest of the work. This dissertation was written for readers who do not necessarily have the biology background, and, as result, a minimal background is provided.

2.1.1 Bioinformatics

Bioinformatics is a compound word that consists of two root words. The first word “bio” indicates we are dealing with problems that originate from biology. The second word “informatics” indicates that the problems are to be solved by the discipline that works on the management of information and the design of computational algorithms, namely, computer science.

2.1.2 The DeoxyriboNucleic Acid (DNA)

The DeoxyriboNucleic Acid (DNA) [16][17] plays the role of the information archive for all organisms. DNA molecules present a double helix structure containing four-letter bases which stand for the 4 nucleotides. These four nucleotides are adenine (A), cytosine (C), guanine (G), and thymine (T). Nucleotides belong to two types: purines and pyrimidnes. Adenine and guanine fall into the first category while cytosine and thymine into the second one.

DNA is not normally present as a single long molecule but as an associated pair of molecules. DNA strands present the shape of a double helix as can be seen in Figure 2.1. Alternating phosphate and sugar elements compose the backbone of a DNA strand. The direction of the nucleotides in one strand is always opposite to the direction from the other. As a result, we consider DNA strands as anti-parallel. These DNA strands are asymmetric and known as 5' and 3' and follow a direction from one to the other. All bases on one strand form a bond with specifically one type of base on the other strand. We call this behavior complementary base pairing. Thus, purines form hydrogen complement with pyrimidines: A with T (A=T or T=A), and C with G (C=G or G=C). When two nucleotides are bound together across the DNA strands form base pair (bp).

The DNA presents a double stranded helix with complimentary information on each strand. This characteristic is of the DNA is vital in the functionality of the DNA for all living beings [18]. DNA has mechanisms for self-replication and for translation of genes into proteins. Furthermore, replications are necessary for the stability of the inheritance. However, some imperfect replications are also necessary for evolution. Genetic information is implemented by the synthesis of mRNA (messenger RNA) into proteins. Proteins are molecules in charge of the structure and activities of living beings. Antibodies, produced by our own organisms, to fight disease, along with some enzymes and muscle tissue belong to the protein category. Proteins can be composed of 200-400 amino acids long, requiring 600-1200 DNA bases. Surprisingly, only a small percentage of the total sequence genome encodes for proteins. There are vast areas of the DNA sequence which account for control mechanisms, and others for which no functionality has been discovered yet, and, therefore, considered as “junk". However, it might not necessarily be junk as there have been discoveries of transposable elements that have a regulatory function in this area, leaving us with more to be discovered in these areas [19][20].

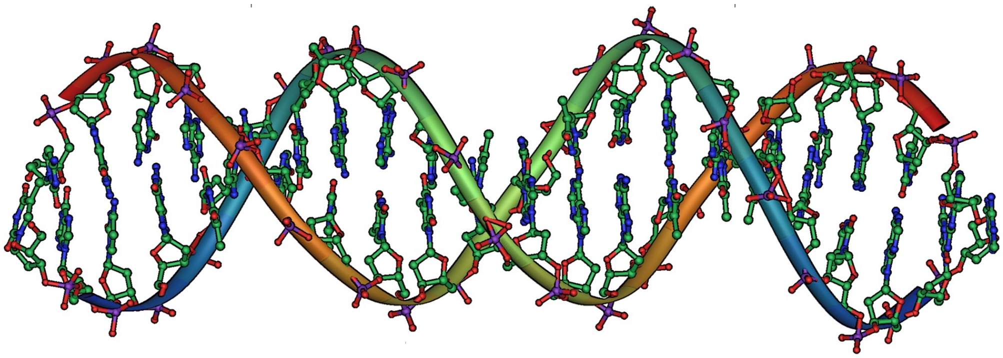


Figure 2.1. Graphical representation of a DNA double helix. [111]

2.1.3 Genes

Our main unit of heredity for all organisms is called a gene. Genes keep the necessary information to build and transfer genetic traits to descendants as well as to maintain cells. Simply speaking, a gene can be seen as a segment of nucleic acid that, when considered as a whole, specifies a trait. Genes can be mapped to multiple biological traits such as such as eye or hair color.

Genes have coding sequences or exons and non-coding sequences or introns. If a gene is active, both exons and introns are copied though transcription, producing as an output an RNA copy of the information from the gene. The output allows for protein synthesis from genetic code. The molecules that are produced from gene expression are called gene products and are the ones that make it possible for organisms to develop proper functionality.. Gerstein et al. [21] provide a precise definition of a gene: “A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products". Besides the regions that code for protein, genes also have regulatory regions as can be seen in Figure 2.2.

Regulatory regions include enhancers that are responsible for compensating for a weak promoter [22]. Small RNAs consist of short sequences roughly between 18 and 25 bases. They act as regulators of stability or availability for translation of RNAs.

2.1.4 Gene expression

When a gene is expressed, the products obtained from it are often proteins. Gene expression is a widely used process for all organisms ranging from the most basic ones to the most complex and developed ones.

Gene regulation allows cell function and control over structure and, as a result, is the base for cellular differentiation and adaptability. Gene expression is the main level where the genotype gives rise to the phenotype. Thus, genetic code can be interpreted.

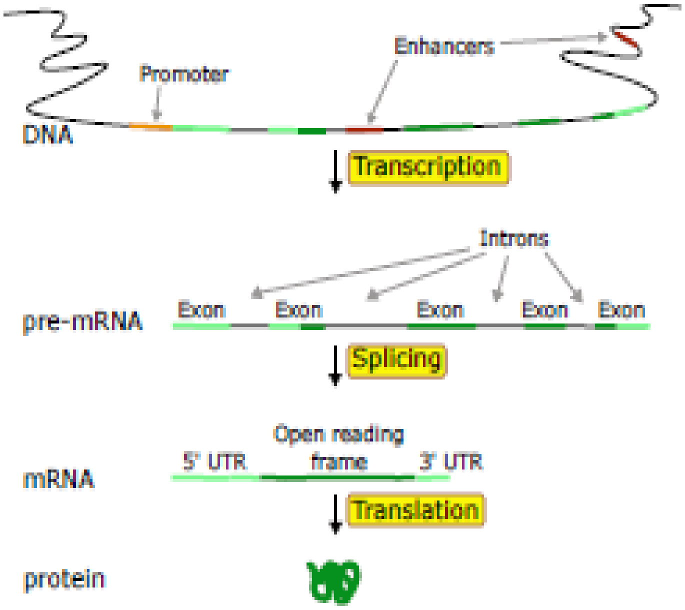


Figure 2.2. Mechanism for gene expression. [15]

Gene expression is divided into 5 steps:

i. Transcription

The gene acts as a blueprint for the production of RNA. We call transcription to the RNA's production copies of the DNA. It is done by RNA polymerase by adding one RNA nucleotide at a time to a growing RNA strand as can be observed in Figure 2.3. RNA complementarity to DNA is present here. For instance, a T in the DNA will produce an A in the RNA. Nevertheless, in RNA the Uracil is obtained instead of Thymine in place of an Adenine in the DNA strand. For instance, the mRNA complement of a DNA strand piece "CAT" would be transcribed as "GUA". Figure 2.4 shows more on this.

ii. RNA processing

Even though, it is not the only, splicing to remove introns is the most common RNA processing.

iii. RNA export

In eukaryotes, mature RNA has to be exported out of the nucleus to the cytoplasm which is where proteins actually work [15].

iv. Translation

Every RNA triplet has a corresponding binding site for amino acids which are chained together by the ribosome. The ribosome produces a structure-less protein out of the by taking the amino acid from each transfer RNA. The second step of the Figure 2.4 illustrates this process.

v. Folding

Here, the recently formed unfolded polypeptide will proceed to fold into the functional 3-dimensional structure.

After this process, the protein working into the cell or transported outside of it as it occurs in digestive enzymes.

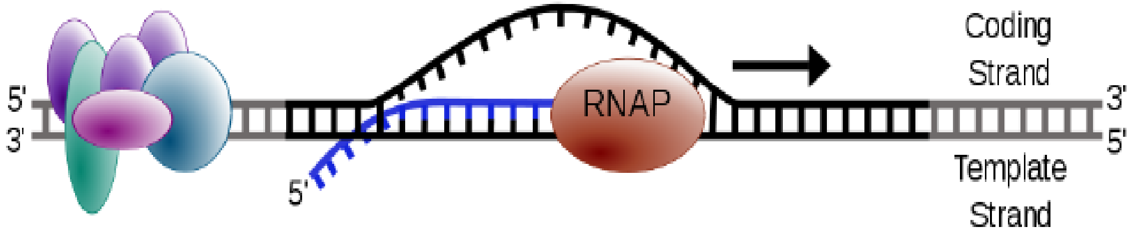


Figure 2.3. Transcription process carried out by DNA polymerase. Blue bases: RNA product.[15]

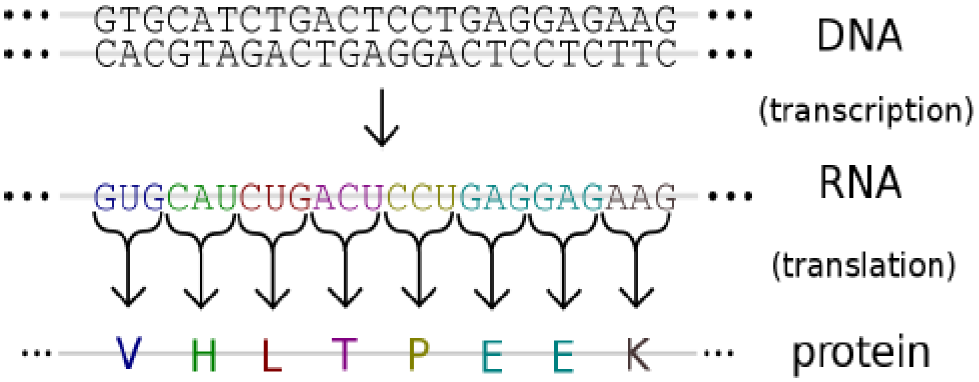

... GTGCATCTGACTCCTGAGGAGAAG ...
... CACGTAGACTGAGGACTCCTCTTC ...
DNA
(transcription)
... GUGCAUCUGACUCCUGAGGAGAAG ...
RNA
(translation)
... V H L T P E E K ...
protein

Figure 2.4. Transcription from DNA to RNA example and translation to protein. [15]

2.1.5 DNA Mutations

We define DNA mutations as the sequences that appear within a genome that do not comply with what is "expected" in that location. The human DNA is very similar from individual to individual with minor differences in the nucleotides that appear in the genome. Nevertheless, when these differences are not minor and become major differences, different implications may occur in the individual affecting his health and/or capabilities.

Germinal mutations are the ones and individual is born with. Somatic mutations happen during one's lifetime and are studied by looking at tumor cells and normal cells from the same patient at different times.

Figure 2.5 and Figure 2.6 show graphically how a mutation occurs within the DNA. This mutation can be of the form of a replacement, deletion or insertion of bases within the genome.

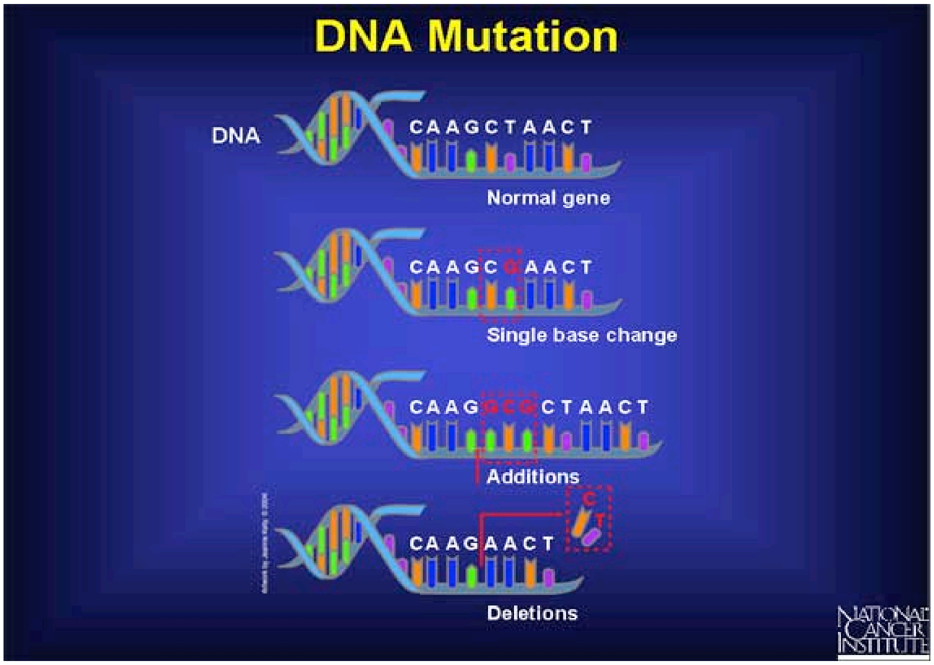


Figure 2.5. DNA Mutation. [23]

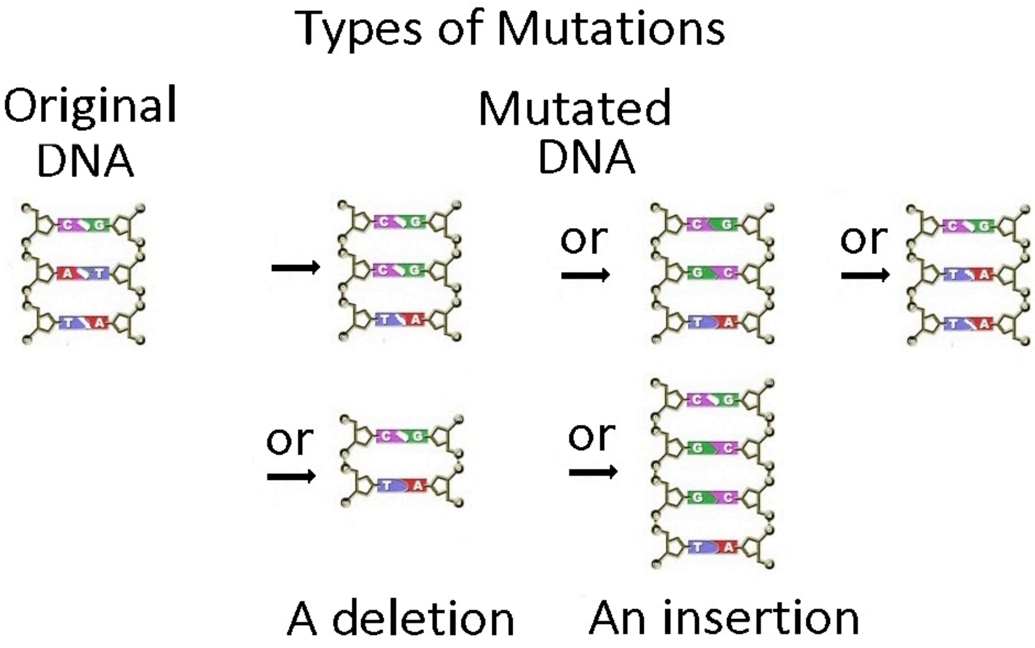


Figure 2.6. Types of DNA mutations. [24]

Many of these mutations are present since birth and many others are acquired during the development of a disease such as cancer.

These mutations have different forms and appear in different locations within the DNA. These mutations act as repeat sequences following a specific pattern or not. The two major types of repeat sequences are tandem repeats (TRs) including micro-satellites, mini-satellites, and satellites, and transposable elements (TEs), which are dispersed within genomes with a moderate to high degree of repetitiveness.

i. **Tandem Repeats (TRs)**

A common feature for both eukaryotic and prokaryotic genomes [25][26][27] is the repetition of both large and short sequences (2 to several hundred nucleotides long), known as tandem repeats. Many experts suggest that as much as 50% of the human genome is composed of repetitions [28].

An example of a tandem repeat is illustrated in Figure 2.7:

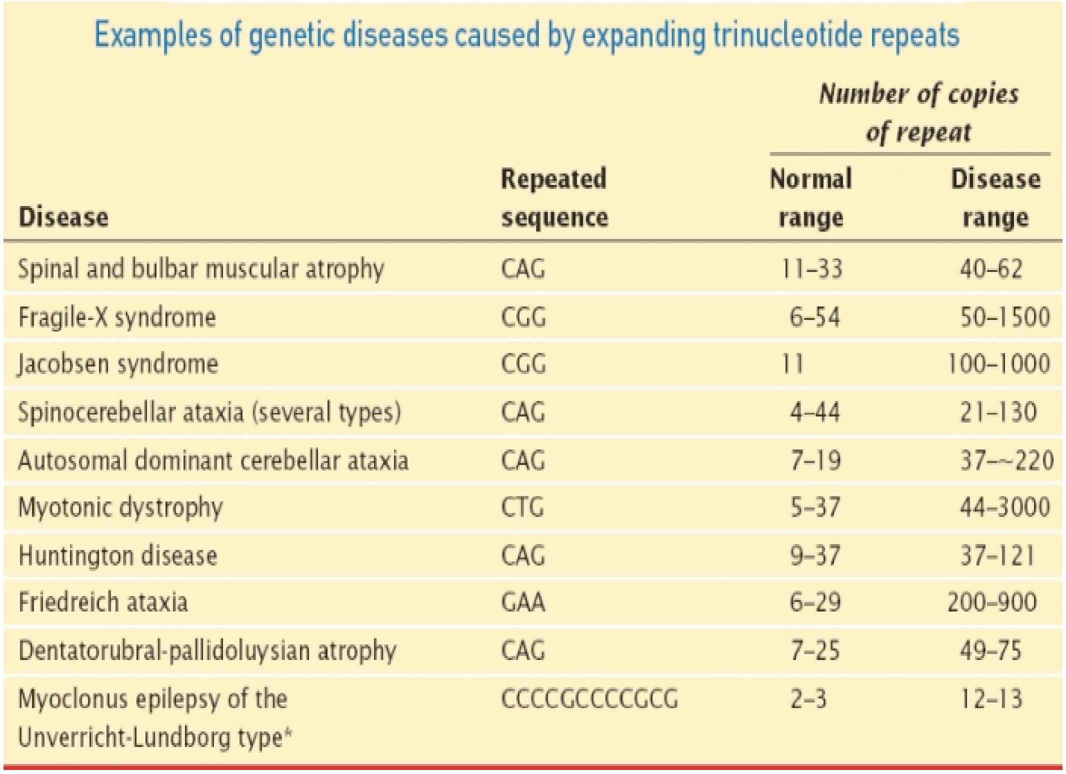
ACTACTACTACT
Tandem repeat of 3 bases and length 4

Figure 2.7. Tandem repeat example.

The biological role of repeated sequences has been researched by a number of scientists who have determined that they are linked to evolutionary mechanisms in prokaryotic organisms [29]. Tandem repeats [30][31] are common in eukariotyc organisms, as well. It has been reported that tandem repeats are responsible for over 30 inherited diseases in humans. Tri-nucleotide expansions have been associated with fragile X syndrome, myotonic dystrophy, Huntington's disease, various spino-cerebellar ataxias, Friederich's ataxia and others. Tandem repeats can exist in coding as well as in non-coding DNA. Expansion found in coding regions can affect the function of the genes involved by disrupting the normal synthesis of the proteins. When these repeats appear inside non-coding regions, it has been shown to affect gene expression via alterations in the transcription levels of the gene when the affected area correspond to either a promoter region, or a region containing factors that affect protein translation. Table 2.1 clearly illustrates genetic diseases related to Tandem Repeats.

Unfortunately, individuals suffering from inherited diseases do not have many choices in terms of treatment. However, in some cases an early diagnosis may help to ameliorate the devastating effects of the disease. Current tools available in clinics, as well as in bioinformatics allow the medical community to carry out diagnostics as early as in the pre-implantation embryo, to determine whether the embryo carry the expansions associated with a given disease or not.

Expansion of DNA repeats are not limited to inherited disorders only. Depending on the length of the repeat and the area it spans, tandem repeats are also known as "satellite, mini-satellite, and microsatellite" DNA. Blanes and Diaz-Cano [32], present a complementary analysis of microsatellite tumor profile in which they observe that changes in the number of repeats occur depending on the progression in the cancer cell, compared with the normal tissue. Multiple studies have shown micro- and mini-satellite instability during cancer progression in a wide variety of tumors, both familiar and sporadic. Breast and ovarian cancers show many more tandem duplications than other tumor types [33].



Examples of genetic diseases caused by expanding trinucleotide repeats

		Number of copies of repeat	
Disease	**Repeated sequence**	**Normal range**	**Disease range**
Spinal and bulbar muscular atrophy	CAG	11–33	40–62
Fragile-X syndrome	CGG	6–54	50–1500
Jacobsen syndrome	CGG	11	100–1000
Spinocerebellar ataxia (several types)	CAG	4–44	21–130
Autosomal dominant cerebellar ataxia	CAG	7–19	37–~220
Myotonic dystrophy	CTG	5–37	44–3000
Huntington disease	CAG	9–37	37–121
Friedreich ataxia	GAA	6–29	200–900
Dentatorubral-pallidoluysian atrophy	CAG	7–25	49–75
Myoclonus epilepsy of the Unverricht-Lundborg type*	CCCCGCCCCGCG	2–3	12–13

Table 2.1. Examples of genetic diseases caused by expanding tri-nucleotide repeats. [34]

ii. **Transposable Elements (TEs)**

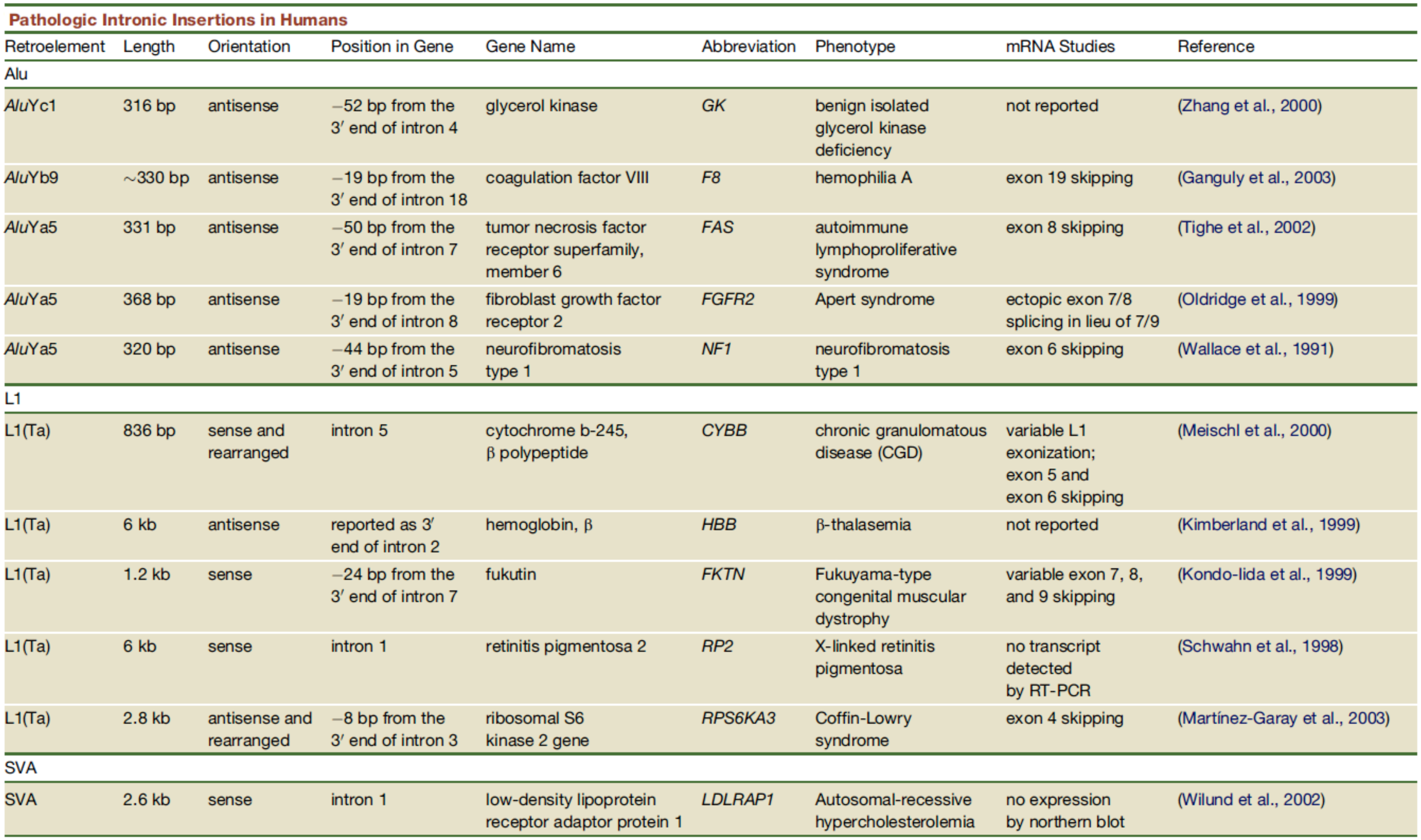
Transposable elements are mobile within the genome, and as such, they can be inserted anywhere, whether within the coding region of a gene, or within intronic regions. As a result of their insertion the expression of a gene is affected by changing the protein coded by the gene where insertion occurred, or by disrupting the natural context of the gene if the insertion occurred within an intronic region. The term "normal" context refers to the regions that surround the gene, meaning upstream and downstream regions that contain controlling elements, such as promoters, enhancers and inhibitors of gene expression. If the insertion occurred within any of those controlling elements, the gene expression would change. An example of a transposable element is illustrated in Figure 2.8.

...ACTCCTTAAACTCGGTACTGGGGC...
Transposable element: ACT

Figure 2.8. Transposable element example.

We have to keep in mind that whether an insertion or modification of sub-sequence within a genome is considered a mutation event, whether one base or 300 bases, or more.

Table 2.1 shows the result of repeats affecting coding regions. Additionally, there are also reports of inherited and non-inherited diseases, including various types of cancer, associated with repeats in non-coding regions. Table 2.2 illustrates such scenarios as well and shows reports regarding transposable repeats, or transposable elements repeats related to genome mutations.



Pathologic Intronic Insertions in Humans

Retroelement	Length	Orientation	Position in Gene	Gene Name	Abbreviation	Phenotype	mRNA Studies	Reference
Alu								
*Alu*Yc1	316 bp	antisense	−52 bp from the 3′ end of intron 4	glycerol kinase	*GK*	benign isolated glycerol kinase deficiency	not reported	(Zhang et al., 2000)
*Alu*Yb9	~330 bp	antisense	−19 bp from the 3′ end of intron 18	coagulation factor VIII	*F8*	hemophilia A	exon 19 skipping	(Ganguly et al., 2003)
*Alu*Ya5	331 bp	antisense	−50 bp from the 3′ end of intron 7	tumor necrosis factor receptor superfamily, member 6	*FAS*	autoimmune lymphoproliferative syndrome	exon 8 skipping	(Tighe et al., 2002)
*Alu*Ya5	368 bp	antisense	−19 bp from the 3′ end of intron 8	fibroblast growth factor receptor 2	*FGFR2*	Apert syndrome	ectopic exon 7/8 splicing in lieu of 7/9	(Oldridge et al., 1999)
*Alu*Ya5	320 bp	antisense	−44 bp from the 3′ end of intron 5	neurofibromatosis type 1	*NF1*	neurofibromatosis type 1	exon 6 skipping	(Wallace et al., 1991)
L1								
L1(Ta)	836 bp	sense and rearranged	intron 5	cytochrome b-245, β polypeptide	*CYBB*	chronic granulomatous disease (CGD)	variable L1 exonization; exon 5 and exon 6 skipping	(Meischl et al., 2000)
L1(Ta)	6 kb	antisense	reported as 3′ end of intron 2	hemoglobin, β	*HBB*	β-thalasemia	not reported	(Kimberland et al., 1999)
L1(Ta)	1.2 kb	sense	−24 bp from the 3′ end of intron 7	fukutin	*FKTN*	Fukuyama-type congenital muscular dystrophy	variable exon 7, 8, and 9 skipping	(Kondo-lida et al., 1999)
L1(Ta)	6 kb	sense	intron 1	retinitis pigmentosa 2	*RP2*	X-linked retinitis pigmentosa	no transcript detected by RT-PCR	(Schwahn et al., 1998)
L1(Ta)	2.8 kb	antisense and rearranged	−8 bp from the 3′ end of intron 3	ribosomal S6 kinase 2 gene	*RPS6KA3*	Coffin-Lowry syndrome	exon 4 skipping	(Martínez-Garay et al., 2003)
SVA								
SVA	2.6 kb	sense	intron 1	low-density lipoprotein receptor adaptor protein 1	*LDLRAP1*	Autosomal-recessive hypercholesterolemia	no expression by northern blot	(Wilund et al., 2002)

Table 2.2. Pathologic intronic insertions in humans. [35]

2.1.6 Cancer Genomics Study

The understanding of the genetic basis of disease in general, and of cancer in particular, has progressed dramatically in the past few decades. Several factors have been involved in that progress: some, purely of biological nature, and some at the intersection between biology and computational sciences. The science of genetics began in the 1860's when Gregor Mendel studied inheritance in pea plants (Pisum sativum) [36]. In the 1940's Oswald Avery, Colin MacLeod, and Maclyn McCarty showed DNA was the genetic material [37][38]. In 1953 James Watson and Francis Crick proposed the double-helix model for the structure of DNA [39]. The Human Genome Project (HGP) began in 1990 and was completed in 2003 by the International Human Genome Sequencing Consortium [40][41][42]. The reference genome produced by the HGP came from a single anonymous male donor from Buffalo, New York [43]. Since then, a number of international projects have started to elucidate the genomic sequence of various individuals. For instance, the HapMap Project which used DNA samples from 270 individuals, the 1000 Genomes Project, the Cancer Genome Atlas, the Cancer Genome Anatomy Project, and the Cancer Genome Characterization Initiative [36].

The Cancer Genome Atlas (TCGA) is a program supported by the National Institutes of Health (NIH). It will help to understand what turns a normal cell into a cancer cell. Its main approach as stated in its web site (cancergenome.nih.gov) is to compare DNA from normal and cancer tissue to find what are the differences. By utilizing such approach scientists working within the TCGA Project have learned that there are certain areas of the genome commonly affected in several types of cancers. Often these changes affect genes that control pathways in cells that cause cells to divide and survive when they normally would die. Another finding is that specific changes –also called signatures-allow scientists to tell one type of cancer from another. These signatures help doctors identify specific "types of cancer" which may respond differently to various treatments or have different prognosis. In order to better understand these concept I would like to define some of the specialized terminology (see cancergenome.nih.gov)

The genome is given by all the DNA that exists within a cell. For the majority of cells, the genome is wrapped into two sets of chromosomes inherited from both parents. These chromosomes consist of six billion DNA letters. The genetic alphabet consists of 4 letters: A, C, G and T. The area of Genomics consists of studying how these letters appear in a sequence and how every string of letters passes the information to the proteins, the real building blocks in the organism. DNA stands for deoxyribonucleic acid and instructions encoded in the string of DNA are passed to the proteins.

For cancer cells, changes in the bases (A,C,T, and G) can cause a modify the meaning of a genomic "word" or "sentence". These changes in the bases can make the cell to produce a protein that does not allow the cell to work as expected. These proteins could make the cell grow faster than normal and affect neighboring cells. The accumulation of these changes, or mutations, within a cell leads to the development of cancer. Mutations are classified as germ line and somatic. Germ line mutations affect germinal cells (oocytes or spermatocytes) and are transmitted to the offspring. Somatic mutations, on the contrary, affect somatic cells and are not transmitted to the offspring. The study of the genome in a cancer cell is known as cancer genomics. The field of cancer genomics has had a profound impact in the understanding of cancer progression at a molecular level. The field itself has been dramatically impacted by the application of the so called next-generation sequencing, or NGS, technologies, which has accelerated the pace of discovery, while reducing the cost. NGS technologies greatly depend on computer power to acquire, process, and analyze the data coming from sequencing the genomes of healthy, as well as cancer cells [44]. The importance of studies made at the intersection of cancer biology and computer sciences can be never be understated. As data generated from multiple sequencing projects grows, the computer capabilities should keep up with the challenges of analyzing the terabyte amounts of data. The strongest contribution of NGS on cancer genomics has been the ability to re-sequence, analyze and compare the matched tumor and normal genome of a single patient [44].

Finally, the sequence of the human genome and the discoveries brought about by it, or related to it, have allowed the faster integrations of genomics into the medical practice [45]. As cancer is defined as a genetic disorder due to the accumulation of mutations, tumor genome sequencing has been used to guide treatment in oncology, as well as to develop new therapeutic anti-cancer targets [45]. It is known, that cancer that look identical through the microscope may have very different underlying genetic changes and may respond differently to specific drugs. By identifying the mutational profile, or signatures, specific for a certain type of cancer, it would be possible to classify individual groups of patients who may have better responses to an specific treatment, and rule out patients that do not belong to such group, avoiding the secondary effects of an unnecessary therapy.

2.2 Related Work.

We will analyze and compare our research from multiple perspectives since it involves work in different areas. We will start reviewing the different tools used for analyzing and comparing genomes and specific sequences and later we will do an extensive comparison of our research versus others which aimed at studying cancer and the genome relationships.

Numerous methods have been developed for analyzing the repeat structure of genomic sequences [46][47][48][49], most of which scan for a specific type of repeat such as short sequence repeats, palindromic repeats [50], tandem repeats [51][52][53][54], or highly periodic short repeat elements [55][56]. In most cases, such methods are unable to detect repeats that do not match a predefined pattern and intra or inter-genomic analyses are usually very difficult. Other methods [57][58], however, such as the use of Fourier transforms for repeat identification, do not search for a specific repeat pattern but rather try to locate occurrences of highly correlated periodic repeats. Nevertheless, this approach typically only identifies very strong genome-wide correlations such as those due to the triplet nature of the genetic code.

Among the tools we have in Bioinformatics, we identify the ones that look for sequences within another sequence or genome and the ones that look within established patterns (motifs).

2.2.1 Fasta [59]

First fast sequence searching algorithm for comparing a query sequence against a database. It is a tool that works for comparative analysis of DNA against DNA searches.

One of its limitations is that it can only find one gapped region of similarity and performs relatively slow. However, on the positive side, it does not require specially prepared, preformatted databases.

2.2.2 Blast [60][61]

The Basic Local Alignment Search Tool (BLAST) was developed at the National Center for Biotechnology Information (NCBI) [62]. It is not normally used for DNA against DNA searches without translation because of optimization issues. This tool pre-filters repeat and "low complexity" sequence regions and it is capable of finding more than one region of gapped similarity. BLAST's implantation consists of a very fast heuristic and parallel algorithm with the restriction of precompiled, specially formatted databases.

2.2.3 RepeatMasker [63][64]

This program searches DNA sequences for repeats that present gaps between them. Sequence comparisons in RepeatMasker are handled by multiple search engines such as cross_match [65], ABBlast/WUBlast [66][67], RMBlast [68] and Decypher [69].

2.2.4 Masker Aid [70]

When RepeatMasker was created, scientists started to find repeat sequences for large genome sequences; however, they also found that it took fairly huge amount of time to the tune of hours to get the process completed on the fastest machines with high end processors and memory. MaskerAid was henceforth proposed to take the added benefit of WU-BLAST along with cross_match programs. As a result, this patch was applied to the

RepeatMasker program in a 4 step process. The new proposed algorithm parses the outputs of BLAST programs and finds next the best alignments to find the repeats quickly. MaskerAid was run on all the existing conditions which cross_match satisfied and showed significant reduction in time taken across modes. Also the miss rate of repeat sequences was much reduced. The tool was also seen as lesser sensitive to user-defined/external parameters than RepeatMasker.

2.2.5 Tandem repeat finder [71]

Scientific studies show that tandem repeats have something special about them and they can be connected to important biological consequences observed in living beings. Hence, it is very important to find all Tandem repeats in DNA sequence quickly and efficiently regardless of the size of the genome sequence. Before the Tandem Repeat Finder appeared, the existing algorithms worked fairly well even though they had various disparities. Moreover, many had unacceptable worst case running times for large sequences. Few considered large and small sequences alike and to predict a pattern took unreasonable time for smaller sequences. Others were bounded by the inclusion of approximate repeats and due to the concept of substitution by a fixed number, it made little sense to apply the same algorithm to small and large sequence at the same time.

Tandem Repeat Finder takes all these factors into consideration and comes up with a solution which comprises of the following features and more: k-tuple matching and no size limit of repeats. This solution also finds hidden tandem repeats using a probabilistic model and, finally, aligns these repeats to obtain a consensus pattern. To end the process, these repeats are clustered together for further analysis.

2.2.6 Reputer [72]

Exact repeat sequences are the ones that appear at least twice or more in the entire genome nucleotide sequence. Next, maximal repeats are basically exact repeats extended in both directions without loss of mismatch. These can be classified as Forward Repeats or Reverse Complement Repeats. A Reverse Complement of sequence AACCTTGG is CCAAGGTT.

Repeats are stored in the suffix tree for finding the pattern efficiently. Storage wise this algorithm is far superior to existing ones. In order to enumerate all repeats existing in the parsed sequence, a Depth-First search of the entire tree is required. This enlists one by one all the repeats with their coordinates and lengths stored in the end nodes. The algorithm is highly optimal from space and time considerations and can be further optimized for space constraints.

2.2.7 The Vmatch large scale sequence analysis software [73]

This work uses suffix arrays [74][75] to pre-process genomes, creating indexes which allow quick searching of probes. This brings the limitation given by the size of RAM memory present in the executing environment. At this time Vmatch in 32 bit computers has a limitation of genomes of 400 million bases.

2.2.8 Amadeus [76]

Amadeus is a software platform for genome-scale identification of known and novel motifs (recurring patterns) in DNA sequences, applicable to a wide range of motif discovery tasks. Amadeus can be used to identify binding site motifs from Protein

Binding Microarray data. A dataset includes its measured binding intensities for each probe sequence covering together all possible 10-mers. In a number of competitions Amadeus ranked first (tied with one other group) in identifying the binding site motifs of 66 TF datasets. Running time is a few seconds per dataset.

2.2.9 Allegro [77]

Allegro is a software tool for simultaneous discovery of cis-regulatory motifs and their associated expression profiles. Its inputs are DNA and genome-wide expression profiles. Its output is the set of motifs [78] found, and for each motif the set of genes it regulates (its transcriptional module). Allegro is highly efficient and can analyze expression profiles of thousands of genes, measured across dozens of experimental conditions, along with all regulatory sequences in the genome. Allegro has a user-friendly graphical user interface.

2.2.10 Our work

Our work works with mutations in general not necessarily following an exact or approximate match to any known signature. All mutations that we want to predict are made at the molecular level, and cover not only one specific gene nor chromosome but the whole human genome.

2.2.11 Comparative chart

Finally a comparative chart of our approach is presented together with other methods mentioned in related work in table 2.3. We can observe that our framework works on any type of sequences just like other applications; however, our work is oriented towards the prediction of new mutations over time in the human genome.

Table 2.3. Comparative chart of our work versus other approaches.

Program	**Alignment**	**Mismatch**	**Exact Match**	**Search against**	**Prediction**
Fafsta	YES	YES	YES	Any Sequence	NO
BLAST	YES	YES	YES	Any Sequence	NO
Repeat Masker	YES	YES	YES	Any Sequence	NO
Masker Aid	YES	YES	YES	Any Sequence	NO
Tandem Repeat Finder	NO	YES	YES	Any Sequence	NO
Reputer	NO	NO	YES	Any sequence	NO
VMATCH	NO	NO	YES	Any sequence	NO
Amadeus	NO	YES	YES	Predetermined database	NO
Allegro	NO	YES	YES	Predetermined database	NO
Our work	NO	YES	YES	Any Sequence	**YES**

2.2.12 Other studies on cancer genomics

One out of eight deaths is caused by Cancer [79]. A more in-depth understanding of cancer is a must in order provide better treatment for patients. Nowadays, genome data increases exponentially, which leads to more opportunities to learn insights of this terrible disease. David von Hansemann and Theodor Boveri examined strange chromosomal aberrations of cancer cells by dividing them under a microscope [80]. These findings indicated that cancer might be related to abnormalities at the chromosome level [81][82].

The molecular level analysis can be dated back to the 1980s, Reddy et al. [83] first associated a single base substitution of G > T of the HRAS gene with the activation of that specific oncogene function in T24 human bladder carcinoma cells. That is the first study of identifying a certain mutation as being highly correlated to cancer. Currently, a generalizable concept of cancer states that malignancies result from accumulated mutations in genes that increase the “fitness” of a transformed cell over the cells surrounding it. The transformed cells sometimes acquire advantageous mutations which enable them to proliferate unlimitedly and sometimes allow them to spread to distant sites, leading to metastases.

Application of Wavelets

Wavelets transforms have become an integral part of mathematical analysis with an ever increasing range of applications, including genomic sequences. Prasad and Iyengar, 1997, in a publication titled Wavelet Analysis with Applications to Image Processing [84], describe several methods on the application of wavelets as a tool for feature extractions and other important methods in interdisciplinary applications. More recently, Meng, Chen, Iyengar and others [85], have published a paper on the applications of wavelets for cancer related genome sequences.

Wavelet Analysis

In order to conquer cancer, there is a high demand to delve into these sequences and mine useful information. From the signal processing point of view, biological sequences such as DNA sequence can be viewed as one dimensional signal. Accordingly, signal processing approaches can be adopted. The classic approach of Fourier transform suffers from the loss of temporal information. It does not give access to the signals' spectral variations during different time intervals. In other words, the time and frequency information cannot be seen at the same time, and thus a time-frequency representation of the signal is needed.

Wavelet analysis which is capable of decomposing time series into time-frequency space gets increasing amount of attention as a potential tool to study cancer genomic data. Figure 2.9 illustrates the general framework of the analysis procedure.

The origin of the wavelets theory dates back to the Fourier analysis developed by a French mathematician, Jean Baptiste Joseph Fourier (1768-1830) [86]. He came up with the idea of representing each signal as a weighted sum of cosine and sine functions, i.e., Fourier Trigonometric series. Most real-world signals' characteristics are non-stationary signals and vary in both time and space. FT is trying to capture frequency content. The calculation for FT is in Equation (1) from [84].

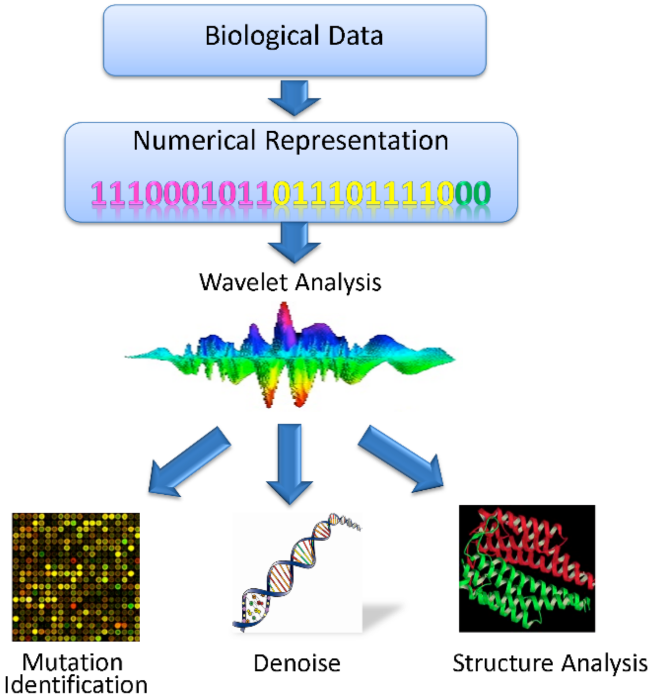


Figure 2.9. Sample procedure of applying wavelet transform in cancer genome analysis.

The FT defines the global representation of the frequency content of a signal over a total period of time. However, it does not represent to the signal's spectral variations during this interval of time. In other words, people cannot view the time and frequency information concurrently. Nevertheless, it would be helpful if both pieces of information were available.

To achieve this, Dennis Gabor proposed the Short-Time Fourier Transform (STFT) to study a small fraction of the signal at a time by segmenting the signal using windows [87]. This obtains the specific contents of each of the analyzed sections separately. The segment of signals in each section is assumed stationary.

However, an important question is raised, which is how to determine the window size. As is shown in experiments, a narrow window results in a poor frequency resolution, whereas a wide window leads to poor time resolution. In addition, one cannot determine the time intervals where a certain frequency exists. Therefore, the wavelet transform was initially proposed as an alternative approach to STFT to overcome the resolution problem. All other windows are the dilated, compressed, and shifted versions based on the mother wavelet.

In terms of history, Haar Wavelets family was developed by Alfred Haar in 1909 [88]. The simplicity of this wavelet transform gives it broader applications. It is generally used to analyze a given signal in terms of functions which are more finite in time than the harmonic functions used in the Fourier analysis. In 1980s, Jean Morlet replaced the Gabor window used in Short-Time Fourier Transform (STFT) [89] by stretching and compressing oscillating windows. Using this technique, he got more reliable and accurate results. This approach is later named as the Morlet Wavelets. In 1989, the idea of multi-resolution was proposed [90]. This idea forms the base theory of versatile wavelets families. Based on this concept, the well-known and frequently used Daubechies wavelets family was invented [91].

To sum up, wavelet analysis techniques have the following advantages over the traditional FT [92]: (a) wavelets are capable of analyzing both stationary and non-stationary signals, while FT gives less information in analyzing non-stationary signals; (b) wavelets have a god localization of frequency and time domains while the standard FT is only hgandled in frequency domain; (c) the base functions of wavelets can both be scaled and shifted and therefore are more flexible, while the FT can only be scaled; and

(d) wavelets have a much wider range of applications than FT. For example, they could be used in non-linear regression and compression. A brief summary of the comparison is shown in Table 2.4.

Properties	Fourier	Wavelet
Stationary signal	Yes	Yes
Non-stationary signal	No	Yes
Time domain	No	Yes
Frequency domain	Yes	Yes
Scale	Yes	Yes
Shift	No	Yes

Table 2.4. Comparison of Wavelet Analysis and Fourier transform (FT). [85]

Renato Dulbecco raised the point that the complete sequence of the human genome would be an essential tool for systematically discovering the genes that cause cancer [93]. The recent advance of next-generation sequencing technology makes the whole cell genome analysis possible for individual cancer case. In recent years, data about epigenome and transcriptome on a genome-wide scale of cancer grow exponentially [94], which offers great opportunity to unlock the mystery of cancer. The main streams of existing work of applying wavelet analysis in cancer genome research are summarized as follows.

Cancer is caused by mutations, Generally speaking, there are four types of common mutations in the cancer genome, which are substitution, insertion or deletion (indel), copy number altercations, and translocations. They are illustrated in Figure 2.10.

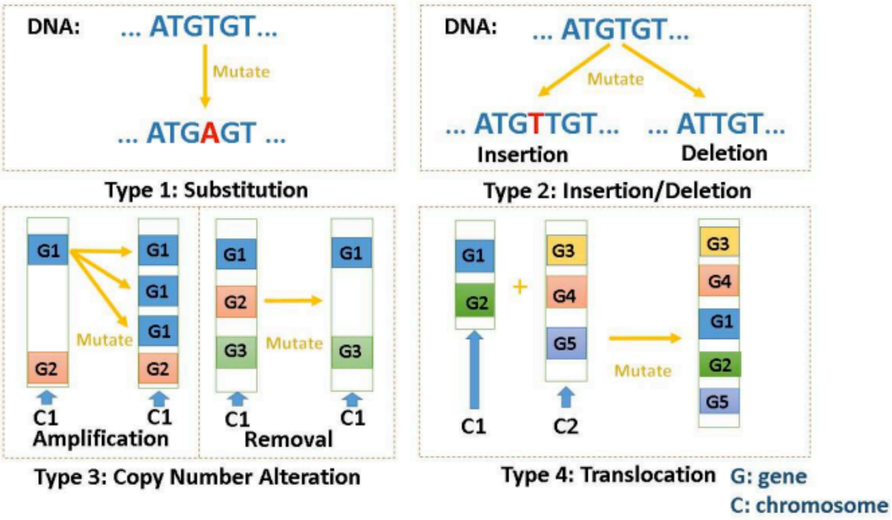


Figure 2.10. The four types of mutations. [92]

Wavelet Analysis in Insertion/Deletion Mutations in Cancer Genome

As described in previous section, cancer is caused by different mutations in the cancer genome. Recent studies find that the form and rate of mutations depend on the context of the mutation point [95]. Wavelet analysis finds its application in this scenario since it provides a multi-scale analysis on the sequence without pre-defined prioris. Therefore, it is suitable for detecting spatial patterns of the sequences around the mutation point without any prior knowledge. In a previous study [96], the authors targeted at identifying the spatial distributions of seven types of mutation related motifs, such as deletion hotspots, DNA pol pause/frameshift hotspots, etc, with respect to insertion/deletion break points. The authors first computed the motif frequency to generate the mo- tif frequency profile. Because of the computational simplicity, Haar wavelet analysis was applied to de- compose the frequency profile. The coefficients' sec- ond raw moments on a multi-scale basis were computed and they were used to measure the size of the difference between motifs occurrence patterns in insertion/deletion flanks vs. control regions. Their study identified the significant spatial distribution patterns of mutation motifs. The identified motifs could be utilized as targets for some cancer medicine. In another study presented in [97], the authors collected 1625 spontaneous base-pair substitutions in the MutL2 strain of Escherichiacoli and analyzed the spatial distribution of these mutations across the Ecoli genome. In order to accommodate the total number of mutations and describe the data clearly, the researchers generated 46 bins, each of which contains 100kb nucleotides, starting at the origin of replication. A histogram was generated to show the distribution of missense mutations. Next, the Fourth- order Daubechies wavelet transforms were applied since it is able to remove jumpy appearance of the Haar averaged signals. The analysis found that these mutations are not distributed at random but, instead, fall into a wave-like spatial pattern that is repeated almost exactly in a mirror image in the two separately replicated halves of the bacterial chromosome. These findings give some insight on different mutations occurred in cancer genome.

Wavelet Analysis in Copy Number Alterations

Copy number alterations are a common pattern of structure variations in cancer genome. This sort of mutations is one hot research area in cancer genome research nowadays. Generally speaking, the main methods of detecting this sort of mutations, fall into two categories, which are the microarray-based gene expression analysis approach and the next generation sequencing-based approach. There have been studies utilizing wavelet analysis in both approaches. In terms of the first approach, the study [98] is an exemplar.

In this paper, the expression level values of each gene are on the vertical dimension and locations of genes on the chromosome are the on the horizontal dimension. The problem of detecting copy number variations was modeled as extracting salient information on the transformed curve, such as sharp peaks and drops of the signal under the high noisy situation. As mentioned previously, although Fourier transform is helpful in signal processing to transform the time series to frequency domain, it loses information regarding position of signal changes, which is important information gain from this study. Therefore, wavelet transforms which capture information in both the frequency and time domains, are suitable for discovering sharp discontinuities on the genome locations. Specifically, the signal profile is first decomposed into a family of multi-resolution subbands using Haar wavelet. For each subband, p-values are assigned to Haar coefficients based on a null-distribution estimated from normal reference samples. Following that, significant coefficients are selected by assigning thresholds for false discovery rate. The selected coefficients are used to identify copy number alterations. The reason to select the Haar wavelet transform is that it is good for analyzing piecewise constant copy number signals [99]. Other similar studies following this direction include [100] and [101].

The next generation sequencing-based methods provide a novel and effective solution for analyzing the copy number alterations in a relatively high resolution. Nevertheless, it suffers from the relatively high noise issue [102]. One benefit that the wavelet analysis provides is its capability of decomposing the signal into a spectrum of different frequencies, where many high frequency components are corresponding to noise. Therefore, one direct application of the wavelet decomposition is to do the noise reduction. In [103], a CNAseg algorithm is proposed to identify copy number alterations from the second-generation sequencing data. In this study, the counts of copy number along the chromosome coordinates were converted to the discrete signals and an undecimated discrete wavelet transform was utilized to smooth the counting data. The number of decomposition levels was adapted to the length of the window counts for each chromosome. The reconstructed signal then went through the Hidden Markov Model (HMM) for segmentation and the Chi-square statistics based segment merging step. Experimental results demonstrated that those proposed approaches decrease the number of non-copy number alteration induced HMM segments and reduce the unevenness in read depth. Such reduction helps improve the performance of the system. The wavelet decomposition-based noise reduction is commonly used in studies in this research direction, such as in [100] and [102].

Wavelet in Machine Learning Research Frame- work of Cancer Genome

Nowadays, cancer genomics and proteomics data increase faster and faster. Meanwhile, more and more machine learning based approaches are applied by researchers for cancer genome analysis. We assume here that mathematical models could be built to learn patterns and could make predictions for unseen data by mining the patterns from the existing data. Firstly, the raw data are to be converted into a related compact and meaningful representations for this propose. We usually call this process as feature extraction. It could be utilized to extract features from a series because wavelet analysis could capture the global and local characteristics of sequence data. My applications used wavelet analysis as feature extraction approach. As an example, in [104], researches utilize wavelet analysis to extract features from DNA microarray data in order to extract important features for classification. Another example is that Liu et al. [105] proposed a framework which utilizes wavelet to extract features from thousands of protein markers in survival analysis in the field of colorectal cancer. Daubechies wavelet db7 is utilized by Liu to perform the continuous wavelet transform to extract the coefficients from the protein marker expression data. These coefficients contain information at different scales of the original biomarker signal. Afterwards, they are been used as features for cancer classification.

Cancer is a very terrible disease which caused by mutations and one of the most problem human faced in this generation. It is viewed as a genetic disease. In conclusion, a good understanding of the mechanism of mutations is mandatory to combating this disease. Wavelet analysis has found its application in many areas of cancer genome research, such as mutation identifications and cancer bio-marker identifications. It can perform multi-scale analysis and capture the local and global information of a time series.

Similar to the Wavelet technique previously discussed, our work also utilizes a machine-learning approach in the analysis of the cancer genome. Specifically, speaking, our objective of using such approach is being able to make predictions of mutations within the genome of a patient.

2.2.13 Clinical perspective

From the medical point of view, our analysis focuses mainly on cancer, a disease that is caused by mutations and affects the genome leading to instability and the generation of more mutations. Through numerous studies, scientists have been able to relate certain cancers somatic mutations in different genes in the genome. As a result, classifying cancer based on gene expression has provided much insight into the complex landscape of multiple interactions between gene networks, as well as into possible treatment strategies. Furthermore, the constant appearance of more sophisticated sequencing strategies have helped in the advance of the study in cancer genomics. This has allowed scientists to interrogate cancer-specific genomic variants and compare them with the normal variants in the same patient. Next generation sequencing (NGS) technologies can provide genome-wide coverage at a single nucleotide resolution and at reasonable speed and cost. By having the unprecedented molecular characterization provided by NGS offers the potential for an individualized approach to treatment. These strategies aim to provide the definition of relatively characteristic gene expression profiles, or molecular signatures that may have prognostic implications for targeted therapies.

Consequently, in cancer patients, the objective of NGS is to obtain and compare information about cancer and normal tissue DNAs. By analyzing this, scientists can develop a catalog of somatic variants that exist in tumor genome but not in the normal tissue DNA. The objective of the analysis is to reveal a drug target in the examined cancer which facilitates the selection of therapy, and improves the personalized risk assessment. As a result, determining the molecular signatures of genes mutated in cancer may help to predict the clinical outcome.

Researchers and clinicians can benefit from being able to predict molecular signatures based on the initial events in cancer development in order to, develop therapeutic modifications in treating the patients. Nevertheless, predicting such signatures at the time of the tumor diagnosis is a major challenge. Numerous research groups have reported lists of predictive genes and reported good predictive performance in terms of prognosis and potential for malignancy based on them. However, the gene lists differed widely and had only very few genes in common.

Additionally, those types of prediction were based on computational approaches not involving NGS, such as microarray analysis, qPCR, and others in various types of cancer such as colorectal, lung, prostate, breast cancer and others.

After searching PubMed we found very few articles reporting predictions in cancer gene signatures by utilizing the NGS approach [106][107]. Furthermore, these reports have dealt with predictions from data obtained at one point in time from paired normal and cancer tissues. To the best of our knowledge, there have been no reports on predictions of cancer gene signatures during cancer progression in the same individual.

Thus, the disadvantages of current predictive models in cancer are that they are focused on evaluating mutations that anticipate the risk of progression and its clinical impact on the length of patient's survival. They are not intended to predict mutational events at molecular level, only to detect and classify existing mutations. Thus, they are not suitable for making predictions in terms of the molecular behavior of the cancer genome.

In our hypothesis, by having the full sequence of the genome from both normal and cancer tissue at different points in time, we are able to extract valuable information to validate the predictions generated from genome data before relapse time. After an exhaustive search of the available scientific resources (PubMed, Google Scholar), we could not find reports utilizing such strategies.

In order to fill this gap, in our study, we proposed to search for studies of cancer patients with such identified mutated genes by the use of NGS various points in time. The mutation profile identified earlier in time will serve as the basis for the prediction at later in time. Furthermore, in order to evaluate our prediction results, we compared the observed mutated locations at the time of relapse with the results of our own predictions.

Our main objective in this study was to test a computational framework utilizing published data from a longitudinal study of patients with acute myeloid leukemia (AML) whose DNA from both normal as well as cancer tissues were subjected to next generation sequencing analysis at various points in time. First of all, we processed the sequencing data at the time of cancer discovery. Secondly, we tested our framework by predicting the regions of the genome to mutate at the time of relapse. Finally, we compared our results with the observed mutated regions, identified by sequencing their genomes at the time of relapse and determined that our predictions are in good agreement with the reported data.

CHAPTER 3

RESEARCH PROBLEM AND SPECIFIC AIMS

In this chapter we motivate the problem to be investigated and present a detailed problem statement. Our research focus is in the area of Bioinformatics. In the next section we provide the motivation for this research by emphasizing the need for a prediction framework for longitudinal DNA mutations within the human genome and highlighting the benefits to be gained from conducting the study. In Section 3.2 we concisely describe the problem to be investigated.

3.1 Motivation

Over time, the scientific and industry community have made tremendous efforts on developing computational solutions to aid the research in molecular biology. These computational solutions focus mainly on String algorithms since the DNA can be represented as a sequence of a combination of four characters (ACTG). These characters are basically what we call the bases in the DNA. Nevertheless, when these sequences suffer from alterations in their structures (changes in one or more bases at different locations within the whole sequence), implications in the health of the individual may appear. Many diseases are associated with causing these changes (mutations) within the genome of an individual. In our case, we are concerned specifically with cancer.

In achieving this goal, we built a computational framework that will perform in three main stages. Firstly, the framework will pre-process the files containing the patient's data for feeding the information needed by the prediction stage. Secondly, given the information provided by the previous stage, the framework will perform the prediction of the number of mutations to occur as well as the locations where these will happen through a statistical analysis. Finally, after the framework finishes with the predictions, the predicted number of mutations as well as their locations will be presented in more than one format to the physicians to provide treatment and/or make further discoveries as many predicted locations could occur at areas where no known functionality exists. Additionally our framework has been implemented in Java and also utilizes statistical libraries and tools for accuracy and performance. Thus, our software follows the Object Oriented (OO) paradigm and has been developed on the Eclipse Integrated Software Environment (IDE) with the currently latest Java Development Kit (JDK 7).

3.2 Problem definition

These changes or mutations in the genome can cause various disorders in the individual depending on their number and locations. As a result, knowing how many mutations will occur as well as where they will occur become very important matters.

The effort is divided into the following three sub-problems:

1. Being able to perform statistical analysis on a given patient's genomic data.

2. Making reliable predictions for the mutating behavior of the given sequence.

3. Identifying mutations in regions in the genome that have not been considered functionally important or relevant.

In order to target the solution of these problems we target three specific aims. Each of these provides a solution to every sub-problem previously respectively.

3.3 Specific Aims

3.3.1 Aim # 1: To build a tool for statistical analysis on genomic data sets.

In order to carry out our first contribution, we built a framework which is capable of extracting statistical information from the genomic data from a patient. It is very important to mention that our framework works with the whole genome data from a patient and not just with one chromosome or a couple of them. Thus, given the genomic data from a patient, we proceed with an analysis at the chromosome level of the whole genome. This analysis involves producing statistical data from the mutations present in each chromosome. The input for this first stage is the genomic data from a sick patient who already presents mutations in his genome.

The statistical analysis involves building a series of variables based on the actual genomic data from the patient that later will be used to perform the predictions.

3.3.2 Aim # 2: To utilize this data set and formulate a prediction model.

Our second goal is our primary contribution as it deals with the prediction itself. Our prediction works at two levels. First, we will predict the number of mutations to occur within one chromosome and, second, the locations where these mutations will happen. For this contribution, our framework will make usage of the data obtained at the completion from the previous stage to build a prediction model to determine the number of mutations to occur. Later, the framework will use another algorithm (Section 7, Fig. 7.4) to predict the respective locations for these mutations.

3.3.3 Aim # 3: To facilitate the discovery of new functionalities in areas of the genome that had not been previously implicated in cancer.

Finally, our third goal consists on providing information to the physicians and scientists that would lead to the discovery or redefinition of new genes and or functionality of certain areas of the genome. This happens because when determining the locations of the newly predicted locations, scientists will be able to find out that some of the areas in the genome where these mutations are predicted to appear do not belong to any known gene or known functionality but that they still have repercussions on the patient's health and, as a result, it can be the case that a new gene can be identified or redefined (e.g. the range of the gene should be extended) or that perhaps a new functionality from this area has been discovered. This information can be provided to the experts in the area of genetics since they can perform further analysis in the laboratory to determine what is there or what functionality might be hidden. Additionally, a specialized and personalized treatment will be possible to obtain due to the fact that we genomic data used was specific to the patient.

The importance of our work resides on being able to predict mutations in a patient's genome ahead of time in order to provide information that will potentially assist physicians on the selection of personalized therapies they can give and/or the development of new specialized drugs.

This prediction is done at the gene and chromosome levels. It provides physicians with information to make a reliable estimation of the effects of a mutation. Moreover, it may lead to the discovery of unknown functionality within the human genome

3.4 Significance and impact of our work

Cancer progression is a complex, multistep process, leading to genomic instability and loss of growth control. Scientists are trying to identify the so-called driver mutations and differentiate them from the “passenger” mutations that passively follow the driver mutations. Without knowledge of the biological functions of the genes this process is very difficult to accomplish. Furthermore, a great deal of genes has no known function, which make it almost impossible to analyze by this approach.

Based on the mutational profile of the cancer cells before relapse time, we utilized the framework to make predictions of new mutations at a later time. The most interesting aspect of this work is the possibility of comparing the predictions with real experimental data obtained from the same patients at the time of cancer relapse. Those comparisons allowed us to generate accuracy values (between 80% and 92% accuracy) which are remarkable. To the best of our knowledge, there are no reported studies directed toward the prediction of changes at molecular level. Numerous studies dealing with predictions based on the mutational signatures in cancer cells are limited to the predictions in clinical outcomes, rather than in the changes at the genome level in the cells of a patient.

One of the contributions of our research is the possibility of validating the predicted mutational profile with the experimental data. This situation has been possible as a result of the revolutionary advances in genome sequencing and data analysis. The second contribution of our research is the purely computational approach to make predictions with no need to include data on the underlying biological processes. This is by far the most exciting possibility. Scientist working on the molecular biology and genetics of cancer experience great difficulty in characterizing the function of an unknown gene which is the first step toward its identification whether is critical or not in cancer progression.

By utilizing our approach, there are no requirements to ascertain the function of genes for the prediction to take place. The third contribution is about providing a target for cancer therapy. As a result of the prediction, scientist will obtain putative target genes to identify, characterize, and possibly utilize as targets of pharmacological agents for cancer treatment.

CHAPTER 4

DATA MINING AND STATISTICAL ANALYSIS

4.1 Background

Our framework is implemented in Java and uses internal and external libraries. The external library we employed for the statistical analysis and processing of our data is Weka.

After our framework processes the input genomic data and obtains the initial statistical information for the DNA mutations for the patients at the chromosome level (explained in further detail in Chapter 6), Weka is employed for creating a model at the chromosome level. This external library builds a model to predict the number of mutations to occur per chromosome.

We chose Weka since it is one of the most widely adopted tools in academia for the purposes of machine-learning and statistical analysis.

4.2 Weka [108]

Weka is acronym name that means Waikato Environment for Knowledge Analysis. It is a software library and tool that consists of various machine-learning algorithms written in Java.. Weka complies under the GNU General Public License. Weka provides features for data mining targeting classification, data preprocessing , clustering, regression, , among others. Essentially, the purpose is to have an application that can be trained for machine learning capabilities and obtain meaningful information through trends and patterns. Weka is open source and is written in Java. Consequently, it is cross-platform. Weka brings a user-friendly and graphical interface that allows for quick set up and operation as well as a programmable API. The data we use to feed Weka is given in the form of as a flat file or relation, meaning that every data object has to be described by a specific number of attributes of a specific type which can be either alpha-numeric or numeric values. In essence, this tool allows users to discover information that is normally hidden from database and files in general. Classification is the core of Weka's functionality. All of the newest and older machine-learning (ML) algorithms follow an object-oriented (OO) Java class hierarchy. Additionally, regression, association rules and clustering algorithms are also part of Weka's implementation.

All the following definitions and examples regarding Weka were obtained from [109].

Weka basic concepts

For a better understanding of Weka, certain concepts are to be clarified.

4.2.1 Dataset

A dataset is an essential concept for machine-learning and it can easily be thought of as a two-dimensional array or table. A dataset represents a collection of samples which are of class Instance. Attributes from an instance are nominal, numeric or strings. Externally speaking, the representation of an Instance class is a file in ARFF format. This file contains a header with the attribute types and the data as comma-separated list.

4.2.2 Classifier

Every learning algorithm is a sub-class from the abstract Classifier [110]. The requirements for a classifier are: a routine for the classifier model from a training dataset and another one for evaluating the generated model.

4.2.3 Weka from Java – API

WEKA functionality can be called from Java sub-routines by the utilization of its API. It runs similar to when calling classifiers with *-p 0*. Nevertheless, unlike Evaluation, it produces the full class probability for any classifier. The java API comes in the form of a jar file and can be incorporated into any Java project. The functionality can be invoked at any time from any pat from the source code of the program.

4.3 Statistical Analysis

Numerous statistical analysis have been developed through time in order to model complex systems or problems [111][112][113][114]. Iyengar and Rao present a great analysis on statistical techniques for modeling complex systems [115]. Here they present single and multi-response models that are representative of modeling complex systems. Their work starts with various concepts related to linear and nonlinear models and then examines four representative techniques of model discrimination that utilize non-intrinsic and intrinsic parameters, Bayesian methods, and likelihood discrimination. Additionally, the authors also evaluate multi-response models which deal with issues from design of experiments for parameter estimation and model discrimination.

Similarly, as mentioned in the previous section, Weka allows for a selection of a variety of models and machine-learning algorithms. Depending on the size and type data we are dealing with, one model will adjust better for prediction analysis than another. Thus, in our scenario, linear regression was the model that best fit to our predicting and gave more accurate results. However, we are open to the inclusion of other models in the future as our data size increases and therefore, the usage of the Weka API in our framework in order to achieve future adaptability and scalability.

Linear Regression

The approach of Regression is applied in statistics to model a relationship between one dependent variable and a set of one or more explanatory/independent variables [116]. If the explanatory variable were just one, then we call it simple linear regression. In any other case, it is called multiple linear regression.

Linear regression is used mainly in two types of applications:

1. For prediction. Here we build a predictive model for a given set of observed data. With this model, we can perform predictions for a future point in time based on a set of previous points in time (training data).

2. For measuring the intensity of the relationship among the dependent variable and the explanatory variables. This allows to know which explanatory variables have higher impact on the value of the dependent one and which ones don't.

The least squares approach is the most widely used method to calculate the coefficients from a linear regression model.

The model has the following form:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where all betas are the coefficients and all the Xi are the features. Finally, epsilon is the error or noise to adjust the model.

CHAPTER 5

MODELING CANCER RELATED GENOME SEQUENCES OF PATIENT DATA

This chapter is directly related to our second contribution which is the creation of a model to predict the DNA mutations. Here, we discuss the reasoning behind the selection of our model, that is, the linear regression model.

5.1 Introduction

The importance of modeling of cancer related genomic data has generated much attention from researchers from various interests, primarily from pharmaceutical design companies. The most common means of finding correlation of genomic data is by structuring a statistical model for the system through which we can predict, control and optimize the variables. A detailed exposition of statistical techniques in modeling of complex systems (single and multi-response models) can be seen in a well-known publication [115]. This chapter examines a representative technique for designing experiments and providing a mathematical structure for the analysis of the data that we have collected in our study.

5.2 Statistical analysis for our model

In this study, the prediction values are numerical. From the perspective of data mining, this problem can be modeled as a regression problem. In our research problem, one challenge is that the training data instances are limited since it is costly to gain patient data. As is known in data mining field, fitting a high order model would require learning more parameters, which needs a large number of samples.

If we tried to use a higher order model such as non-linear models, then we would be facing the problem of over fitting. The results in such case would be totally off as we could observe when switching Weka to other algorithms. When using a higher order model, the number of instances has to be considerably larger compared to the number of features in the model; however, in our case, the number of features and the number of instances have a ratio of almost 1 to 1 which obviously would bring the problem of over fitting.

Given this observation, the linear regression model which fits a linear equation to features and the target variable is utilized. Specifically, three selected numerical features (explained in Section 7.3) are normalized and fit to the linear regression model. The general form of the model is represented as $Y = a + \boldsymbol{bX}$, where Y is the predicted value and $\boldsymbol{X}$ is the normalized feature vector. $\boldsymbol{b}$ is the weights and a is the error term. In addition, our prediction task has a temporal dimension. Assuming there are N observation time points, which are $T_{1,}$ $T_{2,}$ …, $T_{N,}$, suppose we want to predict the mutation number at time point T_N utilizing the features from T_1 to $T_{N-1.}$ A natural assumption is that the mutation number is highly correlated with features at time point which are close to $T_{N.}$ However, it is very difficult to measure this numerically. The linear regression model provides a nice way of quantifying the contribution of each feature. The significance of the features are represented by the corresponding weights trained in the model under the condition that all features are normalized. For example, if the learned parameter for feature *F1* is 100, and the other parameter for feature *F2* is 0.01. We can conclude that

F1 is more important. This property can give us more insight in our study as we want to know which features are important.

5.3 Mechanisms and Models

In order to fit the linear regression model, the mathematical approach of learning the parameters needs to be defined. In this dissertation, we model the learning of parameters as an optimization problem, in which the object function is the mean square error. Formally, assume for the predicted value y^i for data instance *i* ($1 \le i \le M$), M is the total number of chromosomes in one patient. Three features are represented using F_1^i, F_2^i, and F_3^i. Let the corresponding parameters be β_0 , $\beta_1,$ $\beta_2,$ β_3 . The cost function, which corresponds to the half mean square error can be summarized as follows:

$$J = \frac{1}{2M} \sum_{i=1}^{M} [y^i - (\beta_0 + F_1^i \beta_1 + F_2^i \beta_2 + F_3^i \beta_3)]^2$$

In order to minimize this function, the gradient descent algorithm can be utilized. The partial derivative of the function with respect to β_0 , $\beta_1,$ $\beta_2,$ β_3 are as follows:

$$\frac{\partial J}{\partial \beta_0} = \frac{1}{M} \sum_{i=1}^{M} [y^i - (\beta_0 + F_1^i \beta_1 + F_2^i \beta_2 + F_3^i \beta_3)]$$

$$\frac{\partial J}{\partial \beta_1} = \frac{1}{M} \sum_{i=1}^{M} F_1^i [y^i - (\beta_0 + F_1^i \beta_1 + F_2^i \beta_2 + F_3^i \beta_3)]$$

$$\frac{\partial J}{\partial \beta_2} = \frac{1}{M} \sum_{i=1}^{M} F_2^i [y^i - (\beta_0 + F_1^i \beta_1 + F_2^i \beta_2 + F_3^i \beta_3)]$$

$$\frac{\partial J}{\partial \beta_3} = \frac{1}{M} \sum_{i=1}^{M} F_3^i [y^i - (\beta_0 + F_1^i \beta_1 + F_2^i \beta_2 + F_3^i \beta_3)]$$

In each iteration of updating the parameters, β_0 , β_1, β_2, β_3 are updated as follows:

$$\beta_0 \leftarrow \beta_0 - \alpha \frac{\partial J}{\partial \beta_0}$$
$$\beta_1 \leftarrow \beta_1 - \alpha \frac{\partial J}{\partial \beta_1}$$
$$\beta_2 \leftarrow \beta_2 - \alpha \frac{\partial J}{\partial \beta_2}$$
$$\beta_3 \leftarrow \beta_3 - \alpha \frac{\partial J}{\partial \beta_3}$$

Using the matrix annotation, assuming matrix ***F,*** which is shown as follows:

$$\boldsymbol{F} = \begin{bmatrix} 1 & F_1^1 & F_2^1 & F_3^1 \\ \dots & \dots & \dots & \dots \\ 1 & F_1^i & F_2^i & F_3^i \\ \dots & \dots & \dots & \dots \\ 1 & F_1^M & F_2^M & F_3^M \end{bmatrix}$$

And the parameter ***β*** and the response vector ***y*,** which is defined as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix},$$

$$\boldsymbol{y} = \begin{bmatrix} y^1 \\ \cdots \\ y^i \\ \cdots \\ y^M \end{bmatrix}$$

So the model could be written as:

$$\boldsymbol{y}' = \boldsymbol{F\beta}$$

The cost function can be written as: (M is the total number of chromosomes in one patient):

$$J = \frac{1}{M}(\boldsymbol{y} - \boldsymbol{y}')^T(\boldsymbol{y} - \boldsymbol{y}')$$

If we minimize this function, we can get the solution for $\boldsymbol{\beta}$ as:

$$\boldsymbol{\beta} = (\boldsymbol{F}^T \boldsymbol{F})^{-1} (\boldsymbol{F}^T \boldsymbol{y})$$

CHAPTER 6

MATERIALS AND METHODS

We searched for reports that included whole-genome sequencing strategies in their methodology. In order to select suitable studies, we defined a number of requirements, namely studies which were longitudinal, had reported findings on more than one patient, had data which was completely anonymous and publicly available and downloadable in a suitable format. The studies we reviewed had enrolled several patients. The reported results were publicly available and included multiple data on patient's genome during diagnosis from both healthy and cancer tissues, as well as whole-genome sequencing data from affected tissues during disease relapse. Although genome-wide sequencing strategies have been available for more than a decade, there are still a relatively small number of published studies where patients' genome was sequenced after relapse. Most of the published scientific literature on whole-genome sequencing encompasses data comparing the genome changes at nucleotide level from both healthy and affected tissues at one point in time most likely when diagnostic procedures are carried out. As the technologies become less expensive and more widely available, various research groups have started to carry out longitudinal studies where patients' genome is sequenced several times during the disease's progression.

6.1 Framework

We built a framework which is capable of extracting statistical information from the patient’s genomic data. Our framework works with the whole patient's genomic data from one patient and not with just one chromosome. Thus, given the genomic data from a patient, we proceed with an analysis at the chromosome level of the whole genome. This analysis involves producing statistical data from the mutations present in each chromosome. The input for this first stage is the longitudinal sequencing data from a patient’s tumor where mutations have been identified and characterized.

All data to be used needs to be normalized into a standard format as will be presented in the next sections. As more data is obtained from the patients, more training will be provided for our model. Figure 6.1 below shows this in a graphical manner.

6.2 Patient data

Whole genome sequencing data from eight patients with acute myeloid leukemia (AML) served as a basis to test our framework. The data was available as supplementary material to an article published in 2012 by Li Ding et al. in the journal Nature [13]. These patients were from five different French-American-British hematologic subtypes, with elapsed times of 235 to 961 days between samples (see Table 6.1 for each individual’s information). The goal of the study was to investigate the mutational profiles of primary tumors and determine whether the clonal evolution of mutations contributes to relapse. We provided our framework with longitudinal data for the mutational profiles observed in the genomes of patients in order to generate a prediction of nucleotide changes (mutations), followed by a comparison of those predictions with the actual mutations detected by whole-genome sequencing of the tumors after relapse, thus allowing us to measure accuracy and significance of our approach.

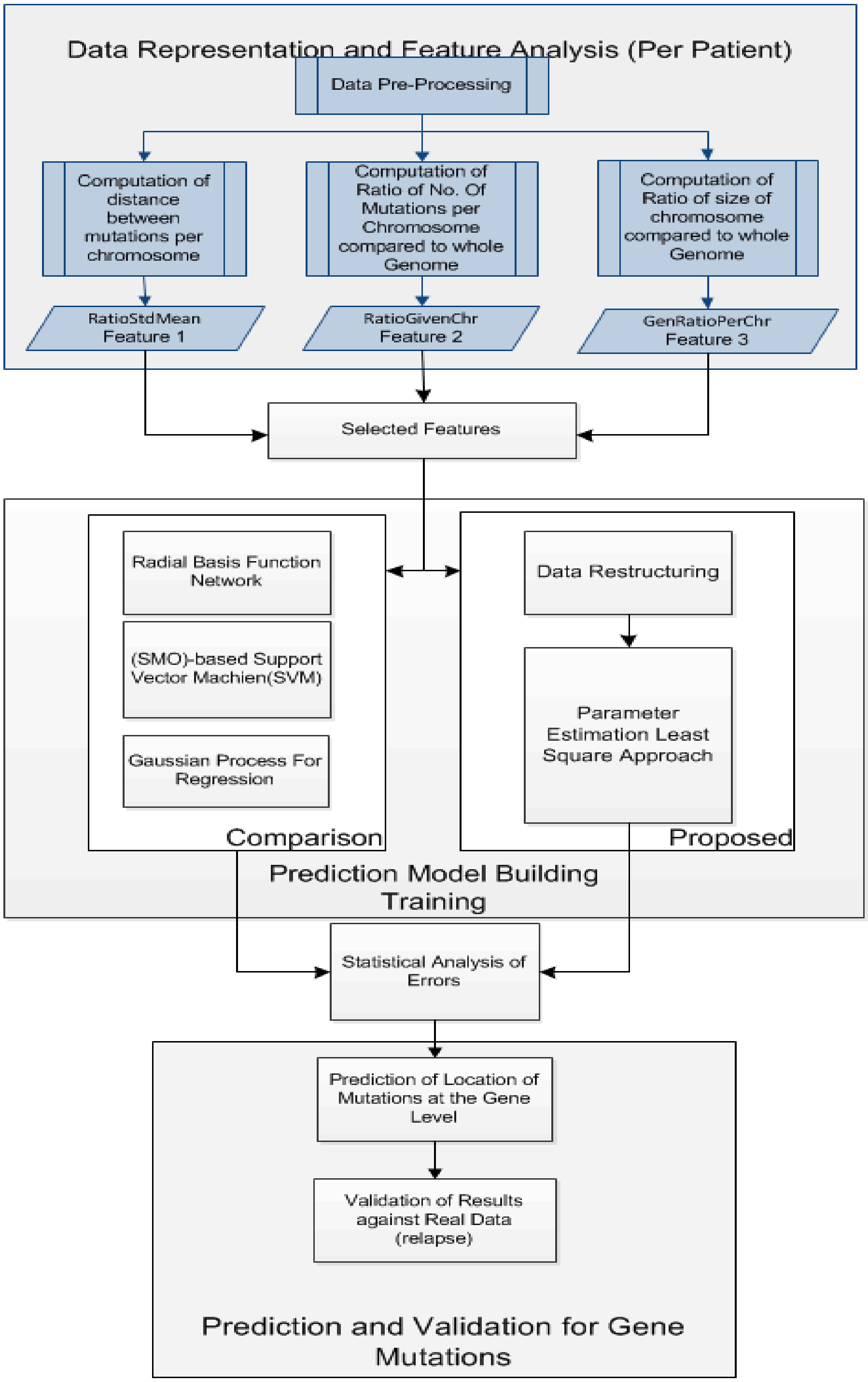

Data Representation and Feature Analysis (Per Patient)
Data Pre-Processing
Computation of distance between mutations per chromosome
Computation of Ratio of No. Of Mutations per Chromosome compared to whole Genome
Computation of Ratio of size of chromosome compared to whole Genome
RatioStdMean Feature 1
RatioGivenChr Feature 2
GenRatioPerChr Feature 3
Selected Features
Radial Basis Function Network
(SMO)-based Support Vector Machien(SVM)
Gaussian Process For Regression
Comparison
Data Restructuring
Parameter Estimation Least Square Approach
Proposed
Prediction Model Building Training
Statistical Analysis of Errors
Prediction of Location of Mutations at the Gene Level
Validation of Results against Real Data (relapse)
Prediction and Validation for Gene Mutations

Figure 6.1. Dependency flow graph for our prediction framework.

Due to different sources of data, and in order to work in a consistent manner, we decided to format all data in a specific manner as explained below. For sequencing data comparisons and evaluations, we aligned the coordinates of input data as well as the coordinates of the resulting predictions to NCBI build 36.3 of human genome assembly. Information corresponding to each patient was labeled with the letters ‘A’ through ‘H’. Data was downloaded as plain text files. Files, called “shared.txt”, contain the locations of mutations identified. Files, called “relapse.txt”, contain the locations of mutations reported after relapse.

UPN 933124	Caucasian female late 50s
UPN 400220	Caucasian female, 34 year old
UPN 426980	Caucasian male, 69 year old
UPN 452198	Caucasian male, 55 year old
UPN 573988	Caucasian female, 67 year old
UPN 869586	Caucasian male, 23 year old
UPN 758168	Caucasian female, 25 year old
UPN 804168	Caucasian male, 53 year old

Table 6.1. Patient designation, gender and age.

In Figure 6.2, we show a partial view of the files previously mentioned in order to get an understanding of the way we have been working with them.

In order to map the coordinates of mutations to genes or genetic regions, we downloaded a file from NCBI build 36.3 of the human genome assembly that contains gene names and location ranges per chromosome in a text format and saved it as geneMap.txt. The file is used to determine which genes will be involved depending on the locations of the predicted mutations.

Limitations of Data

Obtaining patients data is extremely difficult especially because of HIPAA limitations. In our case, we even required data in a longitudinal manner which increased made it even harder. We attended several seminars given by the University of Miami at Jackson's Hospital. During these seminars, medical doctors and other scientists presented their work on cancer research.

Throughout these events we kept asking them for possible sources of these data but their answer was always that the data in a longitudinal manner was very hard to get. Nevertheless, on May 17th, 2013, Dr. James Downing [117], gave a talk titled "The pediatric cancer genome project". Dr. Downing is the Scientific Director for Saint Jude Children's Research Hospital and his talk covered cancer research that was being done by analyzing tumor tissue genomes. After his talk, I asked Dr. Downing for possible sources of longitudinal data based on these genomes, and he suggested me to look for patients with leukemia since it is in this type of disease where most studies had been done in this manner.

He indicated it would still be hard to access the data but that it was the best route to take. Following his advice, we continued our research and found the article previously mentioned were the data was finally available.

Simulating data was never an option, since in the medical domain this practice is strictly forbidden since the conclusions derived from here could be misleading and as a result lead to serious medical conclusions.

When we discussed later our work with a domain expert from the UT Southwestern Medical Center, he indicated that our data was very significant since even high quality journals published in top journals use les or comparable amount of data. More on this can be observed in the conclusion of this dissertation.

Our studies are based in the molecular level, as a result, the demographics of the data (e.g. separation by genders) does not apply. Furthermore, if we were to split the demographic aspects such as genders or ages, it would make our data samples even smaller since we would require grouping. Finally these demographic aspects are nominal features and cannot be fit into a regression model directly.

Figure 6.2 shows the format to be used for the geneMap.txt file previously mentioned.

Chr	Location
1	108346944
1	147392656
1	230352346
2	25310746
2	77165972
2	140418446
3	182087434
3	191103363
3	192532267
3	196997137
4	7394775
4	72616906
4	119477364
….	
22	26004684
X	25476033
X	105826475

Ashared.txt

Chr	Location
1	107357719
1	179307551
1	218018901
1	224145610
2	184186967
2	213886052
3	21310708
3	43135962
3	135449546
….	
22	26004684
X	25476033
X	105826475

Arelapse.txt

Figure 6.2. Sample data files for patient UPN 933124.

chromosome	chr_start	chr_stop	feature_name
1	815	19919	LOC653635
1	42215	43358	OR4G4P
1	52778	53847	OR4G11P
1	58954	59871	OR4F5
1	77385	80096	LOC100132632
…..			

Figure 6.3. Overview of geneMap.txt.

6.3 Variables to consider

Having identified mutations and their locations, our framework will analyze how spaced these mutations are from each other within each chromosome. After obtaining a vector of distances between one mutation and the next, our framework computes statistical analysis such as the mean and standard deviation of these distances to identify how these mutations behave. For example, if the standard deviation is very small compared to the mean, this signifies that mutations occur with a similar distance from one another. If that is not the case, then we can see that some mutations are located very much far apart from each other while others are much closer. Thus, the ratio between the mean and the standard deviation (*RatioStdMean*) of the distances is a critical variable to consider by our framework. A second statistical variable computed by the framework is also the ratio of how many mutations occur within one chromosome when compared to the total number of mutations in the whole genome for a given patient (*RatioGivenChr*). A third variable, we called *GenRatioPerChr*, is a standard variable that has also been computed for all experiments. This variable corresponds to the ratio of the chromosome size in nucleotides to the length of the whole human genome assembly also in nucleotides. This is a standard value as it has been obtained from the reference human genome assembly data and it is not specific to a particular patient. Due to the fact that chromosomes are of different sizes, there is a possibility of finding more mutation in larger chromosomes than in shorter ones.

6.4 Prediction model

The next step in the framework is about prediction. Our prediction works at two levels. First, we will predict the number of mutations to occur within one chromosome, and then the locations where these mutations will happen. For this contribution, our framework will make usage of the statistical data obtained at the completion of previous stage to build a prediction model to determine the number of mutations likely to occur. This prediction model is obtained by utilizing Weka[118] as explained in chapter 4.

Weka uses an ARFF file (Attribute Relationship File Format) as an input, and then by processing this file, creates the models according to the content of the data and variables involved. The weka ARFF file will have the format indicated in Figure 6.4.

```
@RELATION ChromosomeModelling6
@ATTRIBUTE RatioStdMean NUMERIC
@ATTRIBUTE RatioGivenChr NUMERIC
@ATTRIBUTE GenRatioPerChr NUMERIC
@ATTRIBUTE NumMutations NUMERIC

@DATA
…..
```

Figure 6.4. Weka input file.

For our purposes, the framework creates an ARFF file per chromosome containing the data for it along with the associated computed variables previously mentioned.

Since the prediction model works at the chromosome level, there will be 24 files corresponding to 22 somatic chromosomes (1 through 22), and one “X” as well as one “Y” chromosome.

We decided to build a model per chromosome instead of just one model for the whole genome. We carried out a preliminary analysis in which we tried to build a model for the whole genome, but the results were inconclusive. In addition, existing public data is available in a per chromosome basis. Furthermore, a generic model did not fit well for all scenarios due to both the intrinsic varying nature and behavior of each chromosome.

For a given chromosome ‘i’ (1-22, and X as well as Y), our model is of the form:

$$y^i = \beta 1 * F_1^i + \beta 2 * F_2^i + \beta 3 * F_3^i + \beta 0$$

y^i represents the number of mutations we want to predict and F_1^i, F_2^i, and F_3^i represent the RatioStdMean, RatioGivenChr and GenRatioPerChr respectively. Finally, $\beta 1$, $\beta 2$ and $\beta 3$ are the coefficients of the model and $\beta 0$ is the error. Their values are provided by Weka.

These models are obtained by Weka as a String expression and then we parse it and execute each model by inserting the variables of each chromosome for each patient to predict the number of mutations to occur at relapse time.

After generating a model per chromosome, we fit the information obtained from Weka back into the framework. The framework takes this model to execute the prediction of the number of mutations to occur for each chromosome. Once all predicted number of mutations have been computed for each chromosome, the framework takes each individual chromosome and executes the predictions of the locations of the mutations. At this point, since we know the predicted number ‘n’ of mutations likely to occur at each chromosome, we perform the algorithm shown in Figure 6.5.

```
For i=1 to ‘n’
{   for (each existingLocation ‘l’ in chromosome ‘c’)
        if distance from ‘l’ to ‘l+1’ is greater than meanDist
        {       predict finding here at location ‘l’+meanDist;
                findAssociatedGene;
        }
}
```

Figure 6.5 Algorithm for finding mutations locations.

We deal with data from mutations in each chromosome at different points in time. In our algorithm, for instance, we utilize existent longitudinal data on mutations during diagnosis and/or treatment to generate a prediction of mutations at a later relapse time. Thus, each location 'l' represents the location of an existent mutation before relapse time. Next, our algorithm checks whether the distance between mutations is greater than the average distance (meanDist), and if that is the case, a mutation will happen within this distance at approximate position of the location 'l' + meanDist. Having identified a predicted location for a mutation, we proceed to determine whether it belongs to either an annotated gene or an intergenic region.

Due to the fact that we have the actual number of mutations for all chromosomes, as well as their individual locations identified at relapse time, we are able to compare our predicted results with the experimental data. The results of these comparisons serve as our main criteria to estimate the accuracy of the predictions made by the framework.

CHAPTER 7

RESULTS

7.1 Prediction of number of mutations per chromosome.

In order to present, as well as to discuss the results, we selected the data from one of the patients as representative of all (UPN 933124, see Table 7.1 for information on the patient). Data on other patients and their respective predictions are available on the Appendices section. Table 7.1 presents the results for the predictions made by our framework with regards to the number of mutations likely to occur for this given patient. The first column indicates the chromosome analyzed, and the second column indicates the number of mutations observed at the time of relapse (actual number of mutations observed). The third column indicates the predicted number of mutations.

As we can distinguish from Table 7.1, the predicted number of mutations is very close to the observed number of mutations that occurred at the time of relapse in all chromosomes. In order to calculate the accuracy of the prediction, we consider the number of observed mutations versus the number of predicted mutations. We obtained a percent value of the ratio of predicted number of mutations to the total number of observed mutations: 83% (35/42).

Chromosome	**Observed Mutations**	**Predicted Mutations**
1	4	3
2	2	2
3	3	2
4	1	1
5	2	2
6	2	2
7	3	2
8	5	4
9	2	2
11	3	2
12	3	2
13	2	2
14	2	2
16	3	2
17	1	1
20	2	2
21	1	1
X	1	1
Total	42	35
Accuracy	83%	

Table 7.1. Number of mutations per chromosome in patient UPN 933124.

7.1.1 Significance of our experimental data

The following figures show the error distribution of different patients for the predicted number of mutations at the chromosome level. The x-axis is the predicted value minus ground truth. The y-axis is the counts. It can be seen that the proposed approach gives underestimation of number of mutations on each chromosome of each patient.

In addition, it also can be observed that our miss-predictions focus on negative one, which indicates that the proposed framework is very promising.

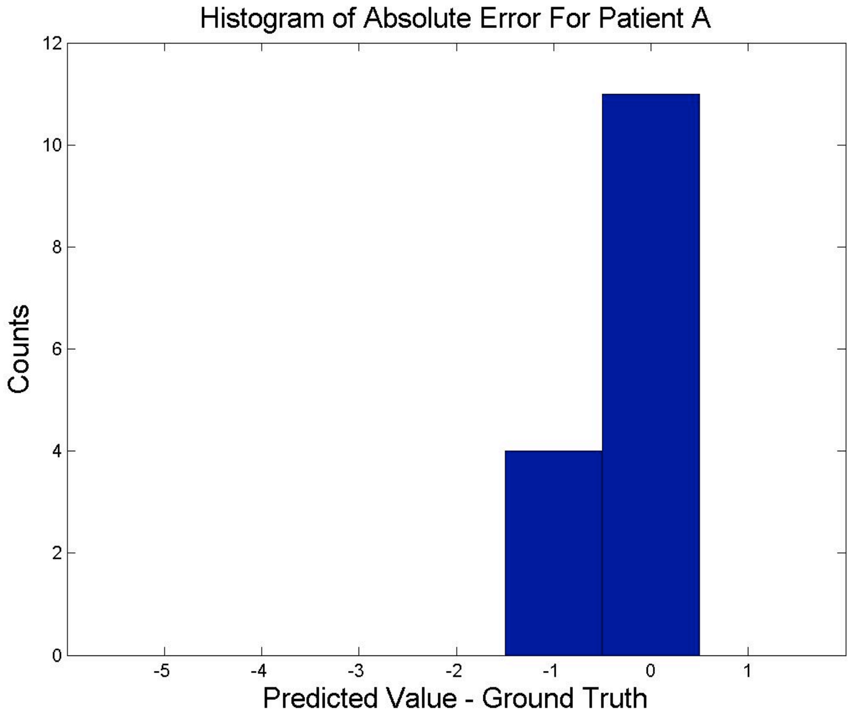


Figure 7.1 Histogram of absolute error for patient A.

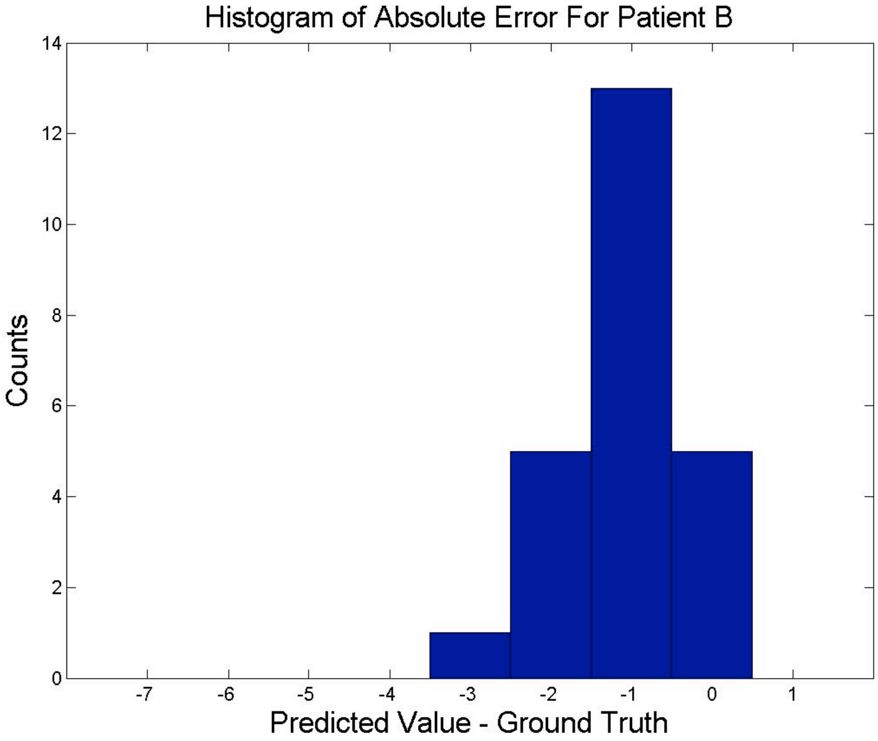


Figure 7.2 Histogram of absolute error for patient B.

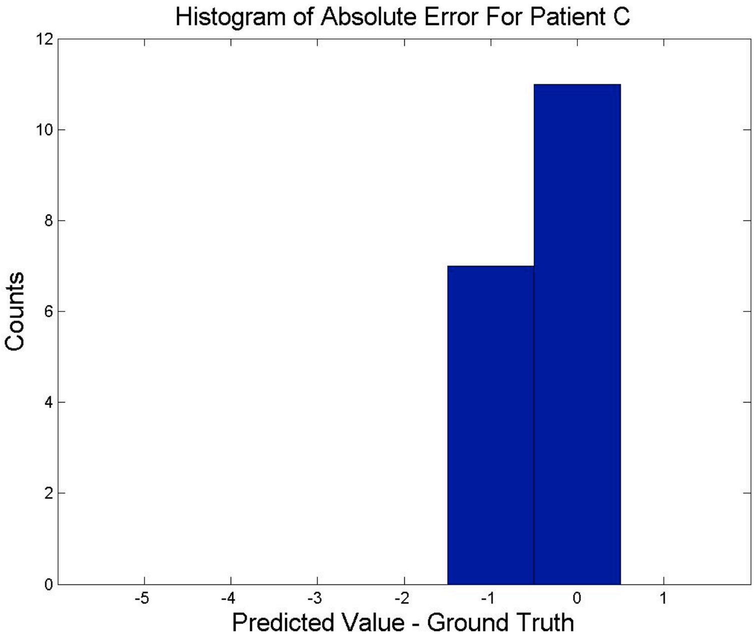


Figure 7.3 Histogram of absolute error for patient C.

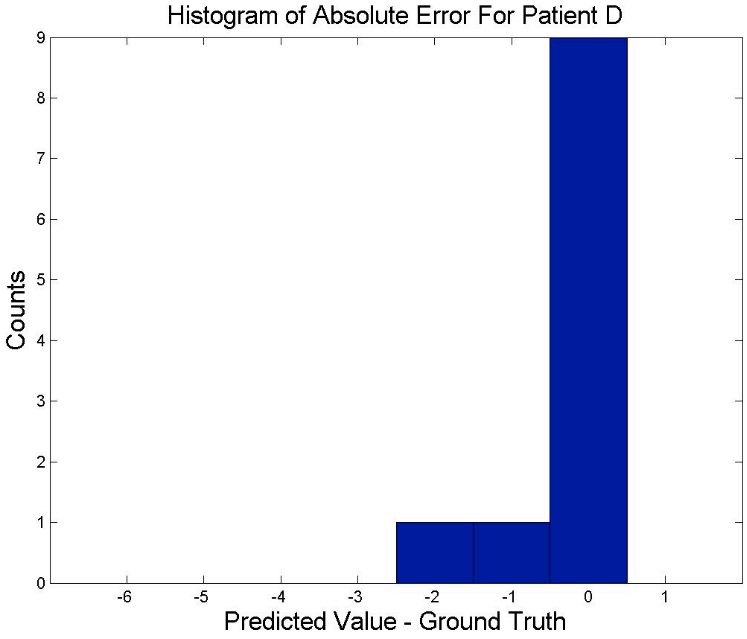


Figure 7.4 Histogram of absolute error for patient D.

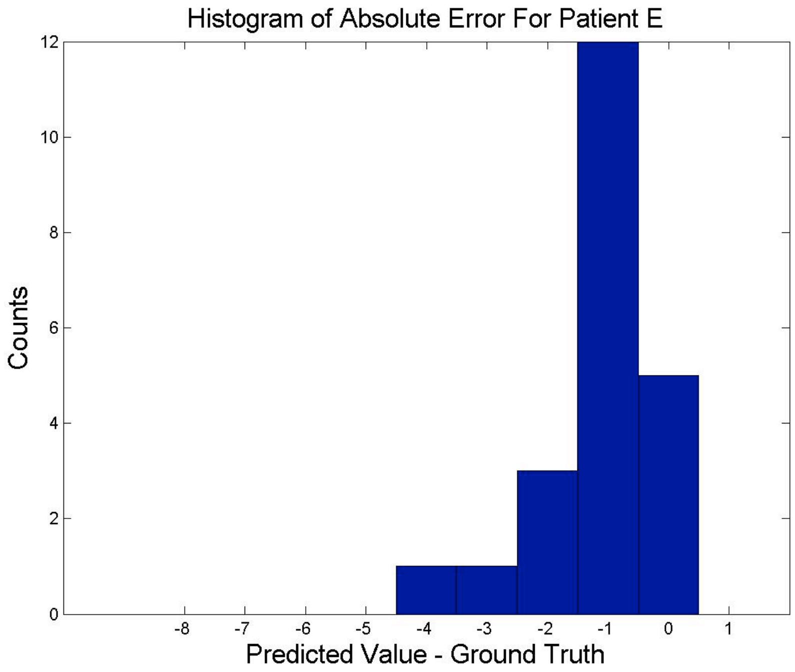


Figure 7.5 Histogram of absolute error for patient E.

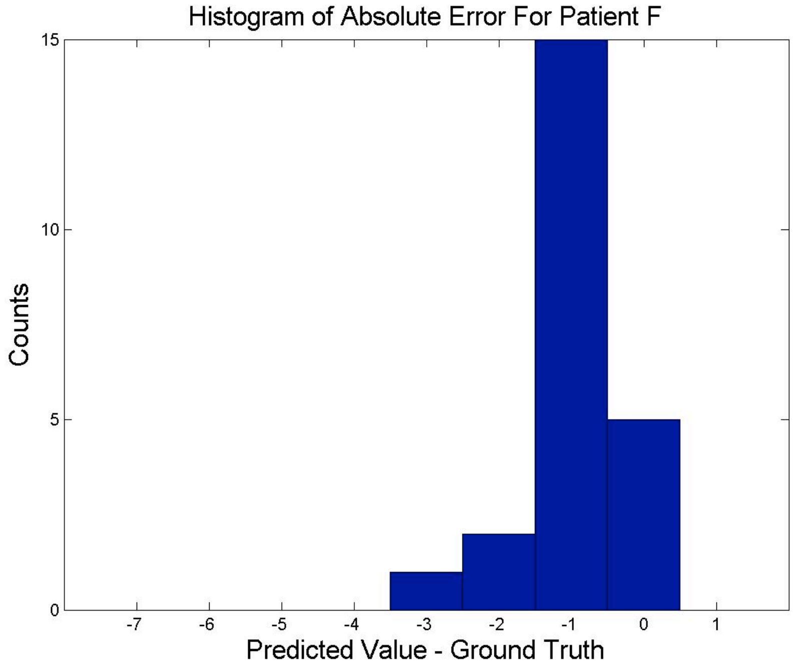


Figure 7.6 Histogram of absolute error for patient F.

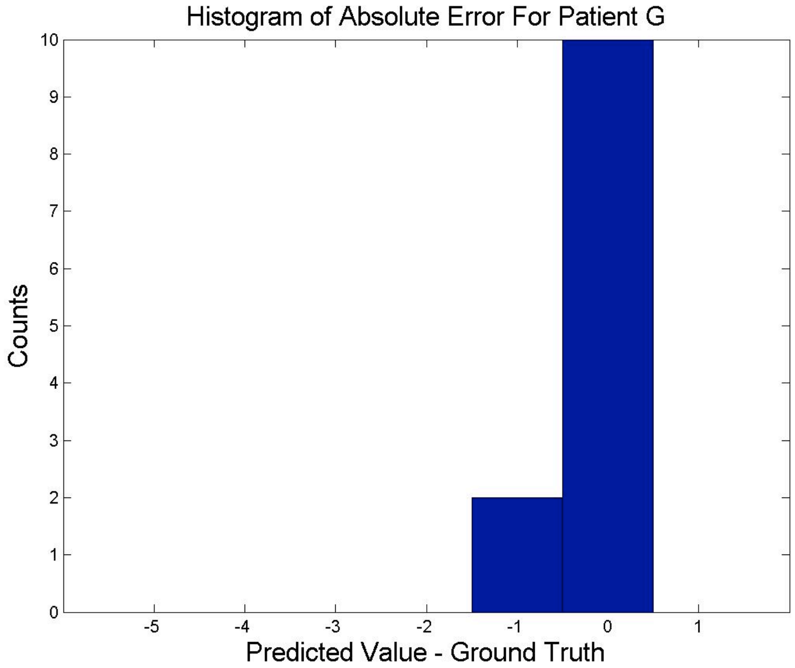


Figure 7.7 Histogram of absolute error for patient G.

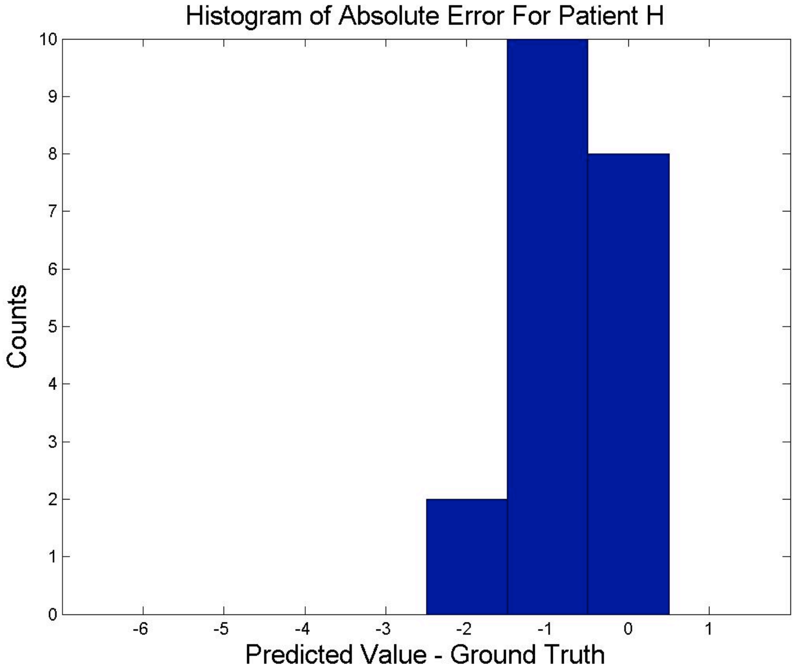


Figure 7.8 Histogram of absolute error for patient H.

The following figures show the histogram of relative errors. The relative error is defined as the error divided by ground truth number.

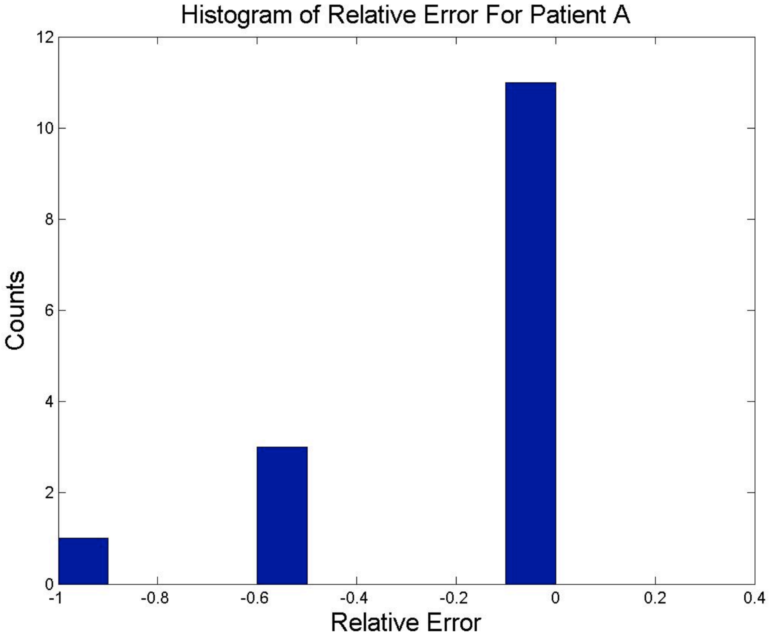


Figure 7.9 Histogram of relative error for patient A.

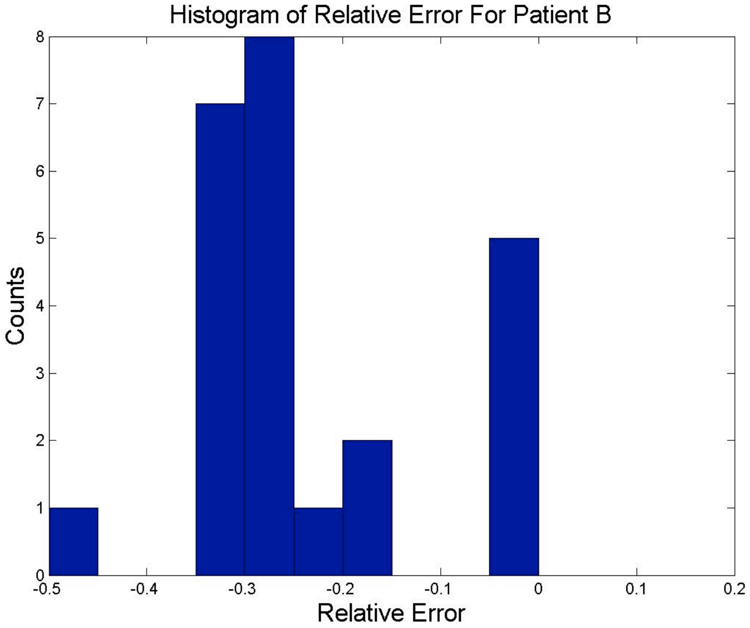


Figure 7.10 Histogram of relative error for patient B.

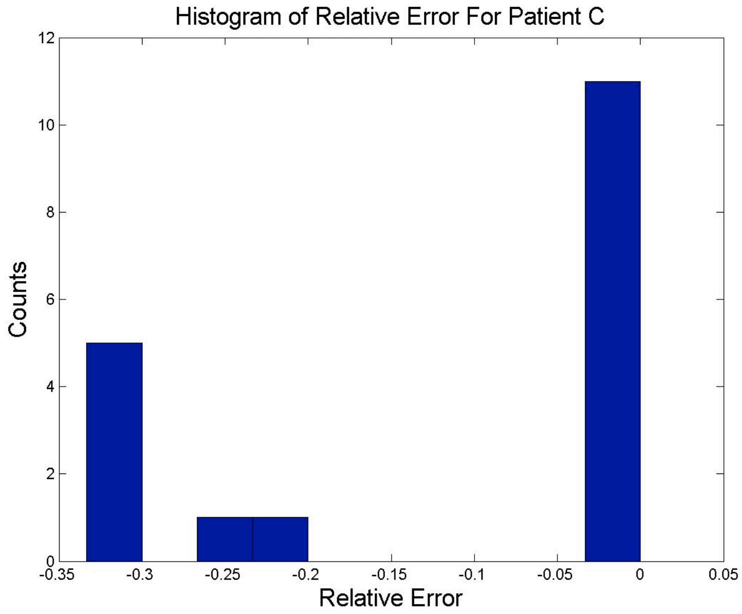


Figure 7.11 Histogram of relative error for patient C.

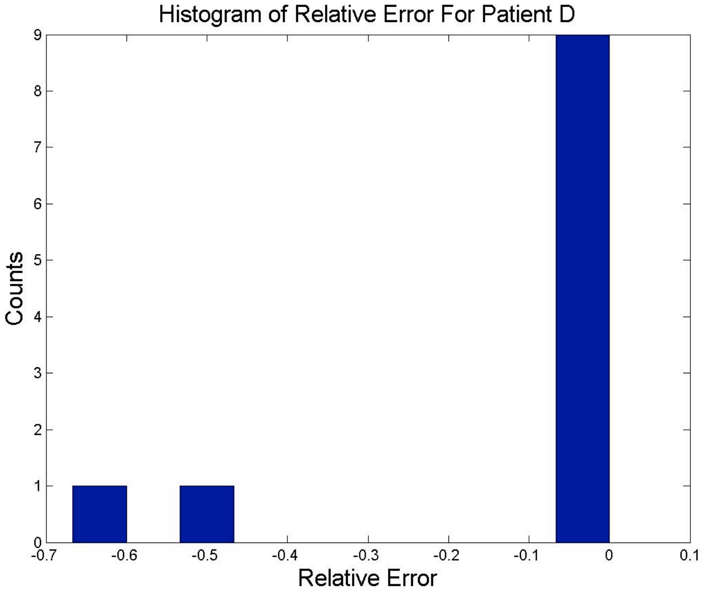


Figure 7.12 Histogram of relative error for patient D.

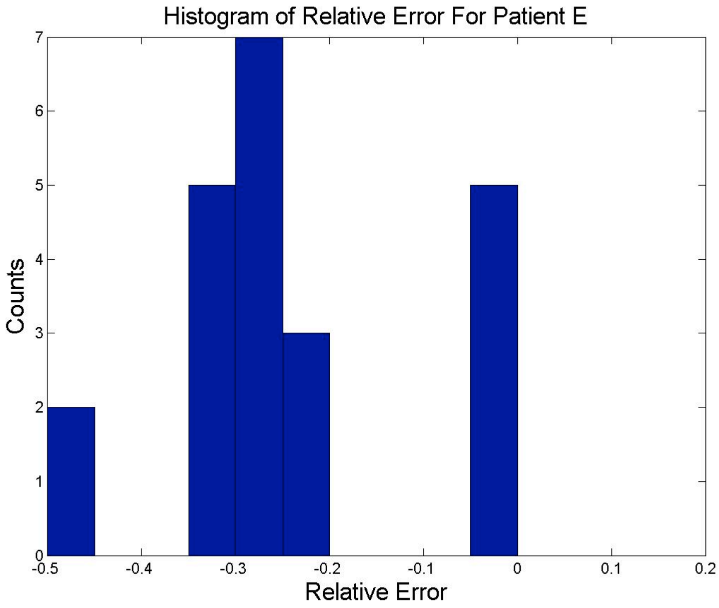


Figure 7.13 Histogram of relative error for patient E.

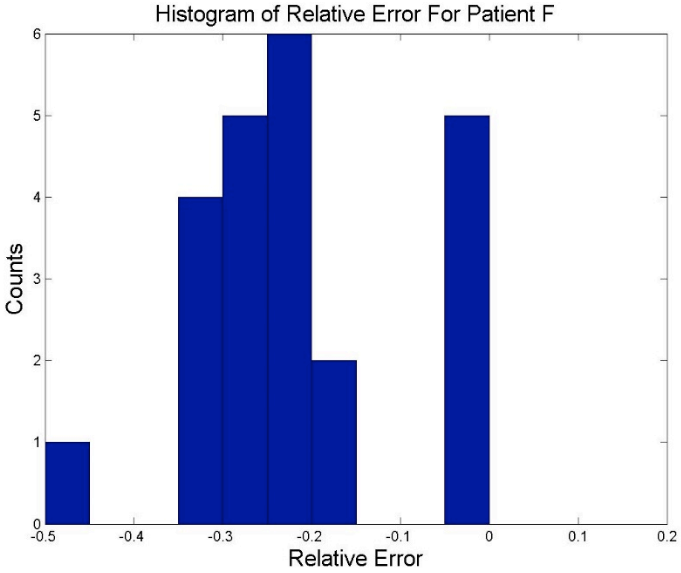


Figure 7.14 Histogram of relative error for patient F.

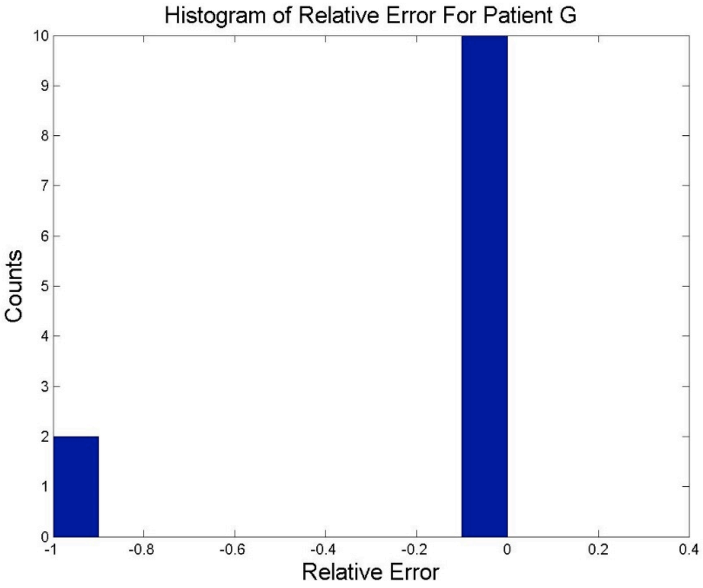


Figure 7.15 Histogram of relative error for patient G.

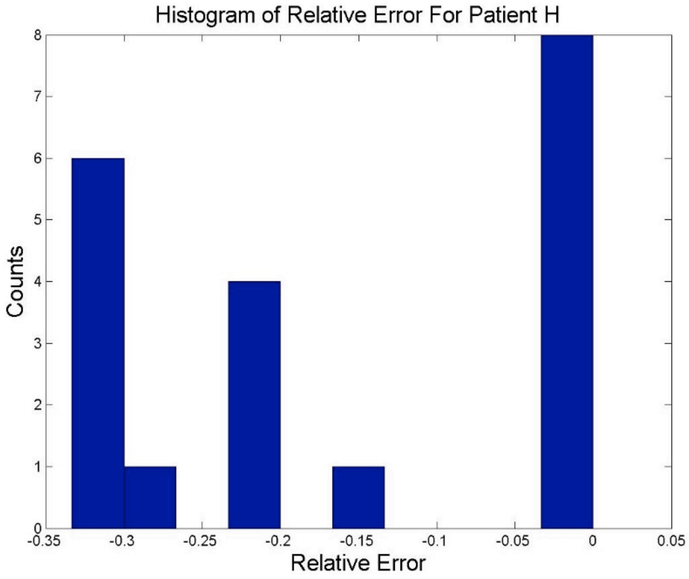


Figure 7.16 Histogram of relative error for patient H.

After a thorough review of our experimental data on the error distribution, we have the following results:

Observation (1): The x-axis is predicted value minus ground truth. The y-axis is the counts. It can be seen that the proposed approach gives underestimation of number of mutations on each chromosome of each patient.

Observation (2): It also can be observed that our mis-predictions focuses on negative one, which indicates that the proposed computational framework is very promising.

7.1.2 Analysis of performance measures to evaluate the success of our prediction

The state of the art evaluation criteria for classification models include accuracy, sensitivity, specificity, precision, true positive value, and true negative values. These metrics are suitable for binary classification problems; however, in our case our research problem is a regression problem and, as a result, the previously mentioned criteria cannot be applied. Consequently, our analysis for accuracy was different as can be observed in the paragraphs below. Detailed statistical analysis of the errors is also provided in Section 7.1.3.

We obtained performance measures to evaluate the success of our prediction. For this, we defined p_1, p_2, p_3, …, p_n as the numeric value of the prediction at the i^{th} instance and we define a_1, a_2, a_3,…, an as the actual values at the i^{th} instance.

The correlation coefficient measures the statistical correlation between a's and p's. The closer we get to 1 for this value, the better correlated results.

Correlation Coefficient = $S_{PA}/\sqrt{S_P S_A}$ where $S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1}$

$$S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1} \quad \text{and} \quad S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$$

We obtained the mean absolute error which represents the average of the individual errors without considering their signs.

Mean Absolute Error = $\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$

The root mean squared error (most widely used measure). It represents the square root of the mean squared error and presents the same dimensions as the predicted value itself.

Root mean squared error = $\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$

“Simple predictor” is the average of all actual values given by the training data.

The relative absolute error computes the total absolute error with the same kind of normalization. Relative errors are normalized by the error of the simple predictor that predicts average values.

Relative absolute error = $\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|}$

Finally, the relative squared error represents the scenario where a simple predictor had been used. It is given by total squared error normalized by dividing it by the total squared error of the default predictor.

Relative squared error = $\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}$

Finally, all these values are shown in table 7.2 for Patient A. The respective tables for other patients can be found at the Appendices.

Chromosome	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
1	0.9412	0.6797	0.9251	27.19%	33.78%
2	0.9999	0.015	0.0181	1.20%	1.22%
3	0.9994	0.1137	0.1144	3.79%	3.41%
4	0.9769	0.4618	0.6322	18.47%	21.37%
5	0.9475	0.6512	0.8757	26.05%	31.98%
6	0.9998	0.0702	0.0752	2.16%	2.15%
7	0.9826	0.2975	0.3562	18.31%	18.55%
8	0.9972	0.1895	0.2338	6.89%	7.51%
9	0.9403	0.5651	0.7798	28.26%	34.03%
10	0.9896	0.5765	0.5954	15.37%	14.36%
11	0.9565	0.4473	0.5969	25.56%	29.17%
12	0.9946	0.2674	0.3315	8.91%	10.36%
13	0.9999	0.0314	0.0314	2.09%	1.48%
14	1	0.0014	0.0016	0.16%	0.15%
15	0.9995	0.122	0.1225	3.49%	3.22%
16	0.9609	0.2272	0.3094	22.72%	27.68%
17	0.9987	0.155	0.1554	5.64%	5.13%
18	0.9914	0.3529	0.3854	12.83%	13.08%
19	0.9885	0.3556	0.3766	15.80%	15.14%
20	0.9702	0.4256	0.5547	21.28%	24.21%
21	0.9959	0.3202	0.376	8.54%	9.07%
22	0.9786	0.4876	0.5916	19.50%	20.60%
X	0.9922	0.2044	0.2841	10.90%	12.47%
Y	0.9992	0.082	0.1047	3.64%	3.91%

Table 7.2. Table for performance measures for our prediction model.

7.1.3 Significance, stability and statistical analysis of model

In this study, we achieved promising results. The correlation coefficient, which indicates the degree of matching of the proposed framework with real data, is 0.9816 ± 0.009 with 95% confidence interval. For mean absolute error, we achieved 0.2958±0.087 with 95% confidence interval.

Here we also provide a statistical analysis for all errors across all chromosomes for

Patient A. The specific algorithm is given as follows. Because the data collected are limited in our case and the population standard deviation is unknown, we use the Student's *t*-distribution to estimate the confidence interval. For one variable *a* such as the correlation coefficient, *n* is the number of observations (*n*=24 in our case). The algorithm to compute the confidence interval *cf* is as follows.

ALGORITHM

```
Begin
Step 1: Compute the average value of $a$, which is $\bar{a}$;
Step 2: Compute the standard deviation of $a$, which is $s$;
Step 3: Set a confidence interval, here we used 95%, which is a golden standard in statistics community.
Step 4: Compute a $T\_multiplier$, which is the inverse of Students' t cumulative distribution function using the confidence interval and the freedom of $n$-1
Step 5: Compute a $cf$ value
        $cf = T\_multiplier * s/\sqrt{n}$
Step 6: The confidence interval is [$\bar{a}$-$cf$, $\bar{a}$+$cf$]
End
```

The table of the statistics of the first patient is in the below table. As mentioned earlier and as it can be observed in Table 7.3, for patient A, we have correlation coefficient as 0.9834±0.0084 with 95% confidence interval. Additionally, for mean absolute error, we achieved 0.2958±0.087 with 95% confidence interval. Tables 7.3 through 7.10 show these measures as well for all other patients. Finally, table 7.11 shows the statistical analysis for all patients.

	mean	**std**	**cf**	**lower bound**	**upper bound**
Correlation coefficient	0.9834	0.0200	0.0084	0.9749	0.9918
Mean absolute error	0.2958	0.2061	0.0870	0.2088	0.3829
Root mean squared error	0.3678	0.2736	0.1155	0.2523	0.4834
Relative absolute error	0.1286	0.0927	0.0391	0.0895	0.1678
Root relative squared error	0.1433	0.1123	0.0474	0.0959	0.1907

Table 7.3. Statistical analysis of errors for patient A.

	mean	**std**	**cf**	**lower bound**	**upper bound**
Correlation coefficient	0.9807	0.0193	0.0081	0.9726	0.9888
Mean absolute error	0.3000	0.2065	0.0872	0.2128	0.3872
Root mean squared error	0.3730	0.2731	0.1153	0.2577	0.4884
Relative absolute error	0.1336	0.0922	0.0390	0.0947	0.1726
Root relative squared error	0.1484	0.1126	0.0475	0.1009	0.1960

Table 7.4. Statistical analysis of errors for patient B.

	mean	**std**	**cf**	**lower bound**	**upper bound**
Correlation coefficient	0.9821	0.0178	0.0075	0.9746	0.9897
Mean absolute error	0.3009	0.2056	0.0868	0.2141	0.3877
Root mean squared error	0.3739	0.2737	0.1156	0.2583	0.4895
Relative absolute error	0.1329	0.0934	0.0394	0.0935	0.1724
Root relative squared error	0.1477	0.1123	0.0474	0.1003	0.1951

Table 7.5. Statistical analysis of errors for patient C.

	mean	**std**	**cf**	**lower bound**	**upper bound**
Correlation coefficient	0.9826	0.0168	0.0071	0.9755	0.9897
Mean absolute error	0.3005	0.2061	0.0870	0.2135	0.3876
Root mean squared error	0.3732	0.2735	0.1155	0.2578	0.4887
Relative absolute error	0.1345	0.0923	0.0390	0.0955	0.1735
Root relative squared error	0.1468	0.1129	0.0477	0.0991	0.1944

Table 7.6. Statistical analysis of errors for patient D.

	mean	std	cf	lower bound	upper bound
Correlation coefficient	0.9802	0.0195	0.0082	0.9720	0.9885
Mean absolute error	0.2997	0.2050	0.0866	0.2132	0.3863
Root mean squared error	0.3725	0.2749	0.1161	0.2564	0.4886
Relative absolute error	0.1336	0.0924	0.0390	0.0946	0.1726
Root relative squared error	0.1476	0.1121	0.0473	0.1003	0.1949

Table 7.7. Statistical analysis of errors for patient E.

	mean	std	cf	lower bound	upper bound
Correlation coefficient	0.9807	0.0197	0.0083	0.9724	0.9891
Mean absolute error	0.3012	0.2065	0.0872	0.2140	0.3884
Root mean squared error	0.3719	0.2740	0.1157	0.2562	0.4876
Relative absolute error	0.1324	0.0919	0.0388	0.0936	0.1712
Root relative squared error	0.1487	0.1125	0.0475	0.1012	0.1962

Table 7.8. Statistical analysis of errors for patient F.

	mean	std	cf	lower bound	upper bound
Correlation coefficient	0.9814	0.0191	0.0081	0.9734	0.9895
Mean absolute error	0.2997	0.2062	0.0871	0.2126	0.3868
Root mean squared error	0.3722	0.2732	0.1154	0.2568	0.4876
Relative absolute error	0.1337	0.0930	0.0393	0.0944	0.1730
Root relative squared error	0.1481	0.1120	0.0473	0.1008	0.1954

Table 7.9 Statistical analysis of errors for patient G.

	mean	std	cf	lower bound	upper bound
Correlation coefficient	0.9818	0.0190	0.0080	0.9738	0.9899
Mean absolute error	0.3019	0.2067	0.0873	0.2146	0.3892
Root mean squared error	0.3721	0.2733	0.1154	0.2567	0.4875
Relative absolute error	0.1343	0.0922	0.0389	0.0953	0.1732
Root relative squared error	0.1479	0.1131	0.0477	0.1001	0.1956

Table 7.10. Statistical analysis of errors for patient H.

	mean	**std**	**cf**	**lower bound**	**upper bound**
Correlation coefficient	0.9816	0.0011	0.0009	0.9808	0.9825
Mean absolute error	0.3000	0.0018	0.0015	0.2984	0.3015
Root mean squared error	0.3721	0.0018	0.0015	0.3705	0.3736
Relative absolute error	0.1330	0.0019	0.0016	0.1314	0.1345
Root relative squared error	0.1473	0.0017	0.0014	0.1459	0.1487

Table 7.11. Statistical analysis of all patients.

8.1.3.1 Observations

Table 7.11 gives a summary of all eight patients. As indicated in the table, our method is relatively stable across different patients. For example, for the eight patients, the confidence interval with 95% for the correlation coefficient is 0.9816 ± 0.009. Since the eight patients are chosen randomly, these experimental results indicate that our model fitting could give around 0.9816 of correlation coefficient measurement. As we know that 1 indicates the perfect performance. This measurement indicates that our model fits the data relatively well.

Below, we can also see the coefficients obtained for the models at each chromosome per patient. We tried to find some patterns among across different chromosomes within the same patient and then also tried to find patterns across patients and we could not identify any significant patterns.

For the purpose of showing how the models look like, we show all tables that contain the coefficients for the chromosome models for each patient.

From the model $y^i = \beta 1 * F_1^i + \beta 2 * F_2^i + \beta 3 * F_3^i + \beta 0$,

$\beta 1$, $\beta 2$, and $\beta 3$ are the coefficients for RatioStdMean, RatioGIvenChr and ChrToGenomRatio respectively. $\beta 0$ is the error used to adjust the model.

chr	**$\beta 1$**	**$\beta 2$**	**$\beta 3$**
1	-0.0058	-0.9408	0.00000012
2	0.0044	-1.1347	-0.00000002
3	0.0045	4.3818	0.00000015
4	0.0048	5.3997	0.00000032
5	-0.0124	-6.074	0.00000017
6	0.006	4.4676	0.00000009
7	0.0125	3.0178	0.00000008
8	0.0084	7.6347	0.00000013
9	0.0127	-6.1969	-0.00000014
10	-0.0051	6.0138	-0.00000016
11	0.0095	2.5017	0.00000027
12	-0.0061	4.3137	0.00000006
13	-0.014	-1.5312	0.00000014
14	0.0051	-0.6632	0.00000018
15	-0.0102	1.6083	0.00000032
16	-0.0023	-1.5139	0.00000031
17	-0.0066	6.0243	-0.00000011
18	-0.0037	5.1447	0.00000011
19	-0.014	1.4385	0.00000041
20	-0.0032	4.1677	0.00000021
21	0.0061	4.6595	0.00000012
22	-0.0127	6.8996	0.00000022
X	0.0068	4.4555	-0.00000053
Y	-0.004	4.9267	0.00000061

Table 7.12. Coefficients for chromosome models for patient A.

chr	β**1**	β**2**	β**3**
1	-0.0029	-0.7871	-0.00000014
2	0.0081	-2.3130	0.00000051
3	0.0108	5.0479	0.00000038
4	0.0115	5.2580	-0.00000048
5	-0.0128	-5.3196	0.00000018
6	0.0042	3.9586	0.00000096
7	0.0150	2.7207	-0.00000069
8	0.0083	7.9148	0.00000024
9	0.0143	-7.0051	0.00000015
10	0.0009	5.6727	-0.00000033
11	0.0168	3.8659	-0.00000001
12	0.0031	3.1030	0.00000029
13	-0.0228	-2.5005	0.00000088
14	0.0056	0.5399	0.00000067
15	-0.0089	2.9604	0.00000046
16	0.0018	-0.4690	0.00000035
17	-0.0136	6.4899	-0.00000038
18	0.0059	6.4388	0.00000000
19	-0.0180	1.6259	-0.00000027
20	-0.0051	3.3770	0.00000029
21	0.0040	5.0753	-0.00000072
22	-0.0105	6.8326	0.00000019
X	0.0163	4.8320	0.00000026
Y	0.0053	4.8783	0.00000132

Table 7.13. Coefficients for chromosome models for patient B.

chr	$\beta 1$	$\beta 2$	$\beta 3$
1	-0.0072	-0.2514	0.00000071
2	0.0110	-1.1303	-0.00000086
3	0.0065	4.7241	0.00000039
4	0.0077	4.2140	-0.00000051
5	-0.0145	-7.4037	0.00000050
6	0.0047	4.9642	-0.00000016
7	0.0101	2.7127	0.00000057
8	0.0136	7.7381	0.00000018
9	0.0122	-5.4405	-0.00000064
10	-0.0046	7.0865	0.00000008
11	0.0186	3.1069	-0.00000016
12	-0.0078	4.4835	0.00000026
13	-0.0146	-2.2882	0.00000107
14	0.0047	-0.2342	0.00000085
15	-0.0132	1.0773	-0.00000027
16	-0.0039	-0.9811	0.00000116
17	0.0025	6.0216	-0.00000056
18	0.0048	6.2068	-0.00000063
19	-0.0237	0.0305	0.00000055
20	0.0041	4.9333	-0.00000006
21	0.0017	5.8057	-0.00000013
22	-0.0162	6.9951	-0.00000029
X	0.0050	5.2612	-0.00000036
Y	-0.0020	6.1814	0.00000152

Table 7.14. Coefficients for chromosome models for patient C.

chr	**β1**	**β2**	**β3**
1	-0.0076	-0.7604	-0.00000076
2	0.0067	-0.7116	0.00000096
3	0.0097	5.7364	-0.00000048
4	0.0139	4.0246	0.00000114
5	-0.0202	-4.9885	-0.00000002
6	0.0147	4.4609	-0.00000022
7	0.0042	2.6177	0.00000107
8	0.0128	8.7581	0.00000091
9	0.0192	-5.4743	0.00000070
10	0.0022	6.3766	0.00000037
11	0.0194	1.7699	0.00000021
12	0.0033	3.6173	0.00000102
13	-0.0120	-2.0102	0.00000015
14	0.0064	0.7180	0.00000026
15	-0.0081	2.5469	-0.00000058
16	0.0023	-2.9014	0.00000071
17	-0.0021	5.4218	0.00000084
18	-0.0080	4.2675	0.00000001
19	-0.0200	2.3603	0.00000068
20	-0.0035	3.4451	-0.00000044
21	0.0142	4.8766	-0.00000044
22	-0.0171	7.1129	0.00000082
X	0.0148	4.2595	-0.00000152
Y	-0.0040	5.3937	0.00000005

Table 7.15. Coefficients for chromosome models for patient D.

chr	$\beta1$	$\beta2$	$\beta3$
1	0.0019	0.4059	0.00000012
2	0.0140	-2.2165	0.00000066
3	0.0023	3.8831	0.00000013
4	0.0081	6.7147	0.00000009
5	-0.0197	-5.9144	0.00000043
6	0.0138	4.0495	0.00000075
7	0.0187	2.9437	-0.00000072
8	0.0127	8.1505	0.00000027
9	0.0087	-5.4002	0.00000033
10	0.0033	6.2483	0.00000059
11	0.0191	1.8344	0.00000074
12	-0.0079	3.8368	0.00000102
13	-0.0147	-1.0476	0.00000054
14	0.0080	-0.7254	-0.00000023
15	-0.0154	2.5250	-0.00000002
16	0.0073	-0.6750	0.00000067
17	-0.0016	6.0702	0.00000011
18	-0.0014	6.4145	-0.00000033
19	-0.0041	0.9393	0.00000135
20	-0.0042	3.8852	0.00000075
21	0.0148	5.5249	0.00000020
22	-0.0178	7.6082	-0.00000075
X	0.0144	5.5756	-0.00000133
Y	0.0007	4.0187	0.00000059

Table 7.16. Coefficients for chromosome models for patient E.

chr	$\beta 1$	$\beta 2$	$\beta 3$
1	0.0013	-1.8080	-0.00000024
2	0.0140	-0.8890	-0.00000031
3	0.0025	3.7933	-0.00000064
4	0.0101	6.0363	0.00000093
5	-0.0157	-5.1467	0.00000096
6	0.0131	5.4381	0.00000094
7	0.0067	3.8777	0.00000093
8	0.0183	6.8689	-0.00000070
9	0.0219	-5.3636	0.00000030
10	-0.0082	6.2296	0.00000044
11	0.0021	3.8518	0.00000038
12	-0.0123	3.1948	0.00000086
13	-0.0161	-0.6092	-0.00000002
14	0.0145	-0.4893	0.00000091
15	-0.0103	2.5162	-0.00000055
16	0.0036	-2.8785	-0.00000054
17	-0.0074	6.0751	0.00000016
18	0.0011	4.9096	0.00000031
19	-0.0068	1.4684	0.00000077
20	0.0053	2.9407	0.00000085
21	0.0109	5.0740	-0.00000079
22	-0.0151	7.9485	-0.00000032
X	0.0130	3.6375	0.00000014
Y	0.0007	3.9946	-0.00000030

Table 7.17. Coefficients for chromosome models for patient F.

chr	***β*1**	***β*2**	***β*3**
1	-0.0013	-0.0143	0.00000090
2	0.0104	-2.0696	-0.00000036
3	0.0096	3.3256	0.00000011
4	0.0070	6.3824	0.00000013
5	-0.0026	-5.3594	-0.00000045
6	0.0137	5.4012	0.00000101
7	0.0051	2.8205	0.00000025
8	0.0151	8.2805	-0.00000058
9	0.0165	-5.7032	0.00000071
10	0.0032	6.6396	0.00000035
11	0.0055	3.3530	0.00000095
12	0.0035	5.5225	-0.00000042
13	-0.0128	-0.8615	0.00000077
14	0.0095	-1.3463	-0.00000080
15	-0.0031	0.7591	0.00000131
16	0.0049	-0.7272	0.00000082
17	0.0031	5.9467	0.00000026
18	0.0026	4.4366	0.00000110
19	-0.0190	2.3530	-0.00000035
20	0.0047	4.5874	-0.00000040
21	0.0160	4.3033	-0.00000043
22	-0.0190	7.7892	0.00000057
X	0.0126	3.1060	-0.00000125
Y	0.0007	5.4127	0.00000069

Table 7.18. Coefficients for chromosome models for patient G

chr	**β1**	**β2**	**β3**
1	0.0032	-0.3078	-0.00000059
2	0.0012	-0.1699	-0.00000048
3	0.0069	5.6587	0.00000034
4	0.0071	5.6634	-0.00000021
5	-0.0095	-5.7618	-0.00000061
6	0.0096	5.0768	0.00000005
7	0.0091	4.3699	0.00000022
8	0.0166	7.2473	0.00000020
9	0.0219	-6.1187	0.00000011
10	0.0018	6.7467	0.00000037
11	0.0184	2.8218	-0.00000042
12	0.0037	4.5330	0.00000079
13	-0.0176	-1.8377	0.00000106
14	0.0113	0.1349	-0.00000049
15	-0.0192	1.6409	0.00000095
16	-0.0046	-0.7617	-0.00000046
17	0.0013	6.2843	0.00000011
18	-0.0074	4.9845	0.00000013
19	-0.0075	0.7691	0.00000034
20	-0.0062	3.9471	0.00000032
21	0.0034	4.7239	-0.00000039
22	-0.0194	6.4096	0.00000041
X	0.0141	3.5085	-0.00000091
Y	0.0013	5.1003	0.00000093

Table 7.19. Coefficients for chromosome models for patient H

7.1.4 Comparison of performance of our model versus other algorithms

In order to evaluate our framework better, we compared our proposed framework with three state-of-the-art algorithms which are Support Vector Machines using Sequential Minimal Optimization, Gaussian Process regression model and Radial Basis Function Network. The following paragraphs briefly explain them in more detail.

SVMs (Support Vector Machines) is a well known technique widely used for classification and regression. It is one of the most popular state-of-the-art algorithms which can be utilized to provide a model for linear and non-linear problems. Mathematically speaking, the SVM model can be interpreted as solving the optimization problem below:

$$\min_{w,b,\varepsilon} \quad \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i$$

Subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0$$

Where x_i : training vectors.

w : weight vectors.

$\phi(x_i)$: kernel function

b: the bias.

ξ: parameter to handle over-fitting.

Here we utilized one major implementation of SVM, which is the sequential minimal optimization (SMO) implementation to compare to the proposed framework. The performance is given in figure 7.17. It shows that it performs worse than the proposed algorithm. In addition, there is a lot of variance in the performance of the SVM-based solution. The possible reason is the over fitting problem. More details can be found in [119].

Gaussian Processes for Regression

The Gaussian process for regression is well known model for machine-learning. It is a generalized form of the Gaussian probability distribution. Similar to the correspondent distribution, the Gaussian process for regression is determined by the covariance and mean function. The function here f(x) comes from just one sample from the actual distribution. More details on this method can be found in [120].

Radial Basis Function

It follows an artificial neural network approach which employs functions that are based on a radial basis. Radial basis functions based on neuron arguments are the output of this network. For a given input x, $\mathbf{x} \in \mathbb{R}^n$, the algorithm produces a scalar function, $\varphi : \mathbb{R}^n \to \mathbb{R}$, given by:

$$\varphi(\mathbf{x}) = \sum_{i=1}^{N} a_i \rho(||\mathbf{x} - \mathbf{c}_i||)$$

Where

N : number of neurons (hidden layer)

c_i : center vector of neuron i,

a_i : weight of neuron i

For more details on this algorithm, we can refer to [121].

The comparison results are shown in Table 7.20 and Figure 7.17. For this, we compared the correlation coefficients achieved by each different algorithm and verified that our selected model had a better performance and stability across all chromosomes. We can observe that the Radial Basis Function Network shows the highest variance. This is very dangerous in the medical domain as the predicted values show high variability. The other two algorithms have a better performance when compared to the Radial Basis Function Network; however, they also still lower performance than our selected model and higher variance. Our model, on the other hand, is stable across all chromosomes showing more accurate predictions.

Chr	Linear Regression	Radial Basis Function Network	Sequential Minimal Optimization	GaussianProcess
1	0.9412	0.4533	0.9389	0.9356
2	0.9999	0.8452	0.9253	0.9235
3	0.9994	0.3159	0.9993	0.9235
4	0.9769	0.8627	0.9718	0.9741
5	0.9475	0.4144	0.9448	0.9373
6	0.9998	0.0806	0.9985	0.9909
7	0.9826	0.5022	0.9383	0.9956
8	0.9972	0.0616	0.9972	0.9235
9	0.9403	0.0112	0.8055	0.9307
10	0.9896	0.1312	0.9819	0.9827
11	0.9565	0.4605	0.9192	0.9436
12	0.9946	0.9368	0.9946	0.7254
13	0.9999	0.4473	0.8715	0.9657
14	1	0.1127	1	0.9973
15	0.9995	0.957	0.9993	0.9762
16	0.9609	0.0206	0.9609	0.8614
17	0.9987	0.8142	0.9982	0.9142
18	0.9914	0.969	0.9875	0.9654
19	0.9885	0.9045	0.9829	0.9773
20	0.9702	0.2325	0.9625	0.9611
21	0.9959	0.3255	0.9958	0.9992
22	0.9786	0.1466	0.7901	0.9334
X	0.9922	0.0569	0.9892	0.9728
Y	0.9992	0.8394	0.9986	0.9728

Table 7.20. Comparison table for correlation coefficients for our model versus others.

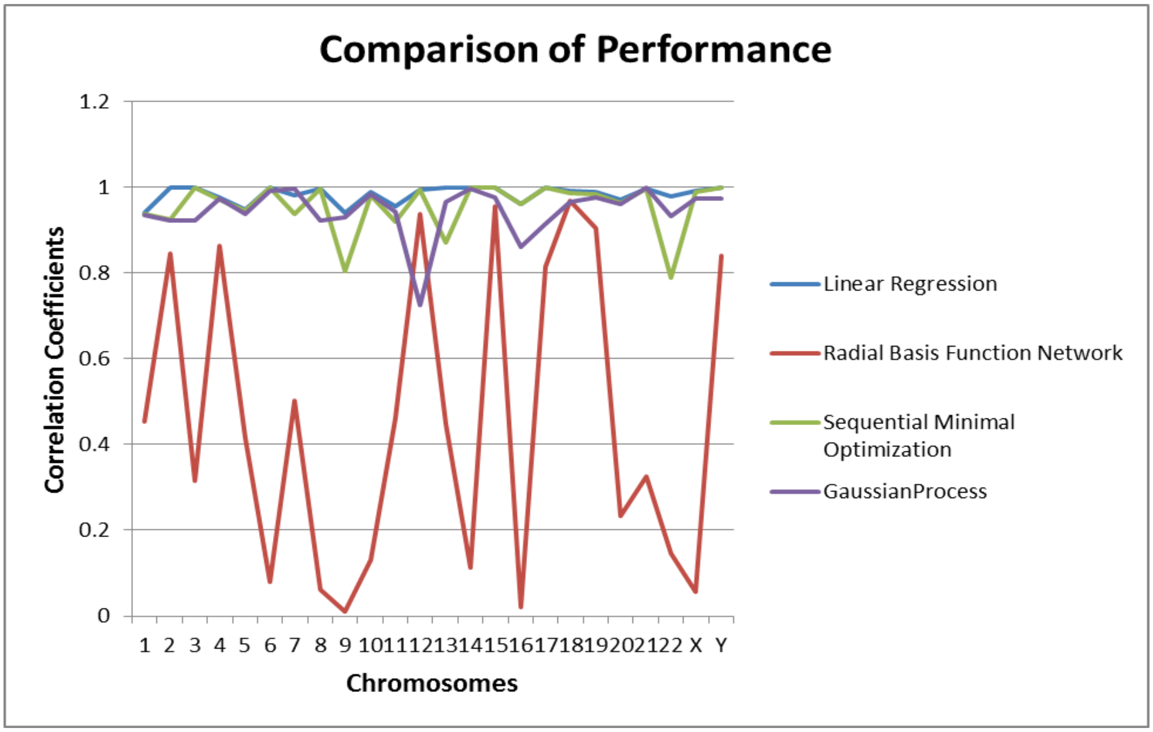


Figure 7.17 Comparison of performance of our model versus others.

7.1.5 Concluding Remarks

The identification and classification of mutations in patients are centric topics in today's cancer research. In this study, we first defined the problem of predicting the number of mutations on a chromosome for a certain patient, which is a fundamental problem in both research and clinical fields. Next, this problem is modeled mathematically as a linear regression problem. After applying a thorough analysis on each chromosome of each patient, we have the following observations:

1. As shown in the detailed statistical analysis in terms of error, the 95% confidence interval of each of the error measurement metrics are computed. For example, the correlation coefficient, which indicates the degree of matching of the proposed

framework with real data, is 0.9816 ± 0.009 with 95% confidence interval. In addition, our collaborator from medical domain thinks that the proposed framework achieves very good performance and opens new research opportunities for bioinformatics researchers and clinical doctors.

2. The proposed framework is relatively stable across patients. By analyzing errors of each patient, it can be seen that the errors are close to each other across different patients. Since patients are selected randomly, this observation gives us the confidence that the proposed framework is relatively stable and robust. It can be applied on new data collected from new patients.

3. In terms of the coefficients learned from different chromosomes on different patients, no significant global patterns are identified. This indicates that each individual patient has unique characteristic and it is hard to generalize a simple rule based strategy for all patients. On the other hand, the proposed framework which trains the linear regression model can adapt to each patient relatively well by representing the different characteristics of patients in the numerical weights gained from the model. This explains the reason why our framework achieves relatively stable performance across different patients. In addition, an observation is that the coefficient for the ChrToGenomeRatio feature is relatively small in all models, which indicate its' contribution is limited. More features are going to be added to enhance the model further in the future.

As we train the model with more longitudinal data, the estimation error should decrease. An example graph shown in the next Figure illustrates this statement. Additionally, as we collect more data points, in the future, we open the possibility to experiment other machine-learning algorithms as well.

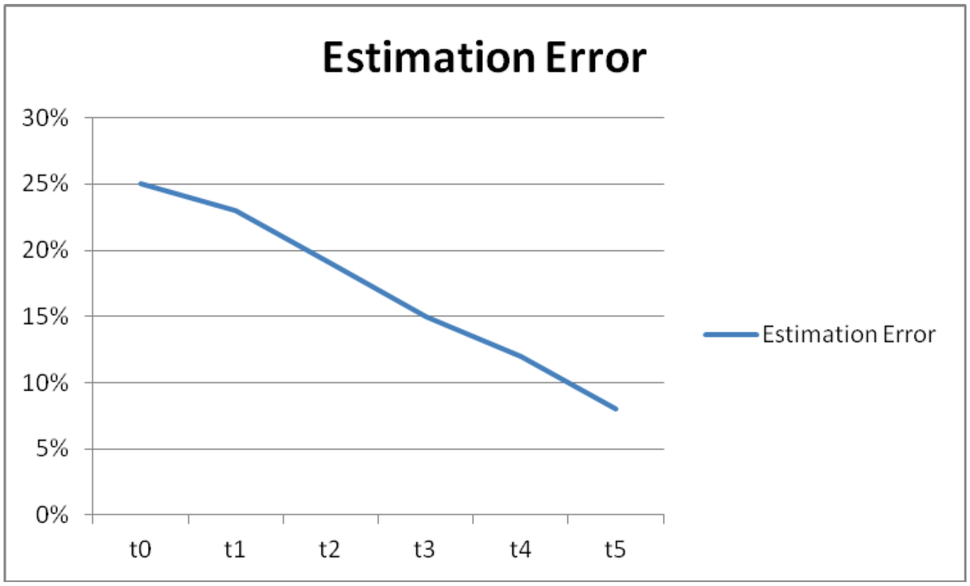


Figure 7.18 Potential estimation error graph as more data is obtained.

7.2 Prediction of location of mutations.

Table 7.2 refers to the locations of mutations, both observed and predicted, for all chromosomes in patient UPN 933124. Column two and three show the nucleotide positions of the mutations. Column four shows the name of the gene, if present in the region, in the case of observed mutations. Similarly, column five shows the name of the gene if present in the region corresponding to the predicted mutations. Not surprisingly, most of the areas containing the mutations (both observed and predicted) lie in the so-called intergenic regions of the genome. It is worth noting that the areas of the genome transcriptionally active are reported to be just 1% of the whole genome [122]. Close to 99% of the human genome does not code for proteins, and correspond to areas with either other known functionalities as regulation of gene expression (enhancers, promoters, inhibitors) or that do not have a defined function yet.

Chr	Observed Location	Location of the predicted mutation	Gene or genetic region observed	Gene or genetic region predicted
1	107357719	107358018	INTERGENIC	INTERGENIC
1	179307551	179309065	INTERGENIC	INTERGENIC
1	218018901	218020628	INTERGENIC	INTERGENIC
2	184186967	184186891	INTERGENIC	INTERGENIC
2	213886052	213884438	SPAG16	SPAG16
3	21310708	21307866	INTERGENIC	INTERGENIC
3	43135962	43138381	INTERGENIC	INTERGENIC
4	26362673	26362567	TBC1D19	TBC1D19
5	22595465	22595335	CDH12	CDH12
5	93857376	93859895	INTERGENIC	INTERGENIC
6	40467367	40464779	LRFN2	INTERGENIC
6	95631430	95629374	INTERGENIC	INTERGENIC
7	85476859	85474442	INTERGENIC	INTERGENIC
7	120917535	120914752	LOC392979	LOC392979
8	25022615	25022103	INTERGENIC	INTERGENIC
8	34167561	34165529	INTERGENIC	INTERGENIC
8	35476742	35478696	INTERGENIC	INTERGENIC
8	51256369	51253560	SNTG1	SNTG1
9	36887024	36889733	PAX5	PAX5
9	137816463	137818783	KCNT1	KCNT1
11	23583427	23584671	INTERGENIC	INTERGENIC
11	40183906	40182763	LRRC4C	LRRC4C
12	9702101	9703657	LOC374443	INTERGENIC
12	22283439	22281122	ST8SIA1	ST8SIA1
13	60889651	60891444	INTERGENIC	INTERGENIC
13	92466026	92467721	INTERGENIC	INTERGENIC
14	47961450	47963585	INTERGENIC	INTERGENIC

14	82608644	82607776	INTERGENIC	INTERGENIC
16	7941529	7943725	INTERGENIC	INTERGENIC
16	58050635	58052689	INTERGENIC	INTERGENIC
17	67861248	67859454	INTERGENIC	INTERGENIC
20	17261178	17262629	PCSK2	PCSK2
20	58861532	58863711	INTERGENIC	INTERGENIC
21	23370033	23372348	INTERGENIC	LOC100130310
X	86570002	86570879	INTERGENIC	INTERGENIC

Table 7.21. Location of mutations, both observed and predicted, for patient UPN 933124.

7.3 Accuracy discussion

Table 7.3 shows the results of the analysis for the patient UPN 933124. It presents the total number of mutations per genome, number of matches, mismatches, and accuracy of the prediction. When the prediction lies within the same region of an observed mutation, it constitutes a match. Conversely, if the location for the prediction does not correspond to the location of an observed mutation, or if the area of the observed mutation is not predicted at all, it is considered a mismatch.

We emphasize that the location of the prediction does not have to match the precise location of the observed mutation. Falling within the same gene boundaries is sufficient to classify it as a match. When the location of the mutation falls in an intergenic region (or a region devoid of a known functional activity) one criterion we use to classify it as a match, is that the prediction lies within less than 3000 nucleotides of the observed one on both directions. Although the distance of 3000 nucleotides may look arbitrary, to the best of our knowledge it is a reasonable assumption when there is no knowledge as to the functional features determined in such an intergenic region.

Matches	32
Mismatches	3
Total Mutations	35
Relative Accuracy	91%
Absolute Accuracy	76%

Table 7.22. Summary of results for patient UPN 933124.

With these criteria in mind, we came up with two accuracies. The first one is the "Relative Accuracy" which measures how many matches we had out of the predicted mutations we accounted for. For instance, in the case of the patient UPN 933124, we found out that even though the number of mutations at time of relapse was 42, we only predicted 35 mutations with 32 of them classified as matches. As a result, the accuracy was 91% (32/35). This accuracy is called "Relative Accuracy" because we are not considering the initial error given by the fact that the number of predicted mutations (35) was lower than the actual ones (42).

The "Absolute Accuracy" incorporates this initial prediction error. Thus, its value is less than the "Relative Accuracy". For instance, for this patient, the initial accuracy for the number of predicted mutations was 83% and the "Relative Accuracy" was 91%, meaning that we really could only predict for this patient 83% (91%) of the total number of mutations correctly, giving us an accuracy of 76%.

We carried out similar analysis for the seven remaining individuals. The accuracies are given as ranges with lower and upper values. This is explained by the fact that chromosomes models behave differently within the same individual, as well as, across patients. The verification of the predicted number of mutations versus the actual number of mutations gave us an accuracy value within a range of 75 to 84%. Likewise, the verification of predicted locations of mutations versus actual locations of mutations resulted in an accuracy value within the range of 69 to 88%.

CHAPTER 8

CONCLUSION

8.1 Research summary

The efforts and techniques used in this thesis represent the best methods we know to date in identifying meaningful mutational changes in leukemias that can help predict the following: 1.) responses to therapy, 2.) patients at increased risk of developing recurrences or progression of disease before and after treatment, 3.) driver mutations, 4.) passenger mutations, 5.) and evaluation of new potential druggable targets.

The n (# of patients) of 8 used in this study is actually a fundamental strength of the proposed study for the following reasons. Most research examining mutational spectra within cancers rely solely on snapshots, i.e. the presence or absence of mutations in tumors at one discrete time point. Though this information is valuable, it tells us nothing regarding the evolution of the disease. In particular, several high profile papers in the last 2-3 years have taken advantage of the TCGA database to mine solid tumors for mutations in these snapshots – work out of robust genome sequencing groups including from Washington University, St Louis (Science, 2012) and Harvard. All of these papers appear in very high end journals but use data from only one time point. A recent paper in Nature (2013) offered similar comparisons of such samples. It is more straightforward to have n's approaching 100-200+ patients for these less complex situations.

The gold standard really should be to get genomic mutational data for tumors from patients across their disease course over multiple time points. This longitudinal information represents the most valuable and relevant data in truly determining how to cure patients of disease. A paper in the very important journal Cell from early 2013 evaluated CLL patient tumor mutational spectra [123]. This work came out of a collaboration between the BROAD Institute of MIT and Harvard and Dana Farber Cancer Center, highly renowned for their leukemia patient volume. They had only snapshots of 140 or so patients for which they were able to identify mutation frequency. However, as the reviewers and authors acknowledge, they had a more valuable data set – 18 patients for which they had longitudinal mutational spectra from 2 time points – before and after chemotherapy. Only sequencing of coding areas were used in this effort. With 18 patients, they were able to make conclusions regarding the evolution of the CLL disease course, helping oncologists further understand the disease process to improve the lives of the patients. All agreed that this 18 patient data set was in many ways more valuable than the 140 patient snapshot mutational database. Additionally, NCI has recently made it a priority to better understand unique findings in smaller data series if the samples are difficult to come by. Longitudinal mutational data is very difficult to come by and perhaps offers the most important clues to disease processes.

In our work, we have been able to collect 8 valuable longitudinal mutational data from leukemia patients before and after multiple therapies. This collection may represent one of the largest series compiled in a robust manner. Our sequencing data is of the entire genome, unlike the coding only analysis from the Cell paper. Finally, our data approximates locations of mutational hot spots, which is very important because many changes in leukemia tumor DNA represent large frame shift alterations, deletions, insertions, or translocations.

8.2 Final statistical remarks

We analyze the mutations in the longitudinal dimension at the chromosome level. From this perspective, one patient actually provides 24 samples. Therefore, instead of 8 samples, we actually test our algorithm using 24x8 = 192 samples. This amount of data is significant in terms of regular statistical and data mining tasks. Compared with peer studies in the cancer research domain as mentioned by our collaborator in biomedical field, our number of patients as well as the measurements of mutations at multiple time shots for each patient are more significant than the recent published papers in Nature and Science. Even though the data samples are actually enough for significance, we still draw the conclusion prudently. In our statistical analysis of errors, the routing approach of estimating the confidence interval is using the normal distribution assumption. If that assumption is used here, the variation of our estimation is even smaller. However, given the current sample size, we still use the student-T distribution assumption to estimate errors. By using this, we draw a conservative conclusion. Given this observation, we would expect the performance of our algorithm on new data sets would be better.

As the cost of performing the whole genome sequencing is relatively expensive nowadays, it is infeasible to carry out the whole genome sequencing for the cancer patient routinely. On the other hand, the information of mutations in chromosomes at the molecular level gives precious information about patients and can guide clinical doctors in practice. The gap exists here should be addressed in order to improve the quality of treatments for patients. In addition, each patient has his/her individual characteristics which need to be considered. The proposed work provides a study from a relatively novel angle. First and foremost, a linear regression model is fit using three input features, which are “RatioStdMean”, “RatioGivenChr” and “ChrToGenomRation”. Since the number of patients’ visitings to hospitals is limited, the number of samples from each patient is sparse. Our model can take care of this issue since the model can be fit using features from as few as two time points. In addition, a model is built on each chromosome for each patient, this model adapts to each patient’s specialty and therefore is relatively robust. From the clinical point of view, the predictions of mutations can be meaningful for clinical doctors from the following two perspectives. First, the mutation frequency and patterns of the chromosomes of patients are useful information for medical doctors to make decision on the effectiveness of treatment and adjust treating strategy. Second, the mutation information can be further utilized as a factor to associate to classifying different types of cancer. A recent study in Nature actually shows that the mutation frequency has correlations with the different types of breast cancer. From this point of view, the mutation frequency predicted of each chromosome and the profile of mutation frequency predicted for each patient will offer the information for doctors to classify the cancer types and adjust treatment strategy accordingly. All these clinical significance motivates us for this research. Based on the success of applying the linear regression model in this study, we will extend the current work to a more broad problem, that is, we will try to fit a model across patients. Using this model, we can make predictions for the possibilities of mutations on a chromosome of a new patient based on the data collected from patients in our database. We believe this model, if successfully constructed, will provide more significant tools for medical practitioners.

8.3 Foundation for further discoveries.

Our concepts may have tremendous impact on a number of important areas relevant to future cancer care of patients. Specifically, our ideas may:

1. Increase our ability to predict which leukemia patients will develop relapses thru treatment.
2. Help predict which high risk patients (those having received chemotherapy for other cancers, patients on immunosuppressive therapy for organ transplantation, etc.) may develop leukemia down the road.
3. Help distinguish driver versus passenger mutations in leukemia.
4. Help identify mutations that may develop during leukemogenesis that are druggable targets.
5. Ultimately help leukemia patients have longer survival.

None of these results are trivial, especially in the setting of leukemias. There is a significant advantage to our work. Our models appear to be quite robust and relevant to leukemia disease states. We have employed longitudinal data sets from actual leukemia patients that we have used to build our models and validate them. The rationale behind our efforts is also convincing. Identification of approximate areas of deletion, insertion, and other mutations will be quite valuable as we begin to realize that most mutations in malignancy may not just be driven by single nucleotide polymorphisms.

8.4 Application of our framework in the treatment of patients by clinical oncologists

It is quite clear that our work will provide high impact understanding of the natural evolution of basal leukemia mutation rates and development before chemotherapy use and more importantly the state of mutations at the time of leukemia relapse. A vital question is how this data will be beneficial or relevant to a medical oncologist's practice on a day to day basis. From the initial applications of the predictive modeling of future mutation rates, we expect physician's to develop key information when they first diagnose a patient's leukemia, i.e. will there be an increase in mutation rate in the specific patient's leukemia at the time of relapse, will there be hotspots in the genome for this greater mutation rate that is associated with druggable targets, will there be at the time of relapse a completely different set of mutations (explaining why the patient's tumor relapses), etc. At the time of relapse, a medical oncologist has several decisions to make: keep the patient on the same leukemia drug but add a second drug, increase the levels of the drug in circulation, or change the drug completely. With the predictive mutation modeling, the medical oncologist can know in advance to potentially be prepared to take one of the above courses based on where the mutations may appear and in what frequency at the time of relapse. This knowledge, in the form of an algorithm of chemotherapy use, would be of great necessity in allowing medical oncologist's to optimize timing and response to patient leukemia relapses.

Finally, the greatest utility that the predictive model from this work may provide in the long run to medical oncologists and cancer specialists is a novel unbiased approach to genomically organize leukemias based on their natural evolution of mutation development. Historically, for many decades, leukemias were characterized (ALL, AML, CLL, CML) purely by histopathologic and cytopathologic analyses. This data was the only information used by physicians to decide how to treat patients with leukemias. This was a limited data set, however. With the identification of mutations associated with each type of leukemia, we are now incorporating this latter information into medical management practices. The use of targeted therapies including Gleevec was predicated on knowing if a leukemia possessed a bcr-abl mutation.

Despite the advances, we are still only making treatment decisions on a static evaluation of mutations at the time of diagnosis. The predictive algorithm provided in this work may open a totally new avenue by which cancer specialists can organize similar behaving leukemias into independent groups based on functional outcomes. Potentially, we would be able to characterize a leukemia as highly aggressive, aggressive, or minimally aggressive based on the prediction of mutation frequency/location at the time of relapse. Highly aggressive leukemias would be predicted to have greater mutation rates and sites at relapse, suggesting an earlier time to relapse. Minimally aggressive lesions would potentially have predicted lower rates of mutation and relapse at later time points. The power of this effort is that we would be taking into account all mutational changes (which may have distinct interactions), not just hypothesized driver mutations, to describe the ultimate output behavior of leukemias, before and after chemotherapy. A medical oncologist would be able to accurately predict how a patient with leukemia will do if he knows from the onset the likelihood of high or low mutation rates, predicting for highly or minimally aggressive disease. At the very least, with the creation of these "clades" or new ways to group leukemias based on functional outcomes, we will get a better sense of how many distinct subgroups or entities of leukemias are seen in the public. At the most, medical oncologists will be prepared to be more or less aggressive with changing chemotherapy regimens based on how the specific patient's leukemia is categorized. Staying ahead of the relapse conditions that promote increased mutational frequency by combining chemotherapies or changing the time of chemotherapy usage may ultimately improve progression free and overall survival, the gold standard for any cancer patient outcome.

8.5 Some potential enhancements.

Our framework has been built using Java technology and utilizes an external library called Weka for its prediction model. At the present time, we are using a linear regression algorithm provided by Weka. As we collect more patients' data, we will be able to perform more sophisticated prediction algorithms such as neural networks. Similarly as Weka evolves and more algorithms are included in this library, our framework will be able to extend its functionality by invoking these new features from the API. Thus, our architecture is flexible for future changes and/or enhancements.

SPECIAL RECOGNITIONS

We would like to show our great gratitude to Puneet Iyengar, MD, PhD for his valuable contribution to this work. He provided strong advice along the process of the writing. He aided us tremendously in performing tasks such as analysis and interpretation of data, and revising the manuscript critically.

APPENDICES

Patients Data before prediction

HIPAA limitations are not involved here as a problem since, as mentioned earlier, we are using the published data from Ding et al publication [13].

Prediction results for predicting number of mutations at each chromosome for every patient.

On the following pages, we can see the results of our framework when predicting the number of mutations to occur at each chromosome for each patient.

Data for Patient: A

Chromosome	**Observed Mutations**	**Predicted Mutations**
1	1	1
2	2	1
5	1	1
6	2	2
7	2	2
8	1	1
10	1	1
11	1	0
12	2	1
13	2	1
15	2	2
18	1	1
19	1	1
22	1	1
X	2	2
Total	**22**	**18**
Accuracy:	**81.82%**	

Data for Patient: B

Chromosome	**Observed Mutations**	**Predicted Mutations**
1	7	5
2	10	7
3	4	3
4	8	6
5	5	4
6	7	5
7	6	5
8	2	2
9	2	2
10	7	5
11	4	3
12	4	3
13	4	3
14	3	2
15	2	2
16	6	5
17	6	4
18	3	2
19	3	2
20	2	2
21	3	2
22	2	1
X	3	2
Y	2	2
Total	**105**	**79**
Accuracy	**75.24%**	

Data for Patient: C

Chromosome	Observed Mutations	Predicted Mutations
1	4	3
2	2	2
3	3	2
4	1	1
5	2	2
6	2	2
7	3	2
8	5	4
9	2	2
11	3	2
12	3	2
13	2	2
14	2	2
16	3	2
17	1	1
20	2	2
21	1	1
X	1	1
Total	**42**	**35**
Accuracy	**83.73%**	

Data for Patient: D

Chromosome	Observed Mutations	Predicted Mutations
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
8	2	1
10	3	1
16	2	2
17	1	1
18	1	1
X	1	1
Total	**15**	**12**
Accuracy	**80.00%**	

Data for Patient: E

Chromosome	Observed Mutations	Predicted Mutations
1	9	7
2	15	11
3	2	2
4	12	9
5	7	5
6	7	5
7	5	4
8	3	2
9	3	2
10	4	3
11	3	2
12	3	2
13	3	2
14	2	2
15	1	1
16	2	1
17	4	3
18	4	3
20	2	1
21	2	2
22	1	1
X	5	4
Total	**99**	**74**
Accuracy	**74.75%**	

Data for Patient: F

Chromosome	**Observed Mutations**	**Predicted Mutations**
1	5	4
2	14	11
3	6	5
4	6	4
5	3	2
6	10	8
7	5	4
8	4	3
9	6	5
10	3	2
11	5	4
12	4	3
13	4	3
14	4	3
15	2	2
17	3	2
18	5	4
19	2	2
20	2	1
21	1	1
22	1	1
X	4	3
Y	1	1
Total	**100**	**78**
Accuracy	**78.00%**	

Data for Patient: G

Chromosome	**Observed Mutations**	**Predicted Mutations**
1	1	1
2	1	1
4	1	1
7	1	1
9	1	1
12	1	0
13	1	1
14	1	1
16	1	0
18	1	1
19	1	1
21	1	1
Total	**12**	**10**
Accuracy	**83.33%**	

Data for Patient: H

Chromosome	**Observed Mutations**	**Predicted Mutations**
1	5	4
2	7	5
3	3	2
4	5	4
5	5	4
6	6	4
7	1	1
8	3	2
9	3	2
10	3	2
11	3	2
12	5	4
14	2	2
15	1	1
16	2	2
17	1	1
18	1	1
20	1	1
22	1	1
X	7	6
Total	**65**	**51**
Accuracy	**78.46%**	

Prediction results with summarized data including actual and predicted locations

Below we have the results for our prediction indicating the actual location of the mutation and the predicted one. Similarly, we show the actual gene involved and the predicted gene involved. We also have a column called "difference" to indicate the error of our estimations.

PATIENT A

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	105905196	105903822	-1374	INTERGENIC-REGION	INTERGENIC-REGION
2	103478832	103478707	-125	LOC728815	LOC728815
5	117968777	117966963	-1814	INTERGENIC-REGION	INTERGENIC-REGION
6	117330941	117333649	2708	RFXDC1	RFXDC1
6	148467197	148465214	-1983	INTERGENIC-REGION	INTERGENIC-REGION
7	120243306	120241310	-1996	TSPAN12	TSPAN12
7	141400429	141399010	-1419	MGAM	MGAM
8	6610082	6611300	1218	INTERGENIC-REGION	INTERGENIC-REGION
10	36300363	36297857	-2506	INTERGENIC-REGION	INTERGENIC-REGION
12	36181843	36180237	-1606	INTERGENIC-REGION	INTERGENIC-REGION
13	77038729	77037998	-731	SCEL	SCEL
15	35693981	35696254	2273	INTERGENIC-REGION	INTERGENIC-REGION
15	96495601	96496583	982	INTERGENIC-REGION	INTERGENIC-REGION
18	67016988	67019842	2854	INTERGENIC-REGION	INTERGENIC-REGION
19	15919303	15919555	252	INTERGENIC-REGION	INTERGENIC-REGION
22	46504211	46505232	1021	INTERGENIC-REGION	INTERGENIC-REGION
X	7449501	7449885	384	INTERGENIC-REGION	INTERGENIC-REGION
X	89181903	89179381	-2522	LOC100130134	INTERGENIC-REGION

PATIENT B

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	18730737	18731528	791	INTERGENIC-REGION	INTERGENIC-REGION
1	22831628	22830967	-661	INTERGENIC-REGION	INTERGENIC-REGION
1	89433568	89435914	2346	GBP4	GBP4
1	119892131	119890585	-1546	INTERGENIC-REGION	INTERGENIC-REGION
1	164527242	164529039	1797	INTERGENIC-REGION	INTERGENIC-REGION
2	4951026	4948442	-2584	INTERGENIC-REGION	INTERGENIC-REGION
2	18760057	18760752	695	INTERGENIC-REGION	INTERGENIC-REGION
2	41983085	41981902	-1183	INTERGENIC-REGION	INTERGENIC-REGION
2	52337923	52337545	-378	INTERGENIC-REGION	INTERGENIC-REGION
2	53366757	53368759	2002	INTERGENIC-REGION	INTERGENIC-REGION
2	114769762	114770766	1004	INTERGENIC-REGION	INTERGENIC-REGION
2	137663285	137664050	765	THSD7B	THSD7B
3	97423737	97422285	-1452	INTERGENIC-REGION	INTERGENIC-REGION
3	148730059	148727347	-2712	INTERGENIC-REGION	INTERGENIC-REGION
3	174114325	174114486	161	SPATA16	SPATA16
4	17783989	17783525	-464	INTERGENIC-REGION	INTERGENIC-REGION
4	18912208	18913676	1468	INTERGENIC-REGION	INTERGENIC-REGION
4	22759171	22761274	2103	LOC643751	LOC643751
4	31779477	31778159	-1318	INTERGENIC-REGION	INTERGENIC-REGION

4	103369618	103367568	-2050	INTERGENIC-REGION	INTERGENIC-REGION
4	132134033	132136548	2515	INTERGENIC-REGION	INTERGENIC-REGION
5	27542961	27545241	2280	INTERGENIC-REGION	INTERGENIC-REGION
5	73771482	73772512	1030	INTERGENIC-REGION	INTERGENIC-REGION
5	95272060	95274071	2011	ELL2	ELL2
5	154221315	154224153	2838	CNOT8	CNOT8
6	11556067	11553580	-2487	INTERGENIC-REGION	INTERGENIC-REGION
6	55539300	55536902	-2398	HMGCLL1	HMGCLL1
6	89788729	89786090	-2639	INTERGENIC-REGION	INTERGENIC-REGION
6	94679810	94677079	-2731	INTERGENIC-REGION	INTERGENIC-REGION
6	98935123	98933684	-1439	INTERGENIC-REGION	INTERGENIC-REGION
7	13405534	13402864	-2670	INTERGENIC-REGION	INTERGENIC-REGION
7	25018068	25018936	868	INTERGENIC-REGION	INTERGENIC-REGION
7	101609893	101612118	2225	CUX1	CUX1
7	101701575	101703838	2263	CUX1	CUX1
7	116559358	116559716	358	ST7	ST7
8	73514564	73511684	-2880	INTERGENIC-REGION	INTERGENIC-REGION
8	76476615	76478306	1691	INTERGENIC-REGION	INTERGENIC-REGION
9	7987216	7987623	407	INTERGENIC-REGION	INTERGENIC-REGION
9	11413168	11415994	2826	INTERGENIC-REGION	INTERGENIC-REGION
10	21774153	21775175	1022	INTERGENIC-REGION	INTERGENIC-REGION
10	24053538	24052997	-541	KIAA1217	KIAA1217
10	37158144	37159499	1355	INTERGENIC-REGION	INTERGENIC-REGION
10	52992328	52992927	599	PRKG1	PRKG1
10	100679446	100679728	282	HPSE2	HPSE2

11	26915201	26913996	-1205	INTERGENIC-REGION	INTERGENIC-REGION
11	27181250	27184040	2790	INTERGENIC-REGION	INTERGENIC-REGION
11	29982645	29982170	-475	INTERGENIC-REGION	INTERGENIC-REGION
12	61776966	61778798	1832	INTERGENIC-REGION	INTERGENIC-REGION
12	72620513	72623219	2706	INTERGENIC-REGION	INTERGENIC-REGION
12	82288736	82290376	1640	INTERGENIC-REGION	INTERGENIC-REGION
13	35326695	35328951	2256	DCLK1	DCLK1
13	53087241	53089305	2064	INTERGENIC-REGION	INTERGENIC-REGION
13	56856089	56856249	160	INTERGENIC-REGION	INTERGENIC-REGION
14	28018404	28019485	1081	INTERGENIC-REGION	INTERGENIC-REGION
14	69464698	69464658	-40	SMOC1	SMOC1
15	39166414	39169210	2796	INOC1	INOC1
15	88432939	88433205	266	IDH2	IDH2
16	26547638	26546235	-1403	INTERGENIC-REGION	INTERGENIC-REGION
16	30758901	30759255	354	LOC100129191	LOC100129191
16	70613693	70614741	1048	DHODH	DHODH
16	76093122	76094288	1166	INTERGENIC-REGION	INTERGENIC-REGION
16	80961258	80960904	-354	INTERGENIC-REGION	INTERGENIC-REGION
17	7334949	7337664	2715	POLR2A	POLR2A
17	10874727	10875744	1017	INTERGENIC-REGION	INTERGENIC-REGION
17	27701282	27698390	-2892	MIRN632	INTERGENIC-REGION
17	34007774	34005119	-2655	SNIP	SNIP
18	23266563	23267169	606	FLJ45994	FLJ45994
18	36665237	36665827	590	INTERGENIC-REGION	INTERGENIC-REGION
19	32989705	32990040	335	LOC642290	LOC642290

19	41696897	41697985	1088	ZNF260	ZNF260
20	17900146	17902446	2300	C20orf72	C20orf72
20	49730834	49729040	-1794	ATP9A	ATP9A
21	26930860	26928939	-1921	INTERGENIC-REGION	INTERGENIC-REGION
21	27215461	27214843	-618	ADAMTS5	ADAMTS5
22	18266604	18269168	2564	TXNRD2	TXNRD2
X	24850996	24849917	-1079	POLA1	POLA1
X	33727134	33727059	-75	INTERGENIC-REGION	INTERGENIC-REGION
Y	8487910	8485327	-2583	INTERGENIC-REGION	INTERGENIC-REGION
Y	16302321	16300318	-2003	INTERGENIC-REGION	INTERGENIC-REGION

PATIENT C

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	107357719	107358018	299	INTERGENIC-REGION	INTERGENIC-REGION
1	179307551	179309065	1514	INTERGENIC-REGION	INTERGENIC-REGION
1	218018901	218020628	1727	INTERGENIC-REGION	INTERGENIC-REGION
2	184186967	184186891	-76	INTERGENIC-REGION	INTERGENIC-REGION
2	213886052	213884438	-1614	SPAG16	SPAG16
3	21310708	21307866	-2842	INTERGENIC-REGION	INTERGENIC-REGION
3	43135962	43138381	2419	INTERGENIC-REGION	INTERGENIC-REGION
4	26362673	26362567	-106	TBC1D19	TBC1D19
5	22595465	22595335	-130	CDH12	CDH12
5	93857376	93859895	2519	INTERGENIC-REGION	INTERGENIC-REGION
6	40467367	40464779	-2588	LRFN2	INTERGENIC-REGION
6	95631430	95629374	-2056	INTERGENIC-REGION	INTERGENIC-REGION
7	85476859	85474442	-2417	INTERGENIC-REGION	INTERGENIC-REGION
7	120917535	120914752	-2783	LOC392979	LOC392979
8	25022615	25022103	-512	INTERGENIC-REGION	INTERGENIC-REGION
8	34167561	34165529	-2032	INTERGENIC-REGION	INTERGENIC-REGION
8	35476742	35478696	1954	INTERGENIC-REGION	INTERGENIC-REGION
8	51256369	51253560	-2809	SNTG1	SNTG1

9	36887024	36889733	2709	PAX5	PAX5
9	137816463	137818783	2320	KCNT1	KCNT1
11	23583427	23584671	1244	INTERGENIC-REGION	INTERGENIC-REGION
11	40183906	40182763	-1143	LRRC4C	LRRC4C
12	9702101	9703657	1556	LOC374443	INTERGENIC-REGION
12	22283439	22281122	-2317	ST8SIA1	ST8SIA1
13	60889651	60891444	1793	INTERGENIC-REGION	INTERGENIC-REGION
13	92466026	92467721	1695	INTERGENIC-REGION	INTERGENIC-REGION
14	47961450	47963585	2135	INTERGENIC-REGION	INTERGENIC-REGION
14	82608644	82607776	-868	INTERGENIC-REGION	INTERGENIC-REGION
16	7941529	7943725	2196	INTERGENIC-REGION	INTERGENIC-REGION
16	58050635	58052689	2054	INTERGENIC-REGION	INTERGENIC-REGION
17	67861248	67859454	-1794	INTERGENIC-REGION	INTERGENIC-REGION
20	17261178	17262629	1451	PCSK2	PCSK2
20	58861532	58863711	2179	INTERGENIC-REGION	INTERGENIC-REGION
21	23370033	23372348	2315	INTERGENIC-REGION	LOC100130310
X	86570002	86570879	877	INTERGENIC-REGION	INTERGENIC-REGION

PATIENT D

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
2	1119223	1118090	-1133	SNTG2	SNTG2
3	183065520	183064291	-1229	INTERGENIC-REGION	INTERGENIC-REGION
4	185168709	185166707	-2002	STOX2	STOX2
5	162177435	162174604	-2831	INTERGENIC-REGION	INTERGENIC-REGION
6	93462438	93461387	-1051	INTERGENIC-REGION	INTERGENIC-REGION
8	3293986	3295797	1811	CSMD1	CSMD1
10	84453341	84452336	-1005	NRG3	NRG3
16	61747374	61747530	156	INTERGENIC-REGION	INTERGENIC-REGION
16	62989048	62988198	-850	INTERGENIC-REGION	INTERGENIC-REGION
17	22045352	22042405	-2947	INTERGENIC-REGION	INTERGENIC-REGION
18	42887822	42889970	2148	HDHD2	HDHD2
X	121434334	121431398	-2936	INTERGENIC-REGION	INTERGENIC-REGION

PATIENT E

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	44944905	44945092	187	MGC33556	MGC33556
1	56758799	56755847	-2952	PPAP2B	PPAP2B
1	62329867	62328205	-1662	INADL	INADL
1	80186452	80184770	-1682	INTERGENIC-REGION	INTERGENIC-REGION
1	162652104	162653254	1150	INTERGENIC-REGION	INTERGENIC-REGION
1	193708003	193707136	-867	INTERGENIC-REGION	INTERGENIC-REGION
1	193987679	193985578	-2101	INTERGENIC-REGION	INTERGENIC-REGION
2	16261487	16259478	-2009	INTERGENIC-REGION	INTERGENIC-REGION
2	79424915	79425833	918	INTERGENIC-REGION	INTERGENIC-REGION
2	81008756	81008969	213	INTERGENIC-REGION	INTERGENIC-REGION
2	83365574	83367876	2302	INTERGENIC-REGION	INTERGENIC-REGION
2	121062191	121062952	761	INTERGENIC-REGION	INTERGENIC-REGION
2	121512897	121514104	1207	INTERGENIC-REGION	INTERGENIC-REGION
2	125156383	125157880	1497	CNTNAP5	CNTNAP5
2	125314902	125317068	2166	CNTNAP5	CNTNAP5
2	132126629	132126964	335	INTERGENIC-REGION	INTERGENIC-REGION
2	136723675	136726085	2410	INTERGENIC-REGION	INTERGENIC-REGION
2	151445176	151443514	-1662	INTERGENIC-REGION	INTERGENIC-REGION

3	1887762	1887661	-101	INTERGENIC-REGION	INTERGENIC-REGION
3	175001625	175001097	-528	NLGN1	NLGN1
4	20084861	20083013	-1848	SLIT2	SLIT2
4	30772025	30773500	1475	INTERGENIC-REGION	INTERGENIC-REGION
4	31398117	31399931	1814	INTERGENIC-REGION	INTERGENIC-REGION
4	93687647	93687017	-630	GRID2	GRID2
4	96814327	96813949	-378	INTERGENIC-REGION	INTERGENIC-REGION
4	105371233	105370352	-881	INTERGENIC-REGION	INTERGENIC-REGION
4	111770497	111769372	-1125	PITX2	PITX2
4	137387938	137387568	-370	INTERGENIC-REGION	INTERGENIC-REGION
4	137620857	137618498	-2359	INTERGENIC-REGION	INTERGENIC-REGION
5	21248739	21250383	1644	INTERGENIC-REGION	INTERGENIC-REGION
5	22482248	22483703	1455	CDH12	CDH12
5	22495984	22494710	-1274	CDH12	CDH12
5	22744459	22743789	-670	CDH12	CDH12
5	45004983	45004149	-834	INTERGENIC-REGION	INTERGENIC-REGION
6	20902367	20899506	-2861	CDKAL1	CDKAL1
6	50040060	50040552	492	INTERGENIC-REGION	INTERGENIC-REGION
6	66897241	66900208	2967	INTERGENIC-REGION	INTERGENIC-REGION
6	87126465	87126336	-129	INTERGENIC-REGION	INTERGENIC-REGION
6	94961835	94961835	0	INTERGENIC-REGION	INTERGENIC-REGION
7	13598141	13595372	-2769	INTERGENIC-REGION	INTERGENIC-REGION
7	47290289	47289020	-1269	TNS3	TNS3
7	68773299	68774264	965	AUTS2	AUTS2
7	147006720	147009358	2638	CNTNAP2	CNTNAP2

8	4788849	4787918	-931	CSMD1	CSMD1
8	23806889	23805931	-958	INTERGENIC-REGION	INTERGENIC-REGION
9	12110080	12111967	1887	INTERGENIC-REGION	INTERGENIC-REGION
9	28701959	28704866	2907	LINGO2	LINGO2
10	52502567	52503933	1366	PRKG1	PRKG1
10	55997651	55997740	89	PCDH15	PCDH15
10	56392820	56391742	-1078	INTERGENIC-REGION	INTERGENIC-REGION
11	61267618	61269091	1473	DAGLA	DAGLA
11	64146939	64144929	-2010	NRXN2	NRXN2
12	83033586	83033124	-462	INTERGENIC-REGION	INTERGENIC-REGION
12	93956009	93955103	-906	NR2C1	NR2C1
13	57139096	57136968	-2128	PCDH17	PCDH17
13	62444093	62442127	-1966	INTERGENIC-REGION	INTERGENIC-REGION
14	46290472	46290215	-257	INTERGENIC-REGION	INTERGENIC-REGION
14	105609184	105611251	2067	INTERGENIC-REGION	INTERGENIC-REGION
15	76418530	76417893	-637	INTERGENIC-REGION	INTERGENIC-REGION
16	26981824	26984046	2222	INTERGENIC-REGION	INTERGENIC-REGION
17	3004064	3003465	-599	OR1P1P	INTERGENIC-REGION
17	13501009	13499899	-1110	INTERGENIC-REGION	INTERGENIC-REGION
17	31379333	31379326	-7	INTERGENIC-REGION	INTERGENIC-REGION
18	5846231	5846301	70	INTERGENIC-REGION	INTERGENIC-REGION
18	29678931	29677804	-1127	INTERGENIC-REGION	INTERGENIC-REGION
18	37355263	37356895	1632	INTERGENIC-REGION	INTERGENIC-REGION
20	12364137	12361276	-2861	INTERGENIC-REGION	INTERGENIC-REGION
21	21114306	21113277	-1029	INTERGENIC-REGION	INTERGENIC-REGION

21	27426278	27427344	1066	INTERGENIC-REGION	INTERGENIC-REGION
22	49192948	49194052	1104	SAPS2	SAPS2
X	23373203	23375572	2369	INTERGENIC-REGION	INTERGENIC-REGION
X	88250191	88250874	683	INTERGENIC-REGION	INTERGENIC-REGION
X	97368810	97370397	1587	INTERGENIC-REGION	INTERGENIC-REGION
X	116745872	116748160	2288	INTERGENIC-REGION	INTERGENIC-REGION

PATIENT F

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	92405867	92406858	991	INTERGENIC-REGION	INTERGENIC-REGION
1	92814472	92813494	-978	EVI5	EVI5
1	96583536	96582384	-1152	LOC100132258	LOC100132258
1	111864509	111865538	1029	ADORA3	ADORA3
2	18769914	18768780	-1134	INTERGENIC-REGION	INTERGENIC-REGION
2	22359545	22361976	2431	INTERGENIC-REGION	INTERGENIC-REGION
2	50857536	50855234	-2302	NRXN1	NRXN1
2	66491986	66494804	2818	INTERGENIC-REGION	INTERGENIC-REGION
2	83614631	83612098	-2533	INTERGENIC-REGION	INTERGENIC-REGION
2	84115889	84116667	778	INTERGENIC-REGION	INTERGENIC-REGION
2	126135788	126137404	1616	INTERGENIC-REGION	INTERGENIC-REGION

2	167238182	167237566	-616	INTERGENIC-REGION	INTERGENIC-REGION
2	172352355	172354666	2311	SLC25A12	SLC25A12
2	173346781	173345732	-1049	RAPGEF4	RAPGEF4
2	184179988	184177573	-2415	INTERGENIC-REGION	INTERGENIC-REGION
3	43881260	43883533	2273	INTERGENIC-REGION	INTERGENIC-REGION
3	46659901	46659510	-391	INTERGENIC-REGION	INTERGENIC-REGION
3	96394238	96395455	1217	INTERGENIC-REGION	INTERGENIC-REGION
3	139135471	139135228	-243	INTERGENIC-REGION	INTERGENIC-REGION
3	180628899	180631315	2416	GNB4	GNB4
4	27863801	27863829	28	INTERGENIC-REGION	INTERGENIC-REGION
4	65007608	65008077	469	INTERGENIC-REGION	INTERGENIC-REGION
4	123096911	123098079	1168	INTERGENIC-REGION	INTERGENIC-REGION
4	132093722	132093932	210	INTERGENIC-REGION	INTERGENIC-REGION
5	23368167	23369740	1573	INTERGENIC-REGION	INTERGENIC-REGION
5	56807873	56807941	68	INTERGENIC-REGION	INTERGENIC-REGION
6	126353	129210	2857	INTERGENIC-REGION	INTERGENIC-REGION
6	9844264	9845543	1279	INTERGENIC-REGION	INTERGENIC-REGION
6	68090304	68090598	294	INTERGENIC-REGION	INTERGENIC-REGION
6	77430774	77430160	-614	INTERGENIC-REGION	INTERGENIC-REGION
6	86029763	86029106	-657	INTERGENIC-REGION	INTERGENIC-REGION
6	94768681	94769618	937	INTERGENIC-REGION	INTERGENIC-REGION
6	109550234	109548825	-1409	C6orf182	C6orf182
6	112245364	112244333	-1031	FYN	FYN
7	13880370	13882510	2140	INTERGENIC-REGION	INTERGENIC-REGION
7	39559272	39559474	202	INTERGENIC-REGION	INTERGENIC-REGION

7	55323684	55322359	-1325	INTERGENIC-REGION	INTERGENIC-REGION
7	104147320	104147939	619	LHFPL3	LHFPL3
8	63596628	63594602	-2026	NKAIN3	NKAIN3
8	72857393	72854439	-2954	INTERGENIC-REGION	INTERGENIC-REGION
8	118381845	118381895	50	INTERGENIC-REGION	INTERGENIC-REGION
9	454456	453609	-847	DOCK8	DOCK8
9	80165699	80166263	564	INTERGENIC-REGION	INTERGENIC-REGION
9	110146992	110149359	2367	INTERGENIC-REGION	INTERGENIC-REGION
9	118884413	118884831	418	ASTN2	ASTN2
9	126126705	126128087	1382	NEK6	NEK6
10	2683387	2682198	-1189	INTERGENIC-REGION	INTERGENIC-REGION
10	6584079	6585226	1147	PRKCQ	PRKCQ
11	10203546	10200736	-2810	SBF2	SBF2
11	40438832	40440545	1713	INTERGENIC-REGION	INTERGENIC-REGION
11	55428174	55430431	2257	OR5W1P	INTERGENIC-REGION
11	87891301	87893310	2009	GRM5	GRM5
12	38297190	38294864	-2326	ABCD2	ABCD2
12	44240246	44239417	-829	INTERGENIC-REGION	INTERGENIC-REGION
12	79813281	79812561	-720	LIN7A	LIN7A
13	55417148	55419652	2504	INTERGENIC-REGION	INTERGENIC-REGION
13	70427767	70426376	-1391	INTERGENIC-REGION	INTERGENIC-REGION
13	81090538	81090305	-233	INTERGENIC-REGION	INTERGENIC-REGION
14	31667574	31668810	1236	ARHGAP5	ARHGAP5
14	33154434	33153004	-1430	NPAS3	NPAS3
14	56091064	56089536	-1528	INTERGENIC-REGION	INTERGENIC-REGION

15	40851227	40850495	-732	TTBK2	TTBK2
15	79991574	79992556	982	INTERGENIC-REGION	INTERGENIC-REGION
17	39843379	39841478	-1901	GPATCH8	GPATCH8
17	45343100	45344560	1460	INTERGENIC-REGION	INTERGENIC-REGION
18	5053185	5051558	-1627	INTERGENIC-REGION	INTERGENIC-REGION
18	5938070	5938026	-44	INTERGENIC-REGION	INTERGENIC-REGION
18	11401121	11401673	552	INTERGENIC-REGION	INTERGENIC-REGION
18	39337589	39338647	1058	INTERGENIC-REGION	INTERGENIC-REGION
19	5893322	5895546	2224	RANBP3	RANBP3
19	43341487	43339217	-2270	SIPA1L3	SIPA1L3
20	13210300	13213050	2750	C20orf82	C20orf82
21	42042422	42039541	-2881	RIPK4	RIPK4
22	47687788	47685465	-2323	INTERGENIC-REGION	INTERGENIC-REGION
X	93658417	93660653	2236	INTERGENIC-REGION	INTERGENIC-REGION
X	104960175	104962733	2558	NRK	NRK
X	111343447	111345610	2163	ZCCHC16	ZCCHC16
Y	7721445	7723250	1805	INTERGENIC-REGION	INTERGENIC-REGION

PATIENT G

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	79586157	79587862	1705	INTERGENIC-REGION	INTERGENIC-REGION
2	147777631	147775238	-2393	INTERGENIC-REGION	INTERGENIC-REGION
4	142853353	142851908	-1445	IL15	IL15
7	19896920	19895976	-944	INTERGENIC-REGION	INTERGENIC-REGION
9	25271855	25270748	-1107	INTERGENIC-REGION	INTERGENIC-REGION
13	91834507	91836278	1771	GPC5	GPC5
14	25426007	25427524	1517	INTERGENIC-REGION	INTERGENIC-REGION
18	6593666	6595162	1496	INTERGENIC-REGION	INTERGENIC-REGION
19	1339089	1339688	599	NDUFS7	NDUFS7
21	30692169	30694843	2674	INTERGENIC-REGION	INTERGENIC-REGION

PATIENT H

Chr	RelapseLocation	PredictedLocation	Difference	RelapseGene	PredictedGene
1	28395117	28396128	1011	INTERGENIC-REGION	INTERGENIC-REGION
1	55348624	55345772	-2852	USP24	USP24
1	80466160	80466553	393	INTERGENIC-REGION	INTERGENIC-REGION
1	104112500	104114609	2109	INTERGENIC-REGION	INTERGENIC-REGION
2	4358173	4355389	-2784	INTERGENIC-REGION	INTERGENIC-REGION
2	14285436	14284784	-652	INTERGENIC-REGION	INTERGENIC-REGION
2	106766704	106769370	2666	INTERGENIC-REGION	INTERGENIC-REGION
2	119250670	119251555	885	INTERGENIC-REGION	INTERGENIC-REGION
2	155081458	155083587	2129	INTERGENIC-REGION	INTERGENIC-REGION
3	1799169	1800181	1012	INTERGENIC-REGION	INTERGENIC-REGION
3	20956204	20956611	407	INTERGENIC-REGION	INTERGENIC-REGION
4	46955181	46952231	-2950	GABRB1	GABRB1
4	149316450	149317894	1444	NR3C2	NR3C2
4	150025553	150024817	-736	INTERGENIC-REGION	INTERGENIC-REGION
4	158256256	158253511	-2745	GLRB	GLRB
5	15823239	15821488	-1751	FBXL7	FBXL7
5	29515287	29512926	-2361	INTERGENIC-REGION	INTERGENIC-REGION

5	104466311	104465825	-486	INTERGENIC-REGION	INTERGENIC-REGION
5	113840798	113840471	-327	KCNN2	KCNN2
6	9855937	9853948	-1989	INTERGENIC-REGION	INTERGENIC-REGION
6	38352606	38355567	2961	BTBD9	BTBD9
6	96112105	96114311	2206	INTERGENIC-REGION	INTERGENIC-REGION
6	98495373	98496129	756	INTERGENIC-REGION	INTERGENIC-REGION
7	141689176	141686364	-2812	TRB@	TRB@
8	2568547	2567553	-994	INTERGENIC-REGION	INTERGENIC-REGION
8	31261001	31259469	-1532	INTERGENIC-REGION	INTERGENIC-REGION
9	75985332	75984229	-1103	INTERGENIC-REGION	INTERGENIC-REGION
9	82308637	82311378	2741	INTERGENIC-REGION	INTERGENIC-REGION
10	22510615	22508285	-2330	INTERGENIC-REGION	INTERGENIC-REGION
10	23382721	23380837	-1884	INTERGENIC-REGION	INTERGENIC-REGION
11	26651719	26654173	2454	SLC5A12	SLC5A12
11	82054955	82055363	408	INTERGENIC-REGION	INTERGENIC-REGION
12	3464622	3467549	2927	INTERGENIC-REGION	INTERGENIC-REGION
12	11883491	11884936	1445	ETV6	ETV6
12	43022207	43021782	-425	TMEM117	TMEM117
12	72136658	72138814	2156	INTERGENIC-REGION	INTERGENIC-REGION
14	20830373	20828361	-2012	RPGRIP1	RPGRIP1
14	43499979	43498964	-1015	INTERGENIC-REGION	INTERGENIC-REGION
15	76806819	76809320	2501	LOC646934	LOC646934
16	59519213	59518116	-1097	INTERGENIC-REGION	INTERGENIC-REGION
16	84039293	84040691	1398	INTERGENIC-REGION	INTERGENIC-REGION
17	22057902	22059434	1532	INTERGENIC-REGION	INTERGENIC-REGION

18	13846987	13844333	-2654	INTERGENIC-REGION	INTERGENIC-REGION
20	43469052	43469585	533	SYS1-DBNDD2	SYS1-DBNDD2
22	24752889	24753260	371	MYO18B	MYO18B
X	6998334	6996308	-2026	HDHD1A	HDHD1A
X	104572042	104570775	-1267	IL1RAPL2	IL1RAPL2
X	114186799	114188858	2059	INTERGENIC-REGION	INTERGENIC-REGION
X	121017574	121014781	-2793	INTERGENIC-REGION	INTERGENIC-REGION
X	127355500	127356514	1014	INTERGENIC-REGION	INTERGENIC-REGION
X	127868848	127869973	1125	INTERGENIC-REGION	INTERGENIC-REGION

BIBLIOGRAPHY

[1] Patel LR, Nykter M, Chen K, Zhang W: Cancer genome sequencing: Understanding malignancy as a disease of the genome, its conformation, and its evolution. Cancer letters 2012.

[2] Campo E: Whole genome profiling and other high throughput technologies in lymphoid neoplasms--current contributions and future hopes. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2013, 26 Suppl 1:S97-s110.

[3] Mardis ER, Wilson RK. Cancer genome sequencing: a review. Hum Mol Genet. 2009 Oct 15;18(R2):R163-8. doi: 10.1093/hmg/ddp396.

[4] Welch JS, Link DC: Genomics of AML: clinical applications of next-generation sequencing. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program 2011, 2011:30-35.

[5] Kilpivaara O, Aaltonen LA: Diagnostic cancer genome sequencing and the contribution of germline variants. Science (New York, NY) 2013, 339(6127):1559-1562.

[6] Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, Bergstralh EJ, Kollmeyer T, Fink S, Haddad Z et al: Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PloS one 2013, 8(6):e66855.

[7] Stretch C, Khan S, Asgarian N, Eisner R, Vaisipour S, Damaraju S, Graham K, Bathe OF, Steed H, Greiner R et al: Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. PloS one 2013, 8(6):e65380.

[8] Bao ZS, Zhang CB, Wang HJ, Yan W, Liu YW, Li MY, Zhang W: Whole-genome mRNA expression profiling identifies functional and prognostic signatures in patients with mesenchymal glioblastoma multiforme. CNS neuroscience & therapeutics 2013, 19(9):714-720.

[9] Sanz-Pamplona R, Berenguer A, Cordero D, Riccadonna S, Sole X, Crous-Bou M, Guino E, Sanjuan X, Biondo S, Soriano A et al: Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. PloS one 2012, 7(11):e48877.

[10] Srivastava M, Khurana P, Sugadev R: Lung cancer signature biomarkers: tissue specific semantic similarity based clustering of digital differential display (DDD) data. BMC research notes 2012, 5:617.

[11] Bedolla RG, Gong J, Prihoda TJ, Yeh IT, Thompson IM, Ghosh R, Kumar AP: Predictive value of Sp1/Sp3/FLIP signature for prostate cancer recurrence. PloS one 2012, 7(9):e44917.

[12] Mustacchi G, Sormani MP, Bruzzi P, Gennari A, Zanconati F, Bonifacio D, Monzoni A, Morandi L: Identification and validation of a new set of five genes for prediction of risk in early breast cancer. International journal of molecular sciences 2013, 14(5):9686-9702.

[13] Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD et al: Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012, 481(7382):506-510.

[14] Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. Nucleic Acids Res. 2013 Jan;41(Database issue):D48-55. doi: 10.1093/nar/gks1236. Epub 2012 Nov 30. Ensembl 2013.

[15] Christian Del Fabbro. Repeated sequences in bioinformatics: assembly, annotation and alignments. PhD Dissertation. Universita Degli Studi Di Udine, 2009.

[16] J.D. Watson and F.H.C. Crick, A Structure for Deoxyribose Nucleic Acid. Nature 171, 1953.

[17] R. Langridge, W. E. Seeds, H. R. Wilson, C. W. Hooper, M. H. F. Wilkins, and L. D. Hamilton.Molecular Structure of Deoxypentose Nucleic Acids. Nature 171, 1953.

[18] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walters. Molecular Biology of the Cell. New York and London: Garland Science, 4th edition, 2002.

[19] Grechko VV. Repeated DNA sequences as an engine of biological diversification. Molecular Biology (Mosk). 2011 Sep-Oct;45(5):765-92.

[20] Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annual Review Genomics Human Genetics 2007;8:241-59.

[21] M.B. Gerstein, C. Bruce, J.S. Rozowsky, D. Zheng, J. Du, J.O. Korbel, O. Emanuelsson, Z.D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? History and updated definition. Genome Res, 17(6):669{681, Jun 2007.

[22] Rister J, Desplan C. Deciphering the genome's regulatory code: the many languages of DNA. Bioessays. 2010 May;32(5):381-4. doi: 10.1002/bies.200900197.

[23] http://www.cancer.gov

[24] http://biologypop.com/dna-mutation-info/

[25] Alex van Belkum, Stewart Scherer, Loek van Alphen, and Henri Verbrugh, “Short-Sequence DNA Repeats in Prokaryotic Genomes,” Microbiology and Molecular Biology Reviews, 62(2), 1998.

[26] Stefan Kurtz, Jomuna V. Choudhuri, Enno Ohlebusch, Chris Schleiermacher, Jens Stoye, and Robert Giegerich, “REPuter: The Manifold Applications of Repeat Analysis on a Genomic Scale,” Nucleic Acids Research, 29(22) (2001) 4633–4642.

[27] Guillaume Achaz, Eric Coissac, Pierre Netter, and Eduardo P. C. Rocha, “Associations between Inverted Repeats and the Structural Evolution of Bacterial Genomes,” Genetics, 164(4), 2003.

[28] E. Lander et al., “Initial Sequencing and Analysis of theHuman Genome,” Nature, 409(6822), 2001.

[29] Eduardo P. C. Rocha, Antoine Danchin, and Alain Viari, “Analysis of Long Repeats in Bacterial Genomes Reveals Alternative Evolutionary Mechanisms in Bacillus subtilis and Other Competent Prokaryotes,” Molecular Biology and Evolution, 16(9), 1999.

[30] Masataka Tsuge, Ryuji Hamamoto, Fabio Pittella Silva, Yozo Ohnishi, Kazuaki Chayama, Naoyuki Kamatani, Yoichi Furukawa, and Yusuke Nakamura, “A Variable Number of Tandem Repeats Polymorphism in an E2F-1 Binding Element in the 5_ flanking region of SMYD3 is a Risk Factor for Human Cancers,” Nature Genetics, 37, 2005.

[31] Subbaya Subramanian, Rakesh Mishra, and Lalji Singh, “Genome-Wide Analysis of Microsatellite Repeats in Humans: Their Abundance and Density in Specific Genomic Regions,”Genome Biology, 4(2), 2003.

[32] A. Blanes and SJ. Diaz-Cano. Complementary analysis of microsatellite tumor profile and mismatch repair defects in colorectal carcinomas. World J. Gastroenterol. 2006.

[33] Lucy R. Yates and Peter J. Campbell. Evolution of the Cancer Genome. Nat Rev Genet. 2012.

[34] http://www.nature.com/scitable/content/examples-of-genetic-diseases-caused-by-expanding-27926

[35] Kathleen H. Burns and Jef D. Boeke. Human Transposon Tectonics. Cell Volum 149, Issue 4, May 11, 2012.

[36] Fang Y, Yao Q, Chen Z, Xiang J, William FE, Gibbs RA, Chen C: Genetic and molecular alterations in pancreatic cancer: Implications for personalized medicine. Medical science monitor: international medical journal of experimental and clinical research 2013, 19:916-926.

[37] Avery OT, Macleod CM, McCarty M: Studies On The Chemical Nature Of The Substance Inducing Transformation Of Pneumococcal Types : Induction Of Transformation By A Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III. The Journal of experimental medicine 1944, 79(2):137-158.

[38] Steinman RM, Moberg CL: A triple tribute to the experiment that transformed biology. The Journal of experimental medicine 1994, 179(2):379-384.

[39] Watson JD, Crick FH: Genetical implications of the structure of deoxyribonucleic acid. Nature 1953, 171(4361):964-967.

[40] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al: Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.

[41] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al: The sequence of the human genome. Science (New York, NY) 2001, 291(5507):1304-1351.

[42] Finishing the euchromatic sequence of the human genome. Nature 2004, 431(7011):931-945.

[43] Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ, de Jong PJ: A bacterial artificial chromosome library for sequencing the complete human genome. Genome research 2001, 11(3):483-496.

[44] Mardis ER, Wilson RK: Cancer genome sequencing: a review. Human molecular genetics 2009, 18(R2):R163-168.

[45] Korf BR: Integration of genomics into medical practice. Discovery medicine 2013, 16(89):241-248.

[46] Jeff Bizzaro and Kenneth Marx, "Poly: A Quantitative Analysis Tool for Simple Sequence Repeat (SSR) Tracts in DNA," Biomed Central Bioinformatics, 4(1), 2003.

[47] AkitoTaneda, “Adplot: Detection and Visualization of Repetitive Patterns in Complete Genomes,” Bioinformatics, 20(5) (2004) 701–708.

[48] Jeff Reneker and Chi-Ren Shyu, “Refined Repetitive Sequence Searches Utilizing a Fast Hash Function and Cross Species Information Retrievals,” Biomed Central Bioinformatics, 6(1), 2005.

[49] Alkes L. Price, Neil C. Jones, and Pavel A. Pevzner, “De novo Identification of Repeat Families in Large Genomes,” Bioinformatics, 21(suppl_1), 2005.

[50] S. Bachellier, J.-M. Clément, and M. Hofnung, “Short Palindrome Repetitive DNA Elements in Enterobacteria: A Survey,” Research in Microbiology, 150, 1999.

[51] Tetsuhiko Yoshida, Nobuaki Obata1, and Kenji Oosawa, “Color-Coding Reveals Tandem Repeats in the Escherichia coli Genome,” Journal of Molecular Biology, 298, 2000.

[52] G. Benson, “Tandem Repeats Finder: A Program to Analyze DNA Sequences,” Nucleic Acids Research, 27(2), 1999.

[53] Adalberto T. Castelo, Wellington Martins, and Guang R. Gao, “TROLL: Tandem Repeat Occurrence Locator,” Bioinformatics, 18(4), 2002.

[54] Roman Kolpakov, Ghizlane Bana, and Gregory Kucherov, “Mreps: Efficient and Flexible Detection of Tandem Repeats in DNA,” Nucleic Acids Research, 31(13), 2003.

[55] D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy, “Spectral Repeat Finder (SRF): Identification of Repetitive Sequences using Fourier Transformation,” Bioinformatics, 20(9), 2004

[56] Alex van Belkuma, Willem van Leeuwena, Stewart Schererb, and Henri Verbrugha, “Occurrence and Structure-Function Relationship of Pentameric Short Sequence Repeats in Microbial Genomes,” Research in Microbiology, 150, 1999.

[57] Natalia Volfovsky, Brian Haas, and Steven Salzberg, “A Clustering Method for Repeat Analysis in DNA Sequences,” Genome Biology, 2(8), 2001.

[58] Mohamed Ibrahim Abouelhoda, EnnoOhlebusch, and Stefan Kurtz, “Optimal Exact String Matching Based on Suffix Arrays,” in Proceedings of the Ninth International Symposium on String Processing and Information Retrieval. Springer-Verlag 2002.

[59] William R. Pearson. Using the FASTA Program to Search Protein and DNA Sequence Databases. Computer Analysis of Sequence Data. Methods in Molecular Biology Volume 25, 1994.

[60] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman. Basic local alignment search tool. Journal of Molecular Biology, Volume 215, Issue 3, 1990.

[61] http://blast.ncbi.nlm.nih.gov/

[62] NCBI National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/

[63] Maja Tarailo-Graovac, Nansheng Chen. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics, 2009.

[64] http://www.repeatmasker.org/

[65] http://nebc.nox.ac.uk/bioinformatics/docs/cross_match.html

[66] http://blast.advbiocomp.com/

[67] http://www.ebi.ac.uk/Tools/sss/wublast/

[68] http://www.repeatmasker.org/RMBlast.html

[69] http://cmgm.stanford.edu/~decypher/dna.html

[70] Bedell Joseph A., Korf Ian and Gish Warren. MaskerAid: A performance enhancement to RepeatMasker.. Bioinformatics 2000.

[71] Benson G. Tandem Repeats Finder: A program to analyze DNA sequences. Oxford University Press, Nucleic Acids Research 1999.

[72] Kurtz S and Chris S. REPuter: fast computation of maximal repeats in complete genomes.. Bioinformatics 1999.

[73] Stefan Kurtz . The Vmatch large scale sequence analysis software. www.vmatch.de, 2007.

[74] Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and EnnoOhlebusch, “The Enhanced Suffix Array and Its Applications to Genome Analysis,” in WABI ’02: Proceedings of the Second International Workshop on Algorithms in Bioinformatics (Springer-Verlag, London, UK, 2002.

[75] UdiManber andGeneMyers, “Suffix Arrays: ANewMethod forOn-Line String Searches,” SIAM Journal on Computing, 22(5), 1993.

[76] Linhart C, Halperin Y, Shamir R. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. Genome Res., 2008.

[77] http://acgt.cs.tau.ac.il/allegro/download.html

[78] Philip Machanick and Timothy L. Bailey. MEME-ChIP: motif analysis of large DNA datasets. Oxford Journals: Bioinformatics 2011.

[79] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[80] E. P. Reddy, R. K. Reynolds, E. Santos, and M. Barbacid, "A point mutation is responsible for the acquisition of trans- forming properties by the t24 human bladder carcinoma oncogene," *Nature*, vol. 300, no. 5888, pp. 149–152, July 1981.

[81] C. Greenman, P. Stephens, and R. Smith, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 1, pp. 153–158, 2007.

[82] D. A. Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acid research*, vol. 39, pp. D32–D37, 2011.

[83] R. R. Voss, "Evolution of long-range fractal correlations and 1/f noise in dna base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992.

[84] L. Prasad and S.S. Iyengar, "Wavelet Analysis with an Application to Image Processing",Chapman and Hall/CRC Press, June 1997, pp. 279.

[85] M. Sifuzzaman, M. R. Islam, and M. Z. Ali, "Application of wavelet transform and its advantages compared to fourier transform," *Journal of Physical Sciences*, vol. 13, pp. 121–134, 2009.

[86] C. Gargour, M. Gabrea, V. Ramachandran, and J. M. Lina, "A short introduction to wavelets and their applications," *IEEE Circuits and Systems Magazine*, vol. 9, pp. 57–68, II Quarter 2009.

[87] D. Gabor, "Theory of communication," *IEEE Radio Communi cation Engineering Journal*, vol. 93, no. 26, pp. 429–457, 1946.

[88] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[89] J. B. Allen and L. R. Rabiner, "A unified approach to short- time fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[90] S. G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *Pattern Analysis andMachine Intelligence, IEEE Transactions on*, vol. 11, no. 7, pp. 674–693, 1989.

[91] I. Daubechies, *Ten Lectures on Wavelets*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1992.

[92] Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S. S. Iyengar, John Yordy, and Puneeth Iyengar, "Wavelet Analysis in Current Cancer Genome Research: A Survey," accepted for publication, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[93] R. Dulbecco, “A turning point in cancer research: sequencing the human genome,” *Science*, vol. 231, no. 4742, pp. 1055–1056, 1986.

[94] L. Chin, W. C. Hahn, and G. Getz, “Making sense of cancer genomic data,” *Genes and Development*, vol. 25, no. 6, pp. 534-555, 2011.

[95] E. V. Ball, P. D. Stenson, S. S. Abeysinghe, M. Krawczak, D. N. Cooper, and N. A. Chuzhanova, “Microdeletions and microinsertions causing human genetic disease: common mech anisms of mutagenesis and the role of local dna sequence complexity,” *Human mutation*, vol. 26, no. 3, pp. 205–213, 2005.

[96] E. M. Kvikstad, F. Chiaromonte, and K. D. Makova, “Ride the wavelet: A multiscale analysis of genomic contexts flanking small insertions and deletions,” *Genome research*, vol. 19, no. 7, pp. 1153–1164, 2009

[97] P. L. Foster, A. J. Hanson, H. Lee, E. M. Popodi, and H. Tang, “On the mutational topology of the bacterial genome,” *G3: Genes, Genomes, Genetics*, vol. 3, no. 3, pp. 399–407, 2013.

[98] L. Song. Computational analysis of genome-wide dna copy number changes. Ph.D. dissertation, Virginia Polytechnic Institute and State University, 2011.

[99] S. Ivakhno, T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham,and S. Tavar˜e, “Cnaseg - a novel framework for identification of copy number changes in cancer from second-generation sequencing data,” Bioinformatics, vol. 26, no. 24, pp. 3051–3058, 2010.

[100] L. M. Tran, B. Zhang, Z. Zhang, C. Zhang, T. Xie, J. R. Lamb, H. Dai, E. E. Schadt, and J. Zhu, “Inferring causal genomic alterations in breast cancer using gene expression data,” *BMC Systems Biology*, vol. 5, no. 121, 2011.

[101] E. Ben-Yaacov and Y. C. Eldar, “A fast and flexible method for the segmentation of acgh data,” *Bioinformatics*, vol. 24, no. 16, pp. i139–i145, 2008.

[102] K. C. Amarasinghe, J. Li, and S. K. Halgamuge, “Convex: copy number variation estimation in exome sequencing data using hmm,” *BMC bioinformatics*, vol. 14, no. Suppl 2, p. S2,2013.

[103] S. Ivakhno, T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham, and S. Tavarẽ , “Cnaseg - a novel framework for identification of copy number changes in cancer from second generation sequencing data,” *Bioinformatics*, vol. 26, no. 24, pp. 3051–3058, 2010.

[104] A. M. Sarhan, “Wavelet-based feature extraction for dna microarray classification,” *Artificial Intelligence Review*, vol. 39, no. 3, pp. 237–249, 2013.

[105] Y. Liu, U. Aickelin, J. Feyereisl, and L. G. Durrant, “Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data,” *Knowledge-Based Systems*, vol. 37, pp. S02–514, 2013.

[106] Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, Wunderlich A, Barmeyer C, Seemann P, Koenig J et al: Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PloS one 2010, 5(12):e15661.

[107] Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A et al: JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics (Oxford, England) 2012, 28(7):907-913.

[108] http://www.cs.waikato.ac.nz/ml/weka/

[109] http://weka.wikispaces.com/Primer

[110] http://weka.sourceforge.net/doc/weka/core

[111] G. E. P. Box, W. G. Hunter and J. S. Hunter. Statistics for Experimenters. New York: Wilet, 1978, ch. 16.

[112] J. R. Kittrell, R. Mezaki and C. C. Watson. Estimation of parameteres for nonlinear least square analysis. Ind. Eng. Chem. vol. 57, no. 12, p 19, 1965.

[113] R. L. Plackett. Principles of Regression Analysis. New York: Oxford, 1960.

[114] P. M. Reilly. Statistical techniques for model-building. PhD dissertation, Univ. Wisconsin, Madison, 1966.

[115] Iyengar and Rao, 1983 S.S. Iyengar, M.S. Rao. Statistical techniques in modelling of complex systems: single and multi-response models. IEEE Trans. Syst. Man Cybernet., SMC-13 (1983), pp. 175–189.

[116] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[117] http://www.stjude.org/downing-iom

[118] Frank E, Hall M, Trigg L, Holmes G, Witten IH: Data mining in bioinformatics using Weka. Bioinformatics (Oxford, England) 2004, 20(15):2479-2481.

[119] Chang, Chih-Chung; Lin, Chih-Jen (2011). "LIBSVM: A library for support vector machines". ACM Transactions on Intelligent Systems and Technology 2 (3).

[120]. Rasmussen, C. E. (2004). "Gaussian Processes in Machine Learning". Advanced Lectures on Machine Learning. Lecture Notes in Computer Science 3176. pp. 63–71

[121]. Lukaszyk, S. (2004) A new concept of probability metric and its applications in approximation of scattered data sets. Computational Mechanics, 33, 299-3004

[122] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE et al: Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009, 461(7261):272-276.

[123] Lessons from the Cancer Genome. Levi A. Garraway, Eric S. Lander. Cell - 28 March 2013 Vol. 153, Issue 1, pp. 17-37.

VITA

JUAN CARLOS MARTINEZ

2005	B.Sc., Informatics Engineering, Pontificia Universidad Catolica del Peru
2006	M.S., Computer Science, Florida International University, Miami, Florida
2013	PhD., Computer Science, Florida International University, Miami, Florida

PUBLICATIONS, ABSTRACTS AND PRESENTATIONS

Journals

Juan Carlos Martinez, Nelson Lopez-Jimenez, Tao Meng, and S. S. Iyengar. Predicting DNA Mutations during Cancer Evolution, accepted for publication. International Journal of Bioinformatics Research and Applications.

Juan Carlos Martinez and Tao Meng. A Modeling Approach for the Prediction of Mutation in Genomic Sequences, accepted for publication. International Journal of Computer Science.

Juan Carlos Martinez, S. S. Iyengar, and Nelson Lopez-Jimenez. Toward the prediction of mutations in genomic sequences from patients with cancer. Journal of Clinical Bioinformatics. (Under review).

Conference/Poster Presentations

Juan Carlos Martinez, S. S. Iyengar, Nelson Lopez-Jimenez. A computational approach toward predicting gene mutations by genome-wide sequencing analysis of both normal and cancer cells from the same individual. Beyond the Genome 2013. San Francisco, CA, October 1-3, 2013.

Juan Carlos Martinez, S. S. Iyengar, Nelson Lopez-Jimenez. A model for predicting gene mutations in the genome of individuals with AML during disease evolution by genome-wide sequencing analysis. The American Society for Cell Biology (ASCB 2013). New Orleans, LA, December 14-18, 2013.

Juan Carlos Martinez, S. S. Iyengar, Nelson Lopez-Jimenez. Predicting gene mutations during cancer evolution: a new tool in searching novel targets for cancer treatment.. 11th Annual Rocky Mountain Bioinformatics Conference. Aspen/Snowmass, Colorado, December 12-14, 2013.

Other Publications

Juan Carlos Martinez, Lixi Wang, S. Masoud Sadjadi, Ming Zhao, Experimental Study of Large-scale Computing on Virtualized Resources. VTDC '09 Proceedings of the 3rd International Workshop on Virtualization Technologies in Distributed Computing.

Hector A. Duran Limon, S. Masoud Sadjadi, Raju Rangaswami, Shu Shimizu, Liana Fong, Rosa M. Badia, Pat Welsh, Sandeep Pattnaik, Anthony Praino, Javier Figueroa, Javier Delgado, Xabriel J. Collazo-Mojica, David Villegas, Selim Kalayci, Gargi Dasgupta, Onyeka Ezenwoye, Khalid Saleem, **Juan Carlos Martinez**, Ivan Rodero, Shuyi Chen, Javier Muñoz, Diego Lopez, Julita Corbalan, Hugh Willoughby, Michael McFail, Christine Lisetti, and Malek Adjouadi. Grid enablement and resource usage prediction of weather research and forecasting. In Proceedings of the Collaborative and Grid Computing Technologies Workshop, page 4, Cancun, Mexico, April 2008.

Yanbin Liu, S. Masoud Sadjadi, Liana Fong, Ivan Rodero, David Villegas, Selim Kalayci, Norman Bobroff, and **Juan Carlos Martinez**. Enabling autonomic meta-scheduling in grid environments. In Proceedings of the 5th IEEE International Conference on Autonomic Computing (ICAC-2008), pages 199-200, Chicago, IL, June 2008.

Norman Bobroff, Liana Fong, Selim Kalayci, Yanbin Liu, **Juan Carlos Martinez**, Ivan Rodero, S. Masoud Sadjadi, and David Villegas. Enabling interoperability among meta-schedulers. In Proceedings of 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid-2008), pages 306-315, Lyon, France, 2008.

S. Masoud Sadjadi, Liana Fong, Rosa M. Badia, Javier Figueroa, Javier Delgado, Xabriel J. Collazo-Mojica, Khalid Saleem, Raju Rangaswami, Shu Shimizu, Hector A. Duran Limon, Pat Welsh, Sandeep Pattnaik, Anthony Praino, David Villegas, Selim Kalayci, Gargi Dasgupta, Onyeka Ezenwoye, **Juan Carlos Martinez**, Ivan Rodero, Shuyi Chen, Javier Muñoz, Diego Lopez, Julita Corbalan, Hugh Willoughby, Michael McFail, Christine Lisetti, and Malek Adjouadi. Transparent grid enablement of weather research and forecasting. In Proceedings of the Mardi Gras Conference 2008 - Workshop on Grid-Enabling Applications, Baton Rouge, Louisiana, USA, January 2008.

S. Masoud Sadjadi, **J. Martinez**, T. Soldo, L. Atencio, R. M. Badia, and J. Ejarque. Improving separation of concerns in the development of scientific applications. In Proceedings of The Nineteenth International Conference on Software Engineering and Knowledge Engineering (SEKE'2007), pages 456-461, Boston, USA, July 2007.

Tao Li, S. Masoud Sadjadi, **Juan Carlos Martinez**, Lokesh Sasikumar, and Manoj Pillai. Data mining for autonomic system management: A case study at fiu-scis. Technical report, Florida International University, 2006. Technical Report: FIU-SCIS-2006-03-01.