10-19-2015

USING MACHINE LEARNING TECHNIQUES FOR FINDING MEANINGFUL TRANSCRIPTS IN PROSTATE CANCER PROGRESSION

Siva Charan Reddy Singi Reddy
University of Windsor

USING MACHINE LEARNING TECHNIQUES FOR FINDING MEANINGFUL TRANSCRIPTS IN PROSTATE CANCER PROGRESSION

by
Siva Charan Reddy Singi Reddy

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

Windsor, Ontario, Canada
2015

USING MACHINE LEARNING TECHNIQUES FOR FINDING MEANINGFUL TRANSCRIPTS IN PROSTATE CANCER PROGRESSION

by

Siva Charan Reddy Singi Reddy

APPROVED BY:

L. Porter
Department of Biological Sciences

A. Ngom
School of Computer Science

L. Rueda, Advisor
School of Computer Science

August 25, 2015

Author's Declaration of Originality

I. Declaration of Previous Publication

This thesis includes one original paper that has been previously published/submitted for publication in peer reviewed journals, as follows:

Thesis Chapter	Publication title/full citation	Publication status
Chapters 5,6 and appendix	Siva Singireddy et al. "Identifying Differentially Expressed Transcripts Associated with Prostate Cancer Progression using RNA-Seq and Machine Learning Techniques." Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE 12th International Conference on. IEEE, 2015.	in proceedings

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyones copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have ob-

tained a written permission from the copyright owner(s) to include such material(s) in my thesis. I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Prostate Cancer is one of the most common types of cancer among Canadian men. Next generation sequencing that uses RNA-Seq can be valuable in studying cancer, since it provides large amounts of data as a source for information about biomarkers. For these reasons, we have chosen RNA-Seq data for prostate cancer progression in our study. In this research, we propose a new method for finding transcripts that can be used as genomic features. In this regard, we have gathered a very large amount of transcripts. There are a large number of transcripts that are not quite relevant, and we filter them by applying a feature selection algorithm. The results are then processed through a machine learning technique for classification such as the support vector machine which is used to classify different stages of prostate cancer. Finally, we have identified potential transcripts associated with prostate cancer progression. Ideally, these transcripts can be used for improving diagnosis, treatment, and drug development.

Dedication

With an overflowing heart of thanksgiving, I wish to dedicate this thesis to my God Gifted Parents Mr. Raja Reddy and Mrs. Shamala. I am also indebted to my loving brother Raja Shekar Reddy, who took me under his wing in the early years of my education.

Acknowledgements

I would like to start by saying thank you to my supervisor Dr. Luis Rueda, the most important person and guidance for my research project. I love the unfailing energy, professionalism, and generosity that you carry in your DNA. The author is profoundly grateful to Dr. Lisa Porter, Dr. Alioune Ngom, Dr. Dora Cavallo-Medved, and Dr. Iman Rezaeian. Thank you for your unending and uncompromising support which has enabled me to finish this thesis. Special thanks to my best friend Abedalrhman Alkhateeb for helping me complete this project. It is a privilege to learn and work with you. Suchet Krishna, Nishanth Singarapu, Amarender Reddy, Vamshi Reddy, Lal Bahadur Sastry, and Abdul Ameer, everyone needs friends like you who will extend love, time, and sacrifice in hard times. I am in debt to you for the rest of my life, thank you all for helping me in my dark times and encouraging to achieve my ambitions. Many thanks to all of you!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A cell, the basic unit of life, is capable of independent reproduction [24]. There are two kinds of cells: eukaryotic and prokaryotic. In eukaryotic organisms, every cell has a nucleus, while the prokaryotic cell is a unicellular microorganism without a nucleus [24]. The human body has eukaryotic cells, each with a nucleus at its centre and a cell membrane for protection [24]. The chromosomes are distributed in the nucleus. Every human cell has 23 pairs of chromosomes, and each chromosome contains many different genes [1]. Deoxyribonucleic acid (DNA) is used to build and maintain the cell and also carries hereditary information within the chromosomes. DNA is composed of nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T) [46].

Cells undergo several ways to transform DNA into proteins. Generally, there are two main steps to convert coding regions of DNA into proteins [24]. In the first step, DNA transcribes to ribonucleic acid (RNA), while in the second step, RNA translates into proteins (see Figure 1.1) [9; 20; 24]. The outcome of transcription is the precursor messenger RNA (pre-mRNA), which undergoes RNA splicing or processing, a process in which exons are retained and introns are removed [24]. The splicing of pre-mRNA occurs in several different

ways. The most common way is that an intron is excluded and an exon is included, which leads to the formation of a different mRNA strand. This process is known as alternative splicing [24].

Figure 1.1 illustrates alternative splicing of a gene. DNA contains exons and introns, also called coding and non-coding regions, respectively. DNA transcribes to RNA, which further translates to proteins. It can be observed from the figure that exon 1, exon 2, and exon 4 are retained to form protein 1; exon 1, exon 3, and exon 4 make protein 2. On the other hand, introns are removed to form the mature mRNA transcript. The study of an entire group of transcripts or RNA for the diagnosis of precise disease conditions is known as transcriptomics [45].

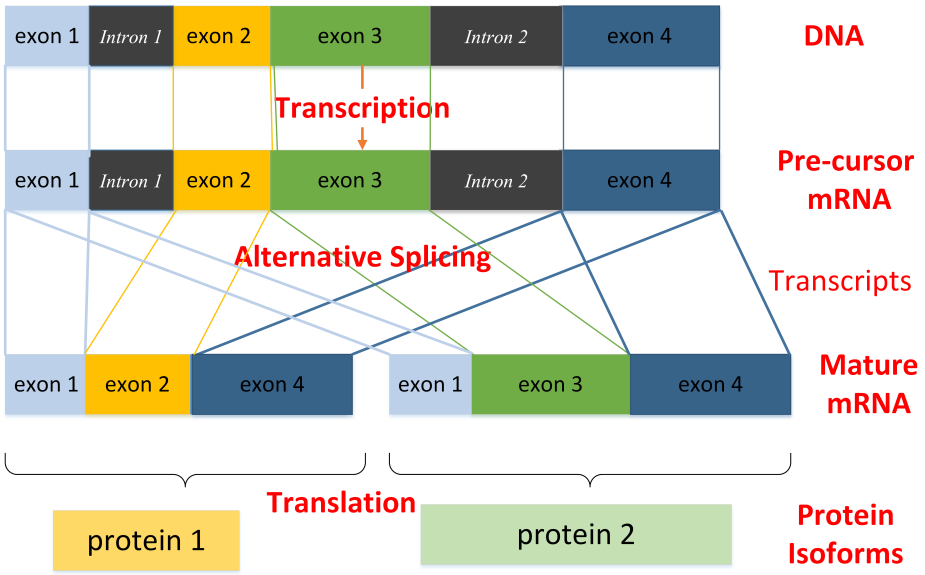


Figure 1.1: Alternative splicing of the gene: DNA translates to RNA; RNA undergoes splicing and forms mature mRNA; mRNA further translates to a protein.

1.1 Prostate Cancer Progression

Prostate cancer is caused by the abnormal and uncontrolled growth of the prostate gland [40]. According to Statistics Canada, one in five Canadian men will be diagnosed with prostate cancer during their lifetimes, and one in four will die from prostate cancer [39]. An estimated 196,900 patients are anticipated to be diagnosed with cancer in 2015 [39]. Approximately 50% of these cases will be lung, breast, colorectal, or prostate cancer [39]. Lung cancer accounts for the majority of the cases, followed by colorectal and prostate cancers [39]. Canadian males are primarily affected by prostate cancer; approximately 24,000 patients are anticipated to be diagnosed with cancer in 2015 [39]. As in other types of cancer, there is a need to conduct research on prostate cancer. In addition, investigating prostate cancer at the molecular level can help determine the structure of tumour initiation, as well as its progression. Prostate cancer is very unlikely to progress; less than one third of the patients will progress to advanced stages. This kind of investigation aids both the diagnosis and treatment of the disease at the earliest possible stage.

The American Cancer Society has categorized prostate cancer into four different stages, each of which is further divided into sub-stages [38; 40]. Table 1.1 provides some information about each stage and sub-stage of prostate cancer. Prostate cancer can be discoverable in the initial stage (T1c) [38]. In the second stage (T2), it spreads to the prostate gland [38]. Stage T2 is divided into three sub-stages (T2a, T2b, and T2c). Cancer grows at a moderate rate in sub-stages T2a and T2b, whereas growth occurs at a higher rate in stage T2c [38].

At stage T3, the cancer spreads to neighboring tissues. This stage is further divided into two sub-stages, T3a and T3b [38]. Both sub-stages are essential in prostate cancer progression, as the cancer spreads to the seminal vesicles in sub-stage T3b [38]. In the final stage, T4, the cancer extends to other organs [38]. Cancer, which starts in one organ,

invading another organ is known as metastasis [38]. It begins to grow cancer cells in the new location, thereby damaging the functioning of that organ [38]. Most cancer patients die when they reach the *metastatic* stage [38]. Estimating the progression helps in detecting and diagnosing cancer, and providing a patient with an appropriate treatment.

Table 1.1: Stages in progression of prostate cancer according to the American Cancer Society [38].

Prostate cancer stage	Description
T1c	The tumour is not detectable via imaging techniques. Cancer is detected using a needle biopsy performed due to an elevated serum prostate-specific antigen (PSA).
T2	The tumour is palpable, but confined to the prostate.
T2a	The tumour is in half, or less than half, of one of the prostate glands two lobes.
T2b	The tumour is in more than half of one lobe, but is not in both lobes
T2c	The tumour is in both lobes but confined within the prostatic capsule.
T3	The tumour has started spreading out of the prostate tissue.
T3a	The tumour has spread through the prostatic capsule on one or both sides, but has not spread to the seminal vesicles.
T3b	The tumour has invaded one or both of the seminal vesicles.
T4	The tumour has spread to other organs.

1.2 RNA-Seq

RNA-Seq is an emerging technology that uses next generation sequencing (NGS) techniques. It helps biologists and clinicians understand the complexity of diseases at the molecular level [11; 45]. It also provides precise information for analysis of alternative splicing events, gene fusions, transcriptions, and post-transcription stages. Recent advances in

RNA-Seq and NGS have made sequencing costs drop drastically, which has led researchers to create many RNA-Seq data sets on prostate cancer [45]. We chose RNA-Seq data sets in our studies for all these reasons, as discussed in more detail in Chapter 3.

1.3 Thesis Motivation

Various studies have found that aberrant splicing of the pre-mRNA yields different kinds of cancers [49]. The discovery of biomarkers is the central step in diagnosis and handling of any kind of disease, especially for cancer. Mayer et al. observed that a differential splice variant, the RON isoform, was upregulated in ovarian cancer [26]. Ren et al. used next generation sequencing technology and discovered that long non-coding RNAs, gene fusions and aberrant splicing influence cell growth [33]. Long et al. worked on 106 malignant samples using RNA-Seq data and extracted 24 genes, of which five genes (BTG2, IGFBP3, SIRT1, MXI1, and FDPS) correlate with prostate cancer [25].

In recent years, researchers have been working to find biomarkers for different types of cancer. They have focused mainly on the genetic level and have found differentially expressed genes. Some researchers have also studied prostate cancer and its progression. However, investigating the transcriptome activity of a cell or organism is more interesting than studying it at the gene level, due to the precise information the activity provides on the disease condition. We examined these kinds of patterns involved in prostate cancer and its progression.

1.4 Main Problem

Researchers face the challenging issue of finding biomarkers for prostate cancer; it is difficult to find them with current approaches [25]. Previous researchers focus on matched normal versus malignant using genes as biomarkers to find differentially expressed genes associated with prostate cancer. We are given data sets of RNA-Seq reads that belong to different samples each associated with particular stage; these samples come from patients or cell lines. We aim to identify differentially expressed transcripts that are associated with different stages of prostate cancer. Ideally, these transcripts can be used for improving diagnosis, treatment and drug development.

To deal with this problem, we applied powerful feature selection and classification algorithms to find discriminative transcripts that are related to prostate cancer and its different stages.

1.5 Contributions

In this work, we introduce a novel model that integrates emerging RNA-Seq technology with machine learning approaches to find the vital discriminative transcripts for the different stages of prostate cancer.

The main contributions are:

- Developing an integrative model that uses feature selection to choose a subgroup of transcripts and classification techniques to find the most relevant transcripts for different stages of prostate cancer.
- Identifying novel transcripts as potential biomarkers for prostate cancer progression.

1.6 Thesis Organization

This thesis consists of seven chapters, starting with an introduction, which provides an overview of the main topics. A literature review is presented in Chapter 2. An overview of RNA-Seq data, workflow, and analysis comprise Chapter 3. In Chapter 4, machine learning techniques for feature selection and classification are discussed. The methods and results are discussed in Chapters 5 and 6, respectively. Finally, Chapter 7 presents the thesis conclusions and the future work derived from this thesis.

Chapter 2

Literature Review

In this chapter, we review the literature that identifies the problems that researchers are currently facing in finding biomarkers for prediction of prostate cancer. This chapter is organized based on the biomarkers used by different scientists. Most of them have used genes, whereas Tavakoli et al. used junctions, as biomarkers to study prostate cancer. Also, Kim et al. studied methylation patterns to investigate prostate cancer.

2.1 Using Genes as Biomarkers

Recently, researchers have found it difficult to predict the progression of prostate cancer. Long et al. worked with genes as biomarkers to estimate the disease development [25]. The authors gathered tissue cores from 106 prostate cancer patients and extracted RNA-Seq data [25]. Initially, the RNA was prepared using a formalin-fixed paraffin-embedded approach and sent for sequencing, which employed the Illumina HiSeq technology to perform 50 base pairs paired-end sequencing [25]. The data set can be retrieved via GEO accession number GSE54460 [25].

Long et al. started their work by following part of the tuxedo suite approach, which utilizes Tophat2 to align the reads from the patients to the reference human genome and uses Cufflinks for transcriptome assembly [25]. They used the DESeq tool to find differentially expressed genes [25]. Subsequently, a set of 24 genes were obtained; 16 were previously associated with prostate cancer, and among them, five genes (BTG2, IGFBP3, SIRT1, MXI1, and FDPS) are typically associated with prostate cancer [25].

Zhai et al. also worked on RNA-Seq data to find differentially expressed genes that are related to prostate cancer [47]. They found protein-coding genes and lincRNAs that are differentially expressed between matched normal and malignant patients [47]. Zhai et al. experimented on 10 matched prostate samples, which were taken from the European Nucleotide Archive with accession number SRP002628 [47]. They performed an analysis that is similar to that of Long et al., except that hierarchical clustering was used for finding differentially expressed genes [47].

Zhai et al. claims that 10 genes and a lincRNA were differentially expressed [47]. The authors claim that the lincRNA that is present in the Cullin-associated and neddulation-dissociated 1 (CAND1) gene expressed high and low between malignant and matched normal samples, respectively [47].

Ren et al. studied prostate cancer in the Chinese population and revealed that long non-coding RNAs influence the prognosis [33]. The authors stated that the non-destructive nature of the prostate cell has no effect [33]. On the other hand, rapidly-advancing cell growth will lead to metastasis, resulting in the death of the patient [33].

The authors generated RNA-Seq data on the 14 matched prostate samples [33]. The RNA was gathered from samples, and oligo(DT) primers were used to separate poly(A) mRNA [33]. In our research, we have used four data sets, namely Kannan's, Kim's, Ren's

and Long's data sets (discussed in Chapter 5). Ren's data set used random hexamer primers, while the other data sets used oligo (DT) primers. The selection of primers is very important, since data set extraction depends on the primer used [37]. The two primers have their own advantages and disadvantages. The choice of primers depends on the mRNA extracted [37]. If the mRNA contains polyA at the end, usually oligo (DT) is preferable, while if the mRNA is too long, it is difficult to cover the whole mRNA strand [37; 8]. In this case, random primers are the best choice, since they are able to extract small pieces of mRNA [8]. Afterwards, they were divided into fragments, and Illumina HiSeq 2000 was employed to sequence the reads [33]. Ren et al. aligned these reads by applying the SOAP2 aligner, and then performed supervised clustering to obtain differentially expressed genes and non-coding RNAs [33].

Ren et al. studied 183 genes that surprisingly mutated to prostate cancer and three gene fusions [33]. They found two new gene fusions, CTAGE5-KHDRBS3 and USP9Y-TTTY15, which are highly linked to prostate cancer, and another gene fusion, TMPRSS2-ERG, which is quite common in prostate cancer [33].

2.2 Using Splice Junctions and Transcripts as Biomarkers

Tavokoli et al. worked on the problem of finding biomarkers for prostate cancer. They proposed splice junctions as biomarkers [41]. The authors started their experiment with Kannan's data set [5], which has 10 matched samples [41]. The RNA-Seq data was generated using the Illumina Genome Analyzer II platform [41].

Tavokoli et al. aligned the data set to the reference genome (GRCh37) with the PASSion tool [41], which outputs splice junctions with cut-off score. They filtered the splice junctions, which were dubious by a 2D peak-finding algorithm [41]. In that algorithm,

a new scoring scheme was proposed for each junction; afterwards, they applied machine learning algorithms to these junctions [41]. Finally, when a support vector machine was used along the junctions, they achieved 100% classification accuracy [41]. They found 10 splice junctions that are highly correlated with prostate cancer [41].

2.3 Using Methylation Regions as Biomarkers

Kim et al. worked on differentially expressed methylated regions that are linked to prostate cancer [15]. The authors produced a data set with four matched normal and seven malignant samples by MethylPlex next generation sequencing technology [15].

The RNA-Seq library was prepared with LNCap and PrEC cells for malignant and matched normal samples, respectively. Kim et al. employed hidden Markov model (HMM) analysis on the data generated [15]. The reads were produced from enhanced portions that include all the genes and accession number [15]. To determine the expression level of the genes, they mapped the reads to the reference genome using the ELAND tool [15].

Kim et al. performed a gene set enrichment analysis (GSEA) to examine the genes that are differentially-methylated regions [15]. GSEA validates and reports differentially expressed genes, provided that two conditions are applied. Lastly, they found 2,481 methylated regions that are expressed differentially, and WFDC2 was found to be a novel tumor-methylated region associated with prostate cancer [15].

2.4 Conclusion

The literature review suggests that researchers have recently depended on RNA-Seq data sets. Most of them utilized a tuxedo suite approach to extract genes for predicting can-

cer. Therefore, we have followed a similar tuxedo suite approach in this work. Previous researchers focus on matched normal versus malignant, while we focus on progression of prostate cancer. Other works mostly focus on genes as biomarkers and depend on statistical tests to find differentially expressed genes. However, we focus on transcripts as biomarkers and use machine learning techniques to identify differentially expressed transcripts.

Chapter 3

Transcriptomics Studies Using RNA-Seq

Analyzing a transcriptome involves determining splice junctions, mRNA, non-coding RNAs, and post transcriptional alterations of transcripts present in a cell for some specific experimental conditions. The study of transcriptomes is called transcriptomics [45].

There are several methodologies available to study transcriptomes [45]. Each technology has its own benefits and drawbacks. Hybridized models, such as microarrays, are applied to analyze gene expression [45]. They provide reliable output and are cost-effective. On the other hand, they have disadvantages, such as weak stability of the signals and low dynamic range of nucleotides [45]. Sanger sequencing was developed to overcome these limitations of microarrays [45]. Sanger sequencing resulted in an extremely expensive and very low throughput method, which created a need to develop new approaches [45]. Tag-based approaches create high-end products [45]. However, they generate short reads that cannot be accurately mapped to the genome. RNA-Seq technology was then developed as a high-throughput methodology to quantify transcripts [45].

3.1 RNA-Seq Technology

Wang et al. proposed RNA-Seq, an emerging technology that utilizes next generation sequencing techniques to investigate RNAs [45]. Figure 3.1 shows the workflow of an RNA-Seq technology experiment. Initially, a library is constructed by extracting the RNA from the underlying samples. Then, the RNA molecules are reverse transcribed to the corresponding complementary DNA (cDNA) fragments. These fragments are defragmented using RNA or cDNA fragmentation by adding adapters to both ends. Subsequently, the resulting fragments are amplified in order to obtain the actual short reads; these reads are then mapped to the reference genome [45].

Small RNA fragments can be sequenced directly. In contrast, mRNA fragments are usually very large, and hence they are split into shorter fragments before sequencing [6; 45]. To reverse transcribe the RNA, a primer such as an oligo (DT) is attached to the fragments and converted to cDNA [6]. Adapters are attached to the 5' and 3' ends of cDNA. Next, the fragments are amplified by a polymerase chain reaction (PCR) procedure [6].

RNA-Seq provides short reads, which can produce highly-informative evidence about the transcripts, and has a particularly superior dynamic range compared to previous approaches, producing more than 9,000-fold range sequences [45]. RNA-Seq has enhanced sequence coverage, and has an amazing resolution at a single nucleotide. As such, RNA-Seq is utilized to discover alternatively spliced RNA, novel isoforms, gene fusions, spliced junctions, and novel microRNA, among others [45].

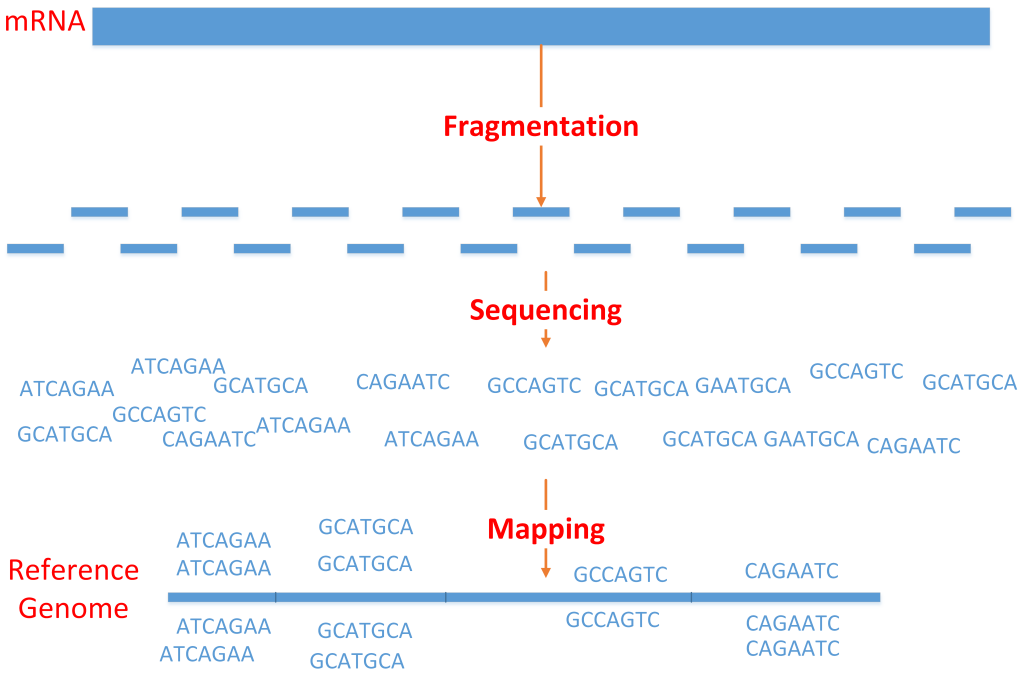


Figure 3.1: Work-flow of an RNA-Seq technology experiment.

3.2 Challenges in RNA-Seq Studies

RNA-Seq technology incorporates multiple steps for library preparation, and manipulation at each step makes complicated for measuring the transcript expression [45]. Initially, the entire RNA content of a cell is gathered, which then will undergo sequencing. Previous surveys indicate that the ribosomal RNA content is high and constitutes more than 80% of the whole RNA [45]. This leads to a decrease in resource usage, along with a reduction in sequencing coverage [45]. Therefore, ribosomal RNA is removed by applying an enzymatic degradation or hybridization-based depletion approach [5; 6].

To achieve higher coverage, deeper sequencing should be performed [45]. This process drives the cost that is directly related to sequence coverage [45]. To obtain sufficient

coverage, a large genome and a more sophisticated transcriptomics technology are required [6; 45]. Obtaining such a coverage of a transcriptome has been burdensome until now, because not all the transcripts are recognize [45]d.

Another challenge of RNA-Seq technology is the representation of gene and exon boundaries of the genome [45]. Finding introns and exons is the most difficult part of RNA-Seq analysis [45]. Aligning the short reads to the reference genome becomes unmanageable for the aligner tools due to the exon-intron boundaries; identifying the start and end of exons and introns is extremely difficult [45]. Problems resulting from background noise make this process very challenging [45].

3.3 Sequencing Technologies

There are several next generation sequencing technologies available in the marketplace. They can be grouped into two types: single molecule-based and ensemble-based [5]. The reads that are generated from these sequencing technologies are ready for use in experiments.

There are two kinds of sequencing in RNA-Seq technology: single-end and paired-end sequencing. In single-end sequencing, the cDNA fragments are sequenced from one end [5]. This sequencing requires very limited DNA and produces high-quality reads [5]. The drawback of this mode is that detecting novel isoforms is extremely difficult [5]. In paired-end sequencing, the fragments are sequenced from both ends to extract the corresponding reads [5]. This type of reads is very useful in finding novel splice variants, due to the reorganization of insertions and deletions [5]. The disadvantage of paired-end sequencing is very expensive compared to single-end sequencing [5].

3.4 Read Alignment

The reads generated by RNA-Seq sequencing technology are aligned to the reference genome to identify splice sites. There are many tools used for this purpose. They fall into two categories: unspliced and spliced aligners.

3.4.1 UnSpliced Aligners

An unspliced aligner maps continuous reads to the reference genome. There are several open source unspliced aligner tools. Bowtie2 is the tool most commonly used by researchers [19]. It is part of the tuxedo approach, being this the reason for which it is applied in our work.

Bowtie2

Langmead et al. proposed Bowtie2, a fast aligner that employs full-text minute (FM) indexing based on the Burrows-Wheeler transform technique [18]. Langmead et al. made advancements in the computation of aligning the reads in Bowtie2. This tool divides the alignment process into four steps. First, reads are divided into seeds with certain base pairs, using FM indexing. These seeds are aligned to the reference genome; seeds that are not aligned due to the presence of insertions and deletions are marked and ranked [19]. The lowest base pair seeds achieve the highest rank, and vice versa. The seeds are then mapped using single instruction multiple data (SIMD) programming until all the seeds are accessed [19].

3.4.2 Spliced Aligners

A spliced aligner aligns the spanned exons boundaries[14]. There are two types of spliced aligner tools: reference-based and *de novo*-based. The reference-based tools work well with known spliced junctions (i.e., providing the tool with annotated splice junctions) [14]. On the other hand, *de novo* tools align the spliced reads without prior knowledge of splice variants of the reference genome [14].

A hybrid spliced aligner tool integrates annotation and *de novo* alignment. Tophat2 is such an aligner, which can perform splice alignment with (and without) knowledge of splice variants, to find novel protein isoforms [14]. We have already discussed in Chapter 2 that tuxedo is the most common approach used by scientists to extract splice variants. Tophat2, integrated with Bowtie2, is used in this research as part of the tuxedo approach

Tophat2

Kim et al. addressed a common problem: alignment of RNA-Seq data to the reference genome to find novel spliced events, which helps in the detection of tumours or cancer [14]. The authors state that other tools fail to perform accurate mapping if there are higher expression levels or more insertions and deletions in the genes [14]. Kim et al. describe a three-step approach for mapping. Initially, Tophat2 uses transcriptome mapping when the annotation is provided; genome mapping is performed, and spliced mapping is done in the last step [14]. This approach produces splice junctions and reads accepted by the reference genome.

3.5 Transcriptome Assembly

Transcriptome assembly involves assembling the reads that have the ability to form potential mRNA or transcripts [43]. The reads that are accepted by aligner tools to the reference genome are ready for transcriptome assembly [43]. When transcriptome annotations are provided, it is a reference-based assembly [43]. On the other hand, *de novo* assembly means the tool does not use reference annotations [43]. Subsequently, transcripts' abundance are calculated in order to be compared within the same samples or with other samples [43]. Differences in the number of reads obtained from each sample or variations in the length of the transcripts will change their abundance [43]. Therefore, a normalized value is needed to compare a transcript with another. There are many ways of computing the normalized value [43]. The usual means of calculating normalized values is fragment per kilo base of transcripts per million reads (FPKM) [43]. The fragment refers to both ends of the cDNA, which is considered one fragment. The per kilo base of transcripts normalizes the number of fragments dividing by the total number of transcripts present in the gene [43]. The calculation per million reads makes the transcripts comparable to different samples [43]. Cufflinks, a reference-based assembler, is used in our research, because we aim to find transcripts that are present in the genes and are already associated with prostate cancer progression [43]. Moreover, Cufflinks is also part of the tuxedo approach. There are other tools for transcriptome assembly and quantification, including iReckon; we briefly discuss Cufflinks and iReckon in this chapter.

Cufflinks

Cufflinks is a transcriptome assembler that also estimates the abundance of the transcripts. It assembles transcripts by building an overlap graph for all the reads that are accepted by

the reference genome. It identifies and filters the reads that are incompatible for assembly. Reads that are compatible must receive at least one splice junction in common and are the constituents of the graph being constructed [43].

Trapnell et al. implemented Dilworth's theorem to cover the minimum path and constructed transcripts from the accepted reads [43]. Initially, reads are first marked for compatibility. The overlap graph is then constructed such that each transcript present in the reference transcriptome is covered [43]. Transcript abundance is calculated as the FPKM value [42]. This value is normalized to verify each transcript with another transcript in the same gene or other samples [42]

iReckon

Mezlini et al. implemented iReckon, a tool for transcriptome assembly and abundance estimation for revealing protein isoforms [27]. iReckon is an implementation of the regularized expectation-maximization (EM) algorithm for construction of transcriptome assembly and estimating abundance. It discovers potentially novel isoforms by integrating the prior knowledge of unspliced pre-mRNA and intron retention. Initially, potential isoforms are identified. The accepted reads from the aligner tools are used to construct the splice graph. These reads are then rearranged to form potential isoforms [27]. For each transcript, a normalized abundance is calculated such that transcripts are comparable with other transcripts.

3.6 Web-Based RNA-Seq tool

There are many open source standalone tools available. Alternatively, recently-developed web-based RNA-Seq tools are also used. Galaxy is the most commonly-used web-based tool.

Galaxy

Blankenberg et al. designed and implemented Galaxy, a web-based open source tool for RNA-Seq analysis [4]. Galaxy offers a wide range of tools to perform analysis on RNA-Seq data; most of the latest tools used for read alignment, and transcriptome assembly are installed in Galaxy. The web interface allows users to store data sets and run the tools using a workflow. Users' data sets and results can be shared with other users in Galaxy. Galaxy provides good visualization of the results, and provides source code and documentation so that the software can be deployed in any server. The advantage of the Galaxy suite works efficiently on small projects. However, Galaxy cannot accommodate large data sets, due to memory and space limitations [4]. We have used Galaxy to run Cufflinks on our data sets.

3.7 Conclusion

In this chapter, we discussed about RNA-Seq technology and its challenges. We have also described different tools that are work on RNA-Seq reads. In the next chapter, machine learning techniques will be discussed that are used in our research.

Chapter 4

Machine Learning

Machine learning is a branch of artificial intelligence that provides various methods and algorithms that are trained on inputs, and a model is extracted from them [44]. Subsequently, that model is tested on a different set of inputs, and then the algorithm performance is measured [44]. Classification and feature selection are two applications of machine learning [44].

4.1 Classification

The objective of classification is to find a discriminant function from the inputs [44]. There are three kinds of learning for classification purposes: supervised, unsupervised, and semi-supervised.

In supervised learning, labeled samples are passed to the classification algorithm, which creates the predictive model. Figure 4.1 represents a two-class supervised classification problem; patients are in the rows, while transcripts are in the columns. The last column contains the class labels: malignant and matched normal samples. We are attempting to de-

sign a model that can find a discriminating function between cancerous and non-cancerous samples. In unsupervised learning, only the samples are given, without the class labels. Semi-supervised learning uses supervised learning class label knowledge as well as an unsupervised method for grouping similar data.

Table 4.1: Example of a two-class classification problem that involves two types of cancer, matched normal and malignant.

Samples	t_1	t_2	t_3	Class
S_1	1	0	1	Malignant
S_2	0	0	0	Matched normal
S_3	0	0	1	Matched normal
S_k	1	1	0	Malignant

In this thesis, we use supervised learning approaches. Each sample has a class label that indicates whether or not that sample is matched normal or malignant, or at a particular stages of prostate cancer. The input vectors are the transcripts that are extracted from the preprocessing stage, which is discussed in Chapter 5. In the literature, transcripts are referred to as features, variables, or attributes.

There are many algorithms that have been designed to work on classification problems. In this thesis, we use support vector machine (SVM), random forest (RF), Naïve Bayes, and decision tree algorithms, as they worked efficiently for our data sets.

4.1.1 Support Vector Machine

SVM is a classifier that is often used to solve biological problems, among others. It works efficiently in finding the discriminant function, which is based on the support vectors. There are two types of classification problems: linearly and non-linearly separable data. Figure

4.1 shows an example of linearly separable data. The data is plotted with transcript 1 on the x-axis and transcript 2 on the y-axis. The green-colored points represent matched normal samples, whereas the red-colored points represent malignant samples. The black-colored points represent the support vectors.

The goal of an SVM is to find a line that separates the two classes. We can find many different lines. The SVM tackles this problem by relying on the samples that are the most difficult to classify, known as support vectors [7]. Initially, the support vectors are identified, and a line is found, which has the maximum distance from the two classes, this model is known as hard margin. In Figure 4.1, D2 has the maximum distance as compared to D1; therefore, line L2 is obtained.

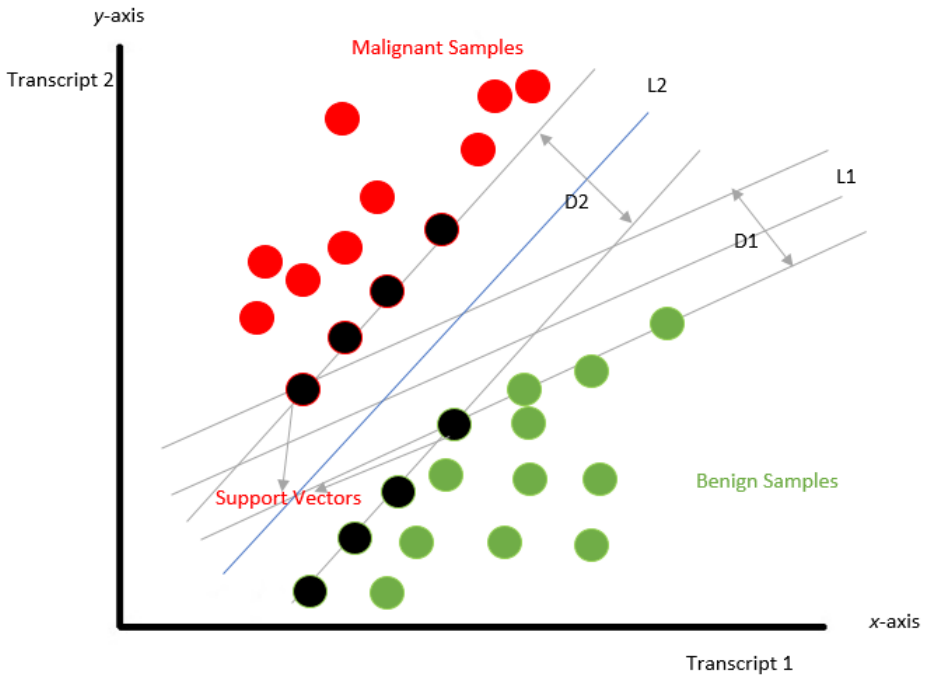


Figure 4.1: SVM for linearly separable data.

Consider when the data is non-linearly separable; a line cannot separate the two classes.

Imagine that, if the data is mapped to a higher dimension, then the line may separate the data. However, transforming from one dimension to higher dimensions is computationally expensive. SVM solves this problem by using the kernel trick. If the data becomes linearly separable, then the line is drawn with the help of the support vectors to separate the classes. The three most popular kernel methods are: linear, polynomial, and radial basis function. All of them are used in our research.

Despite using kernel trick sometimes the data is still non-linearly separable [7]. In this case, SVM uses slack variables on the original data set, which relax the constraints and penalize the misclassified variable with a cost parameter [7]. The cost parameter is directly proportional to the slack variable, and acts as a trade-off between the training error rate in the classification and the maximum width of the margin; this model is termed as soft margin [7].

4.1.2 Decision Tree

A decision tree is a supervised learning algorithm based on Quinlan's algorithm used for classification [30]. The decision tree algorithm builds a tree with a root node and leaves [30]. The root node is selected based on the information gain value. First, the entropies of the classes are calculated, and then the entropy of each feature is calculated [30]. Information gain is the difference between the entropy of the classes and features [30]. The highest information gain attribute acts as a root node, and each node is constructed based on the information gain value. The tree is allowed to grow in this manner [30]. Lastly, patterns are induced by starting from the root, making a decision at each node, following one branch at each step, ending with a leaf node that corresponds to a certain class. One of the advantages of the decision tree is that it is very easy to understand [30].

4.1.3 Random Forest

Liaw et al. proposed the random forest classifier, which is a model that combines multiple decision tree predictors [22]. In this classifier, the data set is divided into training and testing sets. The training set is further divided into two subsets: in the bag and out of the bag. Two thirds of the training samples are in the bag [22]. They are sampled in such a way that the number of training samples is equal to the number of samples in the bag. The sampling is done with replacement, and also known as bootstrapping [22]. The remaining one-third of the data corresponds to the set that is out of the bag (OOB) [22]. In the bag samples are input to the decision trees, which learn the classification rules from the given data used to predict out of bag samples [22].

Figure 4.2 shows how the random forest works, when out of bag samples are given to them as input. Each decision tree in the random forest will predict the class independently, based on the OOB data. Each tree votes to which class each sample belongs. The total vote count is calculated, and the majority-voted class is assigned to that sample. In the figure, decision tree 1 and decision tree 2 voted for the + class; therefore, class + is assigned to the sample. The decision tree is grown to the fullest, and there is no need for pruning. Random forest is really fast and usually achieves very good accuracy for large data sets. For these reasons, random forest was selected as one of our classification algorithms.

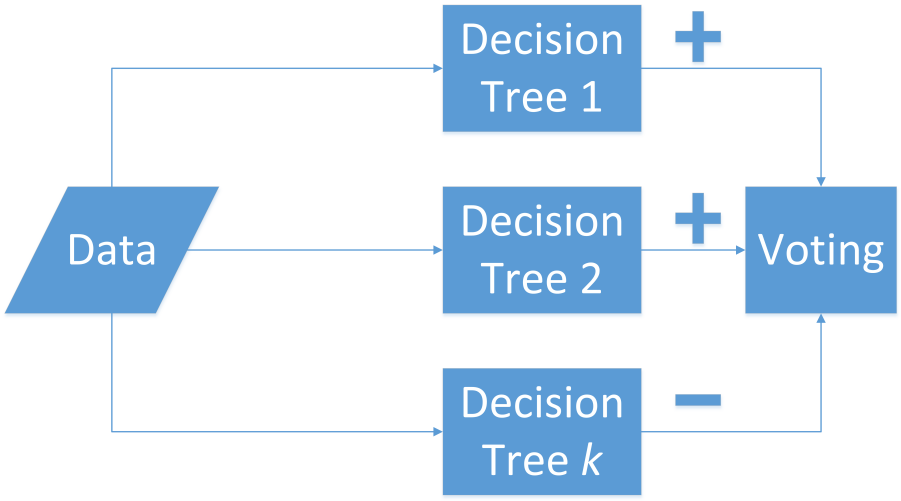


Figure 4.2: Random forest example.

4.1.4 Naïve Bayes

Naïve Bayes is a classification algorithm that uses Bayes' theorem [34]. It performs classification based on prior probabilities and likelihoods. Initially, consider a two-class classification problem [34]. The prior probabilities and likelihoods of both classes are measured with respect to the new input [34]. Finally, the posterior probabilities for the two classes are calculated [34]. The class that has the highest posterior probability value is assigned to that sample [34]. Naïve Bayes classification performance is good when the data is high dimensional; that is the reason for which Naïve Bayes was selected as one of the classification algorithms [34].

4.1.5 Multi-class Classification

In multi-class classification, there are more than two classes. The classification of stages of prostate cancer is an example of a multi-class problem. There are many ways to solve this

problem. Two common approaches are one-against-all and one-against-one [22].

In one-against-all, each classifier is trained and tested on one class versus the remaining classes [22]. If the data set has r classes, then r classifiers are built in this approach. A class is assigned to new samples by the classifier that outputs the highest confidence score [22]. All classifiers solve the one-against-all problem in this way [22].

In one-against-one, the classifier is developed for pair-wise classes [22]. Each classifier classifies a new sample with the class [22]. The class that receives the maximum number of votes is assigned to that sample [22]. We have adopted a special case of the one-against-one approach for our classification problem. The model is discussed in Chapter 5.

4.2 Feature Selection

Feature selection is a way of selecting a subset of features from the given data. It is used to identify and eliminate noisy and redundant features, thereby reducing the dimensionality of the data. Moreover, feature selection makes classification algorithms operate faster and more effectively. The goals of feature selection are to reduce the classifier's complexity and increase classification accuracy as much as possible.

Consider a pseudo example given in Table 4.2, when all the features (t_1, t_2, and t_3) are used by the classification algorithm; classification may not work efficiently. However, if we remove t_2 and t_3, classification might work better as compared to using all features. Alternatively, if those features are removed then classifier may not necessarily be more accurate, due to existing interaction among features. Thus, features that have the capability to discriminate both classes are preserved, while others are removed by feature selection algorithms. There are two types of feature selection techniques: filter and wrapper methods.

In the filter approach, statistical methods are mostly used to score each feature and filter

out irrelevant features. These methods work very fast and ignore any dependencies among the features. Chi-squared is one of those methods that is used in this research.

Table 4.2: Example of a two-class classification problem that involves two types of cancer, matched normal and malignant.

Samples	t_1	t_2	t_3	Class
S_1	1	0	1	Malignant
S_2	0	0	0	Matched normal
S_3	0	0	1	Matched normal
S_k	1	1	0	Malignant

4.2.1 Chi-squared

Chi-squared is a statistical model that calculates a statistical score based on the χ^2 distribution. Initially, features are assumed to be independent, and χ^2 values are calculated for all features [23]. The features are then ranked by their χ^2 value in descending order. We have used chi-squared feature selection in our method, because it operates very quickly and is less computationally-intensive than other methods in filtering features.

In wrapper methods, all the features are mapped into the feature subset space, and the classification algorithm is used to select a subset of features. The major advantage of wrapper methods is that more informative features are selected, because they consider interactions among the features. On the other hand, the disadvantage is that it is very slow when there are a large number of features. Using wrapper methods also incurs a higher risk of over-fitting the data. mRMR is a wrapper method used in our model.

4.2.2 mRMR

Minimum redundancy and maximum relevance (mRMR) is a feature selection method that depends on mutual information values. The mRMR technique is implemented as a wrapper methods, and its main concept is maximum dependency [29]. It selects features in a two-step process. Initially, mRMR selects the most relevant subset of features that have maximum relevance for the target class, that is, mutual information [29]. Consider again the example of Table 4.2 Suppose that we use of features t_1 and t_2 features results in classification accuracy of 100%. If we then remove t_2 and the classification accuracy is 100% with only the feature t_1, then it is useless to include t_2 feature. This approach minimizes redundancy among the selected subset of features. This is the key benefit of mRMR as compared to other feature selection algorithms. Since in the main problem addressed in this thesis we are looking for meaningful transcripts associated with prostate cancer progression, mRMR is used as a feature selection method. In addition, we have to choose a classification algorithm, since mRMR is a wrapper method. An SVM with a linear kernel was used because it yielded good results compared to other classification algorithms, as shown later in the experimental results.

4.3 *k*-Fold Cross-Validation

In this work, *k*-Fold cross-validation is used for classifier validation. This validation method works as follows. Initially, the input data are divided into *k* equal subsets. The classifier is then trained on *k*-1 subsets and tested on the remaining part. Figure 4.3 illustrates 10-Fold cross-validation; we have used 10-Fold cross-validation in our this thesis. The data set is divided into ten equal subsets. Nine subsets are given to the classifier for training, and one

part is used for testing the model. This process is iterated 10 times. Finally, the mean of the desired performance measure is calculated to evaluate the classifier.

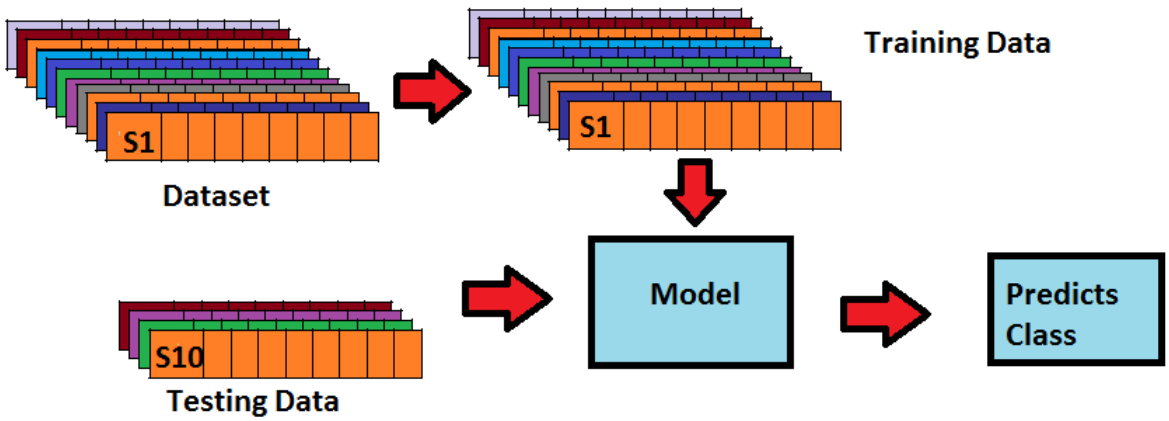


Figure 4.3: Illustration of the k-Fold cross-validation process.

4.4 Performance Measures

Performance measures are required to compare the classifiers' performance on the data sets. Each classifier reports a confusion matrix, which helps in evaluating the performance of the classifier. There are many performance measures that can be used to compare classifiers: accuracy, F-measure, area under the curve (AUC), and the Matthews correlation coefficient (MCC).

Figure 4.4 represents the formulas for calculating various performance measures. Consider a classification problem in which there are two classes, positive and negative. In the case of prostate cancer classification, the positives are malignant samples, and the negatives are matched normal samples. In prostate cancer progression classification, the positives are the highest progressive stage, and the negatives are the lowest progressive stage.

For instance, consider the classification of the T3a and T3b classes. The positives are T3b samples, while the negatives are T3a samples.

Based on Figure 4.4, the actual classes are the labels associated with each original sample, whereas the predicted classes are the classifier-predicted classes for the samples. A true positive occurs when a positive sample is predicted as a positive sample, while a false positive occurs when a negative sample is predicted as a positive sample. Similarly, a false negative occurs when a positive sample is classified as a negative sample. Lastly, a true negative occurs when a negative sample is classified as a negative sample.

Generally, accuracy is a good performance metric in the case of balanced data sets. The higher the accuracy, the better the performance of the classifier is considered (see Figure 4.4). Precision and recall refer to the positive samples. They focus on how well the classifier classifies only the positive samples. Recall is also known as sensitivity or true positive rate. The false positive rate is the difference between 1 and specificity. Precision is the probability of a sample being positive and actually being predicted as positive. Precision and recall are inversely proportional to each other. F-measure is the harmonic mean of both precision and recall. The higher the harmonic mean, the better the classifier is considered.

As shown in the Figure 4.4, the Matthews correlation coefficient (MCC) is considered to be a balanced performance measure to evaluate a classifier. Referring to the formula in the figure, MCC deals with all positives and negatives from the confusion matrix. The MCC value varies from -1 to +1. If the value is close to -1, then the classifier contradicts the actual and predicted classes. If the value is +1, then the classifier is considered the best classifier. If the value is 0, the classifier has performed a random prediction.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN)	P = TP + FN Recall = TP/P
	Negative	False Positive (FP)	True Negative (TN)	N = TN + FP Specificity = TN/N

n = TP + FN + TN + FP Precision= TP/(TP+FP) $\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{\text{n}}$

$$\text{F-measure} = 2\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

$$MCC = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

Figure 4.4: Performance measures used to evaluate the efficiency of a classifier.

Finally, the receiver operating characteristic (ROC) is a graph that is used to measure the performance of a classifier. In Figure 4.5, the false-positive rate is plotted on the x-axis, while the true positive rate is plotted on the y-axis for different thresholds. If the classifier is close to the northwest corner, that classifier is considered the best. If the classifier is close to the southeast corner, that classifier is considered the worst. From the figure, curves A, B, and C corresponds to the best, random, and worst classifiers, respectively. However, a quantitative measure is better suited for comparing classifiers. Thus, the area under the receiver (AUC) operating characteristic is calculated for this purpose. The classifier with the highest AUC is said to be the best classifier.

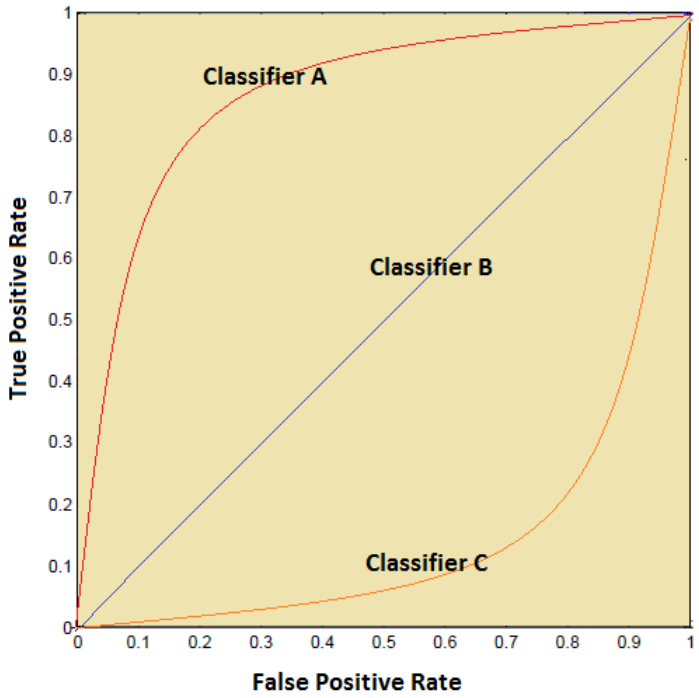


Figure 4.5: Receiver operating characteristic.

4.5 Conclusion

In this chapter, we discussed machine learning algorithms used for classification and feature selection. We have also reviewed cross-validation and performance measures in order to compare classifiers. In the next chapter, the method that we have developed is examined via some computational experiments we performed.

Chapter 5

Methods

In this chapter, we discuss the methodology used to work on RNA-Seq data sets in order to extract transcripts. These transcripts act as potential biomarkers for identifying prostate cancer and estimating progression stages. Moreover, machine learning techniques, such as classification and feature selection, were used on these transcripts to find those that are differentially expressed.

5.1 Datasets

There are many RNA-Seq data sets available for prostate cancer and progression stages [21]. We selected three data sets that deal with matched normal versus malignant prostate cancer classification: Kim's [15], Ren's [33], and Kannan's [13]. The data set from Long et al. [23] was also used, which deals with classification of prostate cancer progression stages with a large number of samples present in each stage. Ren's data set used random hexamer primers, while the other data sets used oligo (DT) primers. All these data sets are in sequence read archive (SRA) file format and are publicly available from the national

center for biotechnology information (NCBI) repository [32]. Details about the data sets are shown in Table 5.1. The second column in the table represents the data set number used by the NCBI repository [32]. Ren et al. researched prostate cancer in the Chinese population using 14 matched prostate samples, whereas Kim et al. studied four matched normal and seven malignant samples. Kannan et al. investigated ten matched prostate samples. Long's data set consists of 106 tumour samples.

Table 5.2 depicts the number of samples present in the various stages of prostate cancer in Long's data set. The first column in the table identifies the cancer stage, while the second column specifies the number of patients per stage. The aligner tool that we use accepts FASTQ/FASTA file formats. All the samples were converted from SRA to FASTQ file format.

Table 5.1: Data sets used in our work.

Reference	Data accession number	Number of samples	Study performed
Long et al. [25]	GSE54460	106 malignant	Identified differentially expressed genes
Ren et al. [33]	ERP000550	14 matched	Identified gene fusions and non-coding RNAs
Kim et al. [15]	GSE29155	four matched normal and seven malignant	Identified methylation patterns
Kannan et al. [13]	GSE22260	10 matched	Identified alternative splicing and gene fusions

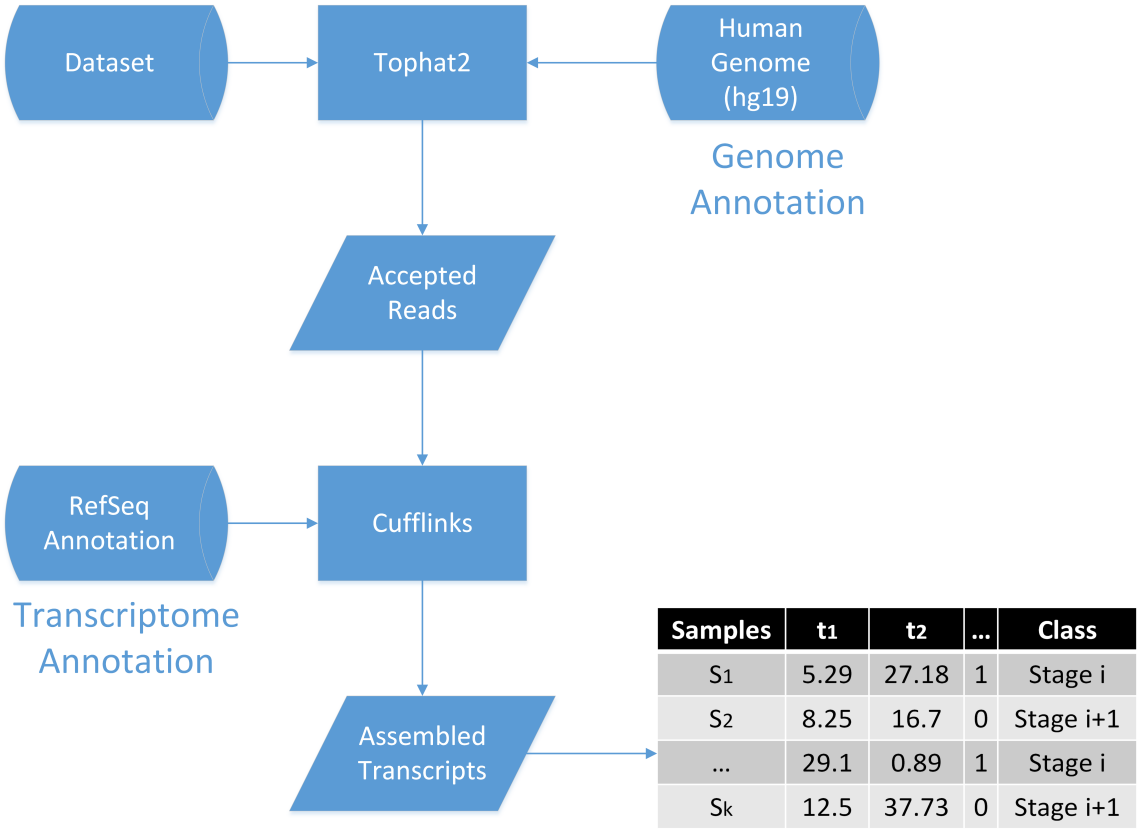
Table 5.2: Long's data set samples in different stages of prostate cancer

Prostate cancer stage	Number of samples
T1c	14
T2	10
T2a	23
T2b	11
T2c	30
T3	2
T3a	6
T3b	8
T4	1

5.2 Data Preprocessing

All SRA file format samples were converted to FASTQ files and sent to the preprocessing stage. Figure 5.1 shows a diagram for the preprocessing step performed on the data sets in order to extract the transcripts. Tophat2 is used to align the reads. The inputs to Tophat2 are the FASTQ files from the patients and human genome (hg19) [17]. This tool outputs the reads that are aligned to the reference genome, which are known as accepted reads. Cufflinks is then used to perform transcriptome assembly. The inputs to this tool are accepted reads and transcriptome annotation (RefSeq) [31]. Cufflinks outputs transcripts that are assembled for which their abundance are calculated using FPKM values. This preprocessing step is repeated for each sample.

For each data set, we constructed a table with transcripts the corresponding FPKM values for each sample. Figure 5.2 shows a sample table, which contains the transcripts and their FPKM values. We already know that each sample belongs to a matched normal or malignant class, or to a different stage of prostate cancer. This is represented in the last column as the class label.



Samples	t1	t2	...	Class
S1	5.29	27.18	1	Stage i
S2	8.25	16.7	0	Stage i+1
...	29.1	0.89	1	Stage i
Sk	12.5	37.73	0	Stage i+1

Figure 5.1: Preprocessing phase of our method: Tophat2 aligns the reads to the reference genome, and Cufflinks assembles the transcriptome and estimates transcript abundance.

NM_000016	NM_000017	NM_000018	NM_000019	NM_000014	NM_001030	Class
4.82855	1.20165	16.3065	7.40305	68.2911	72.4166	T1c
0.878737	0.651515	12.9071	1.09227	8.25032	25.0891	T2
0.876861	0.233907	13.6248	1.19375	29.132	25.0951	T1c
1.22128	0.390938	11.8374	2.86261	12.597	18.0403	T1c
1.16996	0.27308	2.31995	1.52351	27.818	15.3085	T1c
1.10009	0.171302	1.75097	2.83616	16.7828	14.2632	T2
1.50573	0.159752	4.12922	1.56914	9.01934	12.4745	T1c
1.31807	0.438502	3.08118	2.15501	27.2023	25.7941	T1c
0.600567	0.709654	8.47392	1.63298	14.4311	10.1924	T2
1.06015	0.0771271	5.8541	0.462074	22.1183	6.16565	T2
0.209974	0.0294061	4.72282	1.22669	11.0732	5.44355	T1c
2.1172	0.484487	3.98431	1.46205	40.4463	10.9402	T2
0.853989	0.420091	12.2512	1.61572	8.50849	14.6241	T1c

Figure 5.2: A sample input file for the classification algorithm. This file is the output of the preprocessing phase.

5.3 Classification and Feature Selection

We have used Weka, a data mining tool that integrates feature selection and classification algorithms [12]. Weka is an open-source Java tool developed by the University of Waikato and is widely used for data mining in bioinformatics and other fields. Figure 5.3 shows the pipeline of our proposed method. The preprocessing step produces a table that contains transcripts and their FPKM values, as discussed in the previous section. In the figure, there are two tables. The one on the left is for matched normal versus malignant tumour classification, and the one on the right is for prostate cancer progression. These two tables act as inputs to the feature selection algorithm, which will filter out noisy and redundant transcripts.

The filtered transcripts are sent to the classification algorithms. The classifiers classify

the samples based on the classification rules learned in the training phase. Finally, the differentially expressed transcripts are obtained.

10-Fold cross-validation is performed on the classifiers to maintain their generalization capability on the test set. Performance measures are used to evaluate the performance of the classifiers on different data sets.

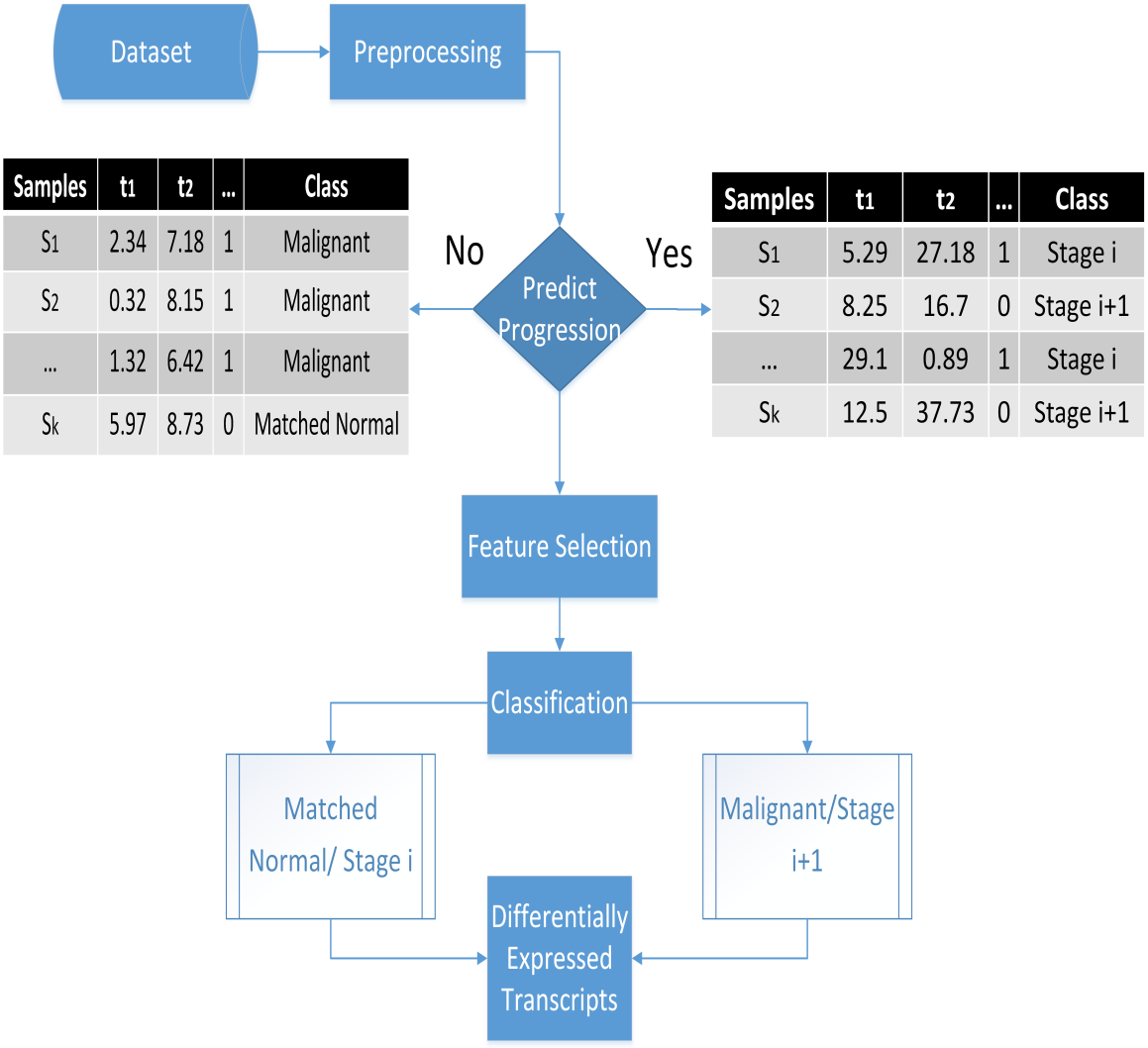


Figure 5.3: Pipeline of our method for matched normal versus malignant and prostate cancer progression classifications.

5.3.1 Multi-class Problem

Since we deal with different stages of prostate cancer, we model the problem as a multi-class classification problem. We have already discussed the multi-class problem in Chapter 4. For this research, we consider a special case of the one-against-one scheme.

All biological processes are continuous. Consider cancer progression, in which the cancer continuously grows at each stage. Moreover, we are interested in finding differentially expressed transcripts between neighboring stages. In Long's data set, we have resolved the multi-class problem by comparing samples between neighboring stages, such as T1c with T2, and T2 with T2a. Another challenge in Long's data set is that there are very few samples in some stages, such as T3 and T4. This may cause a classifier to suffer from the problem of over-fitting, due to a large number of features. To avoid this, we have merged T3 samples with the T3a stage into one class called **T3a** and T4 samples with the T3b stage into a single class called **T3b**.

We are more concerned about finding differentially expressed transcripts from T2 to T3 stages, because they play a vital role in progression of prostate cancer. In the T3 stage, the tumour growths rapidly and aggressively to eventually getting closer to the metastasis stage. Therefore, in addition to considering neighboring stages, we have also added another class as a result of merging all the samples from stages T3, T3a, T3b, and T4; we call this class T34.

5.3.2 Feature Selection

Feature selection has been previously discussed in Chapter 4. We have used feature selection because the preprocessing stage extracted 43,497 transcripts per sample. Applying these features to the classification algorithms produces very poor performance, while being

an intractable problem. Due to irrelevant and noisy transcripts, the classification algorithms performed very poorly. Thus, we have adopted to remove features that degrade classification performance. We have used two feature selection techniques: chi-squared and mRMR.

Chi-squared is a filter method in which we have full control for selecting the number of features we desire. We have selected the top 200 features, because the classifier's performance drastically decreased above 200 features. We send the selected features to the classification algorithms in prioritized order, such as top 1, top (1,2), top (1,2,3), ..., and top (1,2,3,..,200). The mRMR wrapper methods is also used to select the features. In this method, we do not have any control on the number of features. The mRMR method uses a classification algorithm to filter features. We experimented with different classification algorithms. An SVM with a linear kernel worked the best on our data sets. Therefore, we have used SVM with a linear kernel in the mRMR feature selection technique.

5.3.3 Classification

The selected transcripts are sent to the classification algorithms to perform classification. The four classifiers discussed in Chapter 4 are applied: SVM (with linear, polynomial, and radial basis function kernels), random forest, decision trees, and Naïve Bayes.

SVM was selected due to better performance as compared to the other classifiers, especially on biological problems. Random forest is a very fast classifier, particularly on large data sets. We have used 10 decision trees for random forest classifier. Decision trees also work efficiently on similar data sets. Nevertheless, the Naïve Bayes classifier produces very good results on large data sets. All classifiers performed classification with default parameters.

The transcripts filtered by feature selection techniques were added to the Weka tool.

Then, the classification algorithms were selected accordingly and applied with default parameters. These algorithms identify differentially expressed transcripts, Weka provides performance measures to evaluate the learning algorithms.

5.3.4 Performance Evaluation

We have used 10-Fold cross-validation. The positives in the two-class problem are malignant samples, and the negatives are matched normal samples. In the case of progression, which is modeled as a multi-class problem, the top stages are considered positives, while the others as negatives. For example, in T3a and T3b classification, the positives are the T3b samples, and the negatives are the T3a samples. This process is repeated for all other stages. We selected two performance measures, accuracy and AUC, for comparison of classifiers and analysis of results. This is discussed in detail in Chapter 6.

5.4 Biological Significance

Finally, differentially expressed transcripts were obtained from the matched normal versus malignant classification and prostate cancer stage classification. We have found some biological knowledge about of these transcripts. The NCBI website [28] provides information on the transcripts, such as location (locus) in the corresponding chromosome, genes, and other biological information. The transcripts that are common among both matched normal versus malignant or stages were selected for biological significance. A literature review was performed on these selected transcripts for biological significance, any previously-identified relationships to prostate cancer and other types of cancers.

5.5 Comparison with other methods

We have compared our model with CuffDiff, which is a part of the Cufflinks package and a differential-expression tool [3]. CuffDiff performs a statistical test based on Benjamini and Hochberg multiple testing [3]. First, the p-value is determined and then the q-value, which is a multiple test-corrected value calculated for each feature. More details about finding the p-value and q-value can be found in [3]. If the p-value is greater than the q-value, that feature is said to be significant or differentially expressed.

Figure 5.4 shows a workflow for the model we use for comparison. The assembled transcripts from Cufflinks act as input to CuffMerge, which is also part of the Cufflinks package. CuffMerge combines all the transcripts from different samples. We have used CuffDiff with default parameters to find differentially expressed transcripts. The merged transcripts from CuffMerge act as input to CuffDiff with matched normal versus malignant or different stages. To compare the CuffDiff results with our model, these differentially expressed transcripts act as input to the classification algorithms used in our model. Then, performance measures were used to evaluate as discussed in Chapter 6.

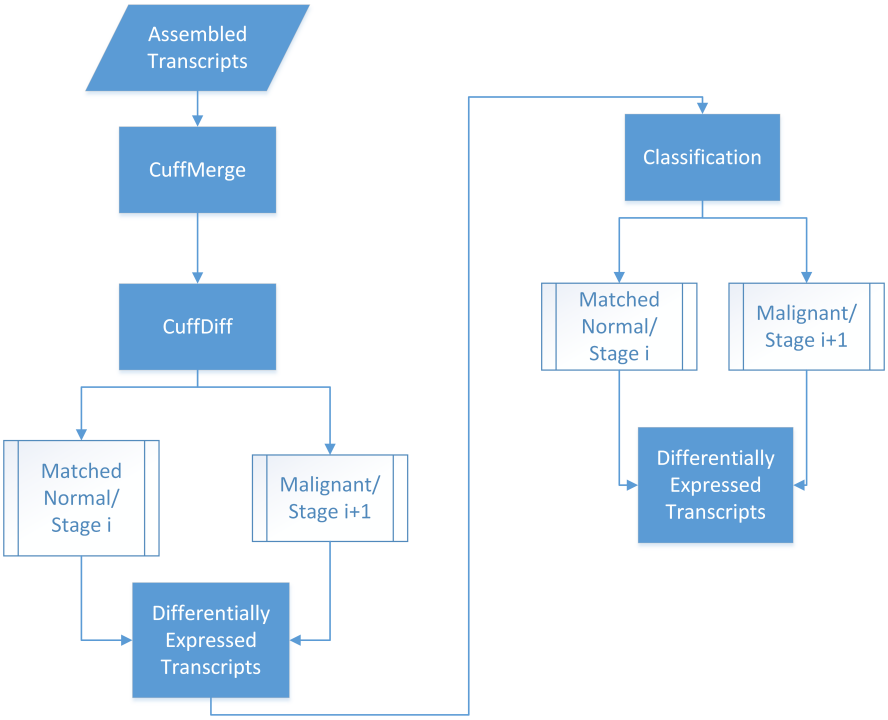


Figure 5.4: Workflow for the model we use for comparison.

5.6 Conclusion

To conclude this chapter we highlight that our method has been described in detail in this chapter. We discussed the data sets used in our model, the preprocessing phase, and the pipeline for the model used in our research. We have evaluated our method with performance measures and compare it with existing methods. In the next chapter, the results are provided and discussed in detail.

Chapter 6

Results and Discussion

In this chapter, the results are organized in matched normal versus malignant and prostate cancer stage classifications. In each section, performance measures and biological significance are discussed.

6.1 Matched Normal Versus Malignant Classification

The matched normal versus malignant classification was performed on three data sets. The transcripts extracted from the preprocessing stage were processed by two feature selection techniques. Subsequently, classification and feature selection algorithms were employed to find differentially expressed transcripts.

6.1.1 Performance Measures

The two feature selection techniques, chi-squared and mRMR, were applied to the transcripts. Several classification algorithms were used on the filtered transcripts: SVM (with linear, and RBF kernels), random forest, decision trees, and Naïve Bayes. Due to its poor

performance, the polynomial kernel was omitted from our model. The accuracy and AUC performance measures were visualized for each classifier.

mRMR Feature Selection

Applying mRMR feature selection to Kannan's, Ren's, and Kim's data sets resulted in the identification of 2, 1, and 5 transcripts, respectively. These transcripts were input to each classifier, and performance measures for each classifier were recorded. Figures 6.1 and 6.2 show the performance measures, accuracy and AUC, respectively. In Figure 6.1, the x-axis represents the three data sets, while the y-axis represents accuracy. In Figure 6.2, the x-axis shows the three data sets, and the y-axis indicates the AUC values that are obtained. The SVM with a linear kernel outperformed the other classifiers for the three data sets.

SVM with linear kernel performed good on the Kannan's data set, because the kernel trick of implicitly mapping of features to another dimension make them linearly separable. On Ren's data set, random forest and Naïve Bayes classifier performance was comparable to the SVM with a linear kernel. In Kim's data set, all four classifiers achieved high accuracy and AUC values, because this data set has seven malignant and four matched normal samples.

Chi-squared Feature Selection

Similarly, chi-squared feature selection was applied to the transcripts obtained from the preprocessing stage. We have selected the top 200 transcripts, because the performance of the classifiers decreased drastically above that number.

The SVM with a linear kernel is visualized for chi-squared feature selection, because it performed the best among the four classifiers. Figure 6.3 illustrates the accuracy of the

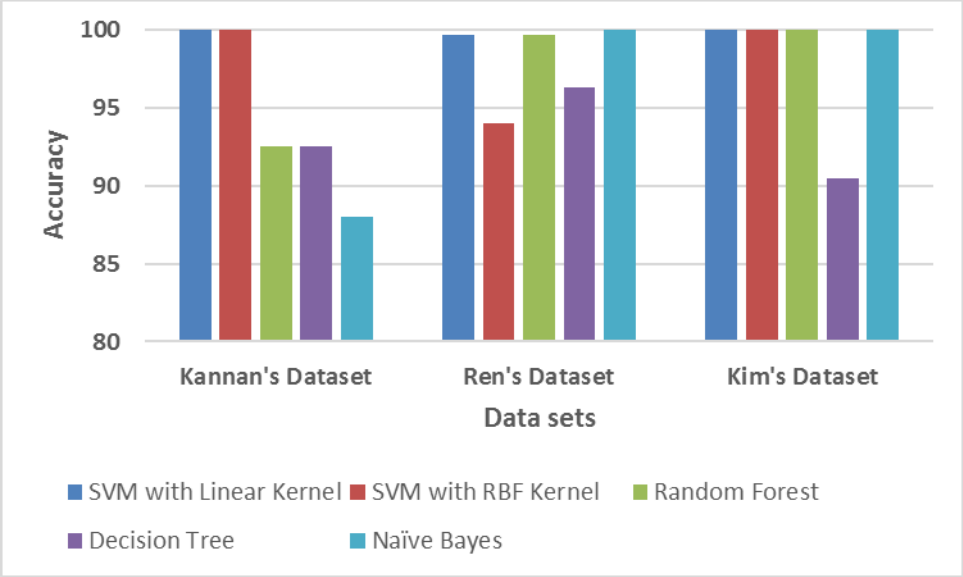


Figure 6.1: Accuracy of classifiers for matched normal versus malignant classification using mRMR feature selection.

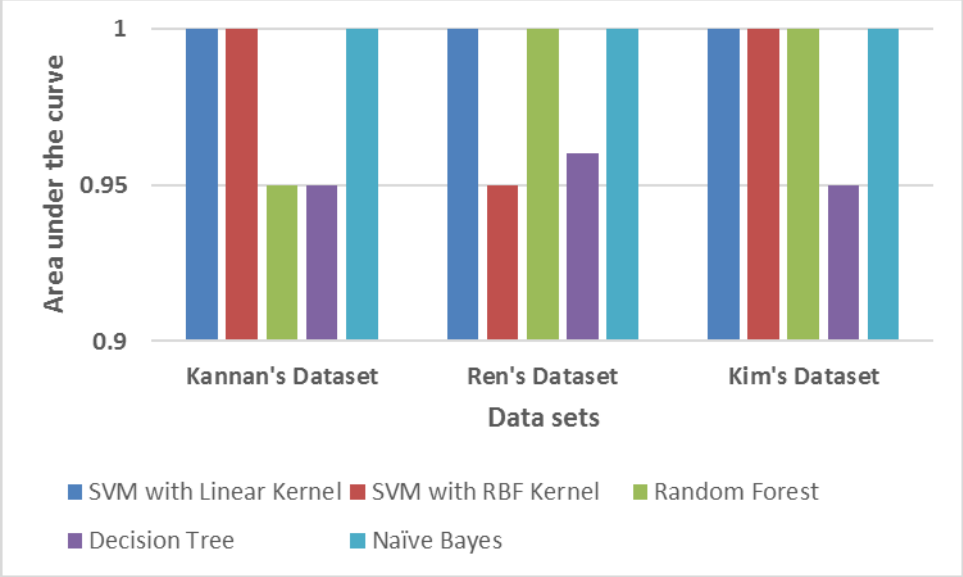


Figure 6.2: AUC of classifiers for matched normal versus malignant classification using mRMR feature selection.

SVM classifier with a linear kernel as a function of the number of features. Similarly, Figure 6.4 depicts the AUC for the SVM with linear kernel as a function of the number of features. In both figures, blue, red, and green-coloured lines represent Kannan's, Ren's, and Kim's data sets, respectively.

The performance of the SVM with linear kernel fluctuated on Kannan's data set until 65 features were used. Performance stabilizes at 110 features, indicating that the top 110 features provide the classifier good discriminative power. The classifier achieved 100% accuracy and good AUC values on Ren's data set when the number of features was greater than 20. Performance using the top 20 features was unstable and made classification difficult to perform, due to noisy features. On Kim's data set, all combinations of features achieved 100% accuracy, because there are fewer samples: four matched normal and seven malignant samples.

6.1.2 Biological Significance

Table 6.1 shows matched normal versus malignant differentially expressed transcripts. Kannan's, Kim's, and Ren's data sets have 4, 4, and 6 transcripts, respectively. We have found three common transcripts (red-colored transcripts in Table 6.1), which were investigated about their biological relevance for prostate cancer. Transcripts NM_019024, NM_001242889 and NR_024490 were present in the "HEAT repeat containing 5B (HEATR5B), Dopa Decarboxylase (DDC), and GABPB1 antisense RNA 1 (GABPB1-AS1)" genes, respectively [28]. HEATR5B and DDC gene transcripts were common between Kannan's and Kim's data sets. The GABPB1-AS1 gene transcript was common between Ren's and Kim's data sets. Additional biological information regarding transcripts such as location (locus) in the corresponding chromosome, genes can be found in appendix.

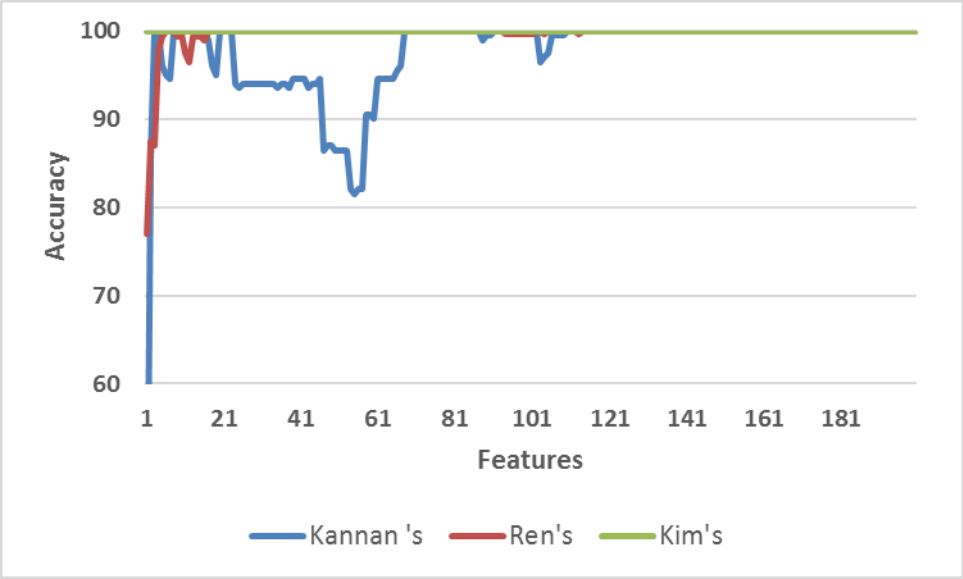


Figure 6.3: Accuracy of SVM with linear kernel for matched normal versus malignant classification using chi-squared feature selection.

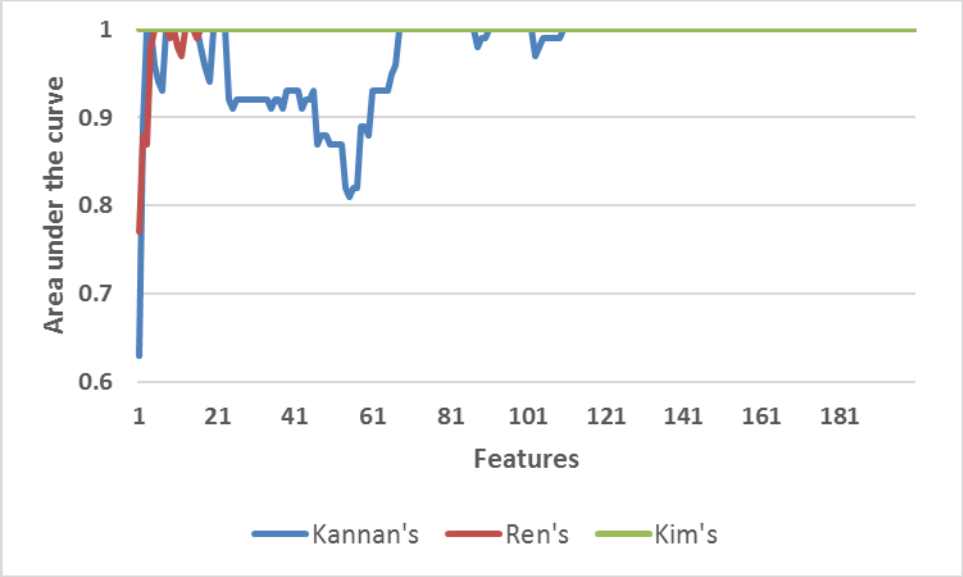


Figure 6.4: AUC of SVM with linear kernel for matched normal versus malignant classification using chi-squared feature selection.

Figure 6.5 shows the average of transcript abundance for matched normal and malignant. The bars represent mean FPKM values for three selected transcripts. The mean FPKM values were calculated for both malignant and matched normal samples in the three data sets. The mean value is considered because it is not biased by the number of samples. We observe from the figure that NM_001242889 is differentially expressed in malignant samples in comparison to matched normal samples. Avgeris et al. researched DDC and found that it was over-expressed in cancer samples as compared to matched normal samples [2]. Similar patterns were observed in our work, which suggests that NM_001242889 present in DDC gene is a relevant biomarker for prostate cancer.

6.2 Prostate Cancer Progression

Likewise, in prostate cancer progression, two feature selection techniques were employed with different classifiers. Long's data set was used to find differentially expressed transcripts. As discussed in Chapter 5, the stages in prostate cancer corresponds to the class labels. The feature vectors were input to the classifiers, and classification performance was graphically visualized. Finally, some of the common transcripts were selected to determine biological significance.

6.2.1 Performance Measures

Two feature selection techniques, chi-squared and mRMR, were used on the data sets followed by the application of the classification algorithms. The performance measures accuracy and AUC are discussed next for each pair-wise stage comparison in Long's data set.

Table 6.1: Matched normal versus malignant differentially expressed transcripts.

Kannan's data set	Kim's data set	Ren's data set
NM_019024	NR_024490	NR_024490
NM_001242889	NM_001242889	NM_000424
NM_152228	NM_019024	NM_001128826
NM_001204401	NM_032415	NM_000494
		NM_000700
		NM_005567

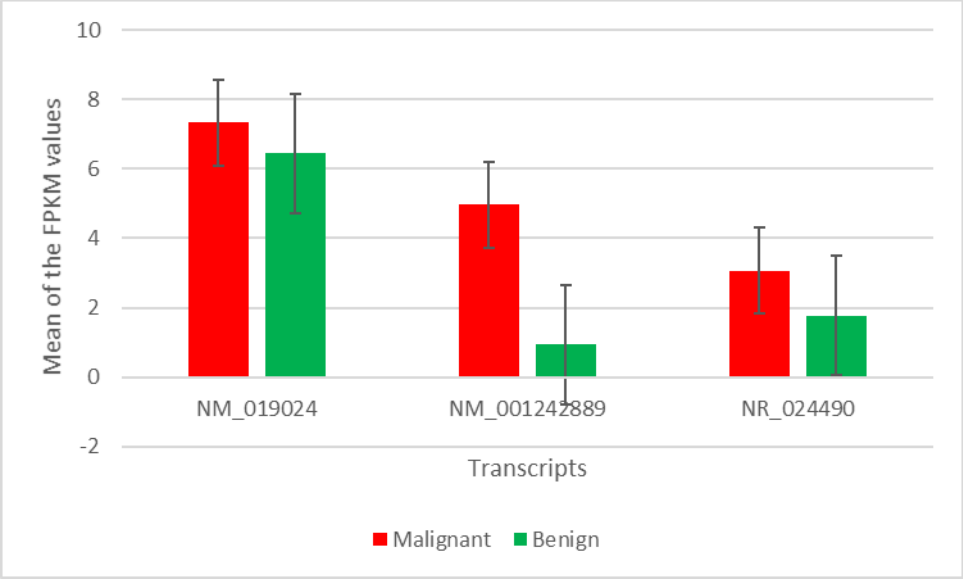


Figure 6.5: Expression trend of matched normal versus malignant transcripts.

mRMR Feature Selection

In the mRMR feature selection technique, there is no control of which features are selected because it is a wrapper method. This technique resulted in the identification of 44 transcripts for the set of all pair-wise stages. Table 6.2 shows the transcripts obtained for different stages of prostate cancer. The pair-wise stages T1c-T2, T2-T2a, T2a-T2b, T2b-T2c, T2c-T3a, T3a-T3b, and T2c-T34 have 6, 7, 6, 5, 5, 3, and 12 transcripts, respectively.

Figures 6.6 and 6.7 depict the performance of the transcripts selected by mRMR for Long's data set. The figures show the accuracy and AUC, respectively, for each classifier on each category of prostate cancer stage. Both figures show that the performance of SVM with a linear kernel is better than those of the other classifiers, especially in the case of T1c-T2 and T3a-T3b. The features of this classifier help find a discriminative function effectively. On the other hand, all other classifiers face noise with their features and are unable to achieve comparable performance.

Chi-squared Feature Selection

In chi-squared feature selection, the top 200 features were selected because of the decrease in classifier performance after that point. We have visualized the performance of the SVM with a linear kernel, because it achieved the highest performance for all pairs of stages compared to other classifiers.

Figures 6.8 and 6.9 show accuracy and AUC, respectively, of the transcripts when the SVM with a linear kernel was applied to the selected features. The x-axis represents the features, while the y-axis represents accuracy and AUC.

For all the pair-wise stages, the performance of the top 25 features is very poor. After that point, the selected features interact with each other and significantly improve the

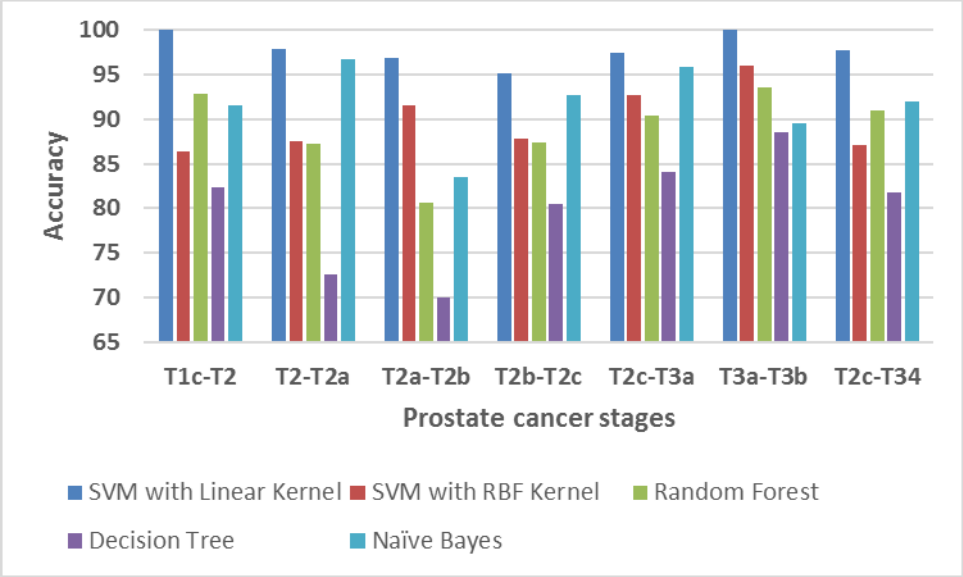


Figure 6.6: Accuracy of classifiers for pair-wise stage classification using mRMR feature selection.

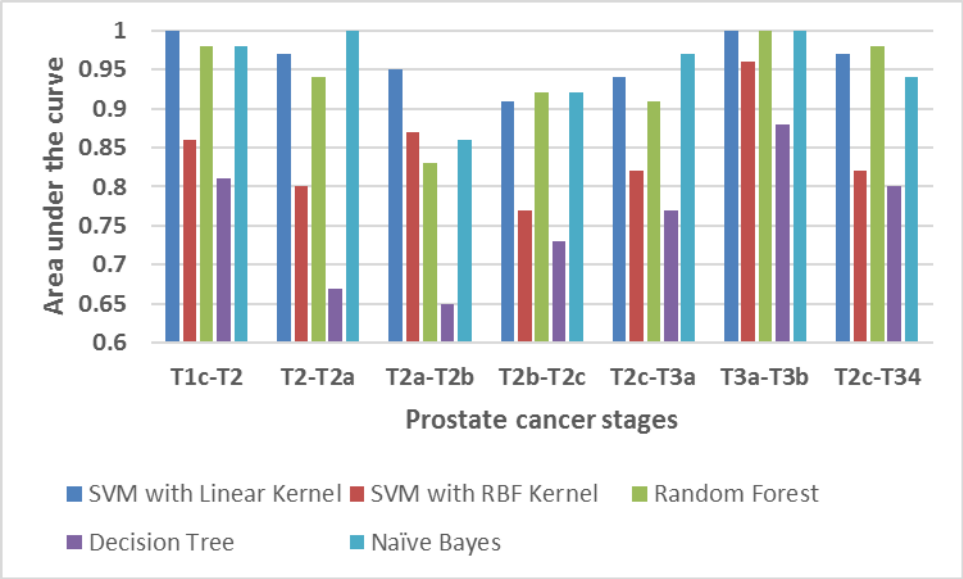


Figure 6.7: AUC of classifiers for pair-wise stage classification using mRMR feature selection.

performance of the classifiers. In the T3a-T3b classification, performance is good until the top 120 features are reached. Beyond that, performance decreases drastically to 95%, instead of increasing. This indicates that feature selection does not always yield good results. Accuracy and AUC remain stable for stage T2-T2a.

6.2.2 Biological Significance

The transcripts shown in the Table 6.2 were obtained by using feature selection techniques. We have found NM_032023 was differentially expressed between T2a-T2b and T2b-T2c pair-wise stages. In T2c-T34 classification, we observed four (red-colored transcripts) transcripts that were already associated with different types of cancer. Therefore, all the five transcripts were filtered for biological significance. Transcripts NM_032023, NR_003004, NM_003940, NM_000959, and NM_017753 were present in the Ras association (RalGDS/AF-6) domain family member 4 (RASSF4), small Cajal body-specific RNA 22 (SCARNA22), ubiquitin specific peptidase 13 (USP13), prostaglandin F receptor (FP) (PTGFR), and lipid phosphate phosphatase-related protein type 1 (LPPR1) genes, respectively.

Previous surveys suggest that the transcripts selected by our method are closely related to different types of cancer. Eckfeld et al. investigated RASSF4, which is a tumour suppressor in human cancer cells, and is found to be downregulated in lung cancer [10]. Ronchetti et al. observed that SCARNA22 was over-expressed in cancer cells [36]. Researchers previously found that the absence of USP13 indicates malignancy in breast cancer. USP13 binds to preserve Phosphatase and tensin homolog (PTEN), which is a tumor suppressor [48]. Romanuik et al. found that PTGFR was previously associated with different types of cancer, particularly breast, ovarian, and renal cancers [35]. Langlois et al. investigated

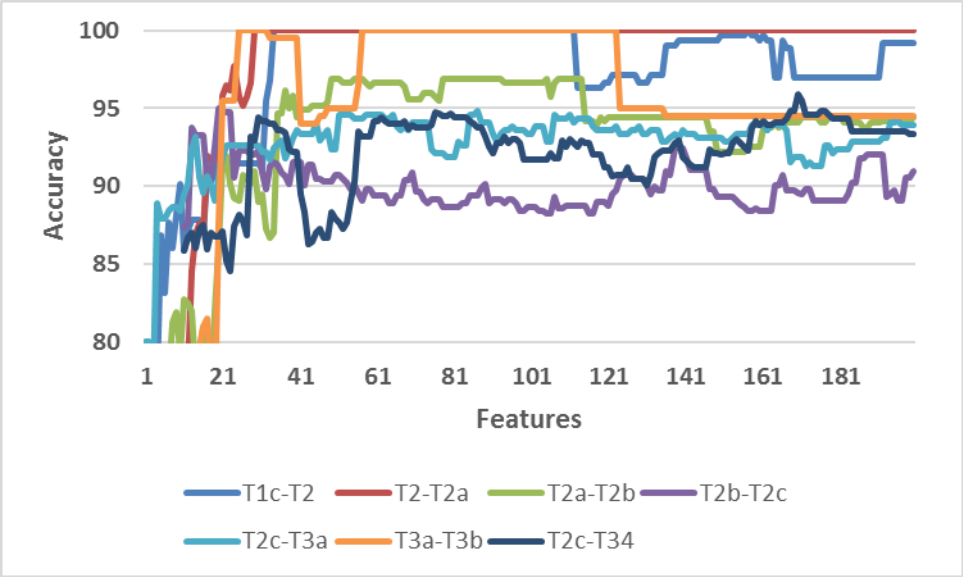


Figure 6.8: Accuracy of SVM with linear kernel for pair-wise stage classification using Chi-squared feature selection.

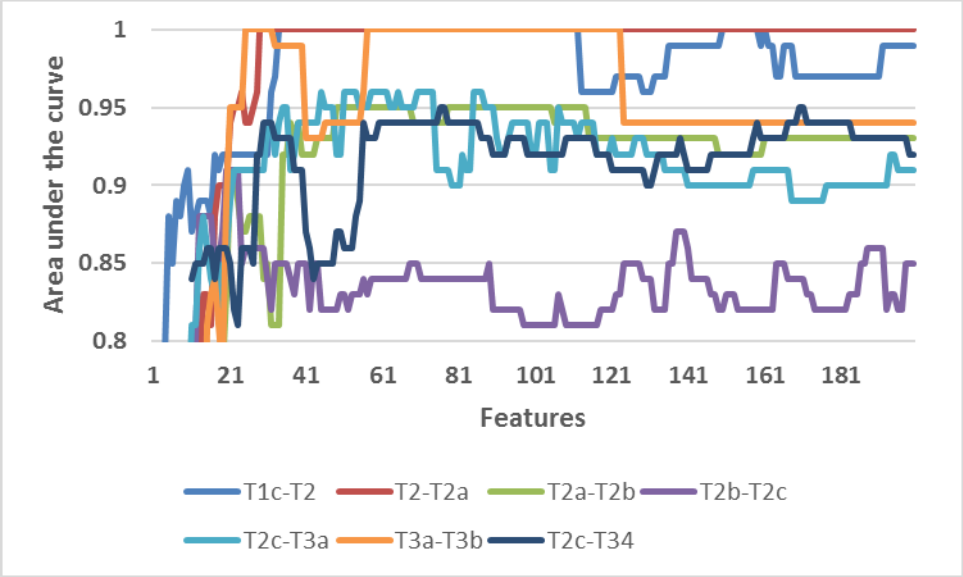


Figure 6.9: AUC of SVM with linear kernel for pair-wise stage classification using Chi-squared feature selection.

LPPR1 and found that it promotes cancer cell growth to metastasis [16].

Figure 6.10 shows the expression trend of Long's data set transcripts. The *x*-axis shows the different stages of prostate cancer, while the *y*-axis shows the average FPKM values. The mean of the FPKM values was used to remove any bias due to the imbalance in the number of samples present in the different stages. We observe that transcript NR_003004 present in SCARNA22 gene is differentially expressed in the T3b stage. NM_003940 present in USP13 gene is differentially expressed in the T2c stage; that transcript behavior was observed in breast cancer tumorigenesis [48]. These observations suggest the potential of the biomarkers found in our study.

6.3 Comparison with CuffDiff

We have used CuffDiff, which is part of the Cufflinks package, for comparison with our method. CuffDiff uses statistical approaches, while our method uses machine learning techniques to extract differentially expressed transcripts. These transcripts act as input to the classification algorithms, and the corresponding performance measures were calculated.

Figures 6.11 and 6.12 illustrate CuffDiff and our model selected transcripts' accuracy performance on different classifiers. The stages are plotted on the *x*-axis, while accuracy values were plotted on the *y*-axis. It can be inferred from the figures that CuffDiff selected transcripts accuracy values were below 90%. However, the accuracy values of our method were above 95% for SVM with linear kernel. The features that CuffDiff selected were not capable of performing adequate classification.

Figures 6.13 and 6.14 illustrate CuffDiff and our model selected transcripts' AUC performance on different classifiers. The stages are plotted on the *x*-axis. AUC values were plotted on the *y*-axis. Both figures depict that AUC values of our method were above 0.9 for

Table 6.2: Long's data set differentially expressed transcripts across different stages.

T1c-T2	T2-T2a	T2a-T2b	T2b-T2c	T2c-T3a	T3a-T3b	T2c-T34
NR_003669	NM_004860	NM_032023	NM_001711	NM_001198979	NR_034169	NM_001257413
NM_001160393	NM_052850	NM_080792	NM_032023	NM_001099285	NM_015380	NM_003940
NM_001161345	NM_001272095	NM_000095	NM_001014443	NM_001198899	NR_046417	NM_001142274
NM_052857	NM_001261390	NM_003102	NM_021724	NM_001130048		NM_001199165
NR_003594_7	NM_153274	NM_080797	NM_012098	NM_000899		NM_052965
NR_033240	NM_001252641	NM_002725				NM_001195283
	NR_038352					NM_001023567
						NM_001143766
						NR_003004
						NM_017753
						NM_000959
						NM_004772

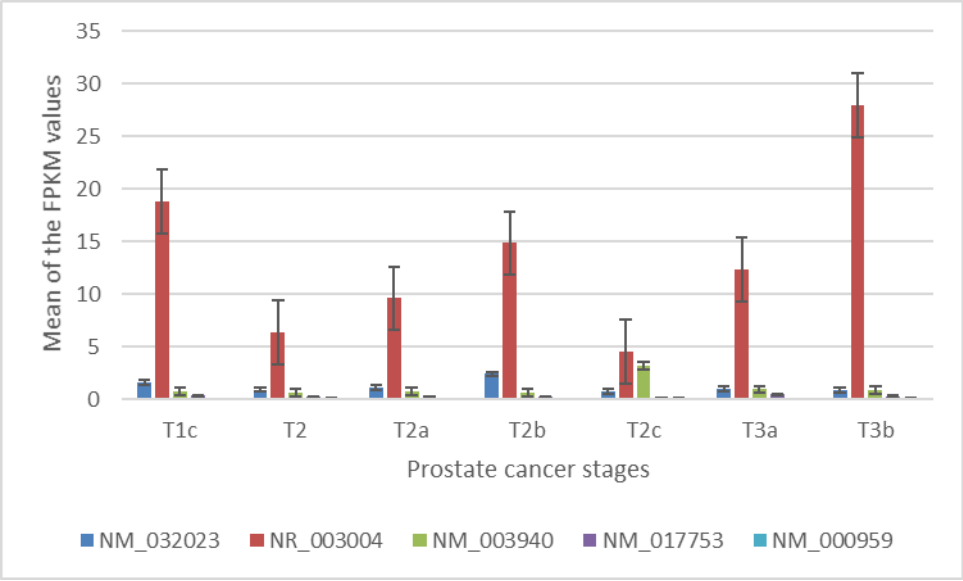


Figure 6.10: Expression trend of Long's data set transcripts.

all the pair-wise stages with SVM (linear kernel). On the other hand, CuffDiff's selected transcripts performed poor classification as AUC values for most of the classifiers were around 0.5. Therefore, this comparison suggests that our method obtained good results when compared to CuffDiff's selected transcripts.

6.4 Conclusion

In this chapter, we have briefly discussed the performance measures across different classifiers for the two feature selection techniques used. We have also compared the existing CuffDiff tool with our method. We found that our method achieved higher performance than CuffDiff. In the next chapter, conclusions and future work are discussed.

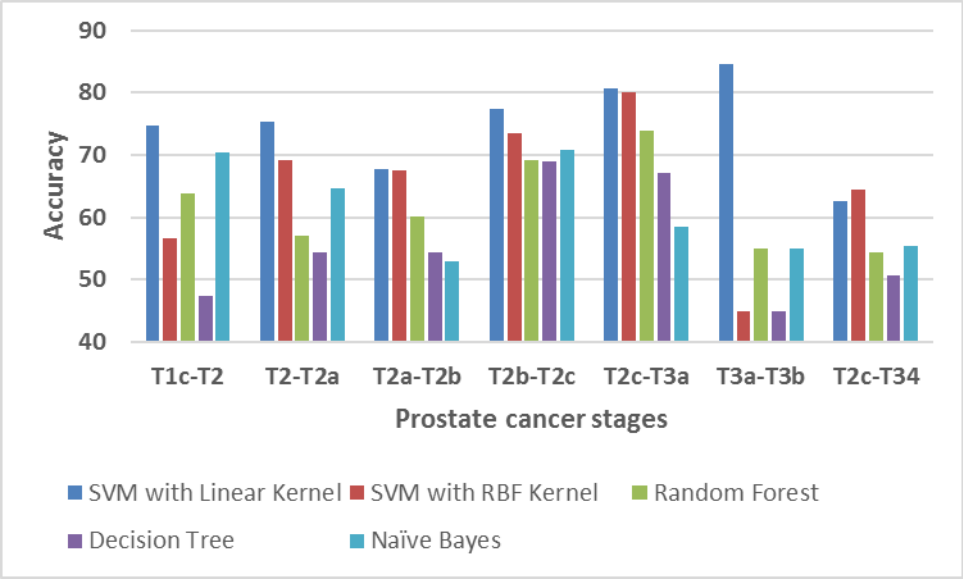


Figure 6.11: Accuracy of classifiers for CuffDiff selected transcripts on Long's data set.

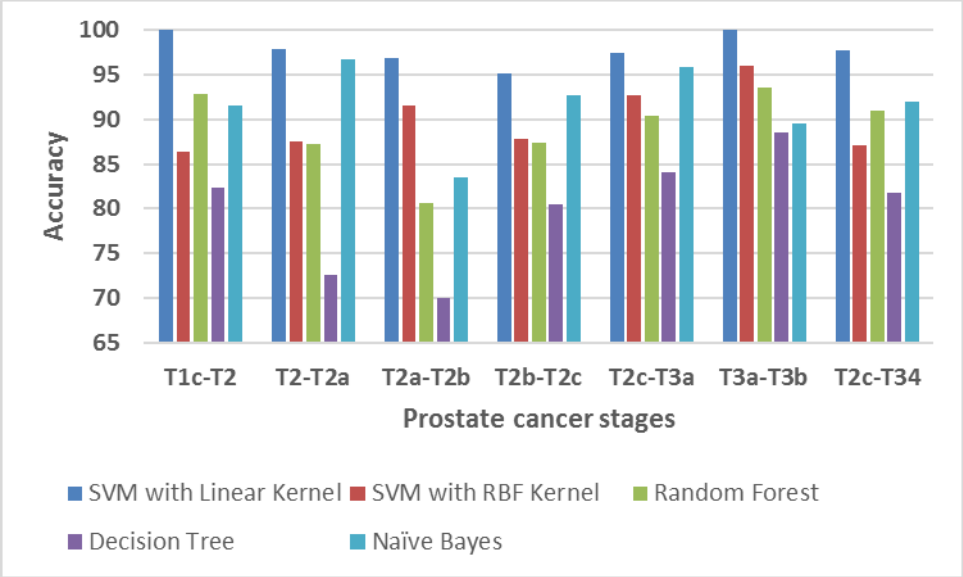


Figure 6.12: Accuracy of classifiers for our method selected transcripts on Long's data set.

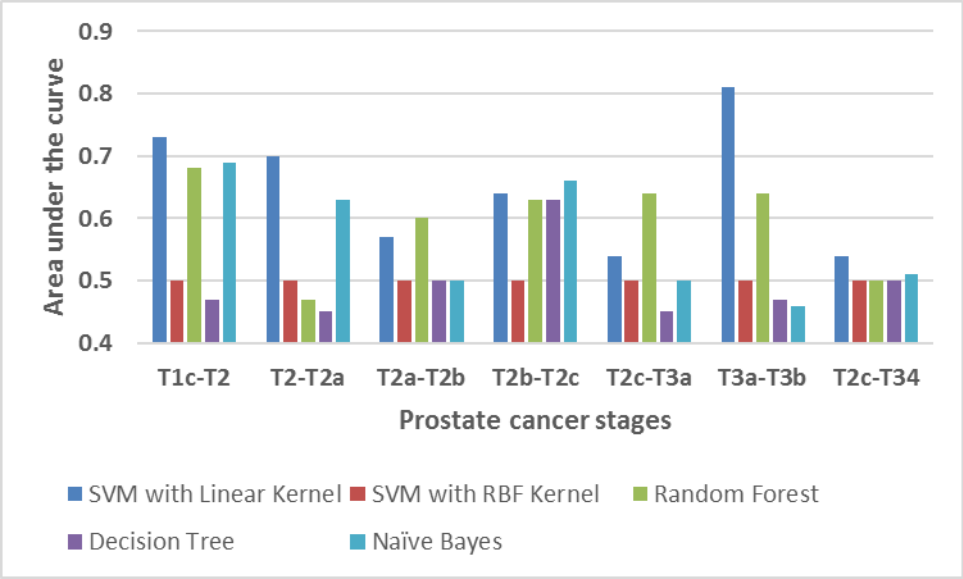


Figure 6.13: AUC of classifiers for CuffDiff selected transcripts on Long's data set.

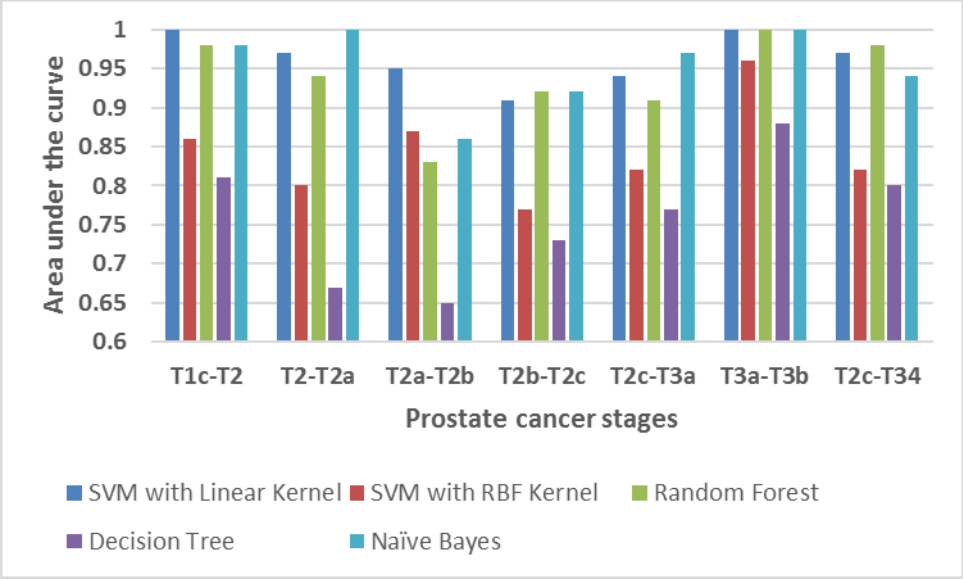


Figure 6.14: AUC of classifiers for our method selected transcripts on Long's data set.

Chapter 7

Conclusions and Future Work

In our research, we are given data sets of RNA-Seq reads that belong to different samples, each associated with a matched normal or malignant sample or a particular prostate cancer stage. We have developed a new method that is used to identify differentially expressed transcripts that are associated with matched normal versus malignant or different stages of prostate cancer. Ideally, these transcripts can be used for improving diagnosis, treatment, and drug development.

To solve the above problem, we extracted transcripts that act as potential biomarkers from RNA-Seq reads, with the help of a tuxedo approach, and applied powerful feature selection and classification algorithms to find discriminative transcripts that are related to prostate cancer and its different stages.

Finally, we found the biological relevance of a few, selected transcripts. All these transcripts tend to be very closely related to prostate cancer and other types of cancers, suggesting them as potential biomarkers for further wet-lab studies. Our method outperformed existing approaches (CuffDiff) for finding differentially expressed transcripts.

7.1 Contributions

In this work, we introduce a novel model that integrates emerging RNA-Seq technology with machine learning approaches to find relevant discriminative transcripts for the different stages of prostate cancer.

The main contributions are:

- Developing an integrative model that uses feature selection to choose a subgroup of transcripts and classification techniques to find the potential transcripts for different stages of prostate cancer.
- Identifying novel transcripts as potential biomarkers for prostate cancer progression.

7.2 Future Work

The results achieved by our model are closely related to prostate cancer and are highly recommended for further biological experiments. Our method has a few limitations. Since prostate cancer data sets were used, we are limiting our model to prostate cancer only. Our model requires a large number of samples for each progression stage. We have used two feature selection techniques and four classification algorithm due to the limited computational resources available. Due to fewer samples in the metastasis stage we could not work on classification of metastasis and non-metastasis.

The future work is as follows:

- The tuxedo suite approach is used for extracting the transcripts. However, different combinations of RNA-Seq tools might yield even better results.

- Different combinations of feature selection techniques, as well as classification algorithms, may result in more potential transcripts.
- We could classify the samples based on metastasis and non-metastasis, which may provide other biological significant transcripts.
- The transcripts obtained by our model can be used for wet laboratory experiments for further biological analysis.
- Our model could also be extended to other types of cancers and their progression stages.
- Gene ontology can be performed on the genes in which our potential transcripts are present.

Appendix A

Documentation to run tools

A.1 SRA Conversion

The data set files are in SRA file format and Tophat2 accepts FASTQ file format, therefore, we need to convert SRA to FASTQ format. We have downloaded SRA toolkit from NCBI website. The following command is used to convert SRA to FASTQ file format.

- `fastq-dump -A SRR057658 --split-3 SRR057658.sra`

A.2 Mapping to Reference Genome using Tophat2

Tophat2 is used for aligning reads to the reference genome. To download and install Tophat2 tool refer to `https://ccb.jhu.edu/software/tophat/tutorial.shtml`.

Bowtie_index can be download from `http://bowtie-bio.sourceforge.net/index.shtml`. The following command is used to run tophat2 tool

- `$tophat2 -o Output -p 8 Bowtie_Index SRR057658.fastq`

A.3 Transcriptome Assembly using Cufflinks

We have used the Galaxy suite for running Cufflinks for transcriptome assembly. Galaxy can be accessed at `http://usegalaxy.org`. First one has to create an account. Then we have to upload the accepted reads, and RefSeq transcript annotation to the Galaxy. We executed Cufflinks for all the samples.

A.4 Differential Expression using CuffDiff

CuffDiff is used to find differential expression transcripts, it was installed on linux machine. CuffDiff tool is part of Cufflinks package. The following command is used to run the Cuffdiff tool.

- `$ cuffdiff -o Output-p 8 --labels Cond A,Cond B Accepted_reads`

Appendix B

Supplementary Results

Table B.1: Biological significance of Long's data set transcripts across T1c-T2 pair-wise stage.

Transcript	Chr	Type	Description	Gene
NR_003669	16	ncRNA	metallothionein 1I, pseudogene (MT1IP), transcript variant 1	MT1IP
NM_001160393	11	mRNA	tRNA phosphotransferase 1 (TRPT1), transcript variant 6	TRPT1
NM_001161345	12	mRNA	checkpoint with forkhead and ring finger domains, E3 ubiquitin protein ligase (CHFR), transcript variant 2	CHFR
NM_052857	17	mRNA	zinc finger protein 830	ZNF830
NR_003594	8	ncRNA	RNA exonuclease 1 homolog (S. cerevisiae)-like 2	REXO1L2P
NR_033240	14	lncRNA	SLC25A21 antisense RNA 1 (SLC25A21-AS1)	SLC25A21

Table B.2: Biological significance of Long's data set transcripts across T2-T2a pair-wise stage.

Transcript	Chr	Type	Description	Gene
NM_004860	17	mRNA	fragile X mental retardation, autosomal homolog 2	FXR2
NM_052850	19	mRNA	growth arrest and DNA-damage-inducible, gamma interacting protein 1	GADD45GIP1
NM_001272095	16	mRNA	syntaxin 4, transcript variant 1	STX4
NM_001261390	17	mRNA	calcium binding and coiled-coil domain 2, transcript variant 1	CALCOCO2
NM_153274	1	mRNA	bestrophin 4	BEST4
NM_001252641	19	mRNA	prefoldin-like chaperone, transcript variant 3	URI1
NR_038352	5	ncRNA	decapping mRNA 2, transcript variant 3	DCP2

Table B.3: Biological significance of Long's data set transcripts across T2a-T2b pair-wise stage.

Transcript	Chr	Type	Description	Gene
NM_032023	10	mRNA	Ras association (RalGDS/AF-6) domain family member 4	RASSF4
NM_080792	20	mRNA	signal-regulatory protein alpha (SIRPA), transcript variant 3,	SIRPA
NM_000095	19	mRNA	cartilage oligomeric matrix protein	COMP
NM_003102	4	mRNA	superoxide dismutase 3, extracellular	SOD3
NM_080797	20	mRNA	death inducer-obliterator 1, transcript variant 3	DIDO1
NM_002725	1	mRNA	proline/arginine-rich end leucine-rich repeat protein, transcript variant 1	PRELP

Table B.4: Biological significance of Long's data set transcripts across T2b-T2c pair-wise stage.

Transcript	Chr	Type	Description	Gene
NM_001711	X	mRNA	Homo sapiens biglycan	BGN
NM_032023	10	mRNA	Ras association (RalGDS/AF-6) domain family member 4	RASSF4
NM_001014443	1	mRNA	ubiquitin specific peptidase 21, transcript variant 3	USP21
NM_021724	17	mRNA	nuclear receptor subfamily 1, group D, member 1	NR1D1
NM_012098	9	mRNA	angiopoietin-like 2	ANGPTL2

Table B.5: Biological significance of Long's data set transcripts across T2c-T3a pair-wise stage.

Transcript	Chr	Type	Description	Gene
NM_001198979	1	mRNA	small ArfGAP2 (SMAP2), transcript variant 2	SMAP2
NM_001099285	2	mRNA	prothymosin, alpha (PTMA), transcript variant 1	TMSA
NM_001198899	1	mRNA	YY1 associated protein 1 (YY1AP1), transcript variant 6	YY1AP1
NM_001130048	13	mRNA	dedicator of cytokinesis 9 (DOCK9), transcript variant 2	DOCK9
NM_000899	12	mRNA	KIT ligand (KITLG), transcript variant b	KITLG

Table B.6: Biological significance of Long's data set transcripts across T3a-T3b pair-wise stage.

Transcript	Chr	Type	Description	Gene
NR_034169	2	ncRNA	family with sequence similarity 133, member D	FAM133DP
NM_015380	22	mRNA	SAMM50 sorting and assembly machinery component	SAMM50
NR_046417	15	ncRNA	olfactory receptor, family 4, subfamily F, member 13, pseudogene	OR4F13P

Table B.7: Biological significance of Long's data set transcripts across T2c-T34 pair-wise stage.

Transcript	Chr	Type	Description	Gene
NM_001257413	17	mRNA	IKAROS family zinc finger 3 (Aiolos) , transcript variant 12	IKZF3
NM_003940	3	mRNA	ubiquitin specific peptidase 13 (isopeptidase T-3)	USP13
NM_001142274	2	mRNA	cytoplasmic linker associated protein 1, transcript variant 3	CLASP1
NM_001199165	17	mRNA	centrosomal protein 112kDa, transcript variant 3	CEP112
NM_052965	1	mRNA	tRNA splicing endonuclease subunit, transcript variant 1	TSEN15
NM_001195283	14	mRNA	feline leukemia virus subgroup C cellular receptor family, member 2, transcript variant 2	FLVCR2
NM_001023567	15	mRNA	golgin A8 family, member B, transcript variant 1	GOLGA8B
NM_001143766	10	mRNA	zinc finger protein 438, transcript variant 1	ZNF438
NR_003004	4	snoRNA	small Cajal body-specific RNA 22	SCARNA22
NM_017753	9	mRNA	lipid phosphate phosphatase-related protein type 1, transcript variant 2	LPPR1
NM_000959	1	mRNA	prostaglandin F receptor (FP), transcript variant 1	PTGFR
NM_004772	5	mRNA	neuronal regeneration related protein, transcript variant 1	NREP

Table B.8: Biological significance of matched normal versus malignant classification transcripts.

Data set	Transcript	Chr	Type	Description	Gene
Kannan	NM_019024	2	mRNA	HEAT repeat containing 5B	HEATR5B
	NM_001242889	7	mRNA	dopa decarboxylase (aromatic L-amino acid decarboxylase)	DDC
	NM_152228	1	mRNA	taste receptor, type 1, member 3	TAS1R3
	NM_001204401	X	mRNA	X-linked inhibitor of apoptosis, E3 ubiquitin protein ligase	XIAP
Kim	NR_024490	15	lncRNA	GABPB1 antisense RNA 1	GABPB1-AS1
	NM_001242889	7	mRNA	dopa decarboxylase (aromatic L-amino acid decarboxylase)	DDC
	NM_019024	2	mRNA	HEAT repeat containing 5B	HEATR5B
	NM_032415	7	mRNA	caspase recruitment domain family, member 11	CARD11
Ren	NR_024490	15	lncRNA	GABPB1 antisense RNA 1	GABPB1-AS1
	NM_001128826	9	mRNA	neuronal calcium sensor 1	NCS1
	NM_000494	10	mRNA	collagen, type XVII, alpha 1	COL17A1
	NM_000700	9	mRNA	annexin A1	ANXA1
	NM_005567	17	mRNA	lectin, galactoside-binding, soluble, 3 binding protein	LGALS3BP
	NM_000424	12	mRNA	keratin 5, type II	KRT5

Appendix C

Copyrights Permission

School of Computer Science

University of Windsor
thinking forward

To whom it may concern:

Mr. Siva Singireddy has defended his MSc thesis in computer science under my supervision, on August 25, 2015. His thesis contains some material obtained from the following joint publication at CIBCB 2015, of which he and I are co-authors:

> S. Singireddy, A. Alkhateeb, I. Rezaeian, D. Cavallo-Medved, L. Porter, **L. Rueda**, "Identifying Differentially Expressed Transcripts Associated with Prostate Cancer Progression using RNA-Seq and Machine Learning Techniques", *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2015)*, Niagara Falls, Canada, 2015.

I give Siva permission to use material from the above-listed publication in his Master's thesis.

Sincerely,

Luis Rueda, PhD
Professor
School of Computer Science
University of Windsor

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland, 4th edition, 2002. ISBN 0815332181.

[2] Margaritis Avgeris, Georgios Koutalellis, Emmanuel G Fragoulis, and Andreas Scorilas. Expression analysis and clinical utility of L-Dopa decarboxylase (DDC) in prostate cancer. *Clinical Biochemistry*, 41(14):1140–1149, 2008.

[3] Yoav Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):405–416, 2010.

[4] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy, a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, pages 19–10, 2010.

[5] Yongjun Chu and David R Corey. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics*, 22(4):271–274, 2012.

[6] David C Corney. RNA-seq Using Next Generation Sequencing, 2015. URL `http:`

//www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html. [Online; Last accessed August 2015].

[7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[8] Davidson. Davidson Education. http://www.bio.davidson.edu/genomics/method/cDNAproduction.html, 2015. [Online; Last accessed August 2015].

[9] Andrew GL Douglas and Matthew JA Wood. RNA splicing: disease and therapy. *Briefings in Functional Genomics*, 10(3):151–164, 2011.

[10] Kristin Eckfeld, Luke Hesson, Michele D Vos, Ivan Bieche, Farida Latif, and Geoffrey J Clark. RASSF4/AD037 is a potential ras effector/tumor suppressor of the RASSF family. *Cancer Research*, 64(23):8688–8693, 2004.

[11] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477, 2011.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[13] Kalpana Kannan, Liguo Wang, Jianghua Wang, Michael M Ittmann, Wei Li, and Laising Yen. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proceedings of the National Academy of Sciences*, 108(22):9172–9177, 2011.

[14] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

[15] Jung H Kim, Saravana M Dhanasekaran, John R Prensner, Xuhong Cao, Daniel Robinson, Shanker Kalyana-Sundaram, Christina Huang, Sunita Shankar, Xiaojun Jing, Matthew Iyer, et al. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Research*, 21(7):1028–1041, 2011.

[16] Benoit Langlois, Gwenn Perrot, Christophe Schneider, Patrick Henriet, Hervé Emonard, Laurent Martiny, Stéphane Dedieu, et al. LRP-1 promotes cancer cell invasion by supporting ERK and inhibiting JNK signaling pathways. *PLoS One*, 5(7): e11584, 2010.

[17] Langmead. hg19 human genome Bowtie index download. `ftp://ftp.ccb.jhu.edu/pub/data/bowtie_indexes/hg19.ebwt.zip`, 2015. [Online; Last accessed August 2015].

[18] Ben Langmead. Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics*, pages 11–7, 2010.

[19] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[20] Yeon Lee and Donald C. Rio. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, 84(1):null, 2015. doi: 10.1146/annurev-biochem-060614-034316. PMID: 25784052.

[21] R Leinonen and H Sugawara. The Sequence Read Archive. *Nucleic Acids Research*, 2011(39):19–21, 2011.

[22] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[23] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pages 388–388. IEEE Computer Society, 1995.

[24] Harvey Lodish, Arnold Berk, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Anthony Bretscher, Hidde Ploegh, and Paul Matsudaira. *Molecular Cell Biology (Lodish, Molecular Cell Biology)*. W. H. Freeman, 6th edition, June 2007. ISBN 0716776014.

[25] Qi Long, Jianpeng Xu, Adeboye O Osunkoya, et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer Research*, 74(12):3228–3237, 2014.

[26] Sebastian Mayer, Marc Hirschfeld, Markus Jaeger, Susanne Pies, Severine Iborra, Thalia Erbes, and Elmar Stickeler. RON alternative splicing regulation in primary ovarian cancer. *Oncology Reports*, 2015.

[27] Aziz M Mezlini, Eric JM Smith, Marc Fiume, et al. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research*, 23(3): 519–529, 2013.

[28] NCBI. National Center for Biotechnology Information. `http://www.ncbi.nlm.nih.gov/gene/`, 2015. [Online; Last accessed August 2015].

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

[30] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[31] RefSeq. RefSeq transcript annotation download. `ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/`, 2015. [Online; Last accessed August 2015].

[32] RefSeq. Refseq transcript annotation download. `http://www.ncbi.nlm.nih.gov/sra`, 2015. [Online; Last accessed August 2015].

[33] Shancheng Ren, Zhiyu Peng, Jian-Hua Mao, Yongwei Yu, Changjun Yin, Xin Gao, Zilian Cui, Jibin Zhang, Kang Yi, Weidong Xu, et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Research*, 22(5):806–821, 2012.

[34] Irina Rish. An empirical study of the naïve Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[35] Tammy L Romanuik, Gang Wang, Olena Morozova, Allen Delaney, Marco A Marra, and Marianne D Sadar. LNCaP Atlas: gene expression associated with in vivo progression to castration-recurrent prostate cancer. *BMC Medical genomics*, 3(1):43, 2010.

[36] D Ronchetti, K Todoerti, G Tuana, et al. The expression pattern of small nucleolar and

small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma. *Blood Cancer Journal*, 2(11):e96, 2012.

[37] Thermo Fisher Scientific. Thermo Fisher Scientific. https://www.thermofisher.com/ca/en/home/life-science/pcr/reverse-transcription/rna-priming-strategies.html, 2015. [Online; Last accessed August 2015].

[38] American Cancer Society. American Cancer Society. How is prostate cancer staged, 2015. URL http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-staging. [Online; Last accessed August 2015].

[39] Canadian Cancer Society. Canadian Cancer Societys Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2015. Toronto, ON: Canadian Cancer Society; 2015., 2015. URL http://www.cancer.ca/~/media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/Canadian-Cancer-Statistics-2015-EN.pdf. [Online; Last accessed August 2015].

[40] Canadian Cancer Society. Canadian Cancer Societys Prostate Cancer Stages , 2015. URL http://www.cancer.ca/en/cancer-information/cancer-type/prostate/staging/?region=on. [Online; Last accessed August 2015].

[41] Ahmad Tavakoli. Finding differential splice junctions in RNA-Seq data as transcriptional biomarkers for prostate cancer. Master's thesis, 2013.

[42] Cole Trapnell, Brian A Williams, Geo Pertea, et al. Transcript assembly and quantifi-

cation by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[43] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, 2012.

[44] Mathukumalli Vidyasagar. Machine learning methods in the computational biology of cancer. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 470(2167):20140081, 2014.

[45] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[46] James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine, and Richard Losick. *Molecular Biology of the Gene, Fifth Edition*. Benjamin Cummings, 5 edition, December 2003. ISBN 080534635X.

[47] W Zhai, XD Yao, YF Xu, B Peng, HM Zhang, M Liu, JH Huang, GC Wang, and JH Zheng. Transcriptome profiling of prostate tumor and matched normal samples by RNA-Seq. *European Review for Medical and Pharmacological Sciences*, 18(9): 1354–1360, 2014.

[48] Jinsong Zhang, Peijing Zhang, Yongkun Wei, et al. Deubiquitination and stabilization of PTEN by USP13. *Nature Cell Biology*, 15(12):1486–1494, 2013.

[49] Ying Zhang, Veenu Tripathi, Katherine Sixt, and Mary Heller. TGF-beta Regulates

Alternative Splicing of CD44 by Inducing Smad3 Binding to CD44 pre-mRNA. *The FASEB Journal*, 29(1 Supplement):562–15, 2015.

Vita Auctoris

Singi Reddy Siva Charan Reddy was born in Hyderabad, Telangana, India. He completed his graduation from Jawaharlal Nehru Technological University, Hyderabad in 2011 with Bachelor of Technology in Computer Science and Engineering. He has graduated with Master's of Computer Science from University of Windsor's School of Computer Science in August 2015.