2016

A Machine Learning Model for Discovery of Protein Isoforms as Biomarkers

Manal Alshehri
University of Windsor

A Machine Learning Model for Discovery of Protein Isoforms as Biomarkers

By

Manal Alshehri

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

Windsor, Ontario, Canada

2016

A Machine Learning Model for Discovery of Protein Isoforms as Biomarkers

by

Manal Alshehri

APPROVED BY:

E. Abdel-Raheem
Department of Electrical and Computer Engineering

A. Ngom
School of Computer Science

L. Rueda, Advisor
School of Computer Science

Nov 30, 2016

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication. I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's 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 obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix. 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 the most common cancer in men. One in eight Canadian men will be diagnosed with prostate cancer in their lifetime. The accurate detection of the disease's subtypes is critical for providing adequate therapy; hence, it is critical for increasing both survival rates and quality of life. Next generation sequencing can be beneficial when studying cancer. This technology generates a large amount of data that can be used to extract information about biomarkers.

This thesis proposes a model that discovers protein isoforms for different stages of prostate cancer progression. A tool has been developed that utilizes RNA-Seq data to infer open reading frames (ORFs) corresponding to transcripts. These ORFs are used as features for classificatio. A quantification measurement, Adaptive Fragments Per Kilobase of transcript per Million mapped reads (AFPKM), is proposed to compute the expression level for ORFs. The new measurement considers the actual length of the ORF and the length of the transcript. Using these ORFs and the new expression measure, several classifiers were built using different machine learning techniques. That enabled the identification of some protein isoforms related to prostate cancer progression. The biomarkers have had a great impact on the discrimination of prostate cancer stages and are worth further investigation.

DEDICATION

I dedicate this thesis to my precious mother Fatimah, my beloved husband Mohammed, and my gorgeous kids Yousef and Sarah. I am also indebted to the wonderful souls Haleemah and Somaiah who made Canada like home.

AKNOWLEDGEMENTS

My deepest appreciation and gratitude are extended to the following persons/organizations who in one way or another have contributed to this study and have made it possible:

- *The King Abdullah Scholarship Program and the Saudi Cultural Bureau in Canada, who made my dream come true;*
- *My supervisor Dr. Luis Rueda, who shared his knowledge and provided guidance and advice;*
- *My committee members Dr. Esam Abdel-Raheem, Dr. Alioune Ngom, and Dr. Jian-guo Lu, who blessed me with their time and effort; and*
- *My colleagues, Dr. Iman Rezaeian and Abedalrahman Alkateeb, who supported me and provided valuable suggestions and comments.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Deoxyribonucleic acid (DNA) is a two strand genetic material that contains all the information that is in living organisms. The information is stored as genetic codes using adenine (A), guanine (G), cytosine (C), and thymine (T). Ribonucleic acid (RNA) consists of the same nucleotides except using uracil (U) instead of thymine (T). The central dogma of molecular biology indicates that the coded genetic information stored in the DNA is transcribed into messenger RNA (mRNA) and then translated to form proteins. Protein molecules are responsible for most of the functions necessary in organisms' lives. Protein synthesis consists of several steps during which transcripts are transformed into amino acid chains. Almost all mRNA molecules undergo alternative splicing during or after transcription. Alternative splicing is an important event in Eukaryotic in which the segments that code for proteins (exons) are kept, and those that do not code for proteins (introns) are removed. Different transcripts are the result of the different inclusion of exons. Regarding post-transcriptional events, alternative splicing is the most important process due to its impact on 95% of mammalian genes [53]. Formation of polypeptide chain occurs through the translation of the actual open reading frame (ORF). ORF is a continuous sequence of codons, and it begins with a start codon and ends with one of the stop codons. Finding ORFs that correspond to a given mRNA transcript is a significant step in reconstructing protein isoforms, which is vital for a better understanding of RNA alternative splicing and its effects on diseases like cancer.

1.1 Transcription

The central dogma of molecular biology refers to transforming the genetic information coded in DNA into proteins through two steps: transcription and translation. In transcription, the first step of gene's expression, a specific segment of a DNA strand is copied into mRNA by the enzyme called RNA polymerase. Each mRNA strand holds the code for the synthesis of a particular protein (or a small number of proteins). Transcription uses complementary language, which means utilizing one strand of DNA as a template to produce a copy of the second strand, called a coding strand. A major characteristic of RNA is that it includes the nucleotide uracil (U) instead of the nucleotide thymine (T) in a DNA complement.

Transcription is divided into initiation, elongation, and termination. In the initiation step, RNA polymerase, along with some general transcription factors, binds to promoter DNA. The attached complex unwinds the two strands of the DNA helix to form a transcription bubble. RNA polymerase and the general transcription factors select a transcription start site in the transcription bubble. In the second step, elongation, RNA polymerase reads the nucleotides in the DNA from the 3' end to the 5' end (3' $\rightarrow$ 5'). During RNA polymerase traveling, the complementary RNA is created in the opposite direction (5' $\rightarrow$ 3') by matching the sequence of the coding strand with the exception of substituting Uracil (U) for Thymine (T). In the third step, transcription termination, the newly synthesized RNA strand is cleaved, and RNA polymerase is released.

Transcription generates primary RNA (pre-mRNA) which undergoes some processes to create mRNA. Post-transcriptional modification, co-transcriptional modification, is the process of converting pre-mRNA into mRNA. The process is vital in Eukaryotes and consists of three essential steps. The first step is processing the 5' end; it takes place before the completion of transcription as the pre-mRNA is being created. Capping involves the addition of 7-methylguanosine to the 5' end. The cap protects the 5' end of the mRNA transcript from degradation while transporting from the nucleus to the cytoplasm for trans-

lation. The second step is processing the 3' end through polyadenylation, which means the addition of a 3' poly-A tail. The poly-A tail is a stretch of several adenosine bases at the end of the mRNA strand. The third step is alternative splicing, which is explained in section 1.2.

1.2 Alternative Splicing

Eukaryotic genes consist of exons (ex-on indicates that they are expressed) and introns (int-ron indicates that they are intervening sequences). Most of the human genes have seven to eight exons and they spliced in three alternative forms. The average length of an exon is 150 nucleotides long, and the average length of an intron is 3,000 nucleotides [54]. Splicing is the process by which introns in a pre-mRNA are removed, and the remaining sequence (exons) concatenate together to form an mRNA sequence. Figure 1.2.1 depicts the alternative splicing process. The figure demonstrates a pre-mRNA that undergo alternative splicing to produce two different mRNAs, and hence two distinct proteins.

All introns in a pre-mRNA are precisely removed before the translation of the mRNA sequence. If the splicing site errs by even a single nucleotide, the reading frame of the joined exons will shift, and the resulting protein will be dysfunctional. Abnormal splicing variants are thought to contribute to the development of cancer [16].

The splicing of pre-mRNAs in Eukaryotes occurs through spliceosomes, and it takes place inside the nucleus either shortly after or during transcription. Spliceosomes are assembled of five small nuclear ribonucleoproteins (snRNAs) and more than 100 other proteins. The main role of spliceosomes in the splicing process is to catalyze RNA cleaving and joining. In order to remove an intron in a given pre-mRNA, spliceosomes recognize the 5' and the 3' splice site, and the branch site, which is in the middle of the intron. They then join the splice sites and exclude the sequence, intron, between them.

There are different types of alternative splicing events; one of them is exon skipping,

in which an exon (a cassette exon) is spliced out of the transcript along with flanking introns. Another example is intron retention in which an intron remains in the mature mRNA transcript.

Changes in the RNA expression machinery might lead to the mis-splicing of transcripts. Even a single nucleotide alteration in a splice site can result in differences in the mRNA produced from a mutant gene's transcripts. A well-known example of a splicing-related disease is cancer [53]. Irregularly spliced mRNAs are detected in a large proportion of cancerous cells [44].

Alternative splicing increases the informational diversity and functional capacity of a gene and provides an opportunity for gene regulation. In other words, it enables one gene to produce a large variety of polypeptides (i.e., proteins) [3]. A single variation in splicing may cause a given exon to be excluded from or included in a transcript, enabling the production of a new protein isoform [54].

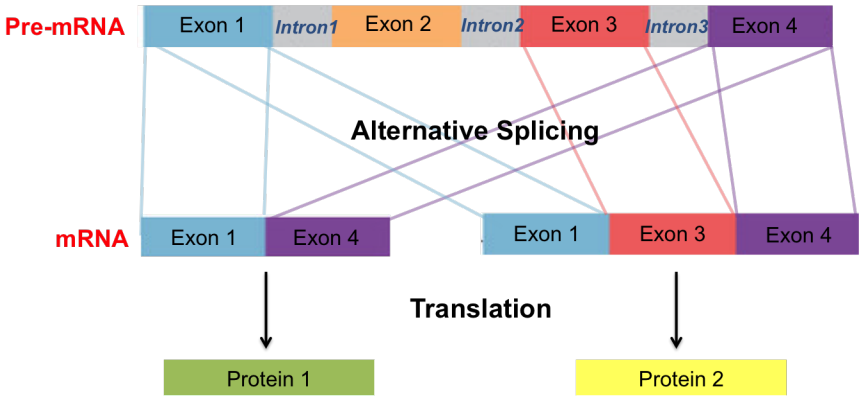


Figure 1.2.1: Schematic representation of alternative splicing.

1.3 Translation

Transcription happens exclusively in the nucleus, and the resulting mRNA exits the nucleus to the cytoplasm through the nuclear pore complex. In the cytoplasm, the mRNA

attaches to the ribosome (rRNA) and the translation process starts. Translation is a synthesis of proteins, in which the nucleotide sequence of an mRNA is translated into an amino acid sequence to yield a protein.

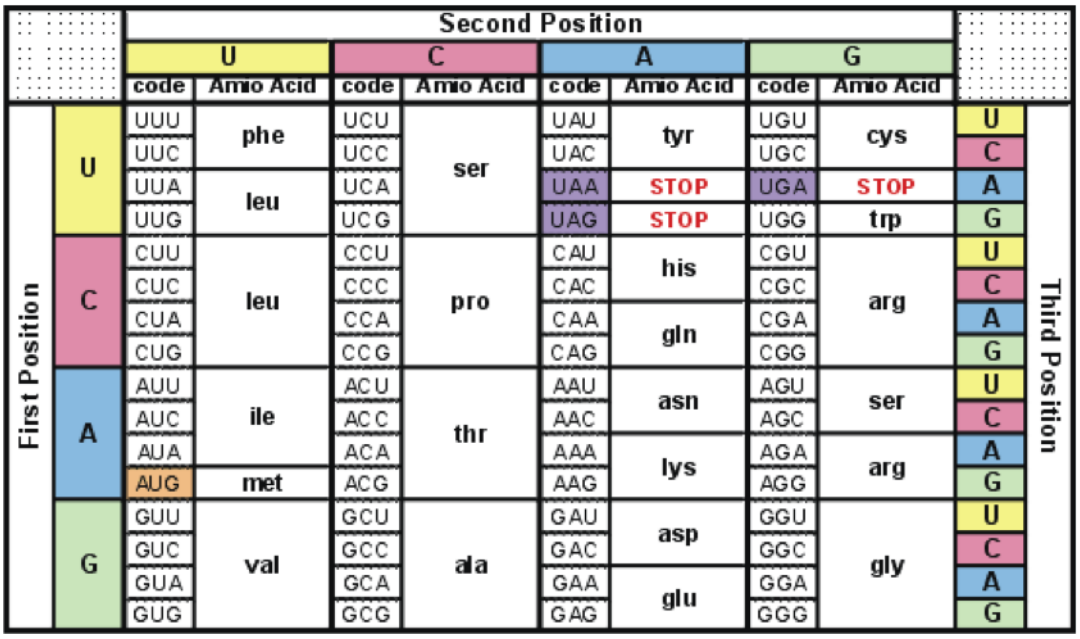
The protein synthesis requires a specialized machinery of high complexity, and it consumes a large part of the energy produced in the cell. The translation process involves a large number of chemical reactions and the participation of additional nucleic acid and protein components. One of these elements, the ribosome, implements the basic machinery for the translation process. The major role of the ribosome is coupling amino acids and an mRNA strand based on the sequence specified by the mRNA. The amino acids are brought to the ribosome by tRNA (transfer RNA) molecules.

The nucleotide sequence of the mRNA is composed of four different nucleotides (A, U, G, and C) whereas a protein is composed of 20 different amino acids. To allow the four nucleotides to specify 20 different amino acids, the nucleotide sequence is interpreted in codons, namely groups of three nucleotides. These codons have corresponding anticodons in the tRNA. Each anticodon is linked to one particular amino acid. Thus, each codon specifies one amino acid. The set of rules by which information is translated from mRNAs into amino acids is demonstrated in the codon table, which is presented in Figure 1.3.1. In addition to the main components of the translational system mentioned above, the translation process also involves many protein factors that facilitate the binding of mRNA and tRNA to ribosome.

1.3.1 Translation Mechanism

There are three major classes of cellular RNAs involved in the translation process: messenger RNAs (mRNAs), ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs). RNA translation can be divided into three distinct steps: initiation, elongation, and termination [54].

To initiate translation, a tRNA charged with methionine (tRNA+methionine) attaches



First Position	Second Position: U code	U Amio Acid	C code	C Amio Acid	A code	A Amio Acid	G code	G Amio Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 1.3.1: Standard codon table for translating mRNA to polypeptide chain [35].

to a ribosome. The combination of the charged tRNA and the ribosome (a small subunit) scans the mRNA from the 5' end until it finds a start codon. When it finds the start sequence AUG, a large subunit joins the small one to form a complete ribosome, and the protein synthesis is initiated.

In the elongation phase, a new tRNA+amino acid enters the ribosome at the next codon downstream of the start codon (AUG). If this tRNA's anticodon matches the mRNA codon, they both attach to each other, and the ribosome links the two amino acids together. If the tRNA with the wrong anticodon enters the ribosome, it cannot base pair with the mRNA; it is rejected. Next, the ribosome moves one triplet forward, and a new (tRNA+amino acid) enters the ribosome; the procedure is then repeated.

Finally, translation termination occurs when the ribosome reaches one of three stop codons (UAA, UAG, UGA); there are no corresponding tRNAs to stop codons. Instead, termination proteins bind to the ribosome and stimulate the release of the polypeptide chain (the protein); the ribosome then dissociates from the mRNA. When the ribosome is released from the mRNA, its large and small subunit dissociate. The small subunit can then

be loaded with a new tRNA+methionine and start translation once again. Some cells need vast quantities of a particular protein; to meet this requirement they make many mRNA copies of the corresponding gene and have many ribosomes working on each mRNA. After translation, the produced protein will usually undergo further modifications before it becomes fully active.

1.3.2 Open Reading Frame

An open reading frame (ORF) is the part of a reading frame that begins with a start codon (AUG) and ends with a stop codon (either UAA, UAG, or UGA). ORFs have the potential to be translated; hence, they have the potential to code for a protein or polypeptide chain. The AUG codon indicates where translation starts and the methionine included in the created polypeptide chain. However, stop codons do not appear inside the polypeptide chain. There are no corresponding amino acids for stop codons, and when the translation process reaches a stop codon, the polypeptide chain is released. Each ORF specifies a single protein that starts and end at internal sites within the mRNA. The two ends of an ORF are different from the ends of its mRNA [54].

Since the DNA code is translated into sets of three nucleotides (codons), a DNA strand has three different reading frames. Each mRNA reading frame is a shift by one nucleotide. The double helix of a DNA molecule has two strands; as a result, there are six possible translations for the three frames in each strand. Figure 1.3.2 illustrates the six reading frames of DNA. The upper DNA strand is called the first, positive, or (5' $\rightarrow$ 3') strand, while the lower DNA strand is called the second, negative, or (3' $\rightarrow$ 5') strand. The first reading frame in the positive strand is aligned with the first nucleotide. The second and third reading frames start from the second and third nucleotides, respectively. Obtaining the three reading frames for the negative strand follows the same principle as the positive strand. For both strands, the direction for obtaining the reading frames is always (5' $\rightarrow$ 3').

Among all ORFs detected in the six reading frames, only one ORF is used when trans-

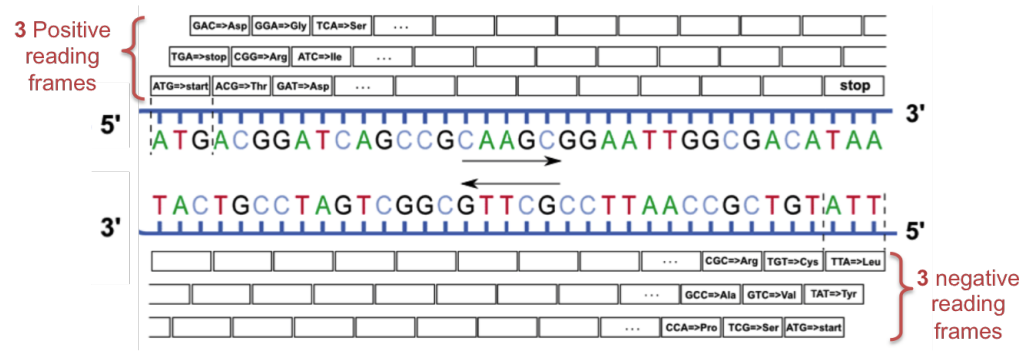


Figure 1.3.2: Six reading frames in DNA [55]

lating a gene in Eukaryotes; it is the longest ORF [54].

1.3.3 Protein Isoforms

A protein isoform is any one of two or more similar proteins that have similar, not identical, amino acid sequences and are encoded by mRNA transcripts from the same gene that has been alternatively spliced. In addition, the term *isoform* is used to discibe highly similar proteins that come from different genes. The creation of protein isoforms is an evolutionary event that allows for different diverse proteins to be created from a small amount of genetic material. Alternative splicing is the main reason behind the formation of protein isoforms. In alternative splicing, the exons of a gene may be included or excluded in processed mRNA. Proteins translated from alternatively spliced mRNA contain differences in their amino acid sequences; as a result they often differ in their biological functions [58], [48]. Many studies investigated the relation between alternatively spliced mRNA transcripts and unusual characteristics; however, it is much more complicated to explore the existence of alternative variants and protein isoforms at the protein level [44]. Isoform can also be defined as follows:

- Any of the proteins with the same function and a similar amino acid sequence, encoded by different genes or by RNA transcript [3].
- The protein products of different versions of messenger RNA created from the same

gene by employing different promoters, which cause transcription to skip certain exons. Since the promoters are tissue-specific, different tissues express different protein products of the same gene [22].

The term isoform causes endless confusion; however, in this thesis the term isoform refers to the various forms of a protein at the amino-acid level, regardless of their genetic origin [13].

1.4 Prostate Cancer

Prostate Cancer is the most commonly diagnosed cancer among Canadian men [36]. Cancerous prostate cells have uncontrolled growth and division, and can invade other parts of the body. Early prostate cancer does not usually cause any symptoms at all. The typical method of screening prostate cancer is the prostate-specific antigen (PSA) test. A PSA is a protein produced by the cells of the prostate gland. The PSA test measures the level of PSAs in a man's blood, and the test is reported as nanograms PSAs per milliliter (ng/mL) of blood. If the PSA level is above the normal range (4.0 ng/mL), it is interpreted as a sign of prostate cancer. After that, further tests such as digital rectal exam (DRE), a pathological validation, and a biopsy will be done. Specialists try to accumulate as much information as possible so that they can determine the exact stage of cancer and subsequently provide the best treatment options. Staging is a way of classifying cancer based on the degree of spreading of cancer in the body. According to the Canadian Cancer Society [6], prostate cancer has four main stages.

- Stage 1, the tumor starts developing in the prostate tissue, but tumour involves half a lobe or less.
- Stage 2, the tumor grows and may occupys more than half a lobe but is limited to the prostate.

- Stage 3, the tumor has spread beyond the prostatic capsule or into the bladder neck or seminal vesicles.
- Stage 4, tumor has invaded other tissues and organs such as the rectum, pelvic wall, lymph nodes, bladder.

1.5 RNA-Seq Technology

The transcriptome is the set of all RNA molecules, including mRNAs, non-coding RNAs, and small RNAs present in a cell at a particular time. Understanding the transcriptome and quantifying transcripts is vital for interpreting gene activity and expression in a specific cell or tissue type, and hence understanding diseases and their developments [52]. RNA-Seq refers to techniques used to determine the primary sequence and relative abundance of the transcriptome within a biological sample. Several studies have uncovered critical information using nucleotide resolution. For example, previous research estimated that roughly 60% of human genes are alternatively spliced [12], while next-generation sequencing data estimates that approximately 95% of all human genes are alternatively spliced [32].

In recent years, RNA-Seq has become the preferred technique for transcriptome studies, including alternatively spliced transcripts and their corresponding protein isoforms [15], [28], [45]. RNA-Seq has the power to provide precise information about alternative transcripts or proteins and the mechanisms that contribute to their formations. RNA-Seq technology has improved researchers' understanding of gene regulations: transcription, post-transcription, and translation; however, its analyses remain challenging.

Studying cancer at the nucleotide level can provide relevant information about the tumor initiation mechanisms and progression; it consequently, benefits diagnosis, prognosis, and treatment. Although several studies have concentrated on finding genes as biomarkers, they have not completely utilized the high-resolution feature that RNA-Seq technology provides.

1.6 Biomarkers

According to the Dictionary of Cancer Terms (NCI), a biomarker is biological molecule found in blood, other bodily fluids, or tissues that is a sign of a normal or abnormal process or a condition or disease [30]. There is a large variety of biomarkers, which can include genes, nucleic acids, and proteins, among other categories. Biomarkers can be identified in blood circulation or excretions such as the urine, and can thus be assessed non-surgically. They can also be identified in tissues, which require imaging to be evaluated.

Genetic biomarkers can be inherited or somatic. Inherited biomarkers are detected as a DNA sequence, and they are isolated from the blood. Somatic biomarkers are recognized as mutations in DNA obtained from tumor tissue. Biomarkers can be used in various evaluations, including investigating the risk of a particular disease, screening for primary cancers, discriminating between benign and malignant tumors, distinguish one stage of cancer from another, monitoring the status of a cancer, and assessing a cancer's response to therapy.

The classic approach for detecting cancer biomarkers is finding biomarkers based on tumor biology and surrounding environment. With the improved knowledge about cancer and the emergence of new technologies, biomarker discovery is often performed using techniques such as RNA-Seq that identify single or multiple biomarkers in a significantly short amount of time.

The search for cancer biomarkers has been for years a hot topic of research. Although many efforts are being devoted to the search for new cancer biomarkers, the discovery of these molecules has been carried out at a slow pace [51].

Several studies have discovered biomarkers as genes for prostate cancer. For example, Xu et al. [45] investigated five cancer samples using RNA-Seq technology. That study founf 116 somatic mutations in 92 genes that are related to prostate cancer. Additionally, Long et al. conducted another study on global transcriptome analysis [27]. They identified 16 genes associated with cancer in which five genes are related to prostate cancer.

1.7 Problem and Motivation

As mentioned above, prostate cancer rates are highest among Canadian men [37]. An appropriate diagnosis of the particular stage of prostate cancer in a patient is critical to ensure the patient is provided the best suitable treatment and his survival rates is increased.

The typical approach for detecting or staging prostate cancer is studying PSA protein levels. However, some studies show that testing PSA levels in the blood is not the ideal marker for discovering or staging prostate cancer [31], since a proportion of men with prostate cancer have PSA values lower than 4.0 ng/mL. Further, a test with PSA level above 4.0 ng/mL does not always prove the existence of prostate cancer because it can be related to other prostatic diseases [24]. Moreover, PSA levels decrease dramatically in advanced stages of prostate cancer. Therefore, one of the most promising tools to improve the use of PSA as a tumor marker is measuring the proportion of PSA's different isoforms [31].

However, in recent years researchers have worked on detecting genetic biomarkers for different types of cancer, including prostate cancer. The focus has mainly been on the genetic level to discover differentially expressed genes. However, studying the transcriptome activity and investigating protein isoforms as biomarkers are more promising than finding biomarkers at the gene level, due to the precise information that protein isoforms provide about tumor condition and progression.

This thesis focuses on identifying a small subset of protein isoforms that can reliably discriminate between stages of prostate cancer with a very high accuracy.

1.8 Contributions

This thesis integrates RNA-Seq data with machine learning techniques to predict prostate cancer progression. First, a tool is developed to find the ORF in all six frames corresponding to a given mRNA sequence and identifying the actual ORF from RNA-Seq

data. The proposed model can find both known and novel protein isoforms corresponding to a given mRNA using RNA-Seq data.

Second, a machine-learning approach is developed that uses protein isoforms as features and differentiates between prostate cancer stages. The model delivers a handful of relevant biomarkers that can be utilized at the protein level and can open new avenues toward better understanding the progression of prostate cancer. Using identified protein isoforms, prostate cancer stages were discriminated between with a very high accuracy.

This thesis consists of six chapters, starting with an introduction, which provides an overview of the main topics. Machine learning techniques are presented in Chapter 2. The literature review is included in Chapter 3. The thesis's methods and results are discussed in Chapters 4 and 5, respectively. Finally, Chapter 6 presents the thesis conclusions and the future work derived from this thesis.

CHAPTER 2

Machine Learning

Machine learning is a branch of the computer science field that combines pattern recognition, artificial intelligence, and computational statistics. Machine learning involves the implementation of computer algorithms that learn from data and make predictions instead of depending on a static program.

Machine learning algorithms are categorized as supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning algorithms construct a model from sample inputs, or a training set, to make predictions for new samples, or a test set. A supervised learning problem could be, for instance, cancer diagnosis in which the aim is to categorize a patient (or sample) as malignant or benign. Unsupervised learning consists of a set of inputs without any corresponding class value, in such a way that the algorithms forms clusters or natural groupings of the samples based on observed patterns. Semi-supervised learning labels some of the data, and a model makes use of both labeled and unlabeled data. Feature selection and classification are two applications of machine learning, and they are discussed next.

2.1 Feature Selection

Feature selection is the process of selecting a subset of relevant attributes to be used in the classification model construction. Feature selection techniques are used to identify and remove irrelevant and redundant attributes that do not improve the accuracy of a classifier

or may, in fact, decrease the accuracy of the model. The primary goals of feature selection are to speed up the classification process and to increase the classification efficiency as much as possible.

The most common aspect of classifying biological data is the high-dimensional nature of the task; an enormous number of features are used with a small group of samples. With a wide range of features, the computational complexity of classification algorithms grows exponentially with each increase in dimension. For this purpose, a subset of features is chosen carefully by feature selection algorithms to build the classification model.

There are two categories of feature selection approaches: filter methods and wrapper methods. Filter methods score and rank the features based on their quality; they are fast because they evaluate each feature independently. Chi-squared [25] is an example of a filter-based feature selection technique. The method evaluates the relevance of a feature by calculating the chi-squared value; the higher the value of chi-squared, the more relevant the feature is.

Wrapper methods are another way to select a subset of features. They automatically determine the number of selected features. They find the best subset of attributes using a predictive classifier to score feature subsets. Unlike filter approaches, wrapper methods consider all relationships among features. They are computationally expensive because they perform an exhaustive search to discover the best subset of features. An example of a wrapper-based methods is the minimum redundancy maximum relevance (mRMR) method [33]. The mRMR method selects features that yield the highest relevance with the target class and are maximally dissimilar to each other. A classification algorithm is needed since mRMR is a wrapper method.

The mRMR feature selection approach was used in this thesis to select the most informative features associated with prostate cancer progression.

2.2 Classification

Classification is one of the primarily supervised learning applications in machine learning. The objective of classification is to assign new instances to one of the known categories, or classes, depending on training dataset that consists of samples with known labels. Classification schemes aim to correctly classify samples such as biological data with the best possible performance.

In this thesis, the input vectors, or features, are the ORFs that are generated by the proposed method. Each patient sample has a corresponding class label that indicates the stage of prostate cancer to which the sample belongs. Various algorithms designed for classification problems such as Naive Bayes, support vector machine (SVM), random tree, and random forest were used. The next section discusses some of these classifiers.

2.2.1 Support Vector Machine

The support vector machine (SVM) is a widely used classifier due to its generalization power and its suitability for small training datasets along that have a large number of features [11]. SVM is a supervised learning model that represents samples as points in space and finds the separating hyperplane with the largest margin. The larger the margin, the better the classifier is. The classifier's numeric input vectors form an n-dimensional space, and the hyperplane is a line that splits the input variable space.

There are two categories of SVM models related to data distribution: linear SVM and non-linear SVM [8]. Figure 2.2.1 shows an example of linearly separable data. The figure shows two separating hyperplanes, *H1* and *H2*, that can be applied to given training samples. Although both hyperplanes separate the two classes linearly, the optimal hyperplane is *H1* because it separates the classes with the maximum distance. Support vectors are the most difficult samples to be classified; they are used to determine the hyperplane's position. A linear kernel function is recommended when a linear separation of the data is straight-

forward; in other cases, the kernel trick is used. The kernel trick concentrates on mapping the input samples onto a higher dimensional feature space. The SVM then tries to choose a hyperplane that separates the classes as clearly as possible with the maximum width of the margin; that case is known as the soft margin.

Different kernel functions need to be tested to obtain the best model in each case, as they each use different algorithms and parameters. Both linear SVM and non-linear SVM were examined. Different kernel functions such as polynomial and radial basis function (RBF) were tested.

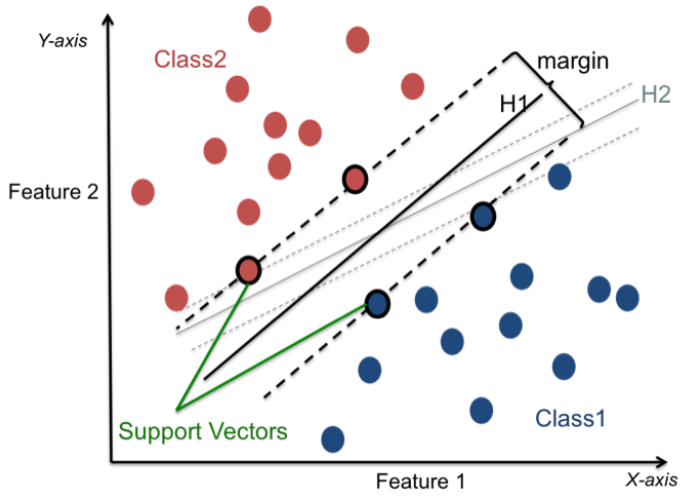


Figure 2.2.1: Separating hyperplane and margin for linearly separable classification problem, *H1* is optimal hyperplane with maximum distance between classes.

2.2.2 Decision Tree

A decision tree is an effective classification method based on Quinlan's algorithm [39]. It consists of a root node, internal nodes, and leaves. Each node indicates a test on a feature, and each branch substitutes the result of the test. The final decision, or class label, is found in the leaves for samples that verify all test paths from the root to the leaves.

Each decision tree is constructed throughout multiple steps using a set of learning samples. In each step, samples are divided into smaller groups based on particular criteria. Some examples for the criteria used in tree design are minimum error rate, minimum or

maximum path length, minimum number of nodes, and maximum information gain [41]. The node that provides the best value for the criteria among all nodes is the root node. New samples are classified starting from the root, and they follow the branch direction based on the test at each internal node until a leaf, or class label, is reached.

An example of a decision tree for playing tennis is shown in Figure 2.2.2. The figure shows a decision tree that provides a decision on playing tennis based on three attributes: outlook, humidity and wind speed. Outlook attribute has the highest value, and so it is used as a decision attribute, or the root node. As shown in the figure, if the outlook is rainy and very windy, the decision to play tennis is a No, while if the outlook is sunny and the humidity is normal, the decision to play tennis is a Yes.

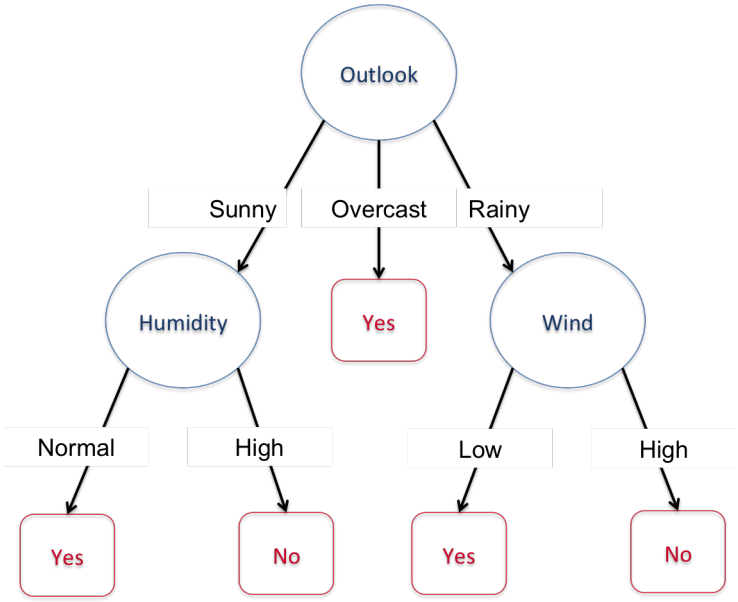


Figure 2.2.2: Decision tree for playing tennis based on outlook, humidity, and wind.

2.2.3 Random Forest

Random forest is a classification method based on constructing a forest of random trees; each tree produces a vote in classifying. [20]. A random subset of the training data is used to build each decision tree. After training the forest, new samples can be passed to the forest to predict their classes. The larger the forest, the higher accuracy is gained without

suffering from overfitting. There are two factors that affect random forest's performance: the correlation between trees and the accuracy of each tree. Increasing the number of features used in each tree increases the tree's performance as well as the correlation; hence, increasing the number of features used in each tree increases the forest's performance. The classification of new samples in the random forest is based on voting. Each tree in the forest votes to one of the classes; the class with the highest number of votes is assigned to the new observation. Figure 2.2.3 shows an example of a random forest containing N trees. For a new sample, X, each tree votes for one of the classes. The final class is determined by the majority voting.

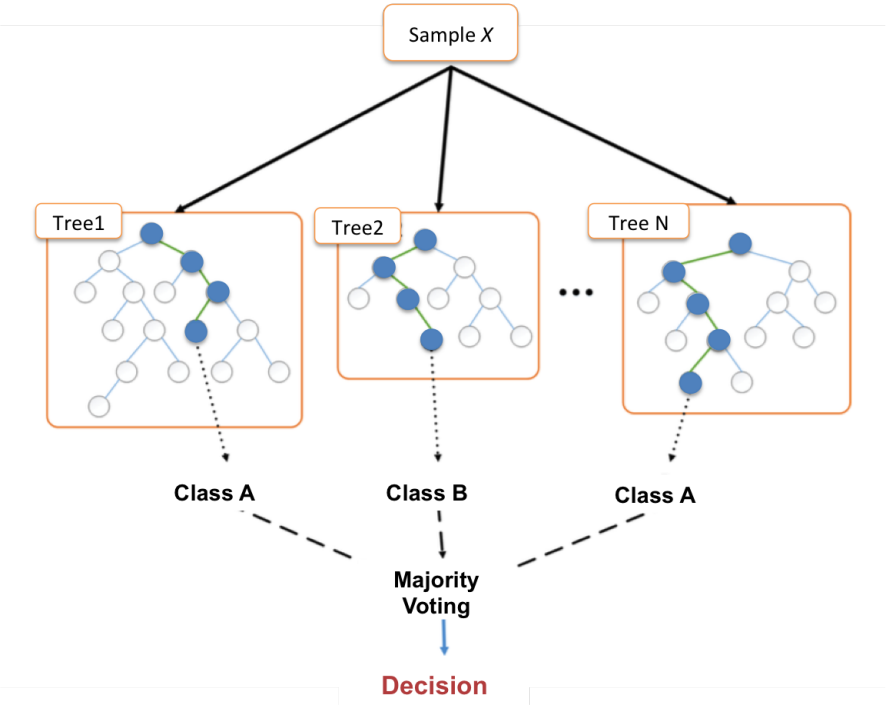


Figure 2.2.3: Random forest.

2.2.4 Naive Bayes

Naive Bayes is a classification algorithm that is based on Bayes' theorem [40]. Classification using Naive Bayes depends on prior probability and likelihood. A classification is produced by combining both sources of information: the prior probability and the like-

lihood of forming a posterior probability using Bayes' rule. This classifier assumes that the presence of a feature in a class is independent from the presence of any other feature. Making predictions involves calculating the probability that a given data instance belongs to each class, and then selecting the class with the largest probability as the prediction.

2.3 Performance Evaluation

Performance measures are a way to evaluate a solution to a problem. They are the measurements for predictions made by a trained model on a test dataset. Some concepts related to performance evaluation are introduced below.

2.3.1 Confusion Matrix

Suppose that there are two classes: positive and negative. The confusion matrix of these two classes is shown in Figure 2.3.1.

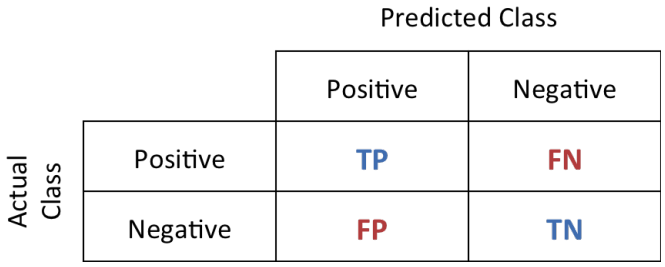


Figure 2.3.1: Confusion matrix corresponding to original samples and classified samples for two classes.

- *TP* is the number of true positives or the number of samples that belong to the positive class and have been classified as positive.
- *TN* is the number of true negatives or the number of samples that belong to the negative class and have been classified as negative.

- *FP* is the number of false positives or the number of samples that belong to the negative class, but have been incorrectly classified as positive.
- *FN* is the number of false negatives or the number of samples that belong to the positive class, but have been incorrectly classified as negative.

Different performance measures can be calculated using the confusion matrix as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.3.1}$$

$$Recall(\textit{Sensitivity}) = \frac{TP}{TP + FN} \tag{2.3.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.3.3}$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.3.4}$$

Area under the ROC curve (AUC) is also used as a performance measurement. Figure 2.3.2 shows an example of an ROC curve. The closer the ROC curve is to the top left corner, the better the performance of the classifier is. If the classifier is performing very well, the true positive rate will increase and the AUC will be close to 1. If the classifier randomly assigns samples to classes, the true positive rate will increase linearly with the

false positive rate and the AUC will be around 0.5.

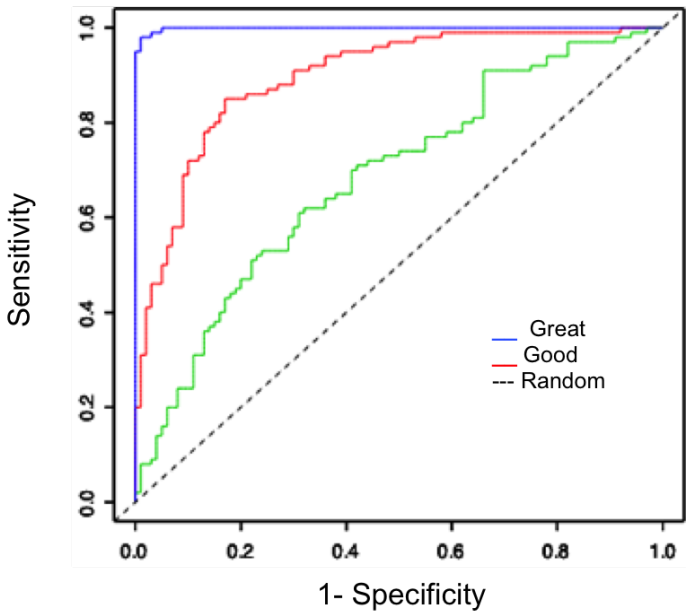


Figure 2.3.2: Example of ROC curve.

2.3.2 Cross Validation

Another approach for evaluating and testing a classification algorithm is cross validation, in which an entire dataset is used to train and test that algorithm. It involves separating a dataset into m equally sized groups of samples (m-folds). The model is trained on all folds except one, and the created model is then tested on the left out fold. This process is repeated, and each fold will be given the opportunity to be left out and act as the test dataset. Finally, the performance measures are averaged across all folds to estimate the model performance on the problem. In this thesis, 10-fold cross validation is used to validate the results.

CHAPTER 3

Literature Review

This chapter reviews previous approaches related to the thesis's. First, some existing online tools used in translating mRNA code and detecting ORF are explained. Some studies that identify proteins as biomarkers for prostate cancer diagnosis are then discussed.

3.1 Tools for mRNA Translation and ORF Finding

Regions of DNA that encode proteins are first transcribed into mRNA and are then translated into protein. By examining a DNA sequence, one can determine the sequence of amino acids that will appear in its final protein. In translation, codons of three nucleotides define the amino acid that will be added next in the growing protein chain. Proteins are formed from ORF, and by analyzing the ORF, we can predict the polypeptide chain that might be produced during translation.

Typically only one ORF is used when translating a gene in Eukaryotes, and it is the longest ORF. Once the ORF is found, the DNA sequence can be translated into its corresponding amino acid sequence. However, discovering the actual ORF is not a straightforward task, especially when dealing with a very large number of long mRNA sequences.

Here, online translation tools and how they address ORFs are discussed. These tools were tested using the transcript `>hg19_chr7_89841000_89866992_+`. The full sequence for the transcript can be found in Appendix A.

ORF Finder [38] is a translation tool available on the National Center for Biotechnology

Information (NCBI) website. The tool identifies all possible ORFs from a sequence that a user inserts in the giving text box. Figure 3.1.1 shows the input screen of ORF Finder, while Figure 3.1.2 shows the query result. The user submits the translation request by inserting the nucleotide sequence into the text area. The tool then returns the range of each ORF along with its protein translation. As shown in Figure 3.1.2 , all sequences that are located between a start codon and a stop codon are listed in descending order, based on their length, along with their position. By choosing the required ORF, the corresponding amino acid sequence appears in the left box. It is up to the user to choose which ORF is the actual ORF, and then download it as a FASTA file. The user can select any detected ORFs among the six reading frames for further information using BLAST.

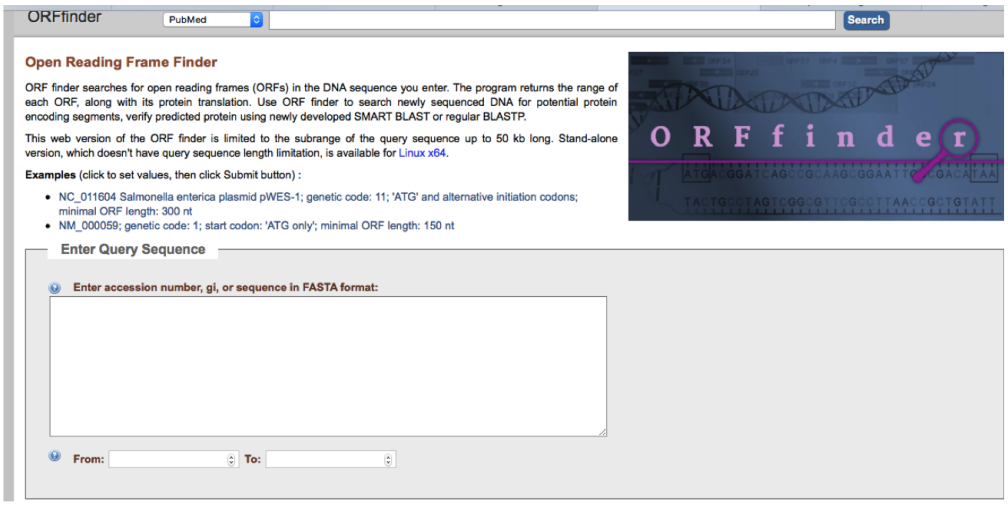


Figure 3.1.1: Input screen of ORF Finder.

There is a shortage in the Web version of the ORF finder; the input sequence length is limited to 50 kb long. Although there is a standalone version for this tool that does not have such a limitation, it has to be requested from the authors. Another shortage related to the tool input is that it accepts only one sequence to be translated and tested for potential ORFs at a time.

The second translation tool is *ExPASY (the Expert Protein Analysis System)* by the

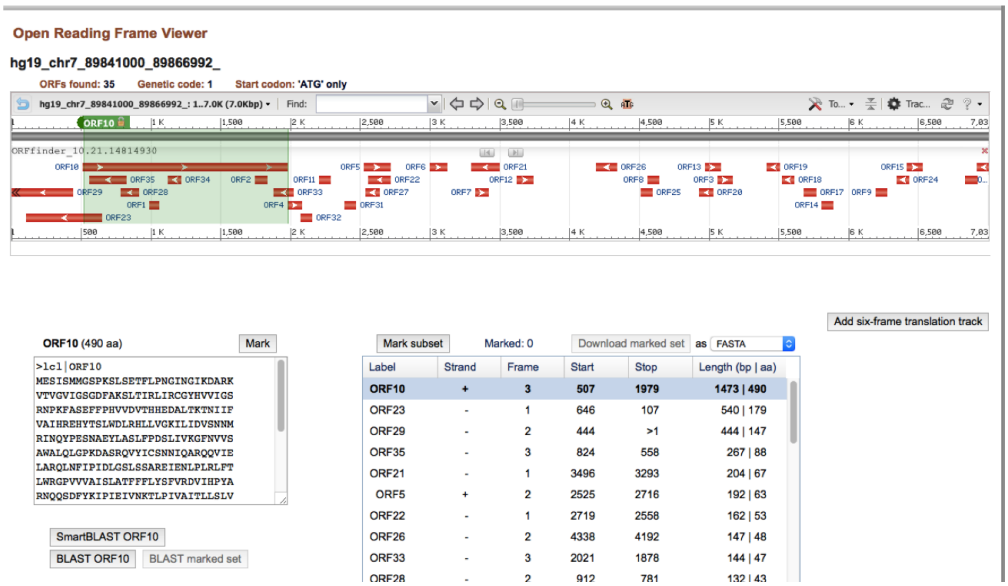


Figure 3.1.2: Output screen of ORF Finder.

Swiss Institute of Bioinformatics [47]. It searches for ORFs in a sequence that a user enters in the search box. Figure 3.1.3 shows the input screen of the tool where the sequence is inserted in FASTA format.

The translation of all six reading frames appears with all potential ORFs. Figure 3.1.4 shows the translation of the sequence; the highlighted regions are potential ORFs. The user has to select a reading frame and then select an ORF from that frame. Then, a new page appears with an option to download the output translation as a FASTA file or to verify the predicted polypeptide chain using BLAST. Similar to ORF finder, ExPASY does not have the opttion of uploading an input file with multiple sequences, and does not identify the actual ORF.

The third translation software is *GetORF* [14]. The tool allows one to enter a sequence of data manually or to upload a FASTA file consisting of multiple sequences. It provides the option to detect ORFs only in the positive strand or in both strands. *GetORF* outputs the results in FASTA format, including the ORF identifier, the translated amino acid sequence, and the ORF start and end sites. The main downside to this tool is that it identifies all

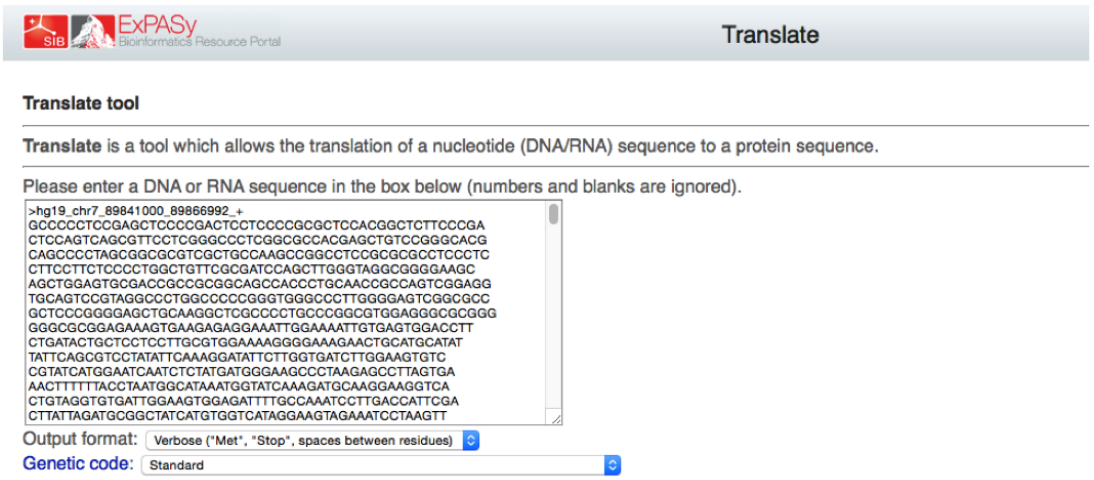


Figure 3.1.3: input screen of ExPASY.

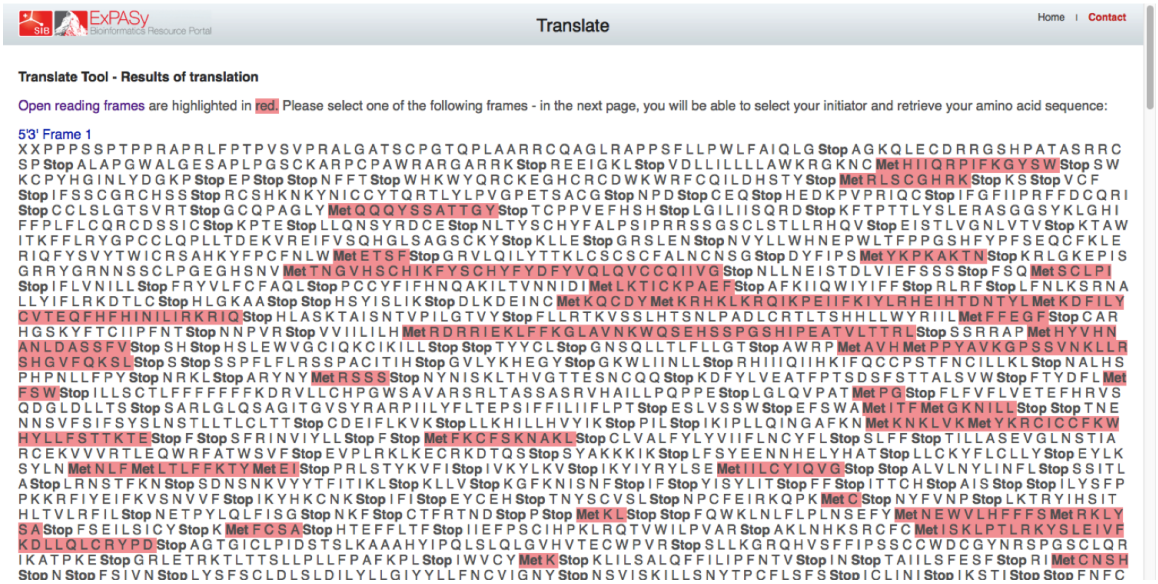


Figure 3.1.4: Output screen of ExPASY.

possible ORFs for a given sequence without discovering the actual ORF. Figures 3.1.5 and 3.1.6 show the input screen and the output result screen of this tool.

In conclusion, some existing platforms for finding ORFs and for translating RNA sequences were investigated. The advantages and disadvantages of each tool were described. Some of the tools only translate one transcript at a time, while others have input length limitations. No tool has the ability to discover the potential ORF; they detect all ORFs instead. It could be very difficult for users to determine the potential ORF manually. For

Figure 3.1.5: Input screen of GetORF.

Figure 3.1.6: Output screen of GetORF.

example, transcript NM_000059 that encodes gene *BRCA2* contains thousands of ORFs in its six reading frames.

In the thesis's tool that was implemented to find the ORF, most of the concepts that these previous tools follow were adopted, such as outputting the results in FASTA format. However, all the disadvantages of each tool have been overcome.

3.2 Using Proteins as Biomarkers

In mass spectrometry, *surface-enhanced laser desorption/ionization (SELDI)* provides a sensitive system that analyzes protein masses within a sample [51]. SELDI is a strong tool for analyzing proteins in a variety of biological materials such as tissue samples, blood, urine, and other clinical samples. The technique is usually used as a diagnostic tool and has been applied to diagnose cancer. Technological advances in mass spectrometry pose challenges in computational methods to process the mass spectral data into predictive models that have clinical and biological significances. The biomarker discovery is achieved by adopting machine learning techniques that allow a comparison of protein levels in serum samples from healthy and diseased patients. Feature selection and classification techniques are powerful tools for building predictive models from protein mass spectrometric profiles and for identifying biomarker candidates. The best way to ultimately confirm the computational predictions results is through biological validation and identification of the underlying proteins [51]. Here, some approaches that discriminate between prostate cancer samples and normal samples using proteins and mass spectrometry technology are reviewed.

Qu et al. claimed that the Prostate-specific antigen (PSA) test is not an accurate biomarker for early diagnoses of prostate cancer [38]. They developed a proteomic approach to discriminating prostate cancer samples (PCA) from benign prostatic hyperplasia samples (BPH) and normal age-matched samples. Their model detects and analyzes multiple proteins to differentiate prostate cancer patients from non-cancer patients. The serum samples used in the experiment were 197 prostate cancer samples (PCA), 92 benign prostatic hyperplasia samples (BPH), and 96 healthy samples. The samples were randomly separated into a training set (325) and a test set (60), and were then analyzed using surface-enhanced laser desorption or ionization (SELDI) mass spectrometry. Peak detection and alignment were performed with the Ciphergen ProteinChip Software. 124 peaks, or features, were detected and used as the training set to construct the classifiers for separating PCA group

from the non-cancer groups. The classification model was then applied to the test set (30 PCA samples, 15 BPH samples, 15 healthy samples).

Two classifiers were implemented: the AdaBoost classifier and the Boosted Decision Stump Feature Selection classifier (BDSFS). The BDSFS classifier is a decision tree classifier with one split. This classifier is constructed on weighted examples and the classification decision is made by voting in each round. 10-fold cross-validation was applied to estimate the accuracy of the two classifiers. The first classifier, AdaBoost, gained perfect results in discriminating prostate cancer samples from normal samples (it had 100% sensitivity and specificity). The second classifier was combined with feature selection and resulted in 21 features out of 500. It achieved a sensitivity and specificity of 97%. The authors claimed that the combination of the SELDI protein profiling system and machine learning algorithms could be used for the early detection or diagnosis of prostate cancer.

Wagner et al. discussed various classification approaches using protein profiles from mass spectrometry experiments to discover protein biomarkers [51]. The authors aimed to define peaks with a high likelihood of being biologically related to given disease state, and identified some specific biomarkers of the disease. The prostate cancer samples were obtained from the Eastern Virginia Medical School using SELDI-TOF mass spectrometry from over 300 patients. The samples were grouped into four categories: benign prostatic hyperplasia (BPH); early ,or localized, cancer; late , or metastasized, cancer; and controls. The feature space was reduced by selecting the smallest set of peaks that yielded reasonable classification results. The input vectors, or peaks, were decreased by using a filter-based feature selection. The features were reduced from 779 to 220, and then a principal component analysis was applied to further reduce the dimension of the data.

For classification, several algorithms were applied, such as quadratic discriminant functions, nonparametric kernels, nearest neighbour methods, and linear support vector machines. The best results were obtained by applying a two-stage linear SVM classifier along

with 13 peaks and cross validation. The average classification accuracy for the four-group classification problem was 87%.

In a recent study, Taskin et al. demonstrated a model that classifies prostate cancer samples using mass spectrometry data [49]. Samples from 322 men were used in the experiment: 69 samples were malignant, 190 samples were benign, and 63 were normal. The authors processed the dimensionality reduction in three stages. The first phase included a filtering method, which employed some statistical tests to evaluate features and assigned a score to each feature. The second phase used two different methods: the wavelet analysis and the statistical moments; they were applied for comparison. In the wavelet analysis, the discrete wavelet transform (DWT) was applied to the mass spectrometry data; approximation coefficients were obtained. The obtained signal had a smoother form for the mass spectrometry data as well as a low dimensionality. The statistical moments used filtering methods for dimensionality reduction. The last phase for reducing the feature space involved a feature transformation method, which was kernel partial least square (KPLS).

After the dimensionality reduction, the prostate data were classified using several algorithms, such as *k*-nearest neighbour and SVMs. The classification was processed in two steps with a 5-fold cross validation. In the first step, prostate samples were classified as normal and cancerous. The cancerous samples were then classified as benign or malignant. The classifiers' performances for the two classification steps were 95.8% and 87.2% respectively.

CHAPTER 4

Materials and Methods

This chapter discusses the methodology used to implement the tool that finds the ORF and reconstructs the potential protein isoform for every transcript. The protein isoforms act as potential biomarkers for estimating prostate cancer progression stages. Furthermore, some machine-learning techniques that were applied to the extracted protein isoforms to find a group of them that are differentially expressed throughout prostate cancer's stages are illustrated.

4.1 The Dataset

The dataset used for this study was Long's dataset, which can be found in NCBI's repository with the GEO accession no. GSE54460 [27]. The dataset involves roughly 490 billion base pair reads and deals with prostate cancer progression stages with a reasonable number of samples in each stage. It contains 105 samples from 100 cancer patients. Table 4.1.1 shows the various stages of prostate cancer progression according to the American Cancer Society, as well as the number of samples in each stage. The samples were grouped according to their stage or substage and were labeled T1C, T2, T2A, T2B, T2C, T3, T3A, T3B, and T4.

Table 4.1.1: Stages of progression of prostate cancer according to the American Cancer Society

Prostate Cancer Stage	Description	Number of Samples
T1c	The tumor is not detectable via imaging techniques. Cancer is detected using a needle biopsy due to an elevated serum prostate-specific antigen (PSA).	14
T2	The tumor is palpable, but confined to the prostate.	10
T2a	The tumor is in half, or less than half, of one of the prostate glands' two lobes.	23
T2b	The tumor is in more than half of one lobe, but is not in both lobes.	11
T2c	The tumor is in both lobes but confined within the prostatic capsule.	30
T3	The tumor has started to spread out of the prostate tissue.	2
T3a	The tumor has spread through the prostatic capsule on one or both sides, but has not spread to the seminal vesicles.	6
T3b	The tumor has invaded one or both of the seminal vesicles.	8
T4	The tumor has spread to other organs.	1

4.2 Preprocessing

Singireddy et al. [43] conducted a study on prostate cancer progression and generated transcripts for each sample in Long's dataset. In their study, the samples files were converted from SRA format to FASTQ format. TopHat2 [23] then was used to align the reads to the reference genome (hg19), which resulted in identifying the accepted reads. The accepted reads were fed to Cufflinks [50] to assemble the transcripts and to calculate their FPKM values (fragments per kilobase of transcript per million mapped reads). FPKM is a measurement unit for RNA-Seq expression.

Reconstructing the transcripts for all samples using Cufflinks generated 43,497 transcripts or features, and some of these transcripts were non-coding RNAs. This study focuses on coding RNAs (mRNA) that have the potential to code for proteins; thus non-coding transcripts were removed and the number of features were reduced to 33,801 coding transcripts. Figure 4.2.1 displays a sample table of the output data from Cufflinks that was used. The columns represent the transcripts and their FPKM values and the rows represent the samples, of which there were 105 samples in total. The last column, the class label, shows the stage of prostate cancer for each sample.

4.3 Our Scheme for Classifying Stages

Even though we deal with several classes (stages of prostate cancer), we perform the classification as a binary classification. Cancer progression means the cancerous tumor continuously grows to the next stage. The progression develops in a specific order starting with stage T1C and ending with stage T4 (T1c $\rightarrow$ T2 $\rightarrow$ T2a $\rightarrow$ T2b $\rightarrow$ T2c $\rightarrow$ T3 $\rightarrow$ T3a $\rightarrow$ T3b $\rightarrow$ T4). This thesis focuses on finding differentially expressed protein isoforms between neighboring stages. For example, it compares samples from T1c with T2, from T2 with T2a, from T2a with T2b, and so on.

Stages T3 and T4 contained a small number of samples, which affects the classifier

	A	B	C	D	E	F	G	H	I	J	K	L
1	NM_000959	NM_001023	NM_001142	NM_001143	NM_001195	NM_001199	NM_001257	NM_003940	NM_004772	NM_017753	NM_052965	class_label
2	0.03345703	1.79315452	1.56316471	0	0.0564578	0.40811998	0.02857689	0.56784502	0	0	1.12121273	T3
3	0.12529229	0.75141786	0.80921513	0	0.09986194	0.41643543	0	0.24781324	0.46687389	0.08560525	0.3367042	T34
4	0.02356818	0	0.86141284	0.30888299	0.03596383	0	0	0.19449671	0.23290054	0.02242972	0	T3A
5	0.02065581	0	0	0.16230421	0.0692501	1.2245753	0	0.15536643	0	0	0.26939435	T2A
6	0.24838791	0	0.65777595	0	0.06972392	1.74658622	0	0.31358756	0.89650617	0	0.38411784	T34
7	0.06587637	0	1.09387617	0	0.08170273	1.48265372	0.02026731	0.28127366	0.73317609	0	0.17562414	T1C
8	0.00563534	0	0	0	0.08871984	0.99550949	0.0076015	0.16240858	0	0	0	T2C
9	0.01551452	0	0	0.36161477	0.10024148	0.2569708	0	0.1997441	0	0.05624634	0.16728705	T2B
10	0.08552552	0	1.64279997	0.25585633	0.02544805	0	0	0.20176823	0	0.01496869	0.06878683	T2C
11	0.03728599	0	0.87853132	0.16356576	0	0	0	0.15767515	0.24280269	0	0	T4
12	0	0	0	0.31045298	0	0	0	0.24528985	0	0.01683941	0.11608786	T2A
13	0.01751776	0	1.15765354	0	0	0	0.00099739	0.27907701	0	0.11447305	0	T1C
14	0	0	0.57749924	0	0.03159413	0	0	0.19055235	0	0.04642895	0	T3
15	0.01591146	0	0.2251909	0.15400443	0	0.82798886	0	0.14794233	0	0.0089734	0.04123765	T2A
16	0	0	0	0.16597152	0	0.23758582	0	0.36059938	0.3081194	0.05270529	0.54148493	T2B
17	0.02003481	0	1.76570834	0	0.1138192	0.36292964	0.06364604	0.23392008	0.45830849	0	0.49979946	T34
18	0.12089653	0	0.61626543	0	0.10944881	0.75057947	0.0215669	0.23194986	0.29780686	0.02383581	0.21269634	T34
19	0.0573615	0	0.70498146	0.10691317	0.03773743	0.86484045	0	0.16546814	0	0.05545739	0	T2C
20	0.01334405	0	0.56491556	0	0.01843545	0.48077293	0	0.12429936	0	0.12191476	0.12066885	T2C
21	0	0	0	0	0	0.6296869	0.01609662	0.11664752	0	0	0	T2A
22	0.00792309	0	1.69101508	0.19868091	0	0.16543013	0	0.30617522	0.4690825	0.14362022	0.06948723	T1C
23	0.00955458	0	1.37048276	0	0	0.15214703	0	0.23771969	0.52793174	0.01212376	0.38698451	T3B
24	0	0	1.9120382	0	0.02788922	0	0	0.34753646	0	0.01866686	0.17163327	T2B
25	0.0495996	0	0	0.32215978	0	0.83680074	0	0.27578481	0.27769094	0	0.13879522	T1C
26	0.16690453	0	0.44721022	0.38124145	0	0.41430168	0.00823905	0.11438617	0.22675491	0	0	T2C
27	0.13952855	0	0.81945052	0	0.03899493	0.72666339	0	0.23447259	0	0	0.17939884	T2A
28	0.06245563	0.68626053	0.76214906	0	0	0.54292263	0.02297805	0.21363857	0	0	0.11204147	T1C
29	0.04075046	0	1.91594828	0	0.04001015	1.29640419	0	0.26620752	0.52950348	0	0.20256383	T34
30	0.0728194	0.57177488	1.22764465	0.10590588	0.05859048	1.04515275	0.13991437	0.28471779	1.15411696	0	0	T34
31	0.0062916	0	0.63898968	0.1681862	0.04813328	1.00323455	0	0.22272151	0.60869682	0	0.17340596	T2C

Figure 4.2.1: Sample file from preprocessing phase, displaying some transcripts names with their FPKM values for all samples, rows, and the class, last column.

generalization power; as a result, stages T3, T3a, T3b and T4 were merged into a new group named T34. The expression variation in protein isoforms between the latter stages of progression of T34 compared to T2c were studied. For the two stages, our experiment involved 47 samples (30 and 17 samples for stages T2C and T34, respectively). The new stage, T34, represented the critical transmission of the tumor from inside the prostate capsule to outside the prostate.

4.4 Finding ORFs

A tool for finding ORFs and corresponding protein isoforms from RNA-Seq data was implemented. A detailed description of the tool mechanism is illustrated here; Figure 4.4.2 shows a schematic view of the proposed method. The proposed tool reads transcripts and performs the whole process on each transcript individually, then subsequently considers the next sequences in the input file. It translates all six reading frames into amino acid sequences using the codon table. Then all possible ORFs are obtained, and only the longest

one in each frame is held. Finally, the longest ORF among the six candidates is considered the potential ORF for the transcript.

The original sequence is read from (5' $\rightarrow$ 3') to obtain the three positive reading frames. In order to obtain the three negative reading frames, one need to obtain the complementary strand of the positive DNA strand. As DNA is antiparallel, we also need the reverse complement sequence to keep the 5' and 3' ends correctly oriented.

Figure 4.4.1 shows an example of obtaining the reversed complementary sequence. In the example, the original sequence is read from (5' $\rightarrow$ 3') and the three positive reading frames are obtained. Two steps are done to obtain the three negative frames. First, the tool finds the complementary sequence for the original sequence by exchanging each (A, T or U, C, G) with (T, U or T, G, C). Second, the complementary sequence is reversed, so it can be read (5' $\rightarrow$ 3'), in the same manner as the first strand.

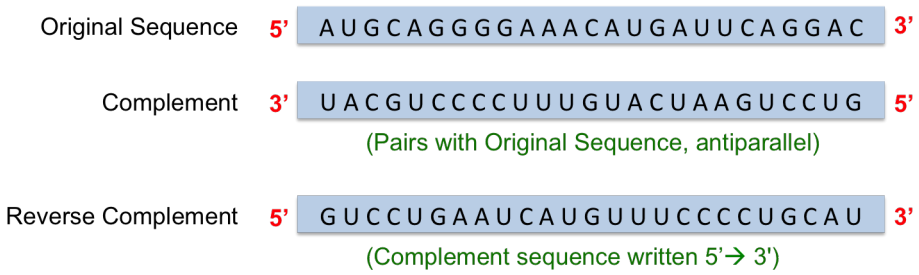


Figure 4.4.1: Obtaining complementary sequence from original mRNA sequence.

After that, the tool scans each reading frame individually and detects all possible ORFs, namely the protein isoforms. The ORF extends from the start codon AUG and continues until the reading frame defined by the start codon is terminated by one of the three translation stop codons UGA, UAA, or UAG. The translation process begins by interpreting the start codon into Methionine, and includes it in the polypeptide chain. When a start codon (AUG) is found in the middle of the ORF, it is translated into Methionine, and the translation process continues until a stop codon terminates it. In the next step, by comparing the length of the detected isoforms in the six frames, the tool selects the longest one as the

most relevant isoform for the given transcript [54].

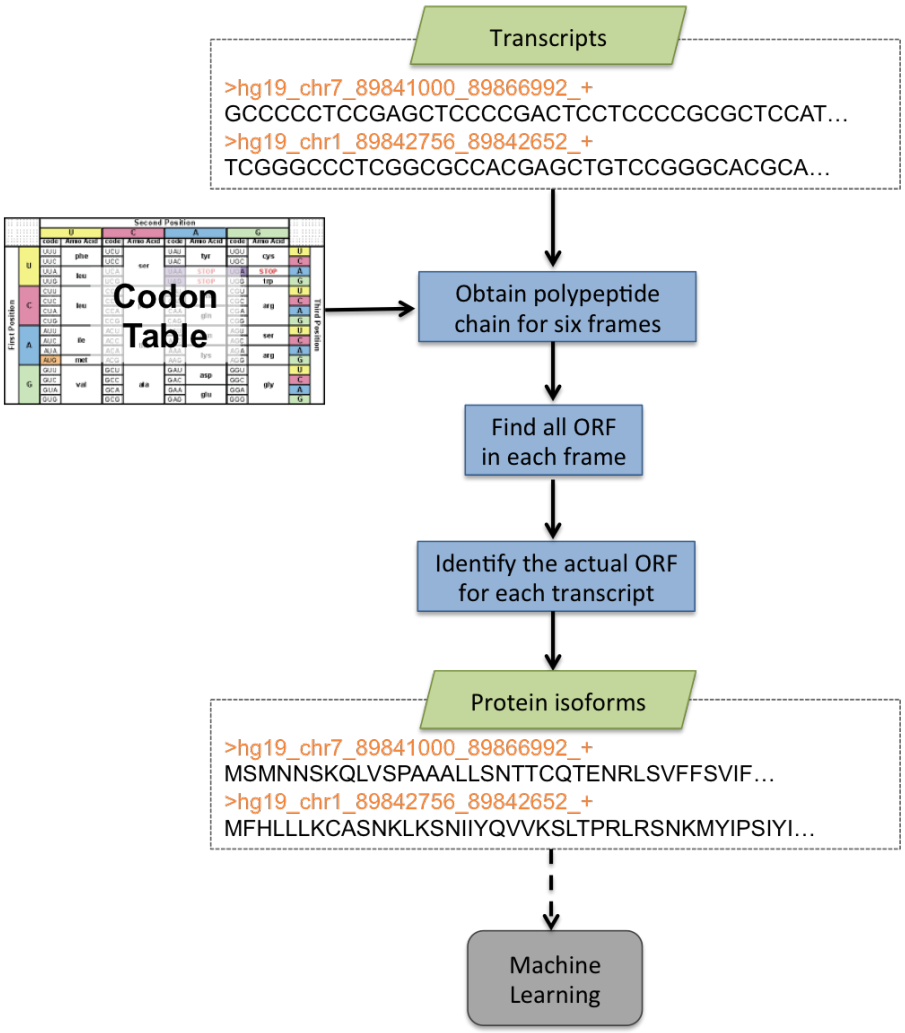


Figure 4.4.2: Schematic view of proposed tool for translating and identifying the actual ORF.

4.5 New Measurement for Quantification

In any given cell, only a fraction of its genes are turned on at a given time; this is the concept of gene expression. Any change in the perfectly controlled process will affect the cell and may cause diseases such as cancer. In other words, irregular expressions in tumors make cancer cells different from healthy cells.

Gene expression can be measured using different methods; RNA-Seq data was used in

this work. the thesis's interest is measuring the expressions of protein isoforms that were constructed using the proposed tool. The expression level is calculated for isoforms in each sample, and by comparing different samples, we can extract some information about prostate cancer progression.

To estimate the number of supporting fragments, a new measurement derived from FPKM value is considered. The new quantification method makes the measurement of abundance more accurately related to only those fragments that are located within that region. FPKM values for all transcripts are computed by Cufflinks, shown in Figure 4.2.1.

The FPKM unit considers all fragments within a transcript, while only the fragments that fall within the actual ORF region need to be taken into the account for this thesis. If *L(transcript)* is the length of a given transcript and *L(ORF)* is the length of its ORF, then

$$AFPKM = \frac{L(ORF)}{L(transcript)} \times FPKMvalue \tag{4.5.1}$$

Figure 4.5.1 shows an example of how one finds abundance with the proposed measure. The new FPKM value is called adapted FPKM (AFPKM). It takes into account the proportion of the transcript that is actually translated into protein and ignores the non-coding part.

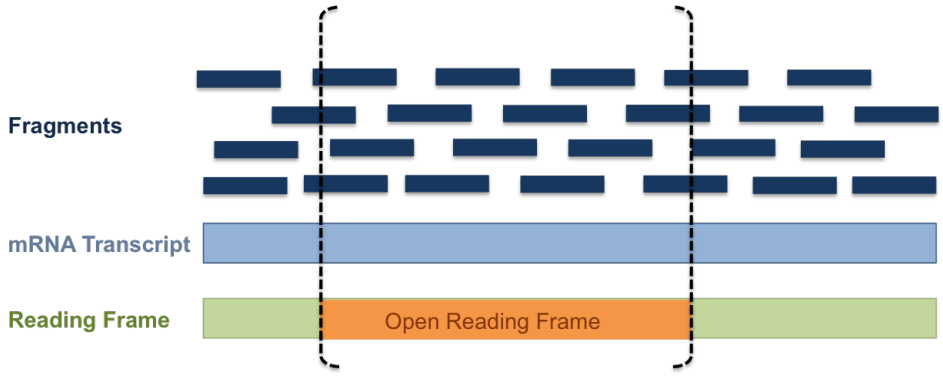


Figure 4.5.1: Use of fragments to measure abundance of protein isoforms.

4.6 Feature Selection and Classification

With 33,801 transcripts or features, it is important to reduce the dimensionality of the problem by applying feature selection to improve the classification performance. For this, the mRMR feature selection method [33] was used. The features here are the protein isoforms with their AFPKM values; and the class labels are the prostate cancer stages.

Figure 4.6.1 shows the pipeline of the machine-learning approach. In the figure, there is a table that involves transcripts and their FPKM values for a pair of stages. The table acts as input to the proposed tool to identify the ORF for each transcript and obtain the corresponding protein isoforms. The feature selection filters out noisy and redundant features. The filtered features are sent to the classification algorithms to detect protein isoforms, or biomarkers, that are differentially expressed.

Weka was used to perform the feature selection and the classification [18]. It is a well-known data mining tool developed at the University of Waikato, and it is commonly used in bioinformatics.

4.6.1 Feature Selection

Among 33,801 protein isoforms or features per sample, there is a significant number of irrelevant and noisy transcripts that affect the classification's accuracy. They make the classification algorithms perform poorly and consume a large amount of memory and time. Thus, feature selection technique was used to reduce the dimensionality of the problem and to remove features that degraded classification performance.

The feature selection technique mRMR [33] was applied; it is a wrapper method used to select a subset of features. MRMR embeds a classification algorithm to choose the most relevant features. The algorithm mRMR incorporates a classifier to determine the best subset of features that can classify the samples. Different classifiers were experimented with, and Naive Bayes gave the best results with default parameters.

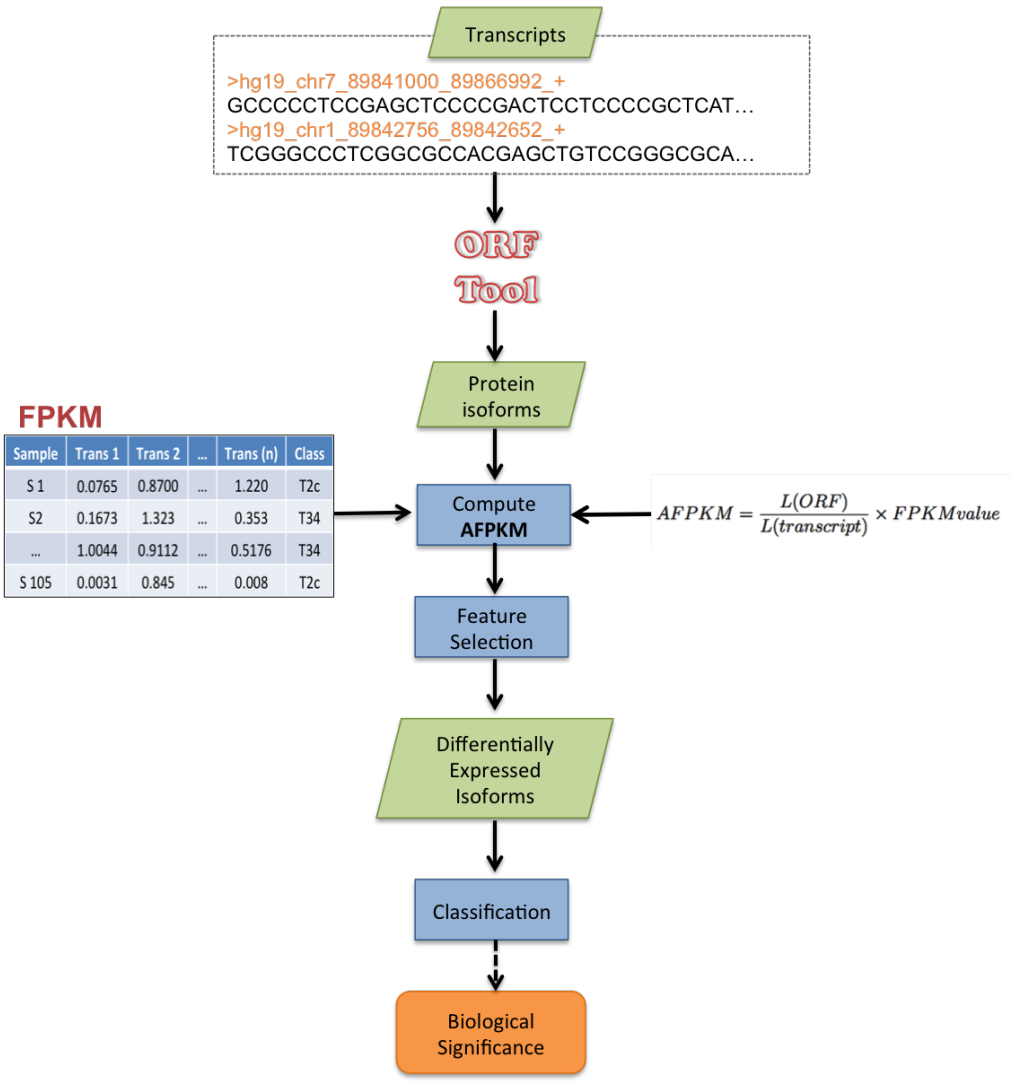


Figure 4.6.1: Pipeline of our method for prostate cancer progression classifications.

4.6.2 Classification

After feature selection, different classification algorithms that discriminate between prostate cancer stages were applied. Various classifiers that classify the samples based on the classification rules learned in the training phase were tested. SVM [11] was used, and the linear, polynomial, and radial basis function kernels were experimented with grid search optimization. Random forest [4] and the J48 decision tree [39] were also tested. Additionally, the Naive Bayes classifier was used to compare its performance with all other classifiers. Finally, the differentially expressed protein isoforms in the desired stages were

obtained. To evaluate the classifiers and maintain their generalization capability, a 10-fold cross-validation was used. Chapter 5 assesses the performance of the classifiers using several performance measures.

4.6.3 Comparison Between Using Protein Isoforms as Features and Using Transcripts as Features

In a previous study [43], Singireddy et al. used the same dataset used in this thesis; the dataset include 105 prostate cancer samples along with their FPKM value. They applied some feature selection algorithms on the pairwise stage T2c-T34 which consists of 47 samples. They found a set of 11 transcripts shown in Figure 4.6.2, that were able to discriminate stage 2 of prostate cancer from subsequent stages with very high accuracy.

Two experiments were conducted to test the significance of using protein isoforms instead of whole transcripts as biomarkers for prostate cancer progression. First, classification results were compared with [43] results for the pairwise stage (T2c-T34). Second, we fed our tool with those 11 transcripts to find the actual ORFs in each. Then AFPKM was computed using the Formula 4.5.1. Finally, different classifiers were run and their performances were evaluated using protein isoforms, compared to transcripts. .

Transcript	Chr.	Gene		Transcript	Chr.	Gene
NM 001257413	17	IKZF3		NM 001023567	15	GOLGA8B
NM 003940	3	USP13		NM 001143766	10	ZNF438
NM 001142274	2	CLASP1		NM 017753	9	LPPR1
NM 001199165	17	CEP112		NM 000959	1	PTGFR
NM 052965	1	TSEN15		NM 004772	5	NREP
NM 001195283	14	FLVCR2				

Figure 4.6.2: Set of 11 transcripts for discriminating stage 2 of prostate cancer from stages 3 and 4 obtained by Singireddy [43].

CHAPTER 5

Results and Discussion

This chapter discusses the results of the approach in finding biomarkers for prostate cancer progression. Protein isoforms extracted using the proposed tool were processed using the feature selection technique. The feature vectors were input to the classifiers, and the classification performance was graphically visualized. Classification and feature selection algorithms were employed to find differentially expressed protein isoforms. Classification performance measures for pairwise stages are discussed. Furthermore, an experiment was conducted to test the significance of using protein isoforms instead of whole transcripts to differentiate between prostate cancer stages. The classification performances were compared using protein isoforms and transcripts. Finally, the biological significances of some selected protein isoforms are demonstrated.

5.1 Generating Protein Isoforms

A tool for finding ORF and reconstructing protein isoforms was implemented. The tool was fed with 33,801 transcripts to identify the potential proteins coded in them. The first prototype of the proposed tool was been written in Python; hence, it is easy to install on many different platforms. Unlike existing tools, the proposed tool can process several transcripts at a time and obtain the actual ORF for each of them. The main module accepts the transcripts in FASTA format as inputs and then outputs two files. One of the files contains sequences in FASTA format, which include an identifier for the sequence, and the

amino acid sequence of the predicted protein isoform. The FASTA file can be searched in a protein sequence database to similar proteins using BLASTP, for example. Figure 5.1.1 shows an example for the first output file. The second file contains more detailed information, such as the sequence identifier; the polypeptide chains for all six frames; identified ORFs for all six frames along with their lengths; and the detected ORF (the potential protein isoform) for a transcript; and its position in the sequence. The amino acid sequences in the generated files have the same order as the input file.

The identified protein isoforms are used as features that incorporate different machine learning techniques to identify a small group of isoforms that are differentially expressed in the various stages of prostate cancer progression.

Figure 5.1.1: Sample of output file containing the ORF.

5.2 Feature Selection Results

The feature selection method mRMR was applied; it resulted in identifying 36 protein isoforms for the set of all pairwise stages. Figure 5.2.1 shows the transcripts IDs for the differentially expressed protein isoforms across different stages. The number of selected protein isoforms for pairwise stages T1c-T2, T2-T2a, T2a-T2b, T2b-T2c, T2c- T3a, T3a-

T3b, and T2c-T34 are 3, 8, 5, 3, 3, 4, and 10 respectively. More information about the selected protein isoforms such as the location (their locus) in the corresponding chromosome and gene name can be found in Appendix B. The classification efficiency using the selected isoforms is discussed for each pairwise stage in the following section.

T1c – T2	T2 – T2a	T2a – T2b	T2b – T2c	T2c – T3a	T3a – T3b	T2c – T3,4
NM_000110	NM_001146192	NM_000681	NM_000029	NM_001198979	NM_002154	NM_001024674
NM_000710	NM_001272095	NM_001134473	NM_001711	NM_003546	NM_006465	NM_001023567
NM_001042574	NM_003540	NM_001145138	NM_032023	NM_032875	NM_024604	NM_001099285
	NM_004860	NM_007225			NM_182485	NM_001198979
	NM_032850	NM_032023				NM_002437
	NM_145892					NM_003387
	NM_153274					NM_001257413
	NM_172350					NM_006214
						NM_016940
						NM_001256

Figure 5.2.1: Transcripts IDs for differentially expressed protein isoforms across different stages.

5.3 Classification Results

The selected features by mRMR were employed to build a classification model for all pairwise stages. Bayes Net, Naive Bayes, JRip, random tree, and random forest were tested. The accuracy range for all classifiers was between 82.98% and 100% with an average of 93.46%. The Bayes Net classifier performed significantly for three pairwise stages out of seven by predicting all samples correctly.

Figures 5.3.1 and 5.3.2 show the recalls and precisions of different classifiers using the selected protein isoforms. The x-axis represents pairwise stages of prostate cancer progression while the y-axis displays the efficiency measurement used. As shown in the figure, the recall of the Bayes Net classifier outperformed other classifiers for most of the pairwise stages. For stages (T1c-T2), (T2-T2a), (T3a-T3b), and (T2c-T34), Bayes Net achieved a perfect recall (its recall=1). Regarding Precision, all the classifiers achieved ideal precision (a precision = 1) for distinguishing between sub-stages T3a-T3b.

The area under the ROC curve (AUC) results was visualized to differentiate between

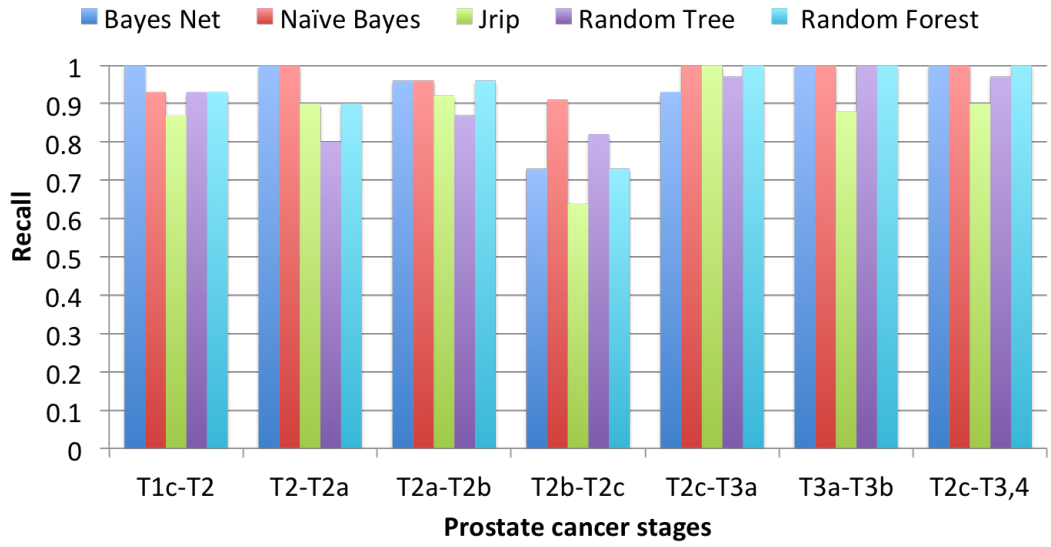


Figure 5.3.1: Classification performance for all pair-wise stages using recall as performance measurement.

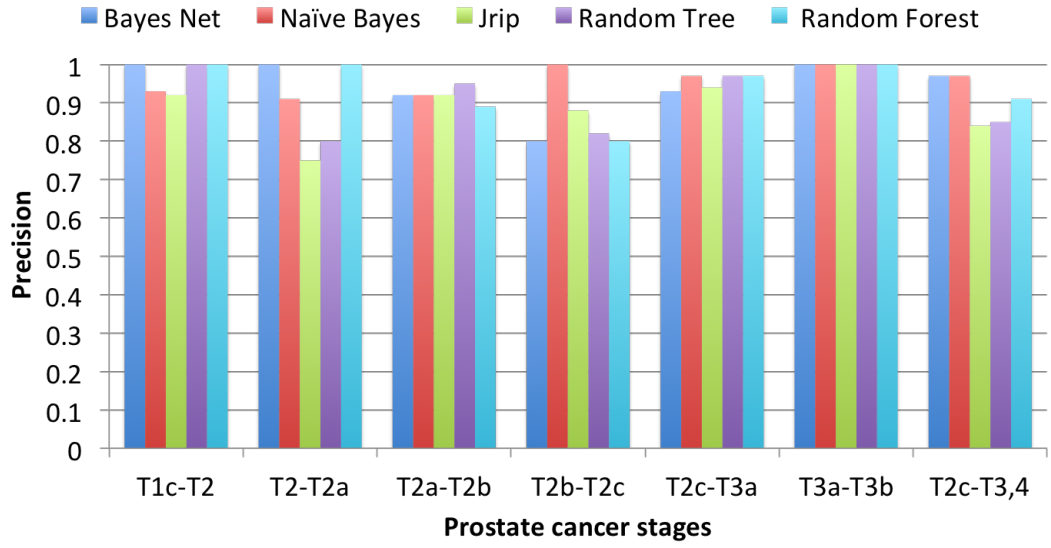


Figure 5.3.2: Classification performance for all pair-wise stages, using precision as performance measurement.

prostate cancer stages. Figure 5.3.3 shows the classifiers' performance considering the AUC value. For T2c-T34, AUC for the Bayes Net and Naive Bayes classifiers reached 0.99, while it reached 0.95 for random forest.

Regarding F-meaure, the classifiers' results can be seen in Table 5.3.1. The F-measure

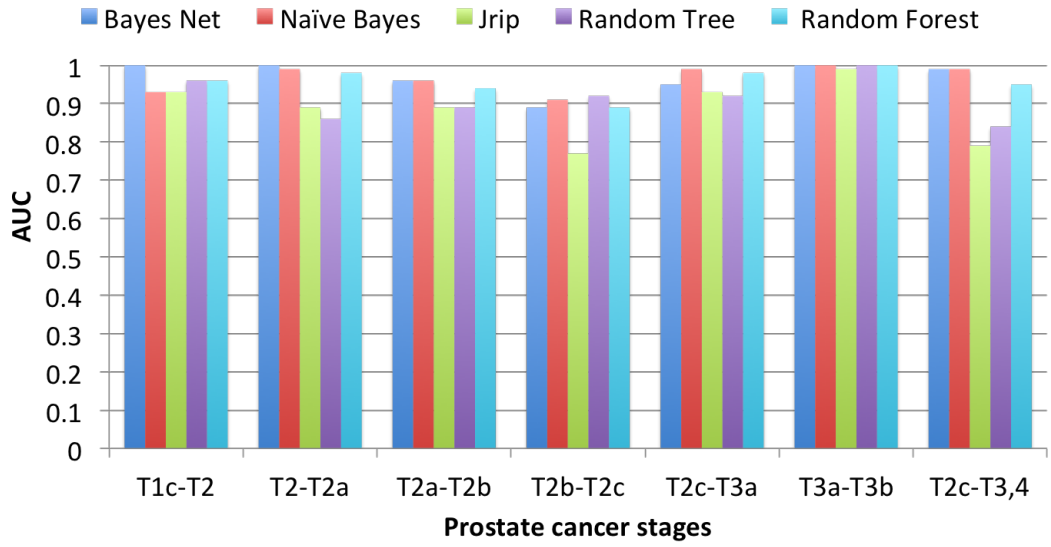


Figure 5.3.3: Classification performance for all pair-wise stages, using AUC as performance measurement.

for Naive Bayes is noticeably high for all pairwise stages (above 0.90). Additionally, Naive Bayes achieved perfect results for the pairwise stages (T1c-T2), (T2-T2a), and (T3a-T3b). (T2b-T2c) were the hardest pair to be classified by all classifiers with an F-measure = 0.82 on average. Similar to all other performance measurements, (T3a-T3b) were accurately classified by most of the classifiers using F-measure.

Table 5.3.1: Classification performance for all pair-wise stages using F-measure as measurement.

Stages	Bayes Net	Naive Bayes	JRip	Random Tree	Random Forest
T1c-T2	1	0.93	0.89	0.96	0.96
T2-T2a	1	0.95	0.82	0.80	0.95
T2a-T2b	0.94	0.94	0.92	0.91	0.92
T2b-T2c	0.76	0.95	0.74	0.90	0.76
T2c-T3a	0.93	0.98	0.97	0.97	0.98
T3a-T3b	1	1	0.93	1	1
T2c-T34	0.98	0.98	0.87	0.91	0.95

In addition, SVM was tested with linear, ploynomial, and RBF kernels. Figures 5.3.4

and 5.3.5 show the accuracy and AUC values of the SVM classifiers for all stages. The accuracy for all classifiers throughout all pairwise stages varied between 79.4% and 97.8% with an average of 91.06%.

The AUC value for the pairwise stage (T2c-T3a) was 0.97 by SVM-linear, SVM-ploynomial, and SVM-RBF. Similarly, (T3a-T3b) achieved AUC=0.94 with all classifiers. The SVM with the RBF kernel achieved the best performance for (T2a-T2b) with an AUC=0.87, while the other kernels' scores were 0.73.

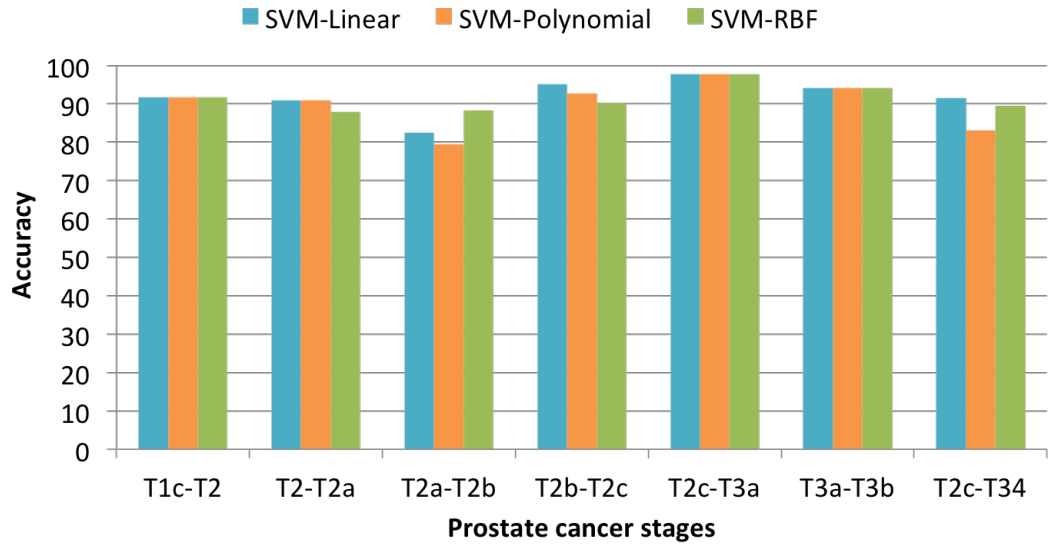


Figure 5.3.4: Performance of SVM classifiers with selected protein isoforms, using accuracy as measurement.

In general, the results illustrate that the selected protein isoform could be used to differentiate samples across all stages with a high level of accuracy. The protein isoforms could be investigated further as potential biomarkers of prostate cancer progression.

5.4 Results for Using Protein Isoforms Versus Transcripts

Figure 5.4.1 shows the classifiers' performances and compares using transcripts identified by [43] and using our protein isoforms. As shown in the figure, for most of the classifiers, the thesis's proposed method provides superior performance while using a smaller

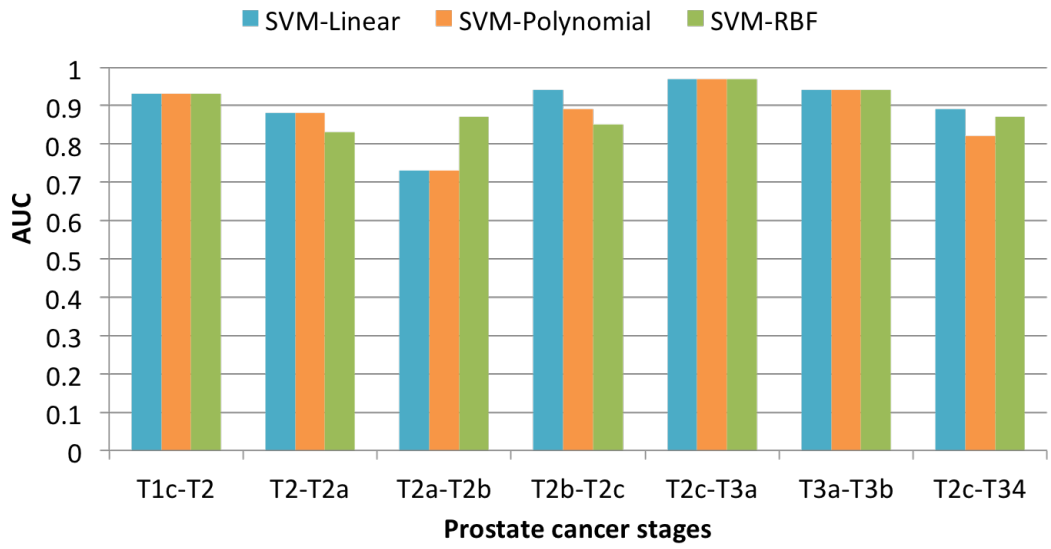


Figure 5.3.5: Performance of SVM classifiers with selected protein isoforms, using AUC as performance measurement.

group of biomarkers. Using ORFs detected by our model for discriminating T2c-T34 stages almost always achieved better results than using original transcripts. J48 performance improved by around 10% when using protein isoforms. Only the SVM-Polynomial achieved higher accuracy when using transcripts instead of using protein isoforms, with 91.2% for the former and 84.5% for the latter.

In the second experiment, the performance of the transcripts was compared to the previous study with their ORFs. The ORF for each transcript was found and applied to different classifiers using them. The performance of the 11 identified transcripts was compared with the performance of their corresponding ORFs for discriminating between stages T2C and T34 of prostate cancer. Figure 5.4.2 shows the classification accuracy for the comparison between using the original transcripts (the FPKM values) and the ORFs detected in them (the AFPKM values). As shown in the figure, using protein isoforms instead of RNA transcripts boosts the classification performance in some classifiers, using default parameters such as SVM-linear and SVM-polynomial.

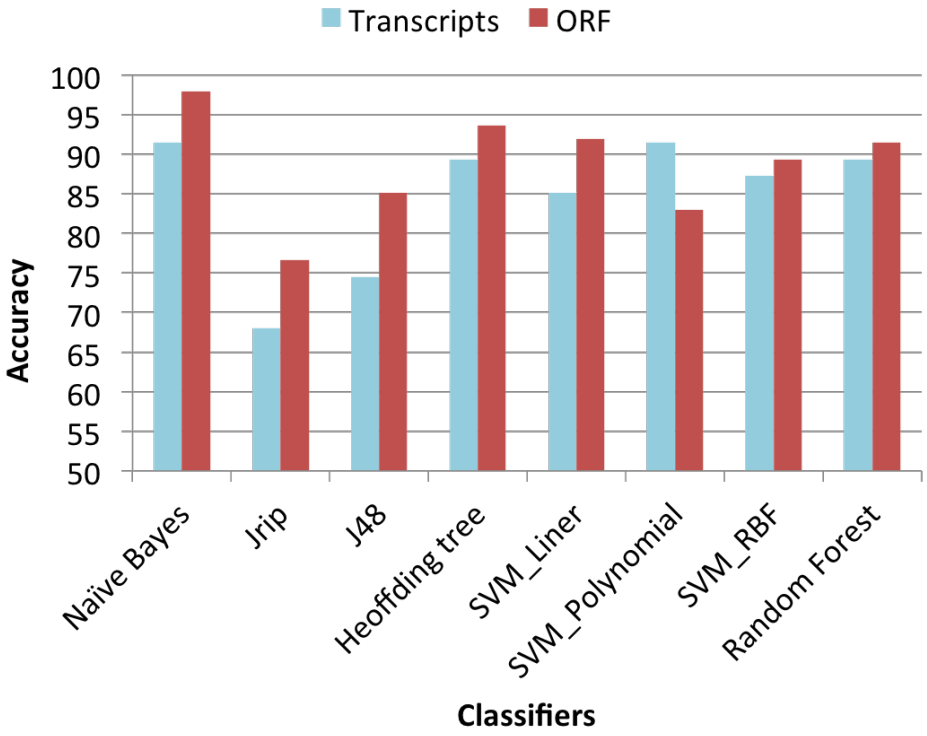


Figure 5.4.1: Comparison between this thesis's protein isoforms and transcripts reported in [43].

5.5 Discussion and Biological Significance

The 36 transcripts shown in Figure 5.2.1 were obtained using the feature selection technique mRMR. It was found that a great proportion of the detected isoforms or genes are associated with cancer or other diseases. In this section, the biological significance of selected biomarkers is highlighted.

Studies [10] and [42] demonstrated that CREB was involved in tumor initiation, progression, and metastasis. Another study focused on utilizing the CREB protein for therapeutic purposes in cancer [56]. The authors stated that CREB was a critical regulator of cell differentiation and proliferation. They suggested that overexpression and over-activation of the protein were noticed in diverse cancer tissues including prostate cancer. However, the down-regulation of CREB in cancer cells inhibits tumor growth and induces apoptosis.

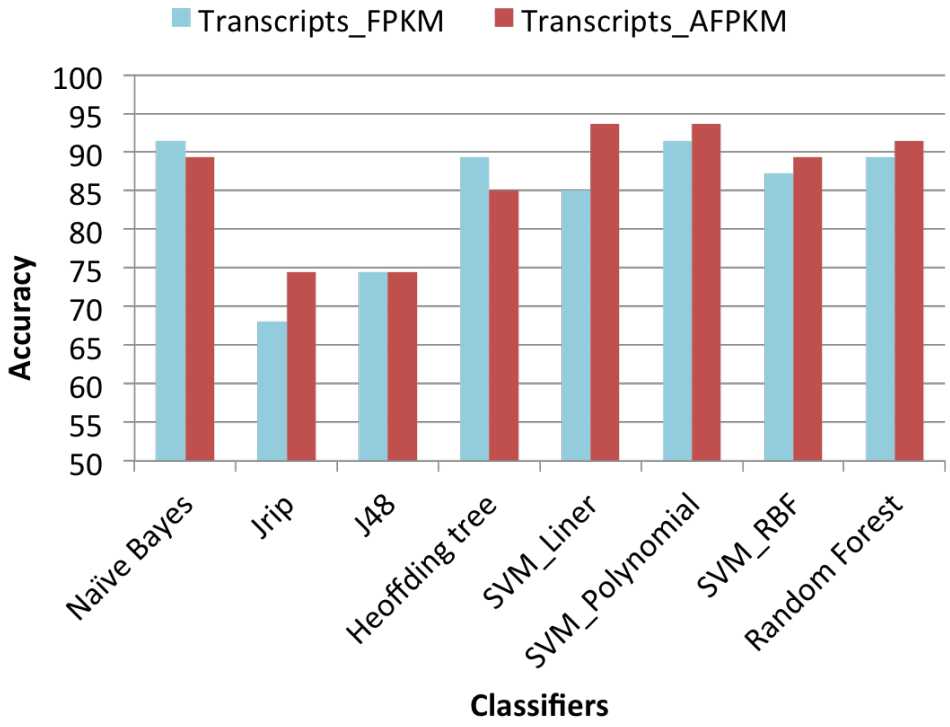


Figure 5.4.2: Comparison between using transcripts with original FPKM [43] and the same transcripts with AFPKM, and its effect in classification performance for (T2c-T34).

The protein isoform detected in the HIST1H4F gene is a member of the H4 histone family. Fraga et al. [17] found that cancer cells had changed forms of histone H4. The changes appeared early in cancer and accumulated during the tumor progression. The authors suggested that the changing of histone H4 is a common characteristic of human tumor cells.

The Gse1 coiled-coil protein (GSE1) is expressed differentially between stages T2a-T2b. Chai et al. state that the GSE1 protein is overexpressed in breast cancer, and the silencing of GSE1 remarkably suppresses breast cancer cells' proliferation, migration, and invasion. They concluded that GSE1 plays an important role in promoting breast cancer progression [7].

In the pair (T2b-T2c), different expression levels for the Biglycan protein were detected. It has been shown that high a expression level of Biglycan is present in various cancers such as endometrial [26], pancreatic [2], and gastric [21]. Xing et al. explored the function of Biglycan in colon cancer progression. Decreasing the expression levels of the protein prevents colon cancer cell migration and invasion. In addition, down regulating Biglycan induces apoptosis in cancerous cells [57].

Payne et al. identified the protein HIST1H4K (which is from the same family as the nominated biomarker HIST1H4L) as a prostate cancer biomarker. The biomarker showed minimal correlation with PSA, which is the target protein in any typical prostate cancer test [42]. The authors suggested integrating HIST1H4K with a PSA test to diagnose prostate cancer in patients.

HSP70 belongs to a class of proteins called heat shock proteins (HSPs). In general, HSPs' expression levels are higher in cancer cells. The HSP70 protein is highly expressed in some cancers; However, its reduction has led to tumor regression [34]. Murphy et al. conducted an intensive study on HSP70 and its role in cancer and concluded that HSP70 gene can function as an oncogene [29]. There are many studies that discuss HSP70's association with cancer [5], [9]. One study focused on investigating HSP70 as a biomarker for prostate cancer [1]. The authors suggested that HSP70 could be a potential biomarker for prostate cancer. Their experimental results showed that HSP70 was overexpressed in patients' plasma, whereas the patients' PSA levels were normal.

It has been shown that changes in the PTMA expression (the protein in the transcript `NM_001099285`) is involved in the development and progression of prostate cancer [46]. Another study demonstrated that lower expression levels of Prothymosin alpha (PTMA) were linked to the inhibition of prostate cancer initiation [19].

In order to obtain a further insight, the hypothesis that the PTMA expression increases with the progression of prostate cancer was investigated for this thesis. Figure 5.5.1 shows the expression trend level of the ten chosen isoforms between stages T2C and T34. The *x*-axis represents the 10 identified protein isoforms, while the *y*-axis represents the median of AFPKM values across different samples from each group. As shown in the figure, the expression level for the protein isoform that we identified in PTMA was by far higher in stage T34, compared to T2c (1.3 and 0.1 respectively).

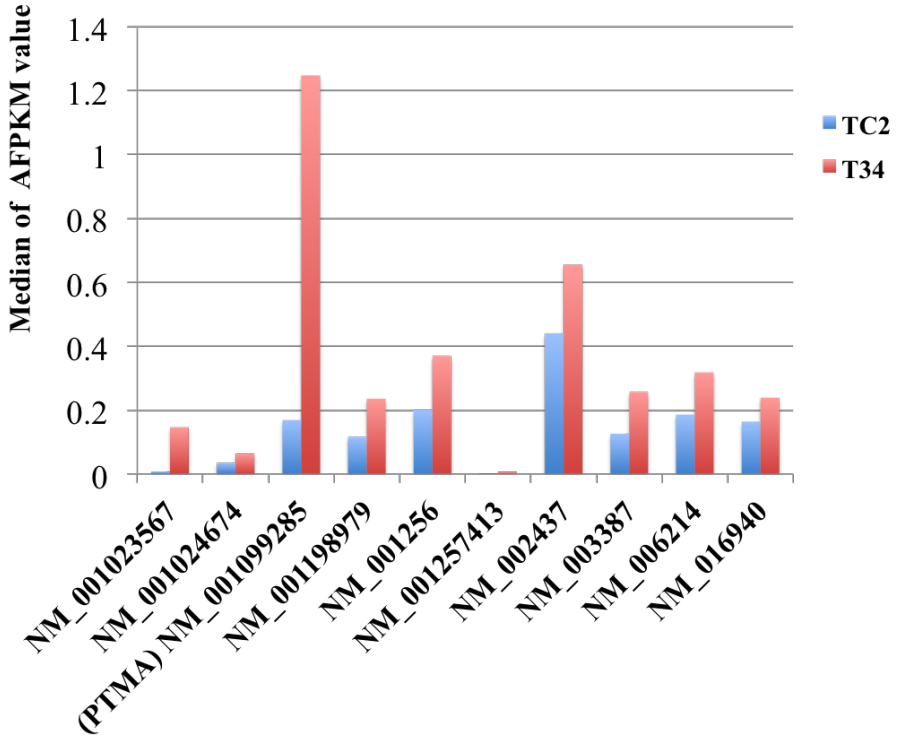


Figure 5.5.1: Comparison between median of AFPKM values for 10 selected protein isoforms in stages T2C and T34 of prostate cancer.

Transcript `NM_001099285` in particular had a significant FPKM variation between T2C and T34. Suzuki et al. found that PTMA's expression was involved in human prostate cancers progression [46]. The authors conducted an experiment *in vitro* and studied autopsy cases of those who died of prostate cancer. The study covered samples of different prostate cancer stages as well as benign samples. The results showed that the PTMA expression level correlated with prostate cancer development; the levels were the highest in the autopsy

cases.

In all of the thesis's samples, a sudden increase in the expression of PTMA was experienced as the tumor progressed across later stages. The literature review about Prothymosin (PTMA) supports the thesis's result and makes the protein a potential biomarker for prostate cancer progression monitoring between stage 2 and subsequent stages. It is worth further investigation.

CHAPTER 6

Conclusion and Future Work

6.1 Contributions

An open reading frame (ORF) is a continuous sequence of codons that begins with a start codon and ends with a codon. Finding ORFs that correspond to a given mRNA transcript is a major step in reconstructing protein isoforms and is vital for a better understanding of RNA alternative splicing effects in diseases like cancer. Prostate cancer is one of the most widespread types of cancer, especially in developed countries. Finding biomarkers is a vital step in the early diagnosis of the tumor and in classifying stages in the progression of the disease.

In this work, a new tool that can find ORFs and protein isoforms for any given transcript was proposed. One of the advantages of the tool over other existing tools is the ability to identify the actual ORF for a transcript. Moreover, it obtains the ORFs for more than one transcript at a time, which eliminates the burden of manually processing of each transcript.

A model that uses RNA-Seq data and machine learning techniques to detect prostate cancer progression was also proposed. The tool was used to find the ORF for a given mRNA sequence and to then identify the corresponding protein isoform. Some machine learning techniques employed the generated protein isoforms, and the model was able to discriminate between stages of prostate cancer with a very high accuracy.

6.2 Future Work

Possible extensions of this work include improving the quantification method by investigating more precise statistical measures for counting only the relevant reads inside the ORF. Several applications of this tool involve a standalone tool for public use and an embedding of integration with a variety of databases such as UniProt.

Further investigations could be done on functional and interactomics analyse of the relevant proteins and their interactions with other proteins and molecules. Such integrations will result in insight about the biological aspects of the disease that are not captured by conventional approaches, which cover up to RNA transcripts or merely genes. Moreover, an investigation of additional classifiers and feature selection algorithms may improve the results.

The results achieved by the proposed model are closely related to prostate cancer. However, having a larger number of samples in each progression stage would improve the model, and further biological experiments are highly recommended. Finally, the model could be extended to other types of cancers and their progression stages.

BIBLIOGRAPHY

[1] Abe, M., Manola, J. B., Oh, W. K., Parslow, D. L., and George, D. J. (2004). Plasma levels of heat shock protein 70 in patients with prostate cancer: A potential biomarker for prostate cancer. *Clinical Prostate Cancer*, 3(1):49–53.

[2] Aprile, G., Avellini, C., Reni, M., Mazzer, M., and Foltran, L. (2013). Biglycan expression and clinical outcome in patients with pancreatic adenocarcinoma. *Tumor Biology*, 34(1):131–137.

[3] Biology Online (2016). Isoform definition.

[4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[5] Calderwood, S. K., Khaleque, M. A., and Sawyer, D. B. (2006). Heat shock proteins in cancer: Chaperones of tumorigenesis. *Trends in Biochemical Sciences*, 31(3):164–172.

[6] Canadian Cancer Society (2014). Stages and sub-stages of prostate cancer.

[7] Chai, P., Tian, T., Zhao, D., and Zhang, H. (2016). GSE1 negative regulation by miR-489-5p promotes breast cancer cell proliferation and invasion. *Biochemical and Biophysical Research Communications*, 471(1):123–128.

[8] Chang, C. C. and Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[9] Ciocca, D. R. and Calderwood, S. K. (2005). Heat shock proteins in cancer: Diagnostic, prognostic, predictive, and treatment implications. *Cell Stress and Chaperones*, 10(2):86–103.

[10] Conkright, M. D. and Montminy, M. (2005). CREB: The unindicted cancer co-conspirator. *Trends in Cell Biology*, 15(9):457–459.

[11] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[12] Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and John, J. S. (2016). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genetics*, 24(4):340–341.

[13] Davey, N. E., Haslam, N. J., Shields, D. C., and Edwards., R. J. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology*, 9(11):1.

[14] EMBOSS: The European Molecular Biology Open Software Suite (2016). GetORF.

[15] Engstrom, P. G., Steijger, T., Sipos, B., Grant, R., and Andr, A. (2013). Systematic evaluation of spliced alignment programs for RNA-Seq data. *Nature Methods*, 10(1):1185–1191.

[16] Fackenthal, J. and Godley, A. (2008). Aberrant RNA splicing and its functional consequences in cancer cells. *Disease Models and Mechanisms*, 1(1):37–42.

[17] Fraga, M. F., Ballestar, E., and Villar, A. (2005). Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics*, 37(4):391–400.

[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H.

(2009). The WEKA Data Mining Software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

[19] Hikosaka, A., Asamoto, M., Hokaiwado, N., and Kato, K. (2004). Inhibitory effects of soy isoflavones on rat prostate carcinogenesis induced by 2-amino-1-methyl-6-phenylimidazo [4, 5-b] pyridine (PhIP). *Carcinogenesis*, 25(3):381–387.

[20] Ho, T. K. (1995). Random decision forests. *Document Analysis and Recognition*, 1:278–282.

[21] Hu, L., Duan, Y. T., Li, J. F., and Su, L. P. (2014). Biglycan enhances gastric cancer invasion by activating FAK signaling pathway. *Oncotarget*, 5(7):96–1885.

[22] Illustrated Glossary from National Centre for Biotechnology Information, NCBI (2016). Isoform definition.

[23] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions, and gene fusions. *Genome Biology*, 14(4):1.

[24] Lilja, H., Ulmert, D., and Vickers, A. (2008). Prostate-specific antigen and prostate cancer: prediction, detection, and monitoring. *Nature Reviews Cancer*, 8(4):268–278.

[25] Liu, H. and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. *ICTAI*, 1(5):388–391.

[26] Liu, Y., Li, W., Tai, Y., Lu, Q., and Yang, N. (2014). Expression and significance of biglycan in endometrial cancer. *Archives of Gynecology and Obstetrics*, 289(3):649–655.

[27] Long, Q., Xu, J., Osunkoya, A., Sannigrahi, S., and Johnson, B. (2014). Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer Research*, 74(12):3228–3237.

[28] Mezlini, A., Smith, E., Fiume, M., Buske, O., Savich, G., and Shah, S. (2013). iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research*, 23(3):519–529.

[29] Murphy, M. E. (2013). The HSP70 family and cancer. *Carcinogenesis*, 34(6):1181–1188.

[30] NCI: Dictionary of Cancer Terms (2016). National Cancer Institute.

[31] Nogueira, L., Corradi, R., and Eastham, J. A. (2010). Other biomarkers for detecting prostate cancer. *British Journal of Urology International (BJUI)*, 105(2):166–169.

[32] Pan, Q., Shai, O., Lee, L., Frey, B., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415.

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

[34] Pick, E., Kluger, Y., Giltnane, M. J., Moeder, C., and Camp, R. L. (2007). High HSP90 expression is associated with decreased survival in breast cancer. *Cancer Research*, 67(7):2932–2937.

[35] Plant and Soil Sciences e-Library (2016). Standard codon table.

[36] Prostate Cancer Canada (2016a). Statistics for prostate cancer initiation and progression in Canada.

[37] Prostate Cancer Canada (2016b). Statistics for prostate cancer initiation and progression in Canada.

[38] Qu, Y., Yasui, Y., Ward, M. D., Cazares, L., Schellhammer, P. F., and Feng, Z. (2002). Boosted decision tree analysis of seldi serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10):1835–1843.

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

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

[41] Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674.

[42] Sandoval, S., Pigazzi, M., and Sakamoto, K. M. (2009). CREB: A Key Regulator of Normal and Neoplastic Hematopoiesis. *Advances in Hematology*, 20(1):5–8.

[43] Singireddy, S., Alkhateeb, A., Rezaeian, I., Rueda, L., Cavallo-Medved, D., and Porter, L. (2015). Identifying differentially expressed transcripts associated with prostate cancer progression using rna-seq and machine learning techniques. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pages 1–5. IEEE.

[44] Skotheim, I. and Nees, M. (2007). Alternative splicing in cancer: Noise, functional, or systematic. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical*, 39(7):1432–1449.

[45] Steijer, T., Abril, J., Engstrom, F., Kokocinski, F., and Hubbard, T. (2013). Assessment of transcript reconstruction methods for RNA-Seq. *Nature Methods*, 10(12):1177–1184.

[46] Suzuki, S., Takahashi, S., Takeshita, K., and Hikosaka, A. (2006). Expression of

prothymosin alpha is correlated with development and progression in human prostate cancers. *The Prostate*, 66(5):463–469.

[47] Swiss Institute of Bioinformatics (2016). ExPASy (Expert Protein Analysis System).

[48] Talavera, D., Vogel, C., Orozco, M., Teichmann, S., and La, X. D. (2007). The (in) dependence of alternative splicing and gene duplication. *IPLoS*, 7(1):1–4.

[49] Taskin, V., Dogan, B., and Olmez, T. (2013). Prostate cancer classification from mass spectrometry data by using wavelet analysis and kernel partial least squares algorithm. *International Journal of Bioscience, Biochemistry and Bioinformatics,*, 3(2):98–102.

[50] Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., and Baren, M. J. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

[51] Wagner, M., Naik, D. N., Pothen, A., and Kasukurti, S. (2004). Computational protein biomarker prediction: A case study for prostate cancer. *BMC Bioinformatics*, 5(1):26.

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

[53] Ward, C. (2010). The pathobiology of splicing. *Canadian Journal of Biochemistry and Cell Biology*, 220(2):152–163.

[54] Watson, J. and Levinthal, C. (2008). Molecular biology of the gene. *Molecular Biology of the Gene.*, 888(1):40–76.

[55] Wikimedia Commons (2016). The six reading frames in DNA.

[56] Xiao, X., Li, B. X., Mitton, B., and Ikeda, A. (2010). Targeting CREB for cancer therapy: Friend or foe. *Current Cancer Drug Targets*, 10(4):384–391.

[57] Xing, X., Gu, X., and Ma, T. (2015). Knockdown of biglycan expression by RNA interference inhibits the proliferation and invasion of, and induces apoptosis in, the HCT116 colon cancer cell lin. *Molecular Medicine Reports*, 12(5):7538–7544.

[58] Xing, Y. and Lee, C. (2006). Alternative splicing and RNA selection pressure: Evolutionary consequences for Eukaryotic genomes. *Nature Reviews Genetics*, 7(7):499–509.

APPENDIX A

Supplementary Information

A.1 Input Sequence for Translation Tools

The following is the sequence for the transcript >hg19_chr7_89841000_89866992_+ used in Chapter 3 as an input for the translation tools.

>hg19_chr7_89841000_89866992_+

GCCCCCTCCGAGCTCCCCGACTCCTCCCCGCGCTCCACGGCTCTTCCC

GACTCCAGTCAGCGTTCCTCGGGCCCTCGGCGCCACGAGCTGTCCGG

GCACGCAGCCCCTAGCGGCGCGTCGCTGCCAAGCCGGCCTCCGCGCG

CCTCCCTCCTTCCTTCTCCCCTGGCTGTTCGCGATCCAGCTTGGGTAGG

CGGGGAAGCAGCTGGAGTGCGACCGCCGCGGCAGCCACCCTGC

AACCGCCAGTCGGAGGTGCAGTCCGTAGGCCCTGGCCCCCGGGTGGGC

CCTTGGGGAGTCGGCGCCGCTCCCGGGGAGCTGCAAGGCTCGCCCCTG

CCCGGCGTGGAGGGCGCGGGGGGCGCGGAGAAAGTGAAGAGAGGAA

ATTGGAAAATTGTGAGTGGACCTTCTGATACTGCTCCTCCTTGCGTGGAA

AAGGGGAAAGAACTGCATGCATATTATTCAGCGTCCTATATTCAAAGGATA

TTCTTGGTGATCTTGGAAGTGTCCGTATCATGGAATCAATCTCTATGATG

GGAAGCCCTAAGAGCCTTAGTGAAACTTTTTTACCTAATGGCATAAATGGT

ATCAAAGATGCAAGGAAGGTCACTGTAGGTGTGATTGGAAGTGGAGA

TTTTGCCAAATCCTTGACCATTCGACTTATTAGATGCGGCTATCATGTGGTC

ATAGGAAGTAGAAATCCTAAGTTTGCTTCTGAATTTTTTCCTCATGTGGTAGA

TGTCACTCATCATGAAGATGCTCTCACAAAAACAAATATAATATTTGTTGCT

ATACACAGAGAACATTATACCTCCCTGTGGGACCTGAGACATCTG

CTTGTGGGTAAAATCCTGATTGATGTGAGCAATAACATGAGGATAAA

CCAGTACCCAGAATCCAATGCTGAATATTTGGCTTCATTATTCCCA

GATTCTTTGATTGTCAAAGGATTTAATGTTGTCTCAGCTTGGGCAC

TTCAGTTAGGACCTAAGGATGCCAGCCGGCAGGTTTATATATGCAGCA

ACAATATTCAAGCGCGACAACAGGTTATTGAACTTGCCCGCCAGTT

GAATTTCATTCCCATTGACTTGGGATCCTTATCATCAGCCAGAGAGA

TTGAAAATTTACCCCTACGACTCTTTACTCTCTGGAGAGGGCCAGTGGT

GGTAGCTATAAGCTTGGCCACATTTTTTTTCCTTTATTCCTTTGTCA

GAGATGTGATTCATCCATATGCTAGAAACCAACAGAGTGACTTTTAC

AAAATTCCTATAGAGATTGTGAATAAAACCTTACCTATAGTTGCCATTA

CTTTGCTCTCCCTAGTATACCTCGCAGGTCTTCTGGCAGCTGCTTATC

AACTTTATTACGGCACCAAGTATAGGAGATTTCCACCTTGGTTGGA

AACCTGGTTACAGTGTAGAAAACAGCTTGGATTACTAAGTTTTTTC

TTCGCTATGGTCCATGTTGCCTACAGCCTCTGCTTACCGATGAGA

AGGTCAGAGAGATATTTGTTTCTCAACATGGCTTATCAGCAGGTTCA

TGCAAATATTGAAAACTCTTGGAATGAGGAAGAAGTTTGGAGAATT

GAAATGTATATCTCCTTTGGCATAATGAGCCTTGGCTTACTTTCCCTCC

TGGCAGTCACTTCTATCCCTTCAGTGAGCAATGCTTTAAACTGGA

GAGAATTCAGTTTTATTCAGTCTACACTTGGATATGTCGCTCTGC

TCATAAGTACTTTCCATGTTTTAATTTATGGATGGAAACGAGCTTT

TGAGGAAGAGTACTACAGATTTTATACACCACCAAACT

TTGTTCTTGCTCTTGTTTTGCCCTCAATTGTAATTCTGGGTAAGATTAT

TTTATTCCTTCCATGTATAAGCCGAAAGCTAAAACGAATTAAAAAAG

GCTGGGAAAAGAGCCAATTTCTGGAAGAAGGTATGGGAGGA

ACAATTCCTCATGTCTCCCCGGAGAGGGTCACAGTAATGTGAT

GACAAATGGTGTTCACAGCTGCCATATAAAGTTCTACTCATGCCA

TTATTTTTATGACTTCTACGTTCAG

TTACAAGTATGCTGTCAAATTATCGTGGGTTGAAACTTGTTAAATGAGAT

TTCAACTGACTTAGTGATAGAGTTTTCTTCAAGTTAATTTTCACAAATGT

CATGTTTGCCAATATGAATTTTTCTAGTCAACATATTATTGTAATTTAGG

TATGTTTTGTTTTGTTTTGCACAACTGTAACCCTGTTGTTACTTTATATT

TCATAATCAGGCAAAAATACTTACAGTTAATAATATAGATATAATGTTAA

AAACAATTTGCAAACCAGCAGAATTTTAAGCTTTTAAAATAATTCAATGG

ATATACATTTTTTTCTGAAGATTAAGATTTTAATTATTCAACTTAAAAAG

TAGAAATGCATTATTATACATTTTTTTAAGAAAGGACACGTTATGTTAGC

ATCTAGGTAAGGCTGCATGATAGCATTCCTATATTTCTCTCATAAAATAG

GATTTGAAGGATGAAATTAATTGTATGAAGCAATGTGATTATATGAAGAG

ACACAAATTAAAAAGACAAATTAAACCTGAAATTATATTTAAAATATATT

TGAGACATGAAATACATACTGATAATACATACCTCATGAAAGATTTTATT

CTTTATTGTGTTACAGAGCAGTTTCATTTTCATATTAATATACTGATCAG

GAAGAGGATTCAGTAACATTTGGCTTCCAAAACTGCTATCTCTAATACGG

TACCAATCCTAGGAACTGTATACTAGTTCCTACTTAGAACAAAAGTATCA

AGTTTGCACACAAGTAATCTGCCAGCTGACCTTTGTCGCACCTTAACCAG

TCACCACTTGCTATGGTATAGGATTATACTGATGTTCTTTGAGGGATTCT

GATGTGCTAGGCATGGTTCTAAGTACTTTACTTGTATTATCCCATTTAAT

ACTTAGAACAACCCCGTGAGATAAGTAGTTATTATCCTCATTTTACACAT

GAGGGACCGAAGGATAGAAAAGTTATTTTTCAAAGGTCTTGCAGTTAATA

AATGGCAGAGTGAGCATTCAAGTCCAGGTAGTCATATTCCAGAGGCCACG

GTTTTAACCACTAGGCTCTAGAGCTCCCGCCGCGCCCCTATGCATTATGT

TCACAATGCCAATCTAGATGCTTCCTCTTTTGTATAAAGTCACTGACATT

CTTTAGAGTGGGTTGGGTGCATCCAAAAATGTATAAAAATATTATTATAA

TAAACTTATTACTGCTTGTAGGGTAATTCACAGTTACTTACCCTATTCTT

GCTTGGAACATGAGCCTGGAGACCCATGGCAGTCCATATGCCTCCCTATG

CAGTGAAGGGCCCTAGCAGTGTTAACAAATTGCTGAGATCCCACGGAGTC

TTTCAAAAATCTCTGTAGAGTTAGTCTTCTCCTTTTCTCTTCCTGAGAAG

TTCTCCTGCCTGCATAACCATTCATTAGGGAGTACTTTACAAGCATGAAG

GATATTAGGGTAAGTGGCTAATTATAAATCTACTCTAGAGACATATAATC

ATACAGATTATTCATAAAATTTTTCAGTGCTGTCCTTCCACATTTAATTG

CATTTTGCTCAAACTGTAGAATGCCCTACATTCCCCCCACCCCAATTTGC

TATTTCCTTATTAAAATAGAAAATTATAGGCAAGATACAATTATATGCGT

TCCTCTTCCTGAAATTATAACATTTCTAAACTTACCCACGTAGGTACTAC

TGAATCCAACTGCCAACAATAAAAAGACTTTTATTTAGTAGAGGCTACCT

TTCCCACCAGTGACTCTTTTTCTACAACTGCCTTGTCAGTTTGGTAATTC

ACTTATGATTTTCTAATGTTCTCTTGGTGAATTTTATTATCTTGTACCCT

CTTTTTTTTTTTTTTTTTTTTTAAAGACAGAGTCTTGCTCTGTCACCCAG

GCTGGAGTGCAGTGGCACGATCTCGGCTCACTGCAAGCTCTGCCTCCCGG

GTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGT

GCCCGCCACCATGCCCGGCTGATTTCTTTTTGTATTTTTAGTAGAGACGG

AGTTTCACCGTGTTAGCCAGGATGGTCTCGATCTCCTGACCTCGTGATCC

GCCCGCCTTGGCCTCCAAAGTGCTGGGATTACAGGTGTGAGCTACCGCGC

CCGGCCTATTATCTTGTACTTTCTAACTGAGCCCTCTATTTTCTTTATTT

TAATAATATTTCTCCCCACTTGAGAATCACTTGTTAGTTCTTGGTAGGAA

TTCAGTTGGGCAATGATAACTTTTATGGGCAAAAACATTCTATTATAGTG

AACTAATGAAAATAACAGCGTATTTTCAATATTTTCTTATTCCTTAAATT

CCACTCTTTTAACACTATGCTTAACCACTTAATGTGATGAAATATTCCTA

AAAGTTAAATGACTATTAAAGCATATATTGTTGCATGTATATATTAAGTA

GCCGATACTCTAAATAAAAATACCACTGTTACAGATAAATGGGGCCTTTA

AAAATATGAAAAACAAACTTGTGAAAATGTATAAAAGATGCATCTGTTGT

TTCAAATGGCACTATCTTCTTTTCAGTACTACAAAAACAGAATAATTTTG

AAGTTTTAGAATAAATGTAATATATTTACTATAATTCTAAATGTTTAAAT

GCTTTTCTAAAAATGCAAAACTATGATGTTTAGTTGCTTTATTTTACCTC

TATGTGATTATTTTTCTTAATTGTTATTTTTTATAATCATTATTTTTCTG

AACCATTCTTCTGGCCTCAGAAGTAGGACTGAATTCTACTATTGCTAGGT

GTGAGAAAGTGGTGGTGAGAACCTTAGAGCAGTGGAGATTTGCTACCTGG

TCTGTGTTTTGAGAAGTGCCCCTTAGAAAGTTAAAAGAATGTAGAAAAGA

TACTCAGTCTTAATCCTATGCAAAAAAAAAAATCAAGTAATTGTTTTCCT

ATGAGGAAAATAACCATGAGCTGTATCATGCTACTTAGCTTTTATGTAAA

TATTTCTTATGTCTCCTCTATTAAGAGTATTTAAAATCATATTTAAATAT

GAATCTATTCATGCTAACATTATTTTTCAAAACATACATGGAAATTTAGC

CCAGATTGTCTACATATAAGGTTTTTATTTGAATTGTAAAATATTTAAAA

GTATGAATAAAATATATTTATAGGTATTTATCAGAGATGATTATTTTGTG

CTACATACAGGTTGGCTAATGAGCTCTAGTGTTAAACTACCTGATTAATT

TCTTATAAAGCAGCATAACCTTGGCTTGATTAAGGAATTCTACTTTCAAA

AATTAATCTGATAATAGTAACAAGGTATATTATACTTTCATTACAATCAA

ATTATAGAAATTACTTGTGTAAAAGGGCTTCAAGAATATATCCAATTTTT

AAATATTTTAATATATCTCCTATCTGATAACTTAATTCTTCTAAATTACC

ACTTGCCATTAAGCTATTTCATAATAAATTCTGTACAGTTTCCCCCCAAA

AAAGAGATTTATTTATGAAATATTTAAAGTTTCTAATGTGGTATTTTAAA

TAAAGTATCATAAATGTAATAAGTAAATATTTATTTAGGAATACTGTGAA

CACTGAACTAATTATTCCTGTGTCAGTCTATGAAATCCCTGTTTTGAAAT

ACGTAAACAGCCTAAAATGTGTTGAAATTATTTTGTAAATCCATGACTTA

AAACAAGATACATACATAGTATAACACACCTCACAGTGTTAAGATTTATA

TTGTGAAATGAGACACCCTACCTTCAATTGTTCATCAGTGGGTAAAACAA

ATTCTGATGTACATTCAGGACAAATGATTAGCCCTAAATGAAACTGTAAT

AATTTCAGTGGAAACTCAATCTGTTTTTACCTTTAAACAGTGAATTTTAC

ATGAATGAATGGGTTCTTCACTTTTTTTTTTAGTATGAGAAAATTATACAG

TGCTTAATTTTCAGAGATTCTTTCCATATGTTACTAAAAAATGTTTTGTT

CAGCCTAACATACTGAGTTTTTTTTAACTTTCTAAATTATTGAATTTCCA

TCATGCATTCATCCAAAATTAAGGCAGACTGTTTGGATTCTTCCAGTGGC

CAGATGAGCTAAATTAAATCACAAAAGCAGATGCTTTTGTATGATCTCCA

AATTGCCAACTTTAAGGAAATATTCTCTTGAAATTGTCTTTAAAGATCTT

TTGCAGCTTTGCAGATACCCAGACTGAGCTGGAACTGGAATTTGTCTTCC

TATTGACTCTACTTCTTTAAAAGCGGCTGCCCATTACATTCCTCAGCTGT

CCTTGCAGTTAGGTGTACATGTGACTGAGTGTTGGCCAGTGAGATGAAGT

CTCCTCAAAGGAAGGCAGCATGTGTCCTTTTTCATCCCTTCATCTTGCTG

CTGGGATTGTGGATATAACAGGAGCCCTGGCAGCTGTCTCCAGAGGATCA

AAGCCACACCCAAAGAGTAAGGCAGATTAGAGACCAGAAAGACCTTGACT

ACTTCCCTACTTCCACTGCTTTTTCCTGCATTTAAGCCATTGTAAATCTG

GGTGTGTTACATGAAGTGAAAATTAATTCTTTCTGCCCTTCAGTTCTTTA

TCCTGATACCATTTAACACTGTCTGAATTAACTAGACTGCAATAATTCTT

TCTTTTGAAAGCTTTTAAAGGATAATGTGCAATTCACATTAAAATTGATT

TTCCATTGTCAATTAGTTATACTCATTTTCCTGCCTTGATCTTTCATTAG

ATATTTTGTATCTGCTTGGAATATATTATCTTCTTTTTAACTGTGTAATT

GGTAATTACTAAAACTCTGTAATCTCCAAAATATTGCTATCAAATTACAC

ACCATGTTTTCTATCATTCTCATAGATCTGCCTTATAAACATTTAAATAA

AAAGTACTATTTAATGATTTAACTTCTGTTTTGAAATGTTGTATACACGT

GGATTTTTTTCTCATTAAATAATAATTCTAGTA

APPENDIX B

Supplementary Results

B.1 Protein Isoforms Selected by mRMR Feature Selection in All Pairwise Stages

Table B.1.1: Details about selected protein isoforms from Long's data set across (T1c-T2) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_000110	1	Homo sapiens dihydropyrimidine dehydrogenase transcript variant 1	DPYD
NM_000710	14	Homo sapiens Bradykinin receptor B1	BDKRB1
NM_001042574	15	Homo sapiens CREB regulated transcription coactivator 3 transcript variant 2	CRTC3

Table B.1.2: Details about selected protein isoforms from Long's data set across (T2-T2a) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_001146192	1	Homo sapiens syntaxin 4, transcript variant 1	ZMYND12
NM_001272095	16	Homo sapiens Bradykinin receptor B1	STX4
NM_003540	6	Homo sapiens histone cluster 1 H4 family member f	HIST1H4F
NM_004860	17	Homo sapiens FMR1 autosomal homolog 2	FXR2
NM_032850	15	Homo sapiens zinc finger FYVE-type containing 19, transcript variant 3	ZFYVE19
NM_145892	16	Homo sapiens RNA binding protein, fox-1 homolog 1, transcript variant 2	RBFOX1
NM_153274	1	Homo sapiens RNA binding protein, fox-1 homolog 1, transcript variant 2	BEST4
NM_172350	1	Homo sapiens CD46 molecule, transcript variant n	CD46

Table B.1.3: Details about selected protein isoforms from Long's data set across (T2a-T2b) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_000681	10	Homo sapiens adrenoceptor alpha 2A	ADRA2A
NM_001134473	16	Homo sapiens Gse1 coiled-coil protein, transcript variant 2	GSE1
NM_001145138	11	Homo sapiens RELA proto-oncogene, NF-kB subunit, transcript variant 2	RELA
NM_007225	17	Homo sapiens neurexophilin 3	NXPH3
NM_032023	10	Homo sapiens Ras association domain family member 4	RASSF4

Table B.1.4: Details about selected protein isoforms from Long's data set across (T2b-T2c) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_000029	1	Homo sapiens angiotensinogen	AGT
NM_001711	X	Homo sapiens Biglycan	BGN
NM_032023	10	Homo sapiens Ras association domain family member 4	RASSF4

Table B.1.5: Details about selected protein isoforms from Long's data set across (T2c-T3a) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_001198979	1	Homo sapiens small ArfGAP2, transcript variant 2	SMAP2
NM_003546	6	Homo sapiens histone cluster 1 H4 family member l	HIST1H4L
NM_032875	17	Homo sapiens F-box and leucine rich repeat protein 20, transcript variant 1	FBXL20

Table B.1.6: Details about selected protein isoforms from Long's data set across (T2c-T3a) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_002154	5	Homo sapiens heat shock protein family A (Hsp70) member 4	HSPA4
NM_006465	15	Homo sapiens AT-rich interaction domain 3B , transcript variant 2	ARID3B
NM_024604	12	Homo sapiens RNA polymerase II associated protein 3, transcript variant 1	RPAP3
NM_182485	4	Homo sapiens cytoplasmic polyadenylation element binding protein 2, transcript variant B	CPEB2

Table B.1.7: Details about selected protein isoforms from Long's data set across (T2c-T34) pair-wise stage.

Transcript ID	Chr	Description	Gene
NM_001024674	14	lin-52 DREAM MuvB core complex component	ADRA2A
NM_001023567	15	golgin A8 family, member B, transcript variant 1	GOLGA8B
NM_001099285	2	Prothymosin, alpha, transcript variant 1	PTMA
NM_001198979	1	small ArfGAP2, transcript variant 2	SMAP2
NM_001256	17	cell division cycle 27, transcript variant 2	CDC27
NM_002437	2	MPV17, mitochondrial inner membrane protein	MPV17
NM_003387	2	WAS/WASL interacting protein family member 1, transcript variant 1	WIPF1
NM_001257413	17	IKAROS family zinc finger 3, transcript variant 12	IKZF3
NM_006214	10	phytanoyl-CoA 2-hydroxylase , transcript variant 1	PHYH
NM_016940	21	RWD domain containing 2B , transcript variant 1	RWDD2B

VITA AUCTORIS

NAME:	Manal Ahmad Alshehri
PLACE OF BIRTH:	Dhahran, Saudi Arabia
EDUCATION:	University of Dammam, B.Sc., Computer Science, Dammam, Saudi Arabia, 2007
	University of Windsor, M.Sc in Computer Science, Windsor, Ontario, 2016