THESIS

MACHINE LEARNING MODELS TOWARDS ELUCIDATING THE PLANT INTRON RETENTION CODE

Submitted by

Swapnil Sneham

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2017

Master's Committee:

Advisor: Asa Ben-Hur

Hamidreza Chitsaz
Christopher Peterson

ABSTRACT

MACHINE LEARNING MODELS TOWARDS ELUCIDATING THE PLANT INTRON RETENTION CODE

Alternative Splicing is a process that allows a single gene to encode multiple proteins. Intron Retention (IR) is a type of alternative splicing which is mainly prevalent in plants, but has been shown to regulate gene expression in various organisms and is often involved in rare human diseases. Despite its important role, not much research has been done to understand IR. The motivation behind this research work is to better understand IR and how it is regulated by various biological factors. We designed a combination of 137 features, forming an "intron retention code", to reveal the factors that contribute to IR. Using random forest and support vector machine classifiers, we show the usefulness of these features for the task of predicting whether an intron is subject to IR or not. An analysis of the top-ranking features for this task reveals a high level of similarity of the most predictive features across the three plant species, demonstrating the conservation of the factors that determine IR. We also found a high level of similarity to the top features contributing to IR in mammals. The task of predicting the response to drought stress proved more difficult, with lower levels of accuracy and lower levels of similarity across species, suggesting that additional features need to be considered for predicting condition-specific IR.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor Dr. Asa Ben-Hur for his constant and long-enduring support. I have learned a great deal from him through his guidance and supervision. His immense knowledge in the field of Bioinformatics has been of great help during this thesis work.

I want to extend my special thanks to my thesis committee members, Dr. Hamidreza Chitsaz and Dr. Christopher Peterson, for providing me with their insightful comments and encouragement throughout my research work. I would also like to express my sincere thanks to Mr. Michael Hamilton, Ms. Gareth Halladay, Mr. Basir Shariat and all the members of my research lab for their collaboration and continual assistance.

All my friends at CSU have been an indispensable part of my academic life as well as my personal life. I heartily thank them for all their ever-lasting love and care. Last but not the least, I would like to thank my family for having faith in me and supporting my higher studies despite all the odds. I would not have become who I am today without their love and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Understanding life, its underlying biological processes, and how they are regulated is a very complex problem. Although the scientific community has been doing extensive research to understand these processes, it is still far from being completely understood. Integration of biology with physics, chemistry and computer science has helped to solve many of its puzzles, from the discovery of the cell to understanding the structure of DNA (Deoxyribonucleic acid). Computer science, in particular, has played a big role in deciphering complex biological information using systematic computational analysis [4]. Its application in biology ranges from integrating interesting biological data to transforming it into knowledge and information in order to provide mechanistic explanation of biological phenomena.

As we know, DNA is the molecule made up of four bases, Adenine(A), Guanine(G), Cytosine(C) and Thymine(T) and the sequences of these bases encode the instructions for the functioning of all organisms. Some parts of DNA act as switches to turn genes 'on' and 'off'. Some parts of DNA are protein encoding, serving as templates for building proteins. There are also certain parts which have no function or we don't understand their function yet. One of the most intriguing questions in biology is how these genomic instructions are translated into a real and live organism [5]. Biologists didn't have access to a complete version of the genomic code for a long time and had to rely on studying few of these letters at a time. Advancement in experimental and computational technology has now not only made it possible to have faster and cheaper access to the entire genomic sequence but also has helped in the public access of these huge and valuable data. Different machine learning algorithms [6] have been used to analyze these high-dimensional genetic data that has contributed in different areas including cancer research [7] and gene prediction [8]. Thus, understanding how information is encoded and finding patterns in genomic data not only helps from the biological perspective but also has a huge application in designing drugs, optimizing crop productivity and management as well.

1.1 Outline

In Chapter 1, we provide a brief introduction to the area of our research. We discuss the problem that we are trying to solve and the objective of this thesis work.

In Chapter 2, we describe some of the related work done in this area. We give a brief introduction of the methods they used and some of the results they obtained.

In Chapter 3, we describe in detail the steps we used to prepare the data we used for my machine learning models. We also present the details of the features we used.

In Chapter 4, we present the details of the models we used for the task of predicting intron retention. We also present the results we obtained for this task and will draw some conclusions based on the results we obtained

In Chapter 5, we will describe the details of the models we used for the task of predicting differential intron retention. We will also discuss the results we obtained from our models and the significance to the obtained results.

In Chapter 6, we will summarize our work. Based on the results we obtained we will draw some conclusions regarding the predicting intron retention and condition specific intron retention. We will also discuss how this work can be improved and continued in the future.

In the following sections, we briefly describe some of the biological concepts to help outline the problem we are trying to solve and the objective of this study which is presented at the end of this chapter.

1.2 Nucleic acids

Nucleic acids are very large molecules that are responsible for storing and transferring the genetic information that governs the biological processes in all living organisms. They are of two types: DNA which is Deoxyribonucleic acid and RNA which is Ribonucleic acid.

1.2.1 DNA

In eukaryotic cells (cells having a nucleus), DNA is present inside the nucleus, whereas in prokaryotic cells (cells without a nucleus), DNA is found in the cytoplasm. DNA is composed of nucleotides that are made up of three components : a 5-carbon sugar (numbered 1' through 5') called the deoxyribose sugar, a phosphate group, and a nitrogenous base. As discussed earlier, DNA has four types of nucleotides : Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). The sequence of these bases encodes information that is used later in the protein synthesis. These bases are of two different sizes. Adenine and Guanine are bigger, since they have two carbon rings in their molecular structure, whereas Cytosine and Thymine are smaller with only one carbon ring in their molecular structure. Thymine pairs with Adenine and Cytosine pairs with Guanine forming the two strands of the DNA [9]. The direction of DNA chain formation proceeds from the 5' to the 3' direction as shown in Figure 1.1.

1.2.2 RNA

The chemical composition of RNA is very similar to that of DNA. Like DNA, RNA is also made up of nucleotides but the sugar in RNA is the ribose sugar. The bases of RNA differ slightly from DNA with the presence of Uracil (U) in place of Thymine (T). In contrast to DNA, RNA is single stranded and it comes in variety of types. We are primarily interested in messenger RNA (mRNA) which is responsible for carrying information from DNA to ribosome for protein synthesis. RNA can exist in complex secondary and tertiary structures. Secondary structures are the two-dimensional base-pair folding in which the self-complementary regions of local sequence forms base pairs. Some common secondary structures are hairpin loops, bulges and internal loop. These secondary structure elements get associated with different Van der Waals force or specific hydrogen bonds to form the tertiary structure of RNA [10].

1.3 The central dogma of molecular biology

The central dogma of molecular biology describes the flow of genetic information within a biological system. This information transfer comprises of basically three steps: DNA replication, Transcription and Translation as shown in Figure 1.1. Whenever the cell divides, the double stranded DNA splits into two single strands. These separated single strands forms the basis for a new strand of complementary DNA and this is how each cell maintains its own copy of complete genome. This is called the DNA replication. During this process, the incoming deoxynucleotide triphosphates form pairs with the template bases as described by Watson et al. [9], and is regulated by the DNA polymerase enzymes.

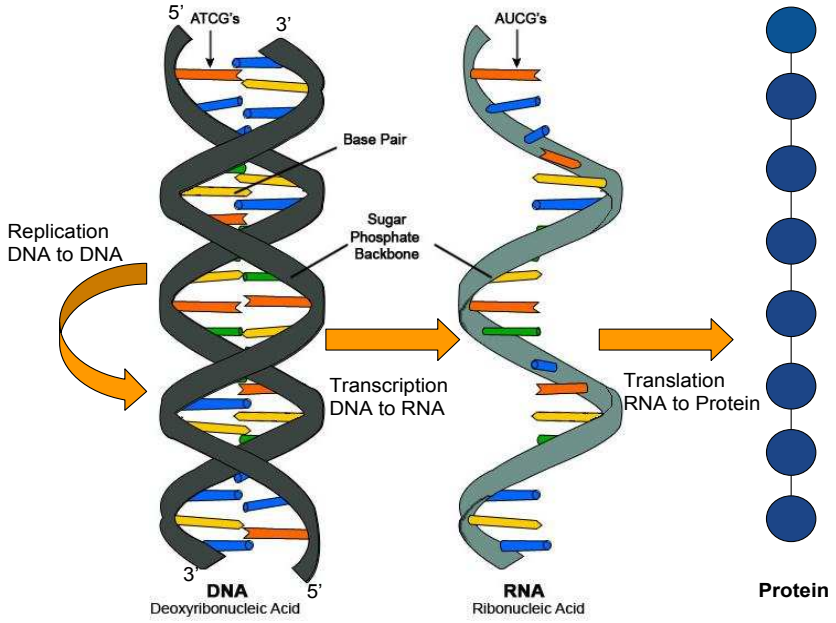


Figure 1.1: Central Dogma of molecular biology showing three steps of genetic information transfer: Replication, Transcription and Translation, adapted from [11]

Transcription is the process by which mRNA is transcribed or copied from the DNA. In the process, first the double helix of DNA is unwound and Ribonucleotide triphosphates (NTPs) form base pairs with the complementary strand of DNA also known as the anti-sense strand. The strand of DNA containing the gene is called the sense strand. These ribonucleotides are

then joined together by an enzyme called RNA polymerase to form the pre-messenger RNA. Hence, the pre-messenger RNA is complementary to the anti-sense strand and is a copy of sense strand. Pre-messenger RNA contains certain sections of sequences that do not take part in protein synthesis called the introns. These introns are removed to form the mature mRNA, and this process in called *RNA splicing*. It is discussed in more detail in Section 1.4.

Translation is the process by which mRNA takes part in protein synthesis with the help of transfer RNA (tRNA). For this, the mature RNA formed during transcription is transported to the ribosome. In eukaryotic cell, however, this process involves transporting mRNA out of the nucleus due to the existence of a nuclear membrane. The triplet of adjacent bases of mRNA is called a codon and each codon is responsible for forming a specific amino acid. The ribosomes reads codons beginning from the start codon which is usually 'AUG' and each codon then forms base pairs with the anticodon of a particular tRNA forming base pairs. Each tRNA has specific amino acid residue which gets incorporated into the protein being synthesized. This process ends when ribosome reads the stop codon which is generally 'UAA', 'UGA' or 'UAG'.

1.4 RNA Splicing

Splicing is the process by which pre-messenger RNA gets edited to form mature messenger RNA. It occurs either concurrently with the transcription process or after the transcription process. During this process, introns get removed from the pre-mRNA and exons are attached together to form mature RNA that can be translated into proteins [12].

1.4.1 The process of Splicing

In a typical eukaryotic intron, the process of splicing involves the 5' end of the intron called the donor site, the 3' end of the intron called the acceptor site and the branch site near the 3' end of the intron. These different parts of introns are shown in Figure 1.2. The intron generally has a 'GU' at its 5' end and a 'AG' at its 3' end. The branch point is always an 'A' which is involved in lariat formation, however the consensus sequence around the branch

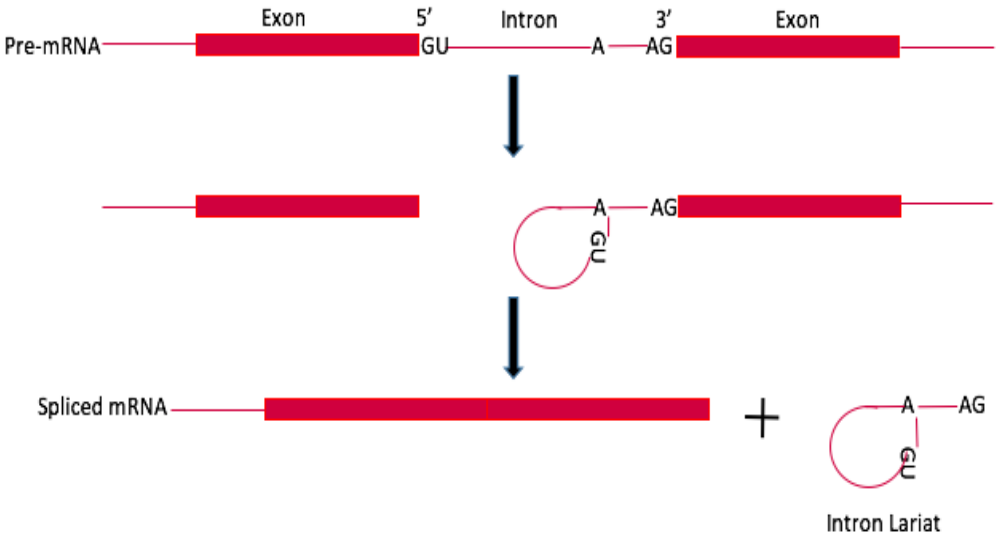


Figure 1.2: The process of RNA Splicing where exons are attached together to form mature mRNA.

point may vary sometimes. Human branch point consensus sequence is yUnAy [13] The upstream intron (5' ward) is rich in pyrimidines ('C' and 'U') and is called the polypyrimidine tract [14]. Figure 1.3 shows the sequence logos of human splice-sites motifs. The term motifs in genetics refers to the widespread pattern of nucleotides that are supposed to have some biological significance. Similarly, it has also been studied that number of nucleotides between the branch point and the 3' acceptor site has an affect on the selection of the splice site [15].

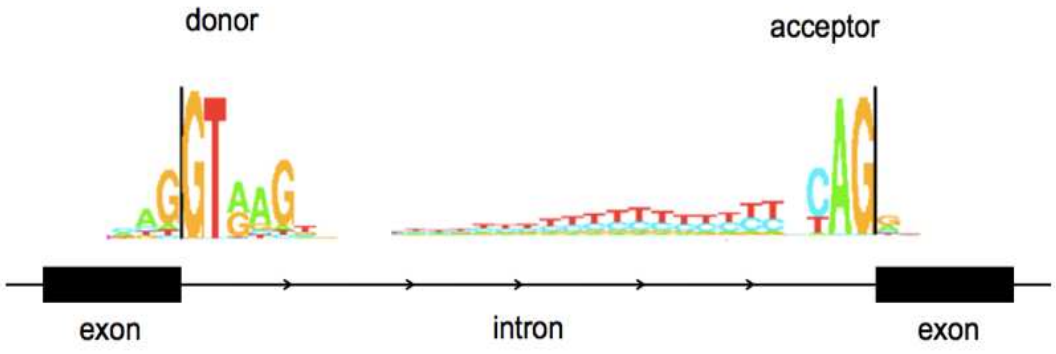


Figure 1.3: Sequence logos of human splice-sites from Stephens et al [1].

The spliceosome is a large protein complex that is involved in the splicing of pre-mRNA. It consists of five small nuclear ribonucleoproteins (snRNPs) U1, U2, U4, U5 and U6 [16]. During the process of splicing, U1 binds to the 5' GU and with the help of the U2AF protein factor and U2 snRNP binds to the branch point 'A' in the intron. The complex thus formed is called the spliceosome A complex. In the next stage, the U4,U5,U6 complex binds together and U6 replaces U1. Then the U4 and U1 snRNPs leave. At this stage, two transesterification reactions occur. In the first reaction, 5' end of intron is detached from the 5' flanking exon and gets attached to the branch point. In the second reaction, the 3' end of intron is detached from the 3' flanking exon and then two exons join together releasing the intron lariat [17].

1.4.2 Splicing Regulatory Elements

Splicing is regulated by the regions in the introns/exons that defines introns/exons to the splicing machinery , called the *cis-acting regulatory sites*, and the corresponding proteins that bind to the cis-acting sites to control the transcription, called the *trans-acting proteins*. There are two types of cis-acting regulatory sites: splicing silencers and splicing enhancers and two types of trans-acting proteins: repressors and activators. Splicing activator proteins bind to the splicing enhancer sites, increasing the probability of the nearby site to be a part of splice junction (intron-exon junction where splicing takes place). In contrast to that, splicing repressor proteins bind to the splicing silencer sites reducing the probability of the nearby site to be a part of splice junction. Both silencer sites and enhancer sites can be present in either exonic regions or intronic regions. Based on their location, they are grouped into four categories: Exonic Splicing Silencer (ESS), Exonic Splicing Enhancers (ESE), Intronic Splicing Silencer (ISS) and Intronic Splicing Enhancers (ISE) [16]. Moreover, recent studies [18] have also shown that splicing is context dependent where certain cellular or environmental condition may govern how splicing occurs, discussed in more detail in Section 1.6.

1.5 Alternative Splicing

Alternative splicing (AS) is a process by which pre-mRNA are spliced in multiple ways to produce multiple mRNA and protein isoforms. Thus, alternative splicing allows a single gene to encode multiple proteins. The protein that is translated from the alternatively spliced mRNA has a different amino acids sequence and thus different function as compared to the protein translated from constitutively spliced mRNA, leading to the diversity of protein family. It has been found that transcripts from approximately 95% of multi-exon human genes are spliced in more than one way and the resulting transcripts in most cases are expressed differentially in different conditions and different tissue types [18].

1.5.1 Forms of Alternative Splicing

There are five basic forms of alternative splicing [19]:

1. **Exon skipping** - This is the form AS in which an exon may be spliced out or included in the resulting transcript as shown in Figure 1.4. Exon skipping is the most prevalent form of AS in animals where it accounts for around 40% of AS events [19], but is a very rare event in plants .

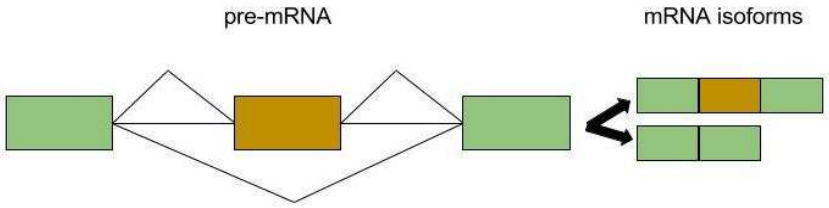


Figure 1.4: Exon skipping

2. **Mutually exclusive exons** - In this form of AS, only one of two consecutive exons is included in the mature mRNA as shown in Figure 1.5. This is less frequent form of AS which generally results from complex events.

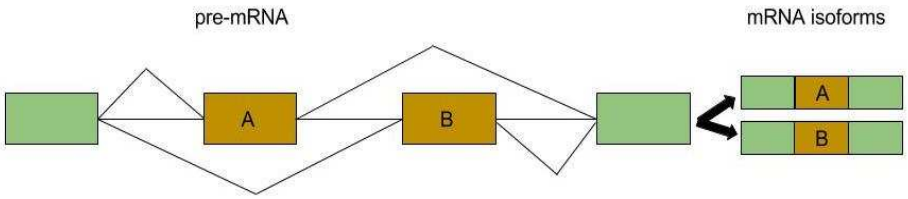


Figure 1.5: Mutually exclusive exon

3. **Alternative donor site** - In this form of AS, an alternative 5' splice site is selected which changes the 3' boundary of the upstream exon. It constitutes around 7.9% of all AS events in animals [19] and around 3.3% of all AS events in plants [20]. It is also known as Alternative 5' site.

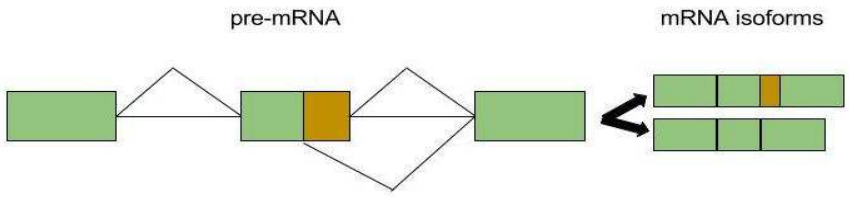


Figure 1.6: Alternative donor site

4. **Alternative acceptor site** - Similarly, here an alternative 3' splice site is selected which changes the 5' boundary of the downstream exon. It constitutes around 18.4% of all AS events in animals [19] and around 6.1% of all AS events in plants [20]. It is also known as Alternative 3' site.

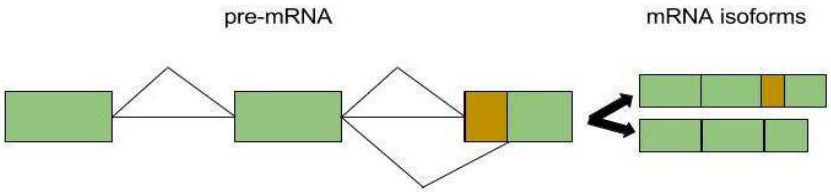


Figure 1.7: Alternative donor site

5. **Intron retention** - This is the fifth form of AS, in which an intron gets retained in the mature mRNA. This is the most prevalent form of AS in plants but is rarest in most vertebrates and invertebrates [19]. It is discussed in more details in the next section.

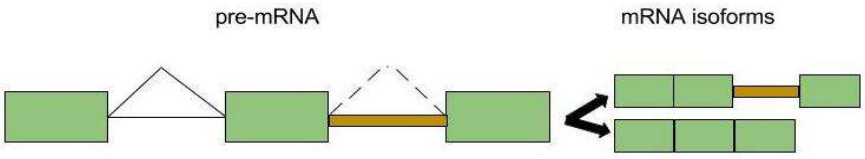


Figure 1.8: Intron retention

1.6 Intron Retention and Differential Intron Retention

Intron Retention (IR) is one of the major forms of AS, in which introns remain as a part of the mature mRNA without being spliced out. IR is the least prevalent form of AS in animals but is the major form of alternative splicing in plants. The difference in the distribution of different forms of AS in plants and animals may indicate the difference in how plants and animals identify introns and exons [21]. In *Arabidopsis thaliana*, the frequency of IR is reported to be as high as 64% of AS events [22–24]. Braunschweig et al. [25] performed a comprehensive study of IR in 40 different human and mouse cell and tissue types and discovered that IR is more frequent in mammals than previously recognized.

As introns become a part of mature mRNA as a result of IR, they directly lead to protein diversity. IR may a introduce premature termination codon in the transcript, which leads to the degradation of mRNA by nonsense-mediated decay (NMD) [26] and it has been found that this process has a regulatory role in controlling gene expression [27]. IR is tissue or condition dependent [25], and the difference in intron retention levels between different experimental conditions or tissue types has been termed as *differential intron retention*. Figure 1.10 illustrates the concept of differential intron retention events as detected by iDiffIR [2]. Certain tissue-specific IR events have also been linked to some rare diseases [28–30]. Eswaran et al. [31] report that a large number of IR events was found in different types of breast cancer tissue as compared to normal breast tissue and similarly, Zhang et al. [32] also report that 2340 IR-affected genes were found in lung carcinomas. Mastrangelo et al. [33] analyzed the splicing process of a cold-regulated gene encoding ribokinase(7H8) protein and suggest that 7H8 cold dependent intron retention is a common characteristic in cereals. Palusa et

al. [34] found that temperature stress (heat and cold) remarkably affect the splicing pattern of several serine/arginine rich (SR) genes in *Arabidopsis*. Since not a whole lot is known on intron retention, particularly in plants, there is currently the need of understanding the mechanism and regulation of IR and differential IR.

1.7 RNA Sequencing

The transcriptome refers to the entire set of mRNA transcripts and their quantity in a particular condition. The study of the transcriptome aims at identifying transcripts, splicing patterns, and also quantifying the changes in the expression levels of each transcript during different developmental stages and conditions. The term gene expression refers to the process by which information in the gene is converted to a functional product, which is usually the protein. The knowledge of the transcriptome helps to interpret the functional element of the genome, and also to understand changes during development and diseases [35].

Different approaches have been developed to analyze and quantify the transcriptome. One of the recently developed method for studying transcriptome is RNA-Sequencing (RNA-Seq) [35, 36]. It utilizes next-generation sequencing platforms like Illumina sequencing [37] to quantify the amount of RNA in a biological sample at a given time. In RNA-Seq, RNAs are converted to a library of complementary DNA (cDNA) fragments with adapters attached to one or both ends. Then they are sequenced using any high throughput sequencing technology, from either one end (single-end sequencing) or both ends (paired-end sequencing). The resulting reads are then aligned to the reference genome or assembled *de novo* without the genomic sequence to produce both transcript structure and/or level of expression of each gene or transcript. The steps used in typical RNA-Seq experiments are shown in Figure 1.9. These experiments are widely used to compare gene expression between different experimental conditions, characterize alternative splicing, look at mutations and to build co-expression networks. These are also used to discover exon-intron boundaries which can be used to verify the gene annotations. If an RNA-Seq experiment is performed between two conditions, then the gene with higher gene expression in the particular condition is called up-regulated and the gene with the lower gene expression is called down-regulated.

Before RNA-Seq methods, hybridization based approaches involving microarrays were used in transcriptomics. Although they have high throughput and are inexpensive, these methods have several limitations. One of the major drawbacks in microarrays is that the sequence had to been known a priori, they are limited in their ability to provide information on transcript structure and abundance, and there were issues with poor quantification of lowly and highly expressed genes [38]. This led to migration of transcriptomics to sequencing based methods.

Since RNA-Seq opens up the possibility to study gene expression at a more detailed level, recent studies are moving towards using RNA-Seq data to understand alternative splicing (particularly exon-skiping) [18]. However, not many studies have used RNA-Seq to understand IR.

Figure 1.9: Steps in RNA-Seq experiments.

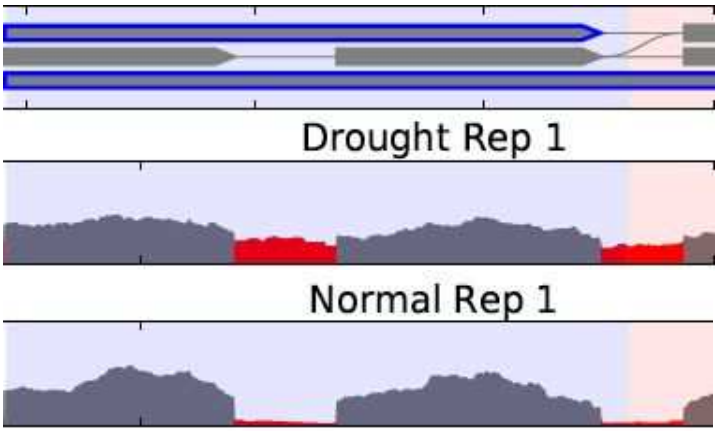


Figure 1.10: Differential Intron Retention as detected by iDiffIR [2]. The IR events in the figure are highlighted in red. The top track in the figure is the gene model and the following two tracks represent the read depths across the gene under drought stress and normal condition, respectively. This figure shows that there is more intron retention under drought stress as compared to the normal condition in this particular gene.

1.8 Problem formulation and objective

It is known that IR is the major form of AS in plants and there is currently an interest in the biology and bioinformatics research communities to understand these events fully. As discussed in Section 1.6, intron retention not only regulates gene expression, but certain tissue specific intron retention events have been linked to some rare diseases as well. So, it is important to understand how these events are being regulated and their implications. However, not a lot is known about it yet [21].

The major objective of this study is to develop accurate machine learning models that can predict predict whether an intron will be retained or not, predict condition specific patterns of intron retention, and provide insight on the factors that drive IR and condition-specific IR and their conservation across different plant species.

Chapter 2

LITERATURE REVIEW

In this chapter, we give a brief review of existing methods for predicting alternative splicing with emphasis on intron retention and exon skipping.

2.1 Predicting exon skipping

Exon skipping is the major form of AS in mammals, and since many studies have been done to predict and understand this process, the methodologies and findings of these studies can be used to understand the mechanism of intron retention as well. Multiple studies have shown that features extracted from exons and the surrounding introns can reliably distinguish alternative and constitutive exons [39, 40], and we begin with reviewing this line of work.

There are different characteristics that distinguish alternatively spliced exons from constitutively spliced exons. Alternatively spliced exons tend to be shorter in length than constitutively spliced exons [41, 42], and their sizes also tend to be a multiple of three. The reasoning is that an exon with a size that is a multiple of three ensures that its skipping does not affect the reading frame of the downstream exons [41]. It has been reported that constitutively spliced exons, however, do not have this property [43]. Furthermore, Sorek et al. [41] studied exon skipping in human and mouse, and found that alternatively spliced exons are flanked by intronic sequences that are more conserved between human and mouse. The authors suggested that higher identity level between human and mouse may be related to the fact that alternatively spliced exons often contain sequences (ESEs and ESSs) that regulate their splicing. Similarly, Clark et al. [44] reported that there is a difference in the nucleotide composition of 5' splice sites between alternative and constitutive exons. The nucleotide composition at the 5' splice site affects the base-pairing of the U1 snRNA and 5' splice site, which is necessary for splicing [45]. This property is also supported by other

research [46] where it is demonstrated that if the 5' splice site sequences are altered, it can result in the transition from alternative to constitutive splicing and vice-versa.

Based on the similar properties of alternatively spliced exons, Sorek et al. [39] presented a set of 228 features, out of which 7 features were drawn from their previous study [41]. These 7 features include exon length, divisibility of exon length by 3, percentage identity to the mouse counterpart, the length of best human/mouse local alignment in the 100 upstream and 100 downstream intronic sequences, identity level in the alignment of this upstream and downstream intron. The rest of the features include tri-nucleotide counts in the exon and the flanking introns, count of each base (A, C, T, G) at the 5' splice sites, and the number of pyrimidines (C and T) in the last 19 bases of the upstream intron. Although, the other features contributed a little to improve the performance of the classifier, the 7 features included in the original study were found to have the highest discriminative power.

Different approaches for including positional information of motifs in DNA sequences for predicting exon skipping have also been studied. Rätsch et al. [40] proposed the weighted degree (WD) kernel for SVM, that is a position-dependent version of the so-called spectrum kernel [47]. One notable contribution about their approach is the successful combination of both the positional information of the motifs and the sequence features like the length of exons and the flanking introns by linearly combining the WD kernels with the linear kernels of these sequence features.

The studies described above do not take into consideration the condition or tissue-dependent changes in alternative splicing. However, recent studies [31, 48, 49] highlight the importance of understanding its mechanism. Xu et al. [49] constructed a list of 46 human tissues with 2 million human transcript sequences to generate a database of human alternative splices and they showed that alternative splicing is indeed tissue dependent. This dataset has been a great resource for understanding the role tissue specific variants in different human diseases.

Barash et al. [18] investigated a "splicing code" that uses a combination of a large number of sequence features to predict tissue-dependent changes in alternative splicing. The input

consisted of exons, their surrounding introns, and the profiling of how these exons are spliced in different tissues. In the tissue profiling, the mRNA expression data was preprocessed so that for each exon there was an estimated percentage inclusion level x_t for each tissue $t \in \{1, ..., T\}$. The percentage inclusion value quantifies the fraction of transcripts that include the exon. The 3,665 cassette-type alternative exons across 27 mouse tissues ranging from whole embryo stage to adult tissues were used. In order to account for variability in different tissues and exons, and to provide the tissue-independent baseline exon inclusion levels, the percentage inclusion value of each exon and tissue type in the dataset was converted into three sets of probabilities q_{inc}, q_{exc} and q_{nc}, where q, represented as a set of three probabilities, is referred to as the "splicing pattern"; *inc* represents increased inclusion levels, *exc* represents increased exclusion levels and *nc* represents no change.

The feature set used by Barash et al. consists of a compendium of 1014 features. These features can be grouped into four types: known motifs, new motifs, short motifs and features describing transcript structure. A total of 171 features derived from *cis*-elements, which have been reported in the previous studies [50–52] to affect alternative splicing or to bind a known splice factor, were used as known motifs. As new motifs, 326 *cis*-elements, that were supported by conservation evidence but have not been linked to tissue-specific context, were used. Similarly, for short motifs, the count of 1-3 nucleotide long k-mers were used, which constituted 460 features. A total of 57 features were used to describe the transcript structure. These features include the internal exon length, the length of the flanking introns, and the ratio of the length of exon with the surrounding introns. In addition to this feature compendium, the SeedSearcher algorithm [53], was used to find nearly 1800 unbiased motifs; unbiased motifs are motifs that are likely to be a part of AS regulation, without biasing the search using previous literature. These motifs are likely splicing silencers and enhancers that regulate splicing. Furthermore, the recursive feature elimination approach was used to select most relevant features. The final assembled code contained nearly 200 features. Among the 200 features, the use of both the feature compendium and the unbiased motifs did not improve the code quality significantly. However, some of the unbiased motifs were selected in the assembled code, that didn't correspond to any feature in the compendium. In

addition to classifying alternatively spliced exons from constitutively spliced exons, they also analyzed how well the splicing code can predict the differences in the percentage inclusion level between pairs of tissues. This study lays a good foundation in studying condition-specific changes in alternative splicing and significance of the splicing code in predicting the direction of changes of splicing in different conditions.

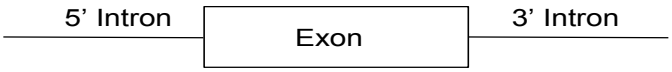


Figure 2.1: An exon with its flanking introns.

2.2 Predicting Intron Retention

As discussed earlier, IR is the least prevalent form of AS in animals which may help explain why it is the least understood form of AS. Indeed, there are only a few studies in the area of distinguishing retained introns from non-retained introns.

The papers [54–56] present extensive studies on the features of retained introns which can be utilized in predictive models. These studies suggest that retained introns are relatively shorter, their sizes tend to be a multiple of three, they are more G/C rich, and have weaker splice sites. According to the intron definition model [57], the splicesome recognizes longer introns for removal and thus shorter introns may remain unspliced in the transcripts. Also, intron length that is a multiple of three ensures that its retention does not affect the reading frame of downstream exons. Furthermore, the studies [54,58] propose that the elevated G/C content in the retained introns may be related to the splice site recognition mechanism. They demonstrate that shorter introns have higher G/C percentage as compared to longer introns. Since retained introns are shorter, they have higher G/C percentage and structurally they are more similar to exons. Thus, they are more likely to be flagged as "exons" by the splicing machinery. This hypothesis is also supported by the results of Goodall et al. [59]. Similarly, introns that are flanked by weaker splice sites are occasionally not properly recognized,

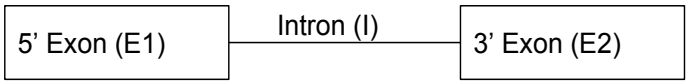


Figure 2.2: An intron with its flanking exons.

leading to their retention [54]. Galante et al. [54] reported that the rate of intron retention is more pronounced in the untranslated regions (UTRs) than in the coding sequences (CDS) and they suggest that the bias may be associated with the process of non-mediated decay (NMD) that triggers the degradation of transcripts that have premature stop codons. Likewise, Wang et al. [60] investigated a set of Exonic Splicing Silencers (ESSs) that affect the retention of introns and proposed that sequences like TAGT, TAGGT, GTARGT (R = A or G) and other G/T rich sequences to have ESS activity. A more recent study done by Cui et al. [61] on *A.thaliana* demonstrated that the presence of a termination codon (TGA and TAA) might be contributor to IR as well.

Some of these features along with some other features have been successfully used to predict IR in human. Braunschweig et al. [25] formed the "IR code", a collection of 136 features, to predict retained introns (RIs) and constitutively spliced introns (CSIs) in human and mouse. In addition to some common features described above, the "IR code" consists of features like the length of flanking exons, relative lengths of intron and flanking exons, their relative log length and the dinucleotide content in different parts of introns and exons. They suggest that although most of these features are highly correlated, together they build a stronger model. They obtained an area under the curve (AUC) of 0.79 with the logistic regression model. Along with confirming that RIs are associated with high G/C percentage, reduced length, and relatively weak 5' and 3' splice sites, their results also identified higher G/C content in the flanking exons, ratios of length of intron and upstream exon and the position of intron within the gene to be important features.

Another key point in their approach is the way they distinguished RIs and CSIs using the Percentage Intron Retention (PIR) metric. PIR is defined as the percentage of the average

number of reads mapping to the 5' and 3' exon-intron junctions over the sum of the average number reads aligning to the exon-intron junction and exon-exon junctions. Mathematically, PIR value for each intron, I, and flanking exons, E1 and E2, is calculated as:

$$PIR = 100 * \frac{average(\#E1I + \#IE2)}{(average(\#E1I + \#IE2) + \#E1E2)}, \tag{2.1}$$

where, #E1I, #E1E2, #IE2 are the normalized read counts for the respective junctions.

There have been a few studies done to build prediction models of intron retention in plants [21, 48, 61]. Mao et al. [48] presented a thorough feature extraction technique to distinguish RIs and CSIs in *A.thaliana* using random forests [62] and support vector machines [63]. Their proposed feature extraction technique is broken down into three components: basic feature extraction, frequent motif extraction, and splice sites and the flanking sequences of intron feature extraction. Some common global features like intron length and G/C content were included in the basic feature extraction; whereas, in frequent feature extraction they used the motifs that were frequent either in retained introns or constitutively spliced introns but not in both are used. One important thing to note in their approach is that they used the label information to find the frequent motifs which indicates that the performance they achieved is over optimistic. As aggg-containing and ggag-containing motifs were found to be more abundant in retained introns, they proposed these motifs as Intronic Splicing Silencers (ISSs). Similarly, AT/TA-rich motifs were more abundant in constitutively spliced introns, so they proposed these motifs as Intronic Splicing Enhancers (ISEs).

In Section 1.6, we discussed the importance of understanding the regulation of differential IR. Interestingly, none of the studies described above take into consideration the tissue or condition-specific changes in IR events in plants. So, in addition to building prediction model for IR, our study also focuses on building reliable prediction model for differential IR with an aim to understand IR at a greater resolution.

Chapter 3

DATA PREPARATION AND FEATURE EXTRACTION

In this chapter, we describe the steps we took to prepare our data and go into the details of the feature we used for the machine learning models.

3.1 Data Preparation

We performed our study on three plant species: *Arabidopsis thaliana*, *Oryza sativa* and *Sorghum bicolor*. *Arabidopsis thaliana* is a widely used model organism to study plant biology. It belongs to the mustard family, has a genome size of nearly 135 Mbp and a chromosome number of 5. *A.thaliana* is well annotated because it is a model organism and thus easier to study compared to other plant species.

Oryza sativa, commonly known as rice, is a model species for monocot plants and many cereals like wheat and maize. It has a genome size of nearly 500 Mbp with 12 chromosomes. It is less annotated than *A. thaliana* plant, but still considered the second best annotated plant species. *Oryza sativa* comes in two varieties: Japonica and Indica. We have used the Japonica variety for our study.

Sorghum bicolor, is another widely grown cereal crop. It has a genome size of nearly 730 Mbp which is much larger than *A. thaliana* and *O. sativa*. It has a chromosome number of 10.

The data preparation pipeline for predicting intron retention and differential intron retention is described in the following sections.

3.1.1 Predicting intron retention

We use the term *reference genome* to refer to the nucleotide sequence of a particular species. Once it is sequenced, the reference genome is annotated with labels that refer to

various functional parts of the genome and the genes within it. The annotations include different parts of genes such as start codons, exons, introns, stop codons and so on. The transcript annotations of a given genome provide the data needed for differentiating retained introns from non-retained introns. From our analysis, shown in Section 5.2, we found that the quality of annotation files used has a large effect on the performance of the models. This is because the more accurate and detailed the annotations are, the better the data is and this makes the model more accurate. Bad quality data introduces unnecessary noise which makes learning much harder. The reference annotation and genome files used for each species are shown in Table 3.1.

Table 3.1: Reference annotation and genome files used for different plant species

Plants	**Annotations**	**Genome**
A.thaliana	TAIR10	TAIR10
O.sativa	MSU v7.0	MSU v7.0
S.bicolor	PacBio based annotation	Sbi v3.0

TAIR10 [64] reference genome and annotation were obtained from TAIR (The Arabidopsis Information Resource) repository for *A. thaliana*. The reference genomes (MSU v7.0 [65] and Sbi v3.0 [66]) for *O.sativa* and *S.bicolor*, respectively were obtained from the Phytozome database, which is the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute. Since *S.bicolor* is not very well annotated, we decided to use better data we already had for *S.bicolor* in which Pacific Biosciences Iso-Seq data was used for creating the annotations [67] .

The annotation files were used to extract the coordinates of the introns and the flanking exons for both retained and non-retained introns. A set of positive examples was created using the sequences of retained introns and their flanking (5' and 3') exons and a set of negative examples was created using the sequences of the non-retained (constitutively spliced) introns and the flanking exons. The number of retained and non-retained examples obtained for the three species is shown in Table 3.2. Table 3.2 shows that as expected, the number of non-retained examples is much higher than retained examples. These examples were randomly shuffled and divided into training and test sets in the ratio of 80% and 20%, respectively.

Table 3.2: Number of retained and non-retained examples.

Species	**Retained**	**Non-retained**
A.thaliana	2,760	132,041
O.sativa	5,001	156,745
S.bicolor	4,817	78,406

3.1.2 Predicting differential intron retention

RNA-Seq analysis [35, 36] utilizes next-generation sequencing platforms like Illumina sequencing [37] to quantify the amount of RNA in a biological sample at a given time. RNA-Seq experiments are widely used to compare gene expression between different experimental conditions, characterize alternative splicing, look at mutations/ SNPs and to build coexpression networks. It is also used to discover exon-intron boundaries which can be used to verify the annotations. As RNA-Seq experiments have the potential for the detection of IR and differential IR events [68], RNA-Seq data was used to study the differences of intron retention in different experimental conditions. It is described in more detail in Section 1.7. The overall pipleline used to obtain data for this task is shown in the Figure 3.1. Each part of the pipeline is described in detail in the following sections and the specific commands used in each step are shown in Appendix A.

Figure 3.1: Data Preparation pipeline

Retrieval of RNA-Seq data

The National Center of Biotechnology Information (NCBI) provides a central repository for sequencing data called the Sequencing Read Archive (SRA) [69]. This database contains sequencing data from different methods (RNA-Seq, CHIP-Seq etc) for multiple organisms which can be downloaded using the SRA toolkit. NCBI accepts data from different kinds of sequencing projects and stores it for public access as SRA studies, each with a unique SRA

study number. These SRA studies may contain biological replicates or technical replicates of the sample experiments which should be downloaded as well for further processing. We made sure that each sample run in the experiment has 1GB or higher number of bases to ensure sufficient read coverage. The details of the data used for this study are shown in Table 3.3. *A.thaliana* and *S.bicolor* were subjected to Abscisic acid (ABA) treatment, which simulates drought stress, whereas *O.sativa* was subject to real drought stress condition.

Table 3.3: Details of the datasets used in this study

Species	**SRA Study**	**Run**	**Treatment**
A. thaliana	SRP056035	SRR1909019	No treatment
		SRR1909020	No treatment
		SRR1909021	No treatment
		SRR1909022	ABA treatment, 6 hours
		SRR1909023	ABA treatment, 6 hours
		SRR1909024	ABA treatment, 6 hours
O.sativa	SRP052306	SRR1761528	Drought
		SRR1761529	Drought
		SRR1761530	No treatment
		SRR1761531	No treatment
S.bicolor	Not yet published	R19	No treatment
		R20	No treatment
		R22	ABA treatment, 6 hours
		R23	ABA treatment, 6 hours

Read alignment

After initial steps of quality control, the reads were then aligned to the reference genome using STAR [70]. STAR is a fast sequence alignment tool for high-throughput RNA-Seq data. The first step for alignment is to generate genome index files using the reference genome file and annotation file. The second step for alignment is to map the reads to the index generated in the first step. It outputs the alignment file, mapping summary statistics and other files. Figure 3.2 shows an example of aligning short reads from RNA-Seq to reference genome. From the mapping summary statistics, the quality of the alignment can be checked. For all the datasets used for this study the percentage of uniquely mapped reads was approximately 90%.

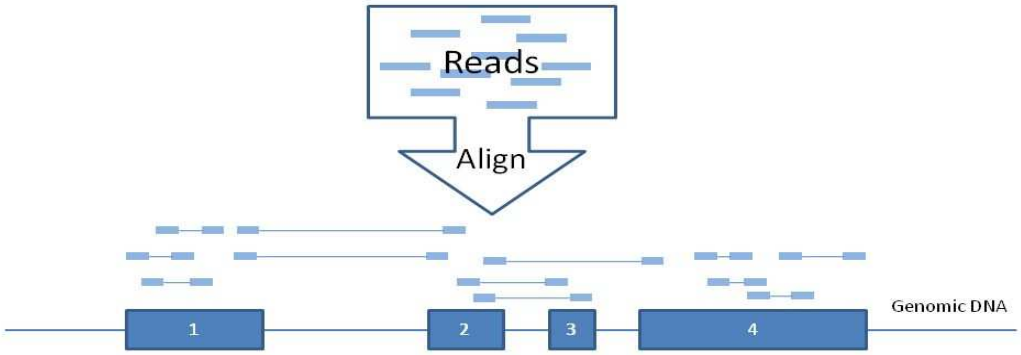


Figure 3.2: Alignment of short reads from RNA-Seq to a reference genome where four exons are shown, numbered 1 through 4. Image adapted from Corney et al. [3].

Alignment filtering

It is a good practice to filter the aligned reads to remove multi-reads (reads that align to more than one genomic location) and reads incorrectly aligned across splice junctions since they may introduce noise in the data and adversely affect the detection of differential IR events. The bash script we used to filter multi-reads is shown in the Appendix A. The filtered reads are then indexed before applying other kinds of filtering which can be done using samtools [71]. To remove potential false-positive splice junctions from unique reads, we used SpliceGrapher's sam_filter.py script [68].

iDiffIR

iDiffIR [2] is a package for detecting differential IR events from high-throughput sequencing libraries. It models the differences in the intronic read coverage across the experimental conditions. Hamilton et al. [2] defines the read coverage of a genomic region I as the mean read depth, which is calculated as

$$\mu_r(I) = \frac{1}{|I|} \sum_{i \in I} r(i), \tag{3.1}$$

where *r(i)* is the number of reads that align at a particular genomic coordinate i. Since there might be differences in the read distributions in exons and introns, the read coverage is

normalized separately for introns and exons. Further, to avoid the detection of differential IR as a result of differential gene expression, the read coverage is normalized across experimental conditions. Differential IR between two conditions is quantified using a log-fold change statistic which is defined as follows:

$$log_{\widehat{FC}}(I) = log_2 \frac{a + \hat{\mu}_r(I_1)}{a + \hat{\mu}_r(I_2)}. \tag{3.2}$$

Here, a is a pseudo-count parameter that controls large fold change values that can occur in low expression regions. $\hat{\mu}_r(I_1)$, $\hat{\mu}_r(I_2)$ are the normalized read coverage in the intron region in condition 1 and 2, respectively. The iDiffIR package was run with the drought treatment data as the first condition and the control as the second condition.

Generating labeled data

For each intron and its flanking exons, iDiffIR generates a log fold-change statistic to quantify differential IR between the conditions, and p-value to indicate how significant that IR event is. Our examples consist of these sequences of introns and their flanking exons. The log fold-change statistic and the p-value are used to label the examples. The examples were divided into three classes based on the following criteria:

1. Up-regulated: log fold-change > 0.5 and p-value < 0.05
2. Down-regulated: log fold-change < -0.5 and p-value <0.05
3. No-change: abs(log fold-change) <= 0.3 or p-value >= 0.1

Thus, the up-regulated class (Figure 3.4a) contains examples that have more intron retention under the stress condition; the down-regulated class (Figure 3.4b) contains examples that have more intron retention in the control condition; the no-change class contains examples that do not have significant difference in the intron retention level between the two conditions. The range of log-fold change from -0.5 to -0.3 and from 0.3 to 0.5, and the p-values from 0.05 to 0.1 are not included in order to create more distinct classes. The number of examples that were generated in each class for the three species is shown in Table 3.4. 80% of total examples were used for training and 20% were held out for testing.

Table 3.4: Number of examples in each class

Species	**Class**	**Number of examples**
A. thaliana	Up-regulated	163
	Down-regulated	178
	No-change	6,527
O.sativa	Up-regulated	55
	Down-regulaed	83
	No-change	1,891
S.bicolor	Up-regulated	331
	Down-regulated	117
	No-change	11,664

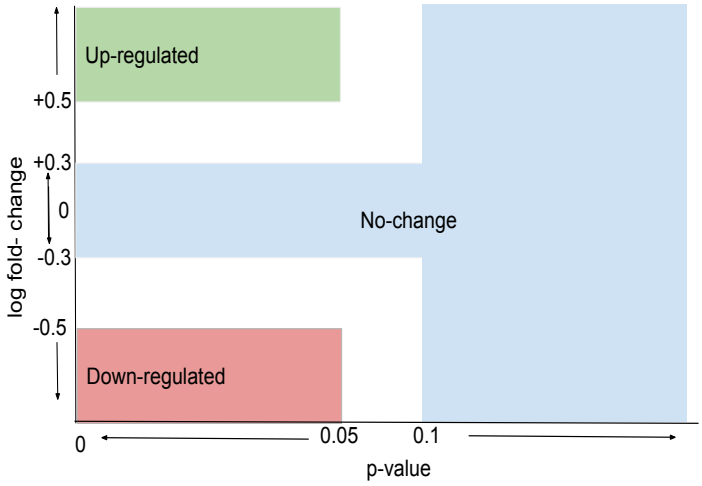
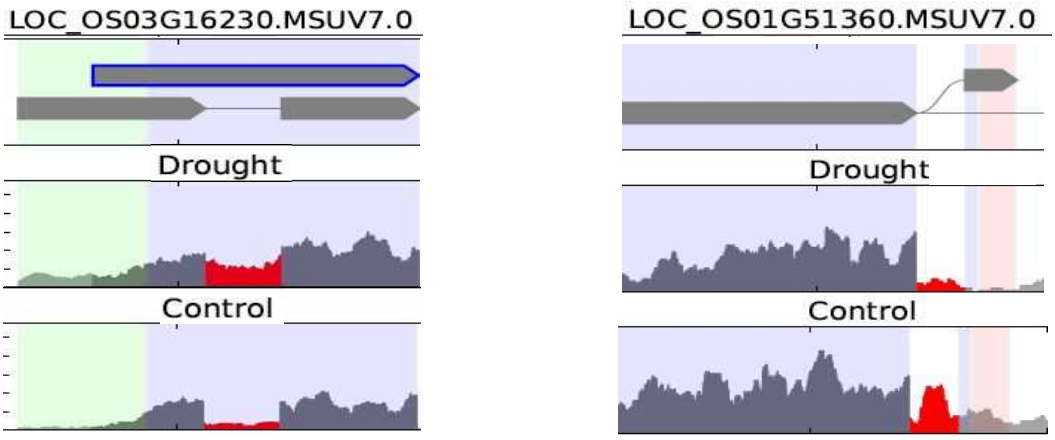


Figure 3.3: Division of data into three classes based on the log-fold change and p-values showing clear separation between the classes.

3.2 Feature Description

As discussed in Chapter 2, retained introns are generally smaller in length, have weaker splice sties and are more G/C rich. Braunchweigh et al. [25] presented an 'IR code' in which they combined these features along with some other features for the task of predicting percentage intron retention. The feature set used by our models is mainly based on the 'IR code', but since it was designed with the human genome in mind, these features were modified to be more suitable for plants. The 137 feature set used in our study is listed in Table 3.5.



(a) Up-regulated (b) Down-regulated

Figure 3.4: Up-regulated and down-regulated differential IR events as detected by iDiffIR [2]. The IR events in the figure are highlighted in red. The top track in the figure is the gene model and the following two tracks represents the read depths across gene under Drought stress and Normal condition respectively. In figure 3.4a, there is more intron retention under the drought stress. Similarly, in figure 3.4b there is less intron retention under the drought stress.

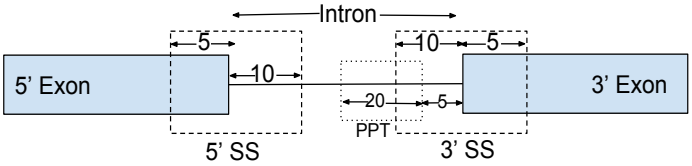


Figure 3.5: Different parts of intron and flanking exons used as features. 5' spice site (5' SS) and 3' splice site (3' SS) consist of 5 nucleotides from exon and 10 nucleotides from intron. Polypyrimidine tract (ppt) consists of 15 nucleotides from -20 to -5 position of the intron.

Each splice site consists of 5 nucleotides from the exon and 10 nucleotides from the intron within the exon-intron boundary as shown in Figure 3.5. The definition of information content is taken from Stormo et al. [72] where information content (IC) at position i is calculated as:

$$IC(i) = \sum_{b} F(b,i) log_2 \frac{F(b,i)}{P(b)}. \tag{3.3}$$

Here, *F(b, i)* is the value of Position Frequency Matrix (PFM) of base b at position i in the motif. *P(b)* is the background nucleotide frequency. This definition of information content does not consider the background probability of each base to be 0.25, instead computes the probability of each base from the input sequences. This gives the value of information

content at position i in the motif. The overall information content of the motif is calculated as the sum of information content at each position of the motif as shown in the equation below:

$$IC = \sum_{i} IC(i). \tag{3.4}$$

Reddy et al. [21] noted that the polypyrimidine tract (PPT) affects splice site selection and also showed that the PPT in plants is rich in Thymine. So, we used the count of Ts in the PPT and we considered the PPT to extend from position -20 to -5 of the intron as shown in Figure 3.5.

Table 3.5: Description of the features used

Feature ID	**Feature**
1	Length of Intron
2	Length of 5' exon
3	Length of 3' exon
4	Relative length of intron and 5' exon
5	Relative length of intron and 3' exon
6	Relative length of 5' exon and 3' exon
7	Relative log length of intron and 5' exon
8	Relative log length of intron and 3' exon
9	Relative log length of 5' exon and 3' exon
10	Information content at the splice donor site (5' SS)
11	Information content at the splice acceptor site (3' SS)
12	Whether the intron is within UTR region
13-28	Dinucleotide content in the intron
29-44	Dinucleotide content in the 5' exon
45-60	Dinucleotide content in the 3' exon
61-76	Dinucleotide content in the first half of intron
77-92	Dinucleotide content in the second half of intron
93-108	Dinucleotide content in the 5' splice sites

109-124	Dinucleotide content in the 3' splice sites
125	G/C content in the intron
126	G/C content in the 5' exon
127	G/C content in the 3' exon
128	G/C content in the first half of the intron
129	G/C content in the second half of the intron
130	G/C content in the 5' splice site
131	G/C content in the 3' splice site
132	contains a premature termination codon (PTC)
133	position of intron among all the introns present in the gene
134	is first intron
135	is second intron
136	is last intron
137	count of 'T' in the polypyrimidine tract (PPT)

Chapter 4

PREDICTING INTRON RETENTION

In this chapter we briefly describe the models we used for the task for predicting retained introns and non-retained introns and discuss the results we got from each model. We also look at the top features that we got from our models for this task and compare them across the different species of plants that we used for our study.

4.1 Methods

We used two machine learning algorithms for this problem which are described next.

4.1.1 Random Forest

Random forest [62] is an ensemble method used for classification and regression tasks. The underlying idea of ensemble methods is to combine "weak learners" to form a "strong learner". It works by building a large number of decision tress (weak learners) by training on random samples of examples and features. An example is then classified by computing the average or mode of the prediction of individual trees.

The interest in random forest for classification problems has been increasing and it has shown excellent performance as compared to other classification algorithms [73–75]. Boulesteix et al. [76] described random forest as a standard data analysis tool in bioinformatics and presented a survey of applications of random forest in bioinformatics and computational biology. Random forests have shown comparable performance to SVM in some recent bioinformatics applications like prediction of miRNA targets [77], identification of DNA-binding proteins [78] and classification of cancer microarray data [79].

Random forests can handle high-dimensional data very well and offers advantages in parameter selection as well. The performance of most machine learning algorithms rely heavily

on selecting good hyper-parameters, but random forests have shown to perform really well without much need for parameter tuning [80, 81].

We used the scikit-learn's implementation of random forest [82]. As our classes (retained and non-retained) are imbalanced, we adjusted the weights of the examples to be inversely proportional to the class frequencies in the input training data. The default value of the parameter 'number of estimators' in skicit-learn's implementation is 10 which is very small. Since random forests are not likely to overfit even with increased number of trees [83] and since a higher number of estimators help in the stability of the model [84], we set 'number of estimators' parameter to be 2000 for our datasets and used the default value of the other parameters.

4.1.2 Support Vector Machine (SVM)

SVM [63] is a supervised learning model used for classification and regression. In the case of classification, it constructs a large margin separating hyperplane, which has been shown to yield good generalization performance [63]. In addition to finding linear decision boundary between classes, SVMs can also efficiently model non-linear decision boundaries between classes using the so-called kernel trick [85]. This allows to implicitly transform data that is not linearly separable into a higher dimensional space where it becomes linearly separable. Although it is directly applicable to binary classification, it also supports multiclass classification by reducing the multiclass problem into several binary classification problem.

SVM is a very robust model and has given excellent performance in different domains [86–88]. It has been applied in various real world problems like classification of images, text, handwritten characters and different biological data as well. In the area of bioinformatics, it has shown to outperform many other classification algorithms like decision trees, k nearest neighbors [89, 90] and even neural networks [91].

Again, we used scikit-learn's implementation of SVM. The parameters for SVM were selected by performing grid-search over the parameter space (kernel = [linear, rbf], C =

[1,10,100] and gamma = [0.001, 0.01, 0.1, 1]) in the training data and the model was evaluated on the test data. To handle the class imbalance problem, we adjusted the weights of the sample to be inversely proportional to the class frequencies in the input training data. This adjusts the parameter C of SVM according to the class-weights.

Table 4.1 shows the total number of retained and non-retained examples we used for both our models. These examples were randomly shuffled and divided into training and test sets in the ratio of 80% and 20% respectively.

Table 4.1: Number of retained and non-retained examples.

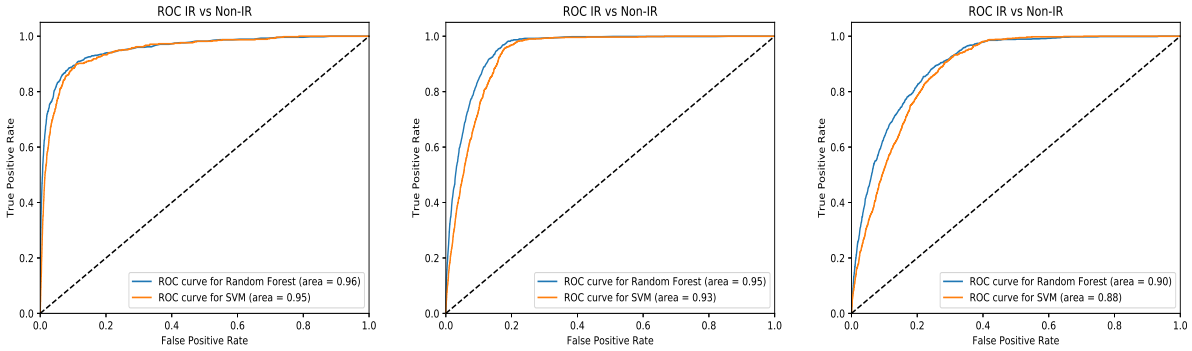
Species	**Retained**	**Non-retained**
A.thaliana	2,760	132,041
O.sativa	5,001	156,745
S.bicolor	4,817	78,406

4.2 Results

To illustrate that there are differences between retained and non-retained introns, we looked at intron length as an example. Table 4.2 shows the average length of introns and the flanking exons in the retained and non-retained examples. We can see that the retained introns are shorter than the non-retained introns in all the three species. In contrast to the introns, the flanking exons in the retained examples are longer than the non-retained examples in all the cases except in 5'exon length of *S.bicolor*. The distribution of intron and exon length in retained and non-retained examples is shown in Appendix B.1

Table 4.2: Average length of intron and flanking exons in retained and non-retained examples.

	Avg. 5'exon length		**Avg. Intron length**		**Avg. 3'exon length**	
	Retained	**Non-retained**	**Retained**	**Non-retained**	**Retained**	**Non-retained**
A.thaliana	247	206	148	171	313	236
O.sativa	284	239	301	424	366	279
S.bicolor	175	177	218	455	232	230



(a) *A.thaliana* (b) *O.sativa* (c) *S.bicolor*

Figure 4.1: Area under ROC of IR vs Non-IR in *A.thaliana*, *O.sativa* and *S.bicolor* respectively.

Next, to evaluate the value of our feature set described in Section 3.2 for the task of predicting retained vs. non-retained introns, we used the models described in Section 4.1. Since the regular accuracy metric is not a good choice for imbalanced data, we used the area under the ROC (Receiver Operating Characteristic) curve as our metric to evaluate our models. Average AUC scores and standard deviation from five runs of each experiment is shown in Table refAUC scores of Random Forest and SVM for predicting IR and non-IR. and Figure 4.1 shows AUC plots from one of the runs as an example.

Table 4.3 shows that our model performs really well to distinguish retained and non-retained introns. This shows that the feature set that we created was useful for differentiating between the two classes. Also, we can observe that random forest performed slightly better than SVM in all three cases.

Table 4.3: Average AUC scores and standard deviation from five runs of experiments of random forest and SVM for predicting IR and non-IR.

Species	**Random Forest**	**SVM**
A.thaliana	**0.958 ± 0.004**	0.95 ± 0.0
O.sativa	**0.95 ± 0.0**	0.93 ± 0.004
S.bicolor	**0.884 ± 0.005**	0.882 ± 0.004

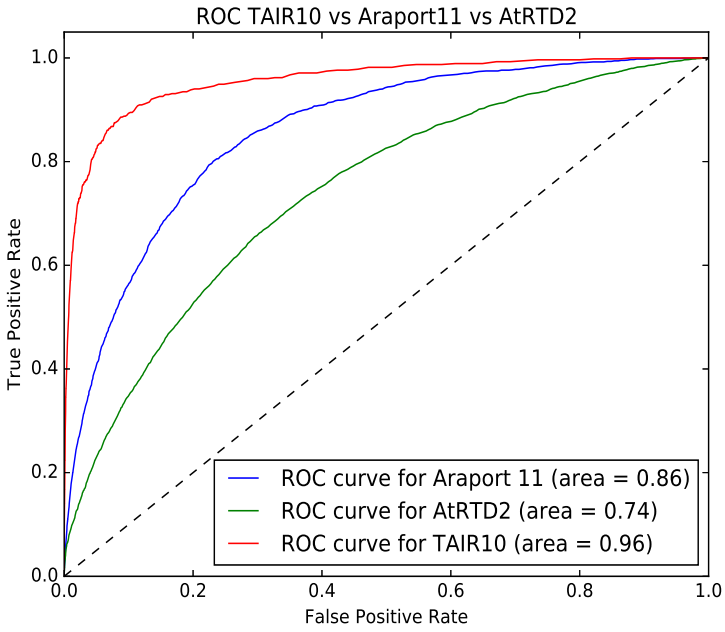


Figure 4.2: Comparison of AUCs for predicting IR vs. non-IR using different reference annotations for *A.thaliana*.

We found that the performance of our model was greatly affected by the reference annotation we used to extract our data. As we can see from Figure 4.2, TAIR10 annotation gave much better performance in predicting IR as compared to the other annotations we tried. Both the Araport11 [92] and AtRTD2 [93] annotations were generated using RNA-Seq data. The short reads of RNA-Seq introduce errors when used to predict transcript annotations, which explains the poor performance using them. Further, we also compared the average length of retained and non-retained introns from the different annotations. From the literature, we know that retained introns on average are shorter than non-retained introns but as we can see from Figure 4.4, retained introns obtained from Araport11 and AtRTD2 annotations were longer than the non-retained introns. This further illustrates the poor quality of these reference annotations.

Table 4.4: Comparison of average length of retained and non-retained introns from different annotations.

	Retained Introns	**Non-retained Introns**
TAIR10	148	171
Araport11	186	172
AtRTD2	196	180

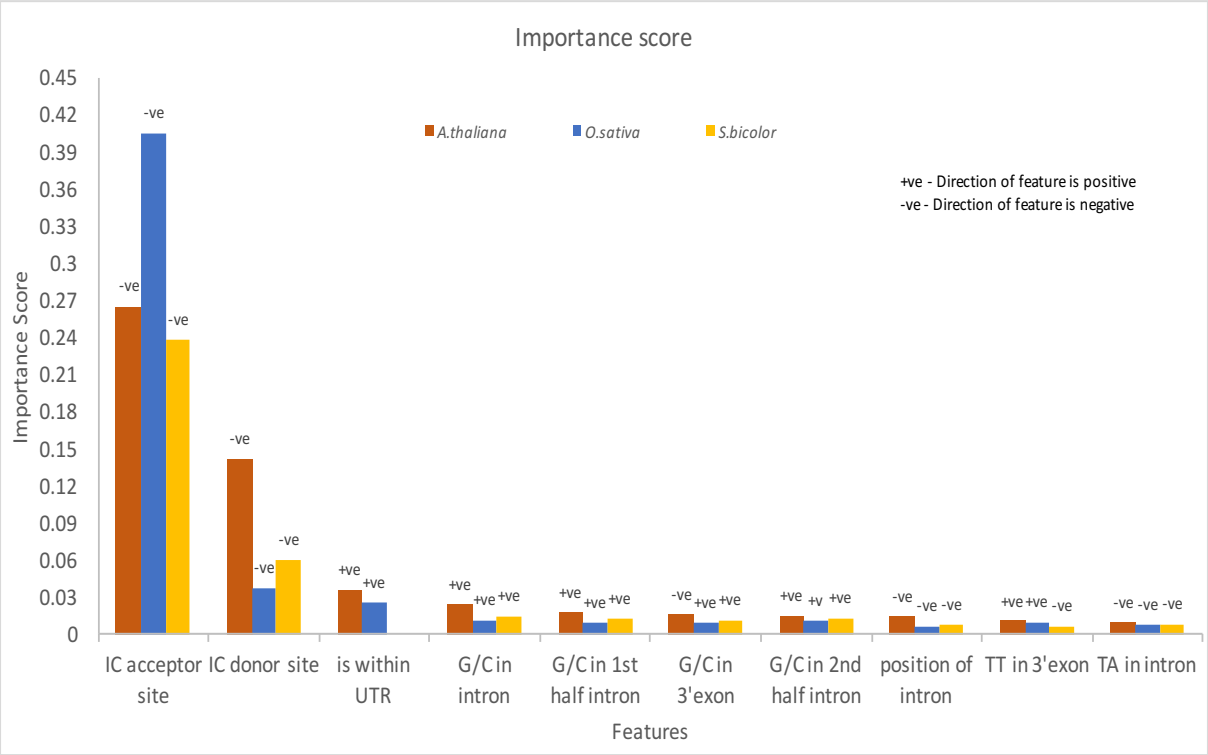


Figure 4.3: Importance score and direction of top features of random forest for predicting retained and non-retained introns for *A.thaliana*, *O.sativa* and *S.bicolor*. The ordering of the top features on x-axis is based on the ranking of *A.thaliana*. The direction in which a feature contributes is indicated by the symbols '+ve' and '-ve' on top of each bar.

Next, we were interested in looking at each model's top features and the direction in which they contribute for each of the three species. The direction of contribution of a feature is defined by whether the mean value of the feature is higher in the retained examples or non-retained examples. The top 10 features of the random forest model and their importance score are shown in Figure 4.3.

In scikit-learn, the importance score in random forest is the "gini importance" (also known as the "mean decrease in impurity") [94]. It is defined as the total decrease in the node impurity (weighted by the probability of reaching that node) averaged over all trees of the ensemble. The gini impurity index of a node, t, is calculated as:

$$Gini(t) = \sum_{i=1}^{n_c} p_i(1 - p_i), \tag{4.1}$$

where n_c is the number of classes in the target variable and p_i is the ratio of this class. If it is a binary classification problem, then decrease in the gini impurity is calculated as:

$$GiniDecrease(t) = Gini(t) - p_{t_L} Gini(t_L) - p_{t_R} Gini(t_R) \tag{4.2}$$

Finally, gini importance is the mean decrease in gini impurity in all trees of the forest.

As we can see in Figure 4.3, the feature "Information content in the acceptor-site" is recognized as the most important feature in all three species. Similarly the features "Information content in the donor-site", "is within UTR", and "G/C content" in the different parts of introns and exons were also consistently predicted to be important. Our results reveal that retained introns have lower information content in the splice-sites and are G/C rich. Thus, our result supports the hypothesis that elevated G/C content in the retained introns affect splice site recognition, in agreement with the literature [58]. Since, retained introns are shorter and are G/C rich they are more likely to be flagged as "exons" by the splicing machinery [54, 58, 59]. Similarly, lower information content in the splice-sites of retained introns implies that these introns are harder to detect by the splicing machinery and thus more likely to be retained.

We compared the top 10 features across the three species to look at the features that are consistently highly ranked in all of them. Figure 4.4 shows the visual comparison of the top 10 features across the three species. From the figure, we can see most of the important features were common across the species. Information content in the acceptor site and Information content in the donor site were found to be the top two features in all three species which indicates that they are very predictive of IR. 8 features out of 10 top features

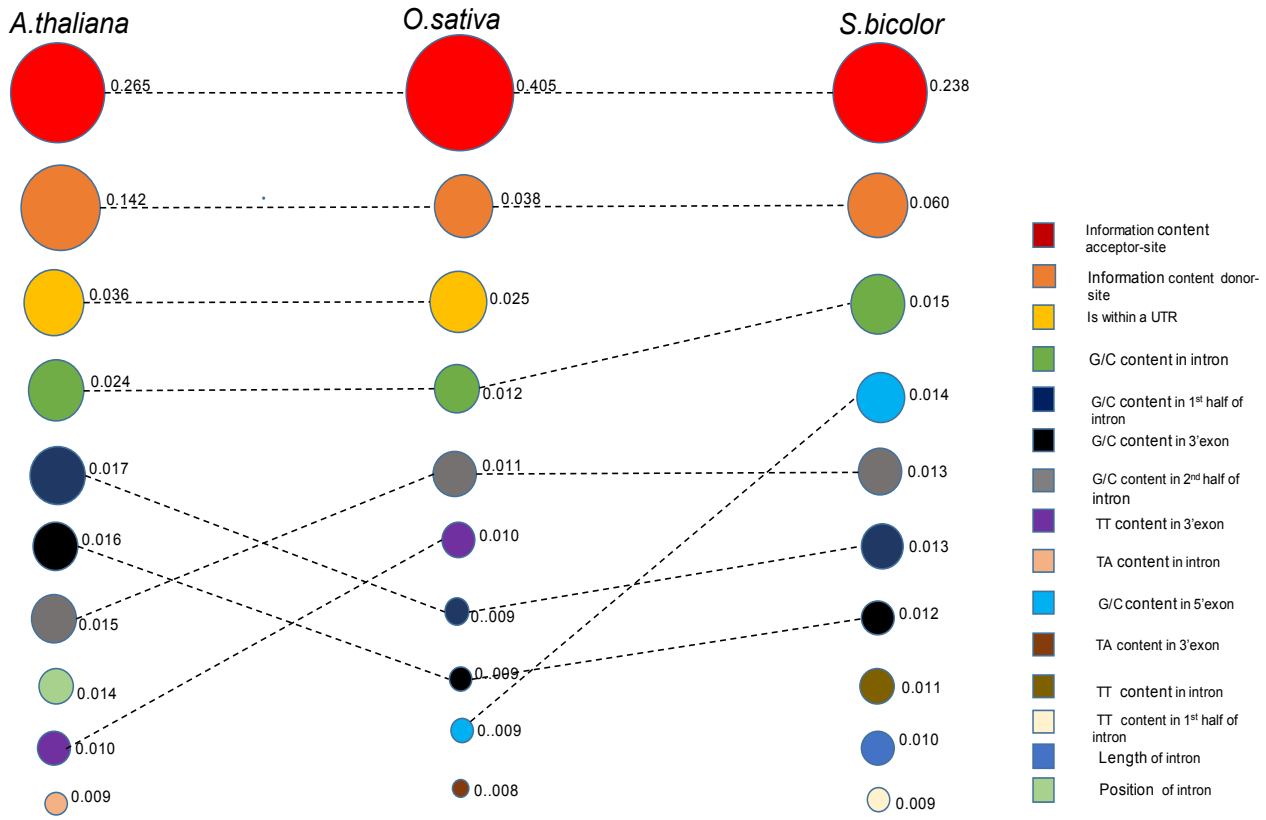


Figure 4.4: Comparison of top 10 features random forest model for predicting retained introns in the three species. Each circle represents a feature and the importance of the feature is indicated by its size. The columns represent the top features of each species and the rows represents the ranking of the features. The dotted lines indicate common features across species.

were common between *A.thaliana* and *O.sativa*. This consistency in the important features across the species support the biological relevance of these features in regulating IR.

Since, our features are mostly based on the "IR code" work of Braunschweig et. al, we compared their top 10 features [25] with the top 10 features from our random forest model for *A.thaliana* and found some overlap in the top features as shown in Figure 4.5. Although some of the important features in the IR code such as "length of intron", "relative length of intron and 3' exon" and "G/C content in 5' exon" were also found to be important in our model as well, they are not included in the overlap in the Figure 4.5 since we compared only the top 10 features of both the models. The study of Braunschweig et. al [25] was based on

mammals, thus the overlap between the top features of their model and the top feature of our model suggests that the process of IR in mammals in quite similar to that of plants.

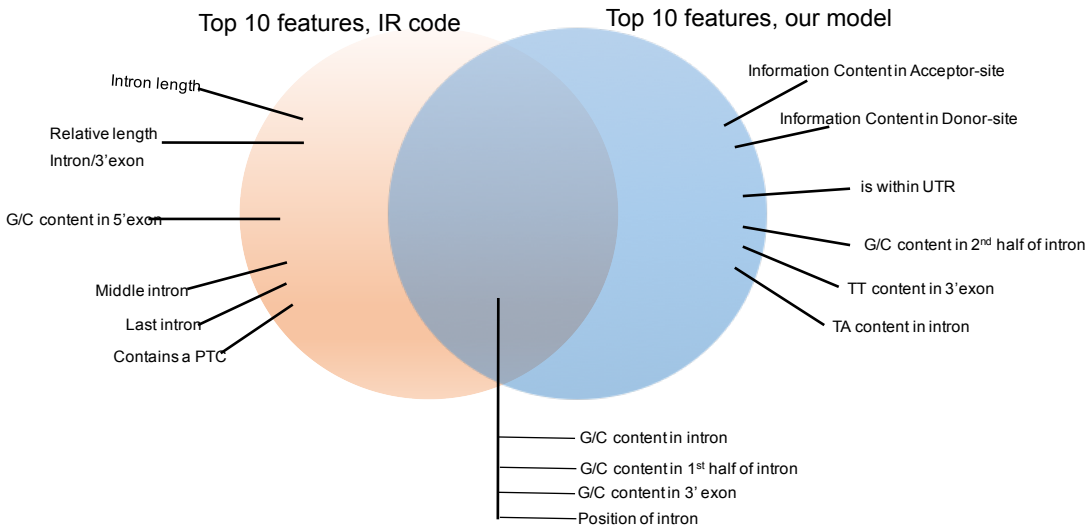


Figure 4.5: Venn diagram comparing the top 10 features of "IR code" and the top features obtained from our model for *A.thaliana*.

Chapter 5

PREDICTING DIFFERENTIAL INTRON RETENTION

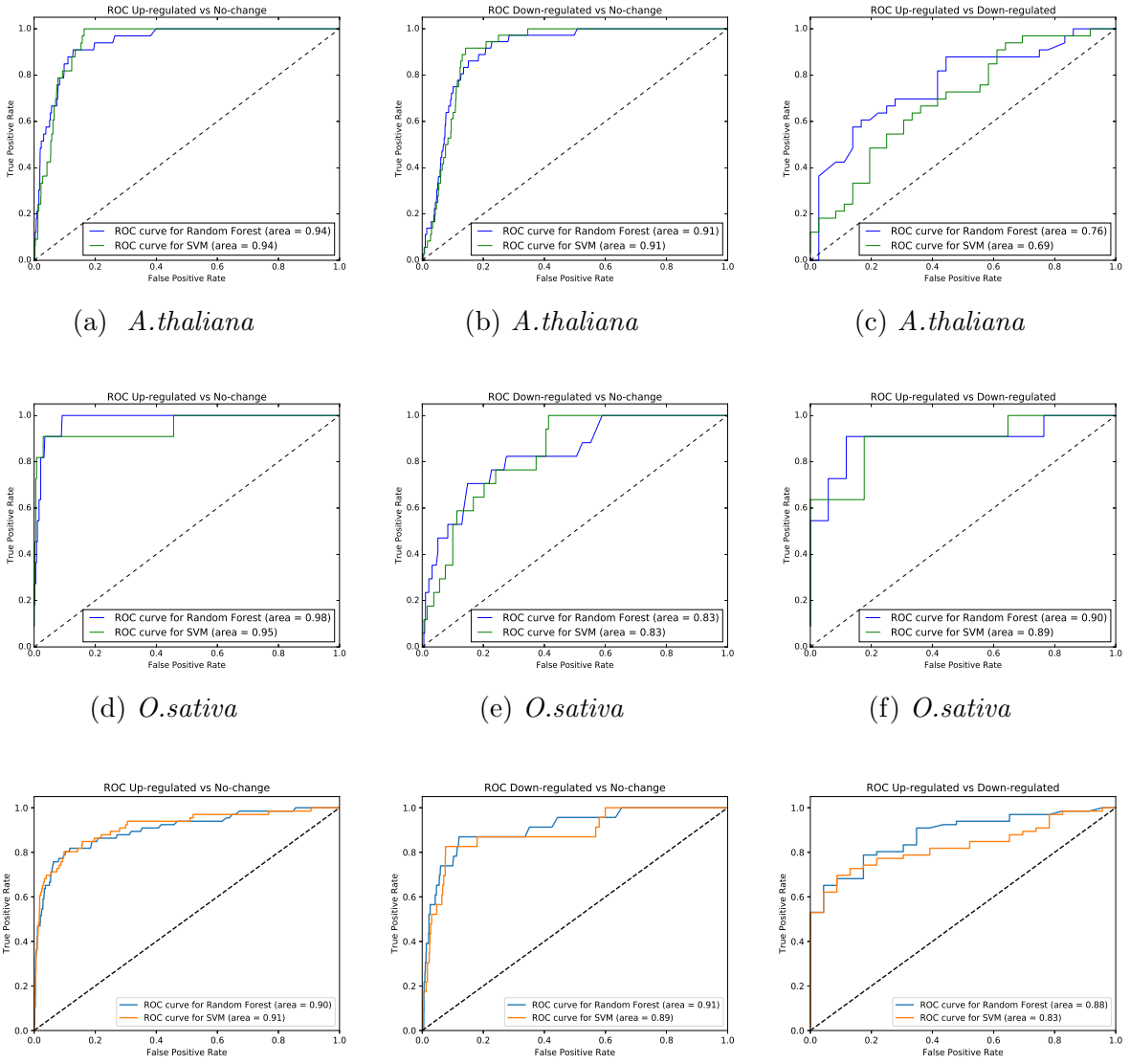
In this chapter, we describe the models we used for predicting condition-specific intron retention in three plant species and discuss the results in each case. We also compare the top features from the models to find any features that are conserved across species.

5.1 Methods

We used random forest and SVM as our model for this task as well. As described in Section 4.1, we used scikit-learn's implementation of random forest and SVM. For the random forest model, for the smaller datasets, 'the number of estimators' was set to 500 and for the larger datasets it was set to around 2000. The default value of other parameters was used for the random forest model. For the SVM we used grid-search over the parameter space (kernel = [linear, rbf], C = [1,10,100] and gamma = [0.001, 0.01, 0.1, 1]) on the training data to select the parameters and evaluated the model on the test data. As discussed in Section 3.1.2, for predicting differential IR, we analyzed the RNA-Seq data of *A.thaliana*, *O.sativa* and *S.bicolor* under drought stress and divided the example sequences into three categories: *up-regulated*, *down-regulated* and *no-change*. Table 5.1 shows the total number of examples in each class we used for both our models. These examples were randomly shuffled and divided into training and test sets in the ratio of 80% and 20% respectively.

5.2 Results

We performed binary classification between each pair of the categories and present the results of both the models for the three species of plants. The models were evaluated using the AUC metric. Table 5.2 shows the average AUC scores and standard deviation from ten runs of each experiment and Figure 5.1 shows the AUC plots from one of the runs as an example.



(a) *A.thaliana* (b) *A.thaliana* (c) *A.thaliana*

(d) *O.sativa* (e) *O.sativa* (f) *O.sativa*

(g) *S.bicolor* (h) *S.bicolor* (i) *S.bicolor*

Figure 5.1: AUC for predicting differential intron retention under drought stress across different species. Rows correspond to different species: *A.thalia*, *O.sativa* and *S.bicolor* respectively. Columns correspond to the different classification problems: up-regulated vs. no-change, down-regulated vs. no-change and up-regulated vs. down-regulated respectively.

Table 5.1: Number of examples in each class

Species	**Class**	**Number of examples**
A. thaliana	Up-regulated	163
	Down-regulated	178
	No-change	6,527
O.sativa	Up-regulated	55
	Down-regulaed	83
	No-change	1,891
S.bicolor	Up-regulated	331
	Down-regulated	117
	No-change	11,664

Table 5.2: Mean AUC scores and standard deviation of ten runs of the experiments of random forest and SVM for predicting differential intron retention in different species.

Species	Up-regulated vs. No-change		Down-regulated vs. No-change		Up-regulated vs. Down-regulated	
	Random Forest	SVM	Random Forest	SVM	Random Forest	SVM
A.thaliana	0.927 ± 0.016	**0.939 ± 0.011**	0.908 ±0.014	**0.924 ± 0.011**	**0.738 ± 0.028**	0.681 ± 0.026
O.sativa	**0.982 ± 0.006**	0.979 ± 0.011	**0.812 ± 0.020**	0.785 ± 0.029	**0.882 ± 0.024**	0.852 ± 0.051
S.bicolor	**0.915 ± 0.019**	0.901 ± 0.019	**0.913 ± 0.036**	0.891 ± 0.045	**0.827 ± 0.068**	0.792 ± 0.060

Our models performed really well for the task of predicting up-regulated vs. no-change and specially in case of *O.sativa* we achieved a very high AUC of approximately 98% with random forest although the number of examples for training the models was very small. However, in case of the second classification problem, down-regulated vs. no-change, we got higher AUC scores for *A.thaliana* and *S.bicolor* as compared to *O.sativa*, but overall the models performed really well to reliably predict down-regulated examples from no-change examples.

These two classification problems (up-regulated vs. no-change and down-regulated vs. no-change) are similar to predicting retained introns vs. non-retained introns since up-regulated and down-regulated introns are basically the retained introns in a particular experimental condition and the introns in the no-change class are the non-retained introns in that condition. So, we were interested in looking at a more harder problem of how accurately our model can predict the direction of regulation (up-regulated or down-regulated). As we can see from the Table 5.2, our models gave pretty good performance in *O.sativa* and *S.bicolor* but did not perform very well in *A.thaliana*. Since both the classes contain differentially

retained introns, this is a more difficult classification problem and incorporating additional information might help improve the performance of the models.

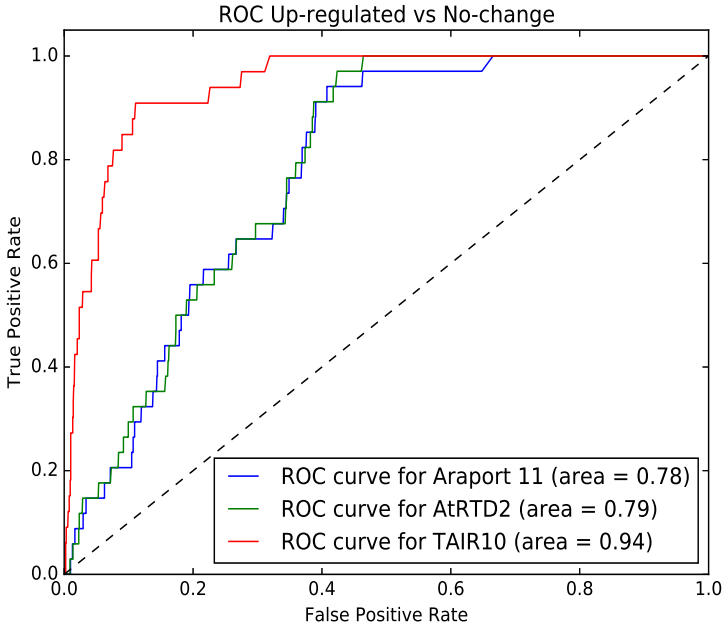


Figure 5.2: Comparison of AUCs for different annotations in *A.thaliana*.

As discussed in Section 4.2, the results of our models were greatly affected by the annotation file we used during the data preparation. We compared the results we obtained for the task of predicting up-regulated vs no-change in *A.thaliana* using TAIR10 annotations to the results we got from Araport11 and AtRTD2 annotations. Similar to the results shown in Figure 4.2 for IR, Figure 5.2 shows that TAIR10 annotations performed much better than Araport11 and AtRTD2 annotation for the task of predicting up-regulated vs. no-change class as well.

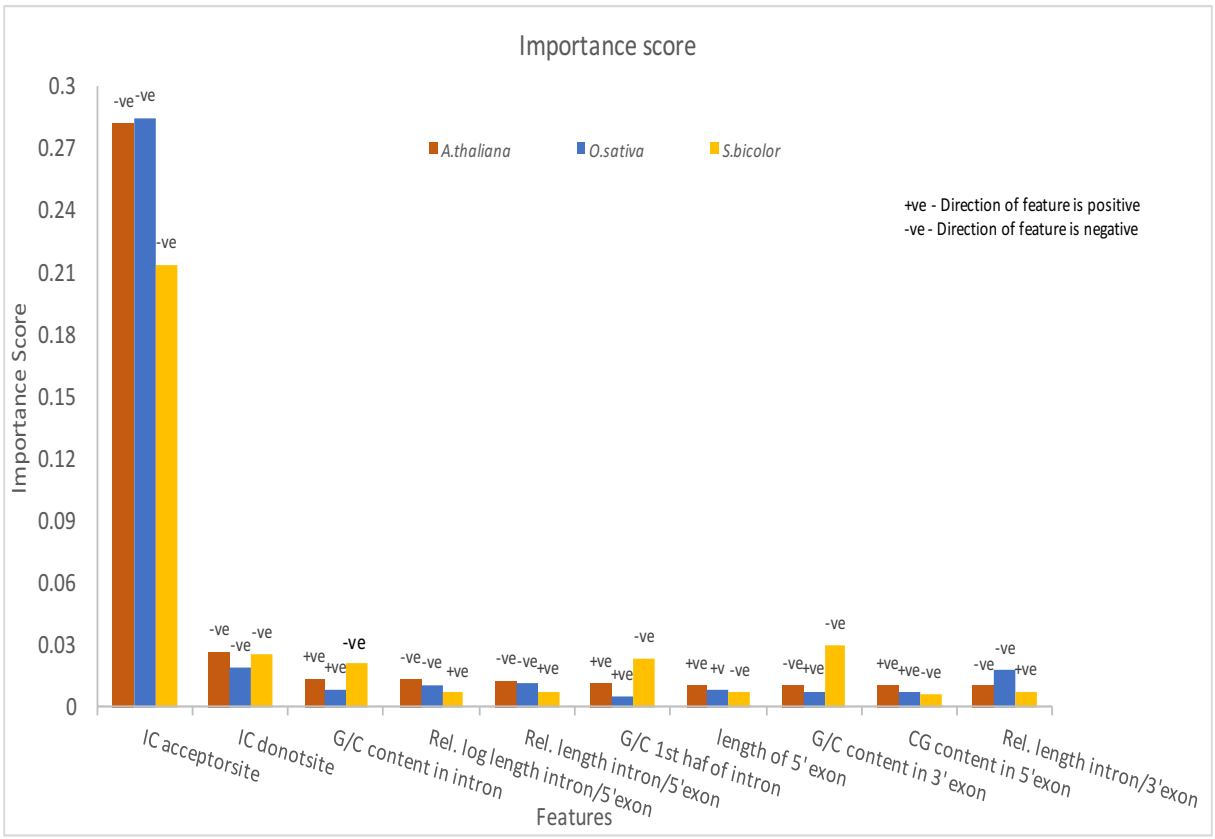


Figure 5.3: Importance score and direction of top features of random forest for predicting up-regulated and no-change examples for *A.thaliana*, *O.sativa* and *S.bicolor*. The ordering of the top features on x-axis is based on the ranking of *A.thaliana*. The direction of features is indicated by the symbols '+ve' and '-ve' on top of each bar.

To obtain further insight on the important features that regulate differential IR, we compared the top features from random forest across the three species. Figure 5.3 shows the importance score and the direction of the top features in the three species and Figure 5.4 shows the comparison of the top 10 features across the three species for the problem of distinguishing up-regulated and no-change introns. Figures for feature comparison for the two other classification problems are shown in Appendix C. From Figure 5.4 we can observe that the feature Information content in the acceptor-site was found to be the most important feature in all three species. Similarly, Information content in the donor-site, G/C content in different parts of intron and flanking exons, and relative length of intron and flanking exons were found to be common in the top features of different species. As discussed in Section 4.2, consistency of these important features across different species increases our confidence in

the regulatory role of these features in differential IR. However, one important thing to note from Figure 5.4, is that there is not as much consistency in the important features across species as we saw in case of IR vs. non IR (Figure 4.4).

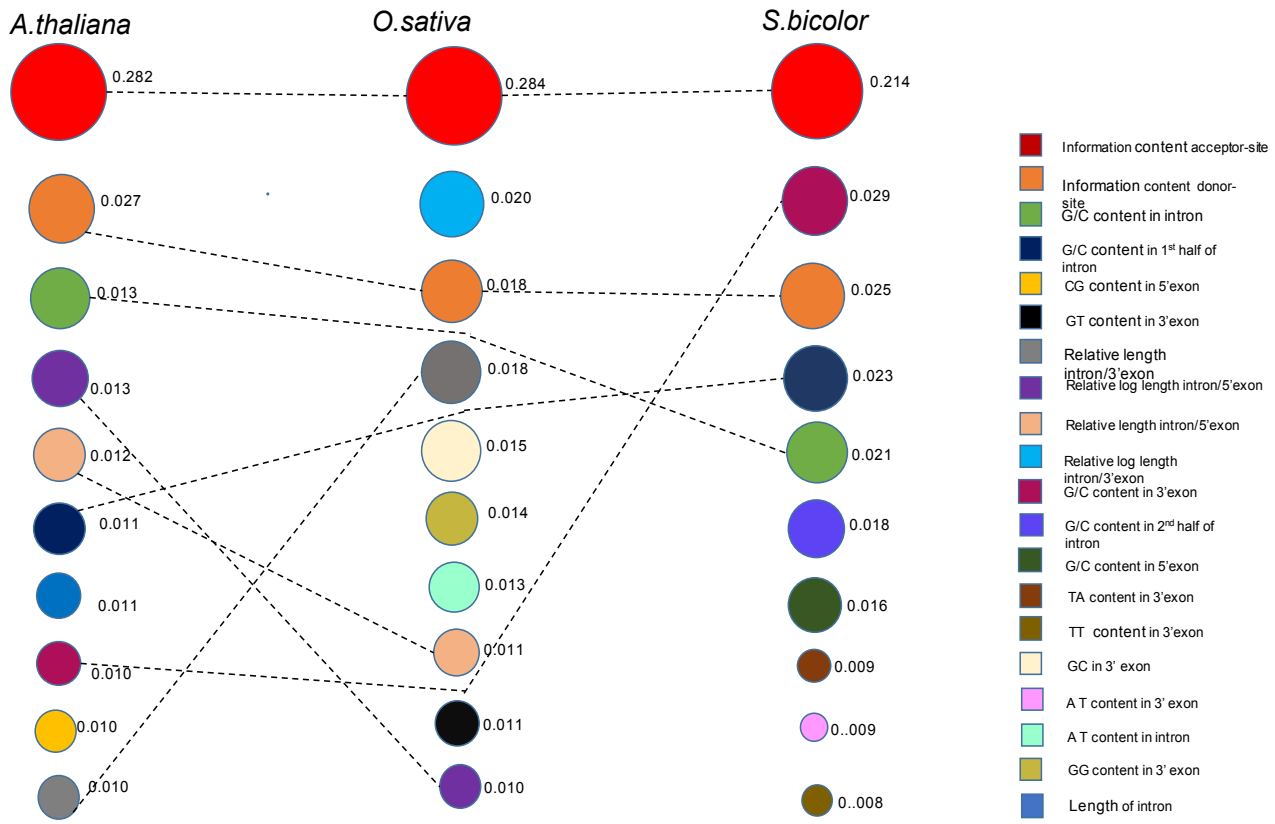


Figure 5.4: Comparison of the top 10 features from random forest models for distinguishing up-regulated introns from no-change introns in the three species. Each circle represents a feature and the importance of the feature is indicated by its size. The columns represent the top features of each species and the rows represents the ranking of the features. The dotted lines indicate common features across species.

Another interesting question is whether there is any evolutionary link between the genes of the three species of the plants we studied with respect to how they are regulated in a particular experimental condition. In order to analyze this, we looked at the gene homologs between the pairs of species that are regulated in the same direction under drought stress. A homologous gene is a gene inherited in two species from a common ancestor. Homologous genes often have high sequence similarity. We used the BioMart tool of the Phytozome platform [95] to find the homologs pairs between *A.thaliana* and *O.sativa*/*S.bicolor* and

considered only those pairs that were differentially retained in the same direction. The result we got from our analysis in shown in Table 5.3. We found very few homologous pairs that were regulated in the same direction under stress.

Table 5.3: List of gene homologs that have the same direction of intron retention regulation under drought stress.

	A.thaliana	***O.sativa/S.bicolor***	**Direction**	**Function**
A.thaliana - O.sativa	AT5G44520	LOC_Os03g56869	Up-regulated	pentose-phosphate shunt, non-oxidative branch
	AT2G47250	LOC_Os03g19960	Down-regulated	RNA splicing, mRNA processing
	AT4G20930	LOC_Os06g46372	Down-regulated	oxidation-reduction process, valine catabolic process
A.thaliana - S.bicolor	AT2G22980	Sobic.002G175500	Up-regulated	proteolysis, secondary metabolic process
	AT4G35940	Sobic.001G440600	Up-regulated	

The results in Table 5.2 analyze differentially retained introns from the same experimental condition of drought stress. In order to further see how well our feature set can predict condition-specific pattern we predicted differentially retained introns in *A.thaliana* from two experimental conditions (i.e up-regulated in condition1 vs. up-regulated in condition2, and down-regulated in condition1 vs. down-regulated in condition2). We used the same drought dataset for condition 1 and NaCl treatment dataset (SRP035234) for condition 2. The NaCl treatment dataset was prepared using the same pipeline described in Section 3.1.2. The details of the dataset used for this task in shown in Table 5.4. Table 5.5 shows that although this is a much difficult classification problem, our feature set could reliably capture the condition-specific patterns of intron retention. However, the AUC score of predicting down-regulated introns is lower than the AUC score of predicting up-regulated introns. This might indicate that the problem of predicting down-regulated introns in different conditions

Table 5.4: The number of up-regulated and down-regulated examples in each SRA study used for predicting deferentially retained introns in two different experimental conditions.

Condition	**SRA Study**	**Up-regulated**	**Down-regulated**
Drought	SRP056035	163	326
NaCl	SRP035234	178	147

is harder and adding more informative features might improve the performance.

Table 5.5: AUC scores of random forest and SVM for predicting differentially retained introns in *A.thaliana* from two experimental condition.

Task	**Random Forest**	**SVM**
Up-regulated (Drought) vs. Up-regulated (NaCl)	**0.89**	**0.89**
Down-regulated (Drought) vs. Down-regulated (NaCl)	**0.82**	0.64

Chapter 6

CONCLUSION AND FUTURE WORK

Although intron retention is the most prevalent form of alternative splicing in plants and differential intron retention has been linked to different rare diseases, there has not been much research to understand the regulation of these processes. Our aim with this project was to design an efficient model that can reliably predict intron retention and differential intron retention in order to better understand the factors that influence these processes. We used three species of plants: *A.thaliana*, *O.sativa* and *S.bicolor*, for our study. We formulated a feature set of 137 features for our models, SVM and random forest.

We showed the usefulness of our feature set in distinguishing retained introns from non-retained introns. Further, we observed a high level of similarity among the top-ranking features across the three different species of plants, suggesting their biological. The problem of predicting differential IR proved to be more difficult. We got lower levels of accuracy and lower similarity in the top-ranking features across species for this task. In the future, we would like to investigate additional features that would represent e.g., splicing regulatory elements (discussed in Section 1.4.2) which could improve model performance. Similarly, chromatin state could be investigated as a feature for this model as it has been shown to have an effect on splice-site selection in exon skipping [96, 97] and, more recently has also been discovered to have an effect in IR [98].

In our study we have defined retained introns looking only at the gene models. Use of RNA-Seq data to identify additional retained introns can help address this limitation. As we know from our results that annotations can often be insufficient. So, we can use introns that are consistently retained/non-retained in different experimental conditions to define retained/non-retained examples to create more confident examples. Similarly, we could also use a different strategy to define the no-change class. In our study we have defined no-change introns as the introns that exhibit no change under a given pair of condition. We could define

no-change introns as the introns that exhibit IR in some conditions but exhibit no change under a given pair of conditions. This can allow us to capture the condition-specific patterns more precisely.

We have performed a thorough analysis of IR/differential IR with our study. But still there exists much room for further investigation and analysis in order to improve our understanding of these processes.

BIBLIOGRAPHY

[1] R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites," *Journal of molecular biology*, vol. 228, no. 4, pp. 1124–1136, 1992.

[2] M. Hamilton, "Predicting differential intron retention with iDiffIR," in *Plant and Animal Genome XXIV Conference*. Plant and Animal Genome, 2016.

[3] D. C. Corney, "Rna-seq using next generation sequencing," *Mater Methods*, vol. 3, p. 203, 2013.

[4] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, p. bbw068, 2016.

[5] C. Gibas and P. Jambeck, *Developing bioinformatics computer skills*. "O'Reilly Media, Inc.", 2001.

[6] H. Adeli and S.-L. Hung, *Machine learning: neural networks, genetic algorithms, and fuzzy systems*. John Wiley & Sons, Inc., 1994.

[7] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, no. 4, pp. 243–268, 2003.

[8] C. Mathé, M.-F. Sagot, T. Schiex, and P. Rouzé, "Current methods of gene prediction, their strengths and weaknesses," *Nucleic acids research*, vol. 30, no. 19, pp. 4103–4117, 2002.

[9] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids," *Resonance*, vol. 9, no. 11, pp. 96–98, 2004.

[10] E. Westhof and P. Auffinger, "RNA tertiary structure," *Encyclopedia of analytical chemistry*, 2000.

[11] V. Thakur, "Why nature preferred DNA over RNA," https://sciencesamhita.com/dna-as-the-genetic-material/, 2015.

[12] S. M. Berget, C. Moore, and P. A. Sharp, "Spliced segments at the 5' terminus of adenovirus 2 late mRNA," *Proceedings of the National Academy of Sciences*, vol. 74, no. 8, pp. 3171–3175, 1977.

[13] K. Gao, A. Masuda, T. Matsuura, and K. Ohno, "Human branch point consensus sequence is yUnAy," *Nucleic acids research*, vol. 36, no. 7, pp. 2257–2267, 2008.

[14] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annual review of biochemistry*, vol. 72, no. 1, pp. 291–336, 2003.

[15] A. Corvelo, M. Hallegger, C. W. Smith, and E. Eyras, "Genome-wide association between branch point properties and alternative splicing," *PLoS Comput Biol*, vol. 6, no. 11, p. e1001016, 2010.

[16] Z. Wang and C. B. Burge, "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code," *Rna*, vol. 14, no. 5, pp. 802–813, 2008.

[17] A. J. Matlin, F. Clark, and C. W. Smith, "Understanding alternative splicing: towards a cellular code," *Nature reviews Molecular cell biology*, vol. 6, no. 5, pp. 386–398, 2005.

[18] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey, "Deciphering the splicing code," *Nature*, vol. 465, no. 7294, pp. 53–59, 2010.

[19] H. Keren, G. Lev-Maor, and G. Ast, "Alternative splicing and evolution: diversification, exon definition and function," *Nature Reviews Genetics*, vol. 11, no. 5, pp. 345–355, 2010.

[20] A. S. Reddy, "Alternative splicing of pre-messenger RNAs in plants in the genomic era," *Annu. Rev. Plant Biol.*, vol. 58, pp. 267–294, 2007.

[21] A. S. Reddy, M. F. Rogers, D. N. Richardson, M. Hamilton, and A. Ben-Hur, "Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements," *Frontiers in plant science*, vol. 3, p. 18, 2012.

[22] H. Ner-Gaon, R. Halachmi, S. Savaldi-Goldstein, E. Rubin, R. Ophir, and R. Fluhr, "Intron retention is a major phenomenon in alternative splicing in arabidopsis," *The Plant Journal*, vol. 39, no. 6, pp. 877–885, 2004.

[23] B.-B. Wang and V. Brendel, "Genomewide comparative analysis of alternative splicing in plants," *Proceedings of the National Academy of Sciences*, vol. 103, no. 18, pp. 7175–7180, 2006.

[24] M. Kalyna, C. G. Simpson, N. H. Syed, D. Lewandowska, Y. Marquez, B. Kusenda, J. Marshall, J. Fuller, L. Cardle, J. McNicol *et al.*, "Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in arabidopsis," *Nucleic acids research*, vol. 40, no. 6, pp. 2454–2469, 2012.

[25] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes transcriptomes," *Genome research*, vol. 24, no. 11, pp. 1774–1786, 2014.

[26] R. Middleton, D. Gao, A. Thomas, B. Singh, A. Au, J. J. Wong, A. Bomane, B. Cosson, E. Eyras, J. E. Rasko *et al.*, "IRFinder: assessing the impact of intron retention on mammalian gene expression," *Genome biology*, vol. 18, no. 1, p. 51, 2017.

[27] K. Yap and E. V. Makeyev, "Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms," *Molecular and Cellular Neuroscience*, vol. 56, pp. 420–428, 2013.

[28] S. Goodison, K. Yoshida, M. Churchman, and D. Tarin, "Multiple intron retention occurs in tumor cell CD44 mRNA processing," *The American journal of pathology*, vol. 153, no. 4, pp. 1221–1228, 1998.

[29] F. J. Blanco, M. T. Grande, C. Langa, B. Oujo, S. Velasco, A. Rodriguez-Barbero, E. Perez-Gomez, M. Quintanilla, J. M. López-Novoa, and C. Bernabeu, "S-endoglin expression is induced in senescent endothelial cells and contributes to vascular pathology," *Circulation research*, vol. 103, no. 12, pp. 1383–1392, 2008.

[30] J. J.-L. Wong, A. Y. Au, W. Ritchie, and J. E. Rasko, "Intron retention in mRNA: No longer nonsense," *Bioessays*, vol. 38, no. 1, pp. 41–49, 2016.

[31] J. Eswaran, A. Horvath, S. Godbole, S. D. Reddy, P. Mudvari, K. Ohshiro, D. Cyanam, S. Nair, S. A. Fuqua, K. Polyak *et al.*, "RNA sequencing of cancer reveals novel splicing alterations," *Scientific reports*, vol. 3, p. 1689, 2013.

[32] Q. Zhang, H. Li, H. Jin, H. Tan, J. Zhang, and S. Sheng, "The global landscape of intron retentions in lung adenocarcinoma," *BMC medical genomics*, vol. 7, no. 1, p. 15, 2014.

[33] A. M. Mastrangelo, S. Belloni, S. Barilli, B. Ruperti, N. Di Fonzo, A. M. Stanca, and L. Cattivelli, "Low temperature promotes intron retention in two e-cor genes of durum wheat," *Planta*, vol. 221, no. 5, pp. 705–715, 2005.

[34] S. G. Palusa, G. S. Ali, and A. S. Reddy, "Alternative splicing of pre-mRNAs of arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses," *The Plant Journal*, vol. 49, no. 6, pp. 1091–1107, 2007.

[35] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[36] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra, "Profiling the hela S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing," *Biotechniques*, vol. 45, no. 1, p. 81, 2008.

[37] M. P. Cox, D. A. Peterson, and P. J. Biggs, "SolexaQA: at-a-glance quality assessment of illumina second-generation sequencing data," *BMC bioinformatics*, vol. 11, no. 1, p. 485, 2010.

[38] K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harbor Protocols*, vol. 2015, no. 11, pp. pdb–top084 970, 2015.

[39] G. Dror, R. Sorek, and R. Shamir, "Accurate identification of alternatively spliced exons using support vector machine," *Bioinformatics*, vol. 21, no. 7, pp. 897–901, 2005.

[40] G. Rätsch, S. Sonnenburg, and B. Schölkopf, "RASE: recognition of alternatively spliced exons in c. elegans," *Bioinformatics*, vol. 21, no. suppl 1, pp. i369–i377, 2005.

[41] R. Sorek, R. Shemesh, Y. Cohen, O. Basechess, G. Ast, and R. Shamir, "A non-EST-based method for exon-skipping prediction," *Genome Research*, vol. 14, no. 8, pp. 1617–1623, 2004.

[42] T. Thanaraj and S. Stamm, "Prediction and statistical analysis of alternatively spliced exons," in *Regulation of Alternative Splicing*. Springer, 2003, pp. 1–31.

[43] A. Resch, Y. Xing, A. Alekseyenko, B. Modrek, and C. Lee, "Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation," *Nucleic Acids Research*, vol. 32, no. 4, pp. 1261–1269, 2004.

[44] F. Clark and T. Thanaraj, "Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human," *Human molecular genetics*, vol. 11, no. 4, pp. 451–464, 2002.

[45] Y. Zhuang and A. M. Weiner, "A compensatory base change in U1 snRNA suppresses a 5' splice site mutation," *Cell*, vol. 46, no. 6, pp. 827–835, 1986.

[46] R. Sorek, G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur, and G. Ast, "Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons," *Molecular cell*, vol. 14, no. 2, pp. 221–231, 2004.

[47] C. S. Leslie, E. Eskin, and W. S. Noble, "The spectrum kernel: A string kernel for SVM protein classification." in *Pacific symposium on biocomputing*, vol. 7, no. 7, 2002, pp. 566–575.

[48] R. Mao, P. K. R. Kumar, C. Guo, Y. Zhang, and C. Liang, "Comparative analyses between retained introns and constitutively spliced introns in arabidopsis thaliana using random forest and support vector machine," *PloS one*, vol. 9, no. 8, p. e104049, 2014.

[49] Q. Xu, B. Modrek, and C. Lee, "Genome-wide detection of tissue-specific alternative splicing in the human transcriptome," *Nucleic acids research*, vol. 30, no. 17, pp. 3754–3766, 2002.

[50] D. Das, T. A. Clark, A. Schweitzer, M. Yamamoto, H. Marr, J. Arribere, S. Minovitsky, A. Poliakov, I. Dubchak, J. E. Blume *et al.*, "A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing," *Nucleic acids research*, vol. 35, no. 14, pp. 4845–4857, 2007.

[51] J. C. Castle, C. Zhang, J. K. Shah, A. V. Kulkarni, A. Kalsotra, T. A. Cooper, and J. M. Johnson, "Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines," *Nature genetics*, vol. 40, no. 12, pp. 1416–1425, 2008.

[52] C. W. Sugnet, K. Srinivasan, T. A. Clark, G. O'Brien, M. S. Cline, H. Wang, A. Williams, D. Kulp, J. E. Blume, D. Haussler *et al.*, "Unusual intron conservation near tissue-regulated exons found by splicing microarrays," *PLoS Comput Biol*, vol. 2, no. 1, p. e4, 2006.

[53] M. Fagnani, Y. Barash, J. Y. Ip, C. Misquitta, Q. Pan, A. L. Saltzman, O. Shai, L. Lee, A. Rozenhek, N. Mohammad *et al.*, "Functional coordination of alternative splicing in the mammalian central nervous system," *Genome biology*, vol. 8, no. 6, p. R108, 2007.

[54] P. A. F. Galante, N. J. Sakabe, N. Kirschbaum-Slager, and S. J. de Souza, "Detection and evaluation of intron retention events in the human transcriptome," *Rna*, vol. 10, no. 5, pp. 757–765, 2004.

[55] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[56] N. J. Sakabe and S. J. De Souza, "Sequence features responsible for intron retention in human," *BMC genomics*, vol. 8, no. 1, p. 59, 2007.

[57] M. Talerico and S. M. Berget, "Intron definition in splicing of small drosophila introns." *Molecular and cellular Biology*, vol. 14, no. 5, pp. 3434–3445, 1994.

[58] M. Amit, M. Donyo, D. Hollander, A. Goren, E. Kim, S. Gelfman, G. Lev-Maor, D. Burstein, S. Schwartz, B. Postolsky *et al.*, "Differential GC content between exons and introns establishes distinct strategies of splice-site recognition," *Cell reports*, vol. 1, no. 5, pp. 543–556, 2012.

[59] G. J. Goodall and W. Filipowicz, "Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants." *The EMBO journal*, vol. 10, no. 9, p. 2635, 1991.

[60] Z. Wang, X. Xiao, E. Van Nostrand, and C. B. Burge, "General and specific functions of exonic splicing silencers in splicing control," *Molecular cell*, vol. 23, no. 1, pp. 61–70, 2006.

[61] Y. Cui, C. Zhang, and M. Cai, "Prediction and feature analysis of intron retention events in plant genome," *Computational Biology and Chemistry*, 2017.

[62] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[63] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[64] P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez *et al.*, "The arabidopsis information resource (TAIR): improved gene annotation and new tools," *Nucleic acids research*, vol. 40, no. D1, pp. D1202–D1210, 2012.

[65] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng *et al.*, "The TIGR rice genome annotation resource: improvements and new features," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D883–D887, 2007.

[66] "Sorghum bicolor," http://phytozome.jgi.doe.gov/, DOE-JGI, [Online; accessed 27-April-2017].

[67] S. E. Abdel-Ghany, M. Hamilton, J. L. Jacobi, P. Ngam, N. Devitt, F. Schilkey, A. Ben-Hur, and A. S. Reddy, "A survey of the sorghum transcriptome using single-molecule long reads," *Nature communications*, vol. 7, 2016.

[68] M. F. Rogers, J. Thomas, A. S. Reddy, and A. Ben-Hur, "Splicegrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data," *Genome biology*, vol. 13, no. 1, p. R4, 2012.

[69] R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," *Nucleic acids research*, p. gkq1019, 2010.

[70] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[71] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[72] G. D. Stormo, "Modeling the specificity of protein-DNA interactions," *Quantitative biology*, vol. 1, no. 2, p. 115, 2013.

[73] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.

[74] A. R. Statnikov and C. F. Aliferis, "Are random forests better than support vector machines for microarray-based cancer classification?" in *AMIA*. Citeseer, 2007.

[75] V. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson, and C. Jeganathan, "Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture," *Remote Sensing of Environment*, vol. 121, pp. 93–107, 2012.

[76] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[77] M. R. Mendoza, G. C. da Fonseca, G. Loss-Morais, R. Alves, R. Margis, and A. L. Bazzan, "Rfmirtarget: predicting human microRNA target genes with a random forest classifier," *PloS one*, vol. 8, no. 7, p. e70153, 2013.

[78] G. Nimrod, A. Szilágyi, C. Leslie, and N. Ben-Tal, "Identification of DNA-binding proteins using structural, electrostatic and evolutionary features," *Journal of molecular biology*, vol. 387, no. 4, pp. 1040–1053, 2009.

[79] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[80] M. Masso and I. I. Vaisman, "Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms," *Journal of Theoretical Biology*, vol. 266, no. 4, pp. 560–568, 2010.

[81] V. Nair, M. Dutta, S. S. Manian, R. Kumari, and V. K. Jayaraman, "Identification of penicillin-binding proteins employing support vector machines and random forest," *Bioinformation*, vol. 9, no. 9, p. 481, 2013.

[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[83] R. Grimm, T. Behrens, M. Märker, and H. Elsenbeer, "Soil organic carbon concentrations and stocks on barro colorado islanddigital soil mapping using random forests analysis," *Geoderma*, vol. 146, no. 1, pp. 102–113, 2008.

[84] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[85] B. Scholkopf, "The kernel trick for distances," *Advances in neural information processing systems*, pp. 301–307, 2001.

[86] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.

[87] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein–protein interactions," *Bioinformatics*, vol. 21, no. suppl 1, pp. i38–i46, 2005.

[88] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2126–2136.

[89] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[90] B. D. O?Fallon, W. Wooderchak-Donahue, and D. K. Crockett, "A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data," *Bioinformatics*, vol. 29, no. 11, pp. 1361–1366, 2013.

[91] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev, "Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 2048–2056, 2003.

[92] C.-Y. Cheng, V. Krishnakumar, A. Chan, F. Thibaud-Nissen, S. Schobel, and C. D. Town, "Araport11: a complete reannotation of the arabidopsis thaliana reference genome," *The Plant Journal*, 2016.

[93] R. Zhang, C. P. Calixto, Y. Marquez, P. Venhuizen, N. A. Tzioutziou, W. Guo, M. Spensley, N. Frei dit Frey, H. Hirt, A. B. James *et al.*, "AtRTD2: A reference tran-

script dataset for accurate quantification of alternative splicing and expression changes in arabidopsis thaliana RNA-seq data," 2016.

[94] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[95] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam *et al.*, "Phytozome: a comparative platform for green plant genomics," *Nucleic acids research*, vol. 40, no. D1, pp. D1178–D1186, 2011.

[96] H. Liu, T. Jin, J. Guan, and S. Zhou, "Histone modifications involved in cassette exon inclusions: a quantitative and interpretable analysis," *BMC genomics*, vol. 15, no. 1, p. 1148, 2014.

[97] Y. Zhou, Y. Lu, and W. Tian, "Epigenetic features are significantly associated with alternative splicing," *BMC genomics*, vol. 13, no. 1, p. 123, 2012.

[98] F. Ullah, A. Reddy, and A. Ben-Hur, "Exploring the relationship between intron retention and dnase i hypersensitivity in plants," in *Integrative RNA Biology Special Interest Group Meeting*, 2016, p. 31.

Appendix A

PACKAGES AND COMMANDS

1. **Retrieval of RNA-Seq data:**

   Package: SRA toolkit

   Version number: 2.8.1-3

   Command:

```
fastq-dump --outdir path_to_output_directory --split-files
   path_to_sra_study
```

2. **Alignment of reads:**

   Package: STAR

   Version number: 2.5.2b

   Commands:

   i. Generate genome index:

```
STAR --runThreadN 8 --runMode genomeGenerate --genomeDir
   path_to_genome_dir --genomeFastaFiles path_to_fasta --
   sjdbGTFfile path_to_annotation --sjdbOverhang 99
```

   ii. Map the reads using:

```
STAR --runThreadN N --genomeDir path_to_genome_dir --
   readFilesIn path_to_fastq_files --
   outFilterScoreMinOverLread 0.95
```

3. **Extract unique reads**:

   Package: samtools

   Version number: 1.8.0

   Command:

```
samtools view -h Aligned.bam | grep -e '@SQ' -e '@HD'  -e '@PG'
    -we 'NH:i:1' | samtools view -Sbh -> unique.bam&
```

4. **Remove false positive splice junctions**:

   Package: SpliceGrapher

   Version number: 0.2.5

   Command:

```
python SpliceGrapher-0.2.5/scripts/sam_filter.py
   path_to_bam_file path_to_classifier -o output_file  -f
   path_to_annotation_file
```

5. **Getting Labels**:

   Package: iDiffIR

   Version number: 0.3.1

   Command:

```
python idiffir/scripts/idiffir.py -l Mutant Wildtype -o
   path_to_output_dir path_to_annotation_file mt_filtered_rep1.
   bam:mt_filtered_rep2.bam  wt_filtered_rep1.bam:
   wt_filtered_rep2.bam
```

   where, mt_filtered_rep1.bam and mt_filtered_rep2.bam are the filtered alignments for the mutant, and wt_filtered_rep1.bam and wt_filtered_rep2.bam are the filtered alignments for the wildtype.

6. **Version of other packages used**:

   scikit-learn: 0.17.1

   scipy: 0.17.1

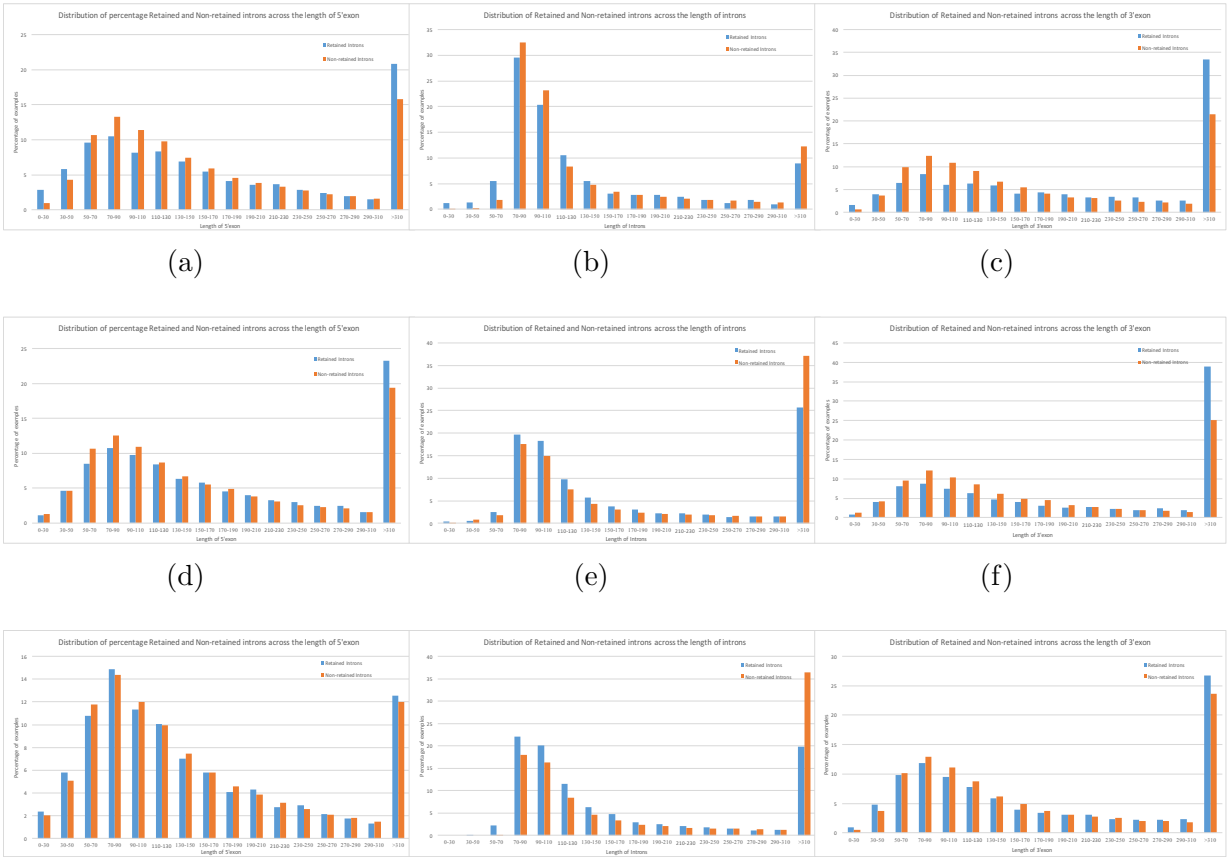
   numpy: 1.11.0

   matplotlib: 1.5.1

   pysam: 0.9.1.4

Appendix B

DATA DETAILS



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure B.1: Comparison of the distribution of retained and non-retained examples across the length of intron and flanking exons in the three species. Rows correspond to different species ($A.thalia, O.sativa$ and $S.bicolor$ respectively). Columns correspond to the distribution across the length of 5'exon, intron and 3'exon respectively.

Appendix C

PREDICTING DIFFERENTIAL IR

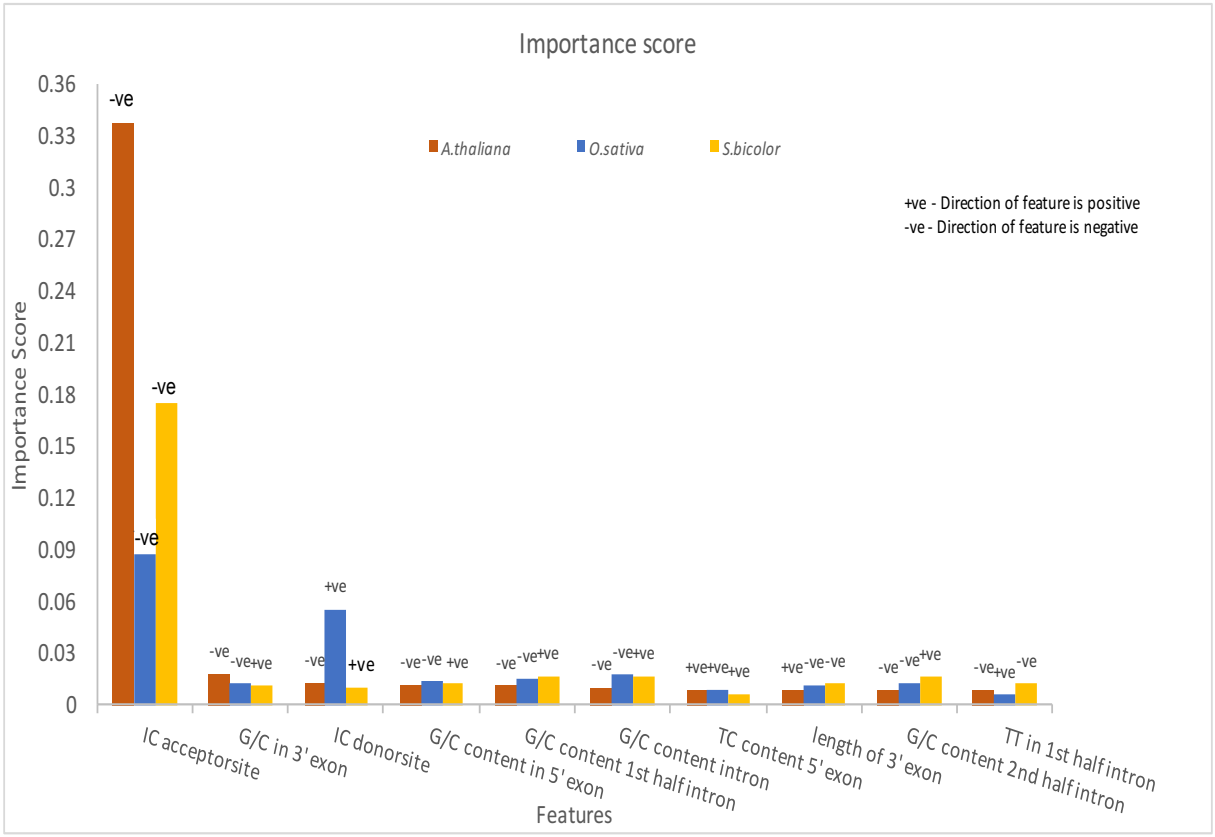


Figure C.1: Importance score and direction of top features of random forest for predicting down-regulated and no-change examples for *A.thaliana*, *O.sativa* and *S.bicolor*. The ordering of the top features on x-axis is based on the ranking of *A.thaliana*. The direction of features is indicated by the symbols '+ve' and '-ve' on top of each bar.

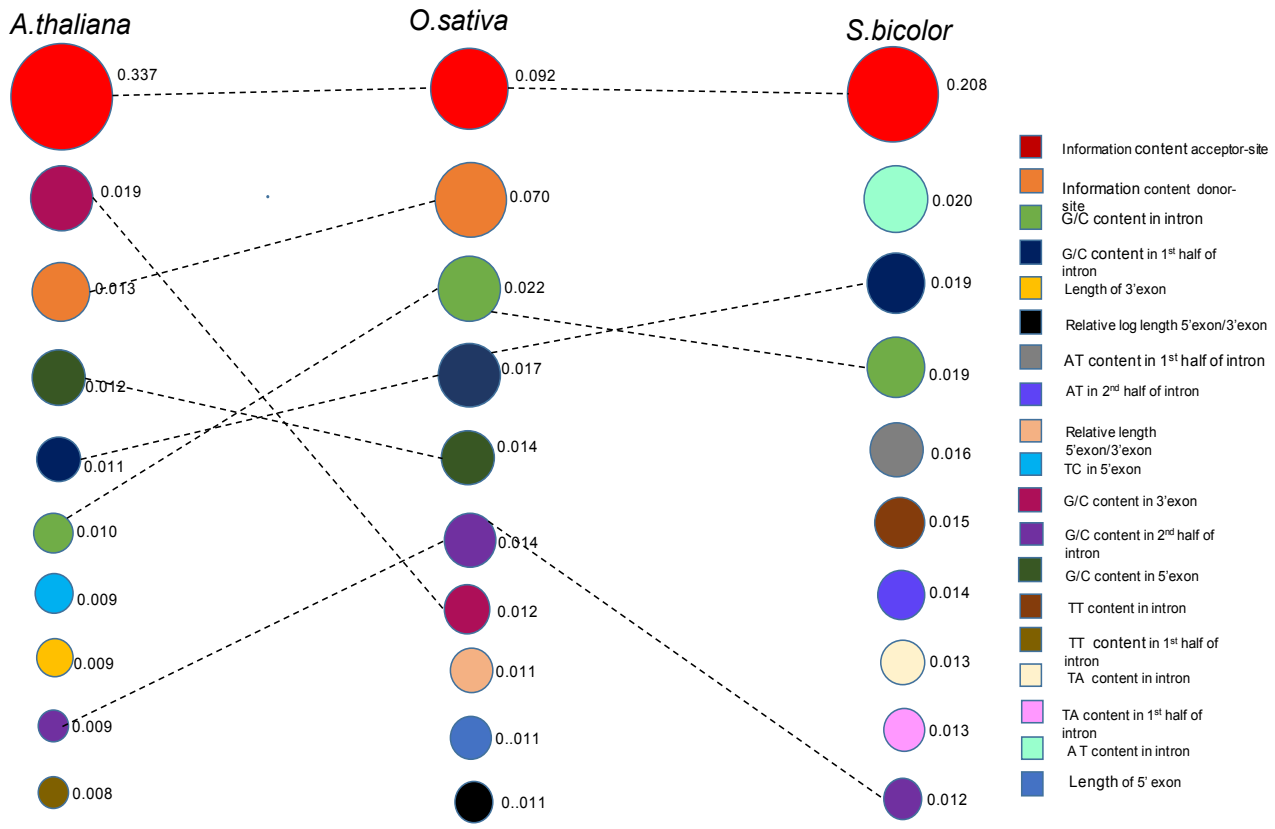


Figure C.2: Comparison of top 10 features random forest model for predicting Down-regulated introns from No-change introns in the three species. Each circle represents a feature and the importance of the feature is indicated by its size. The columns represent the top features of each species and the rows represents the ranking of the features. The dotted lines indicate common features across species.

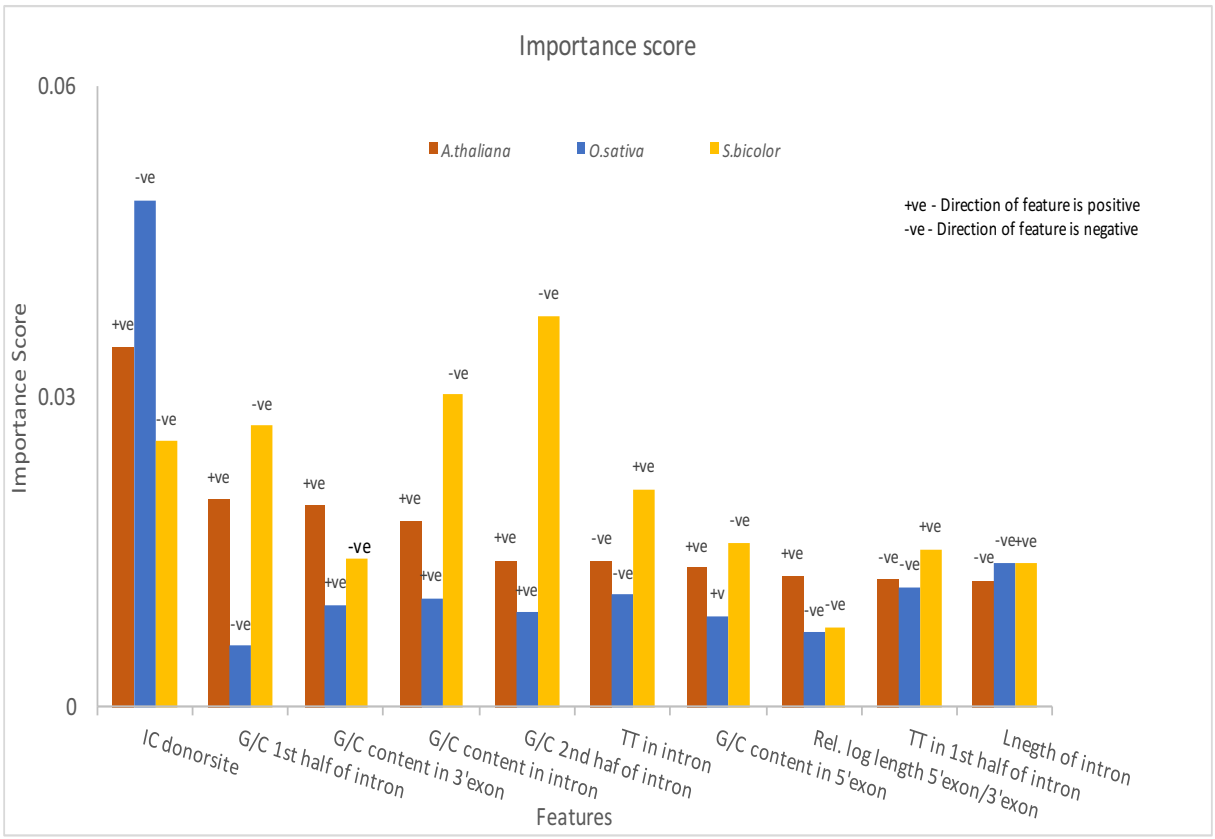


Figure C.3: Importance score and direction of top features of random forest for predicting up-regulated and down-regulated examples for *A.thaliana*, *O.sativa* and *S.bicolor*. The ordering of the top features on x-axis is based on the ranking of *A.thaliana*. The direction of features is indicated by the symbols '+ve' and '-ve' on top of each bar.

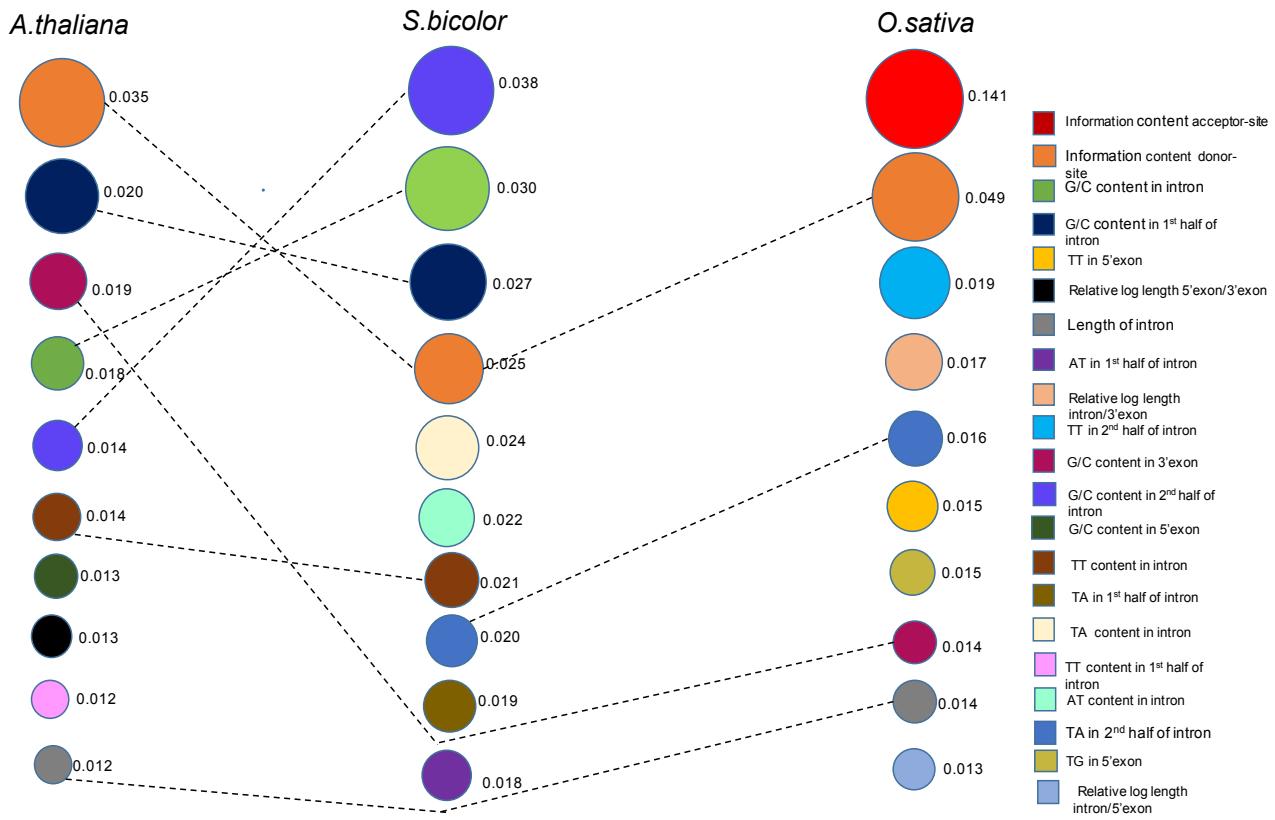


Figure C.4: Comparison of top 10 features random forest model for predicting Up-regulated introns from Down-regulated introns in the three species. Each circle represents a feature and the importance of the feature is indicated by its size. The columns represent the top features of each species and the rows represents the ranking of the features. The dotted lines indicate common features across species.