Identifying Splicing Regulatory Elements with de Bruijn Graphs

Eman Badr

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science

Lenwood S. Heath, Chair
Ruth Grene
Clifford A. Shaffer
Liqing Zhang
Mahmoud M. ElHefnawi

May 6, 2015
Blacksburg, Virginia

Keywords: Alternative splicing, de Bruijn graphs, algorithms, graph mining, splicing regulatory elements

Identifying Splicing Regulatory Elements with de Bruijn Graphs

Eman Badr

(ABSTRACT)

Splicing regulatory elements (SREs) are short, degenerate sequences on pre-mRNA molecules that enhance or inhibit the splicing process via the binding of splicing factors, proteins that regulate the functioning of the spliceosome. Existing methods for identifying SREs in a genome are either experimental or computational. This work tackles the limitations in the current approaches for identifying SREs. It addresses two major computational problems, identifying variable length SREs utilizing a graph-based model with de Bruijn graphs and discovering co-occurring sets of SREs (combinatorial SREs) utilizing graph mining techniques. In addition, I studied and analyzed the effect of alternative splicing on tissue specificity in human.

First, I have used a formalism based on de Bruijn graphs that combines genomic structure, word count enrichment analysis, and experimental evidence to identify SREs found in exons. In my approach, SREs are not restricted to a fixed length (i.e., k-mers, for a fixed k). Consequently, the predicted SREs are of different lengths. I identified 2001 putative exonic enhancers and 3080 putative exonic silencers for human genes, with lengths varying from 6 to 15 nucleotides. Many of the predicted SREs overlap with experimentally verified binding sites. My model provides a novel method to predict variable length putative regulatory elements computationally for further experimental investigation.

Second, I developed CoSREM (Combinatorial SRE Miner), a graph mining algorithm for discovering combinatorial SREs. The goal is to identify sets of exonic splicing regulatory elements whether they are enhancers or silencers. Experimental evidence is incorporated through my graph-based model to increase the accuracy of the results. The identified SREs do not have a predefined length, and the algorithm is not limited to identifying only SRE pairs as are current approaches. I identified 37 SRE sets that include both enhancer and silencer elements in human genes. These results intersect with previous results, including some that are experimental. I also show that the SRE set *GGGAGG* and *GAGGAC* identified by CoSREM may play a role in exon skipping events in several tumor samples.

Further, I report a genome-wide analysis to study alternative splicing on multiple human tissues, including brain, heart, liver, and muscle. I developed a pipeline to identify tissue-specific exons and hence tissue-specific SREs. Utilizing the publicly available RNA-Seq data set from the Human BodyMap project, I identified 28,100 tissue-specific exons across the four tissues. I identified 1929 exonic splicing enhancers with 99% overlap with previously published experimental and computational databases. A complicated enhancer regulatory network was revealed, where multiple enhancers were found across multiple tissues while some were found only in specific tissues. Putative combinatorial exonic enhancers and silencers were discovered as well, which may be responsible for exon inclusion or exclusion across tissues. Some of the enhancers are found to be co-occurring with multiple silencers and vice versa, which demonstrates a complicated relationship between tissue-specific enhancers and silencers.

This work received support from NSF grant DBI-1062472, and the VT-MENA program of Egypt.

Dedication

To my kids, who I love the most, Omar and Adham

Acknowledgments

First of all, I would like to thank the Almighty God,

I would like also to express my gratitude and thanks to my advisor Prof. Heath, for his time, guidance, continuous encouragement, and valuable discussions on my dissertation work through the past four years. Lenny, you have been a great support to me and without you, I would not have been able to stay focused and finish my PhD work. It would take more than few words to express my sincere gratitude to you.

I thank my committee members, Prof. Grene, Prof. Shaffer, Prof. Zhang, and Prof. ElHefnawi for their support, cooperation and comments to improve my work all along the way. Special thanks to Prof. Grene who always found a time for me to meet and discuss despite her very busy schedule. She always supported me and provided me with valuable ideas to verify my computational methods from biological perspective.

I am eternally in debt to my parents, without them I could not be able to complete my PhD. Special thanks to my dear mother for her unconditional love, support, her caring after my kids when I really needed her. She traveled 7000 miles just for helping me. No words can ever give her what she deserves. Thanks to my father for always encouraging me, praying for me and helping in any way he can, my beloved sisters: Nashwa, Nahla and Naglaa for continuous support and encouragement.

My beloved husband Tamer who I can't find words for his support, sacrifices and trying to make it work for me to get my PhD. He was going back and forth between our home country and the United States. So, we can spend a little family time. Not every husband will do that. So, I'm very grateful for having him in my life.

My sincere gratitude to all my friends who I met here in the United States for their unlimited support, love, and help whenever I needed. They are my other family. I owe them so much.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Splicing regulatory elements play a vital role in facilitating accurate splice site recognition through the recruitment of splicing factors. The relationship between SREs and the activation of splicing is complex and incompletely understood. Therefore, the identification of such SREs would greatly enhance our capability to understand regulation and even to predict splicing patterns in specific biological contexts [108]. SREs also have an important role in generating tissue specificity [24, 104].

However, there are some problematic aspects in the computational methods that are used to identify SREs. Most of the work done assumes a predefined size for cis-elements. SREs are usually 4 to 18 nucleotides long [49], but most SRE studies have focused on 6-mers [3, 24, 72, 73, 79, 103]. Some utilized 7-mers instead [60] or 5-mers [111]. In [88, 110], they focused on 7-mers and 8-mers. There is no evidence that SREs should be a fixed size or all of the same length. SpliceAid-F [46] is a recent comprehensive database that includes all known splicing factors and their known binding sites that were experimentally assessed. SpliceAid-F contains binding site sequences for different organisms. That includes human, mouse, chicken, rat, and rabbit. It is clear that the experimentally verified SREs vary in length. Therefore, assuming a predefined size beforehand can lead to inaccurate results, especially when the SRE frequency is a key part in the analysis, as SRE length affects frequency. What is known is that the motifs recognized by SR proteins are short and degenerate [72].

Furthermore, many studies have focused on individual motifs [60]. However, many AS events do not involve only individual regulators. Zhang et al.[109] showed experimentally that knocking out (mutating) from two to four ESEs will affect the splicing efficiency dramatically. It is a more complex process that includes cooperative or competitive interplay between splicing enhancers and silencers. Most tissue-specific alternative splicing events studied so far seem to be regulated by a more complex group of regulators [66, 99]. In general, identifying individual cis regulatory elements are not enough to understand tissue-specific or condition-specific alternative splicing.

Recent models have studied combinatorial SREs in AS regulation [57, 103], but some of the models did not exploit the expression data and focused only on frequently co-occurring SREs and all the methods concentrated on SRE pairs only [43, 57, 87, 103].

Moreover, for simplicity, most of the models have focused on a certain type of alternative splicing. Most of the work done was to define cis-elements for cassette exons only [24, 60, 72, 73, 103, 104, 108] or intron retention only [3, 79] or both [37]. In addition, some models were built to identify specific cis-elements (either enhancers or silencers) in specific regions (either intronic or exonic regions).

There is a need for new methods that can detect SRE sequences and take into consideration all the factors that affect splicing. In Chapter 4, I introduce a new formalism for modeling SREs based on de Bruijn graphs. The proposed model utilizes de Bruijn graphs to identify exonic splicing elements of variable length and entails word count enrichment analysis. Different data sources are incorporated to accurately identify SREs [14]. As a result, my model identifies 2001 putative exonic enhancers and 3080 putative exonic silencers for human genes, with lengths varying from 6 to 15 nucleotides. Many of the predicted SREs overlap with experimentally verified binding sites. My model provides a novel method to predict variable length putative regulatory elements computationally for further experimental investigation.

In Chapter 5, I develop CoSREM, a two-level graph mining algorithm to identify combinatorial SREs and their effect on the splicing process. The proposed method identifies sets of co-occurring splicing elements. It is not limited to discovering SRE pairs. Experimental evidence is utilized here as well to increase the accuracy of the results [15]. CoSREM identified 37 SRE sets that include both enhancer and silencer elements. These results intersect with previous results, including some that are experimental. I also show that the SRE set *GGGAGG* and *GAGGAC* identified by CoSREM may play a role in exon skipping events in several tumor samples.

Chapter 6 reports on a genome-wide analysis I performed to study alternative splicing on multiple human tissues (brain, heart, liver, and muscle). The RNA-Seq data set from the Human BodyMap project [42] was utilized. I used DEXSeq [5] to identify tissue-specific exons. Then, I applied my algorithms, GenSRE [14] and CoSREM [15], to identify both individual and combinatorial regulatory elements responsible for exons that exist in one tissue but not in other tissues. Using this approach, 28,100 tissue-specific exons were identified across the four tissues. I identified 1929 exonic splicing enhancers with 99% overlap with previously published experimental and computational databases. A complicated enhancer regulatory network was revealed, where multiple enhancers were found across multiple tissues while some were found only in specific tissues. Putative combinatorial exonic enhancers and silencers were discovered as well, which may be responsible for exon inclusion or exclusion across tissues. Some of the enhancers are found to be co-occurring with multiple silencers and vice versa, which demonstrates a complicated relationship between tissue-specific enhancers and silencers [13]. This is, to my knowledge, the first analysis to focus only on discovering exonic regulatory elements (individual and combinatorial) across tissues.

Chapter 2

Biological Background

Alternative splicing (AS) is the main reason behind the difference between the estimated 24,000 protein-coding genes in the human genome and the estimated 100,000 different proteins that are synthesized [59, 108]. It is a post-transcriptional mechanism for regulating gene expression and generating proteomic diversity [25, 104]. The eukaryotic genes consist of coding regions, which are usually called exons, that are surrounded by noncoding regions, which are called introns. During transcription, the primary transcript RNA undergoes comprehensive modifications to form mature mRNA. During this process, the introns in different genes are removed (spliced out) from RNA transcripts, and exons are recombined to form one functional mRNA transcript [77]. The primary transcript RNA is called precursor mRNA (pre-mRNA). With the removal of introns, the pre-mRNA eventually becomes mature mRNA, which marks the end of gene transcription. In alternative splicing, genes generate different mRNA isoforms by selecting different combinations of exons [24, 65], as indicated in Figure 2.1.

The RNA splicing process depends on the recognition of specific sequence elements in pre-mRNAs. These elements are called splicing signals (Figure 2.2). Correct recognition of splice sites by the spliceosome (the machinery responsible for assembling the pre-mRNA) is critical for proper excision of introns from the primary transcript [99]. The splicing signals are the consensus sequences that act as signals to the spliceosome to splice out the intron region. There are four main sites: the 5' splice site, the 3' splice site, the branch point sequence (BPS), and the polypyrimidine tract (U and C nucleotides). These sites participate in the splicing process and are present in every intron, and thus are known as the core splicing signals [99]. Most introns have GU at their 5' splice site and AG at the 3' end. The nucleotide at the branch point is always an A, while the consensus around this point somewhat varies [109]. In humans, the BPS is yUnAy [44], where y represents pyrimidines, as they are not as well conserved as A and U. The BPS is followed by a series of pyrimidines, especially uracil, then by AG at the 3' end. The polypyrimidine region is usually 15-20 base pairs long, located about 21-34 base pairs before the 3' end of the intron to be spliced. In plants, there

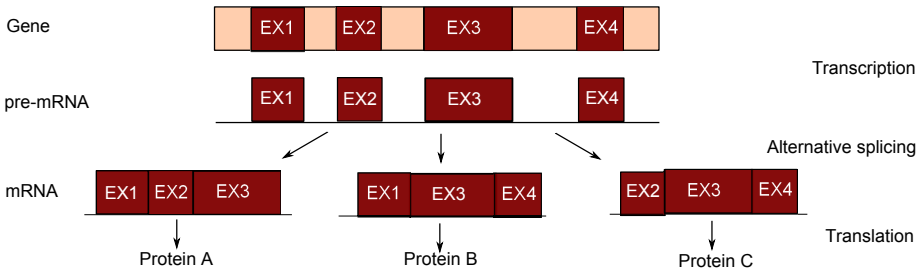


Figure 2.1: In alternative splicing, one gene can produce multiple isoforms.

is another element, the UA-rich tract (U and A nucleotides), which is required for effective splicing of introns and improves splicing of U12 introns [59, 88]. U12 introns are another class of introns that have different consensus dinucleotides at the 5' and 3' splice sites, respectively (AU-AC instead of GU-AG) [83]. These signals are necessary for spliceosome assembly, but often insufficient to determine the splice site strength or to regulate the different types of AS events [105].

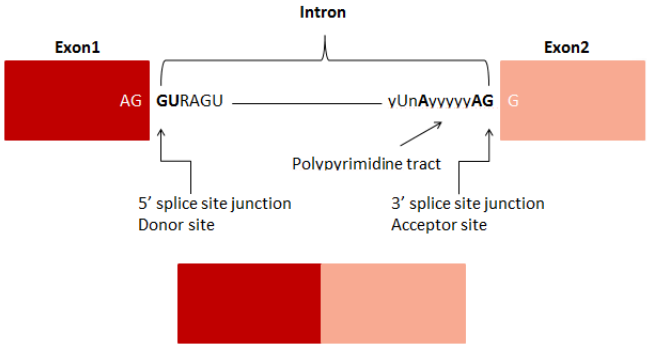


Figure 2.2: Splicing signals.

There are five types of alternative splicing [99], which are illustrated in Figure 2.3:

1. Cassette exon: In this case, an exon may be spliced out of the primary transcript or retained. This is the most common mode in mammalian pre-mRNAs. The impact of alternative splicing on different levels of the eukaryotic evolutionary tree is not the same. It is more prevalent in higher eukaryotes than in lower eukaryotes, and the percentage of genes and exons that undergo alternative splicing is higher in vertebrates than in invertebrates [59].

2. Retained intron: A sequence may be spliced out as an intron or retained. Intron retention is most common in plants, lower metazoans, and fungi [59].

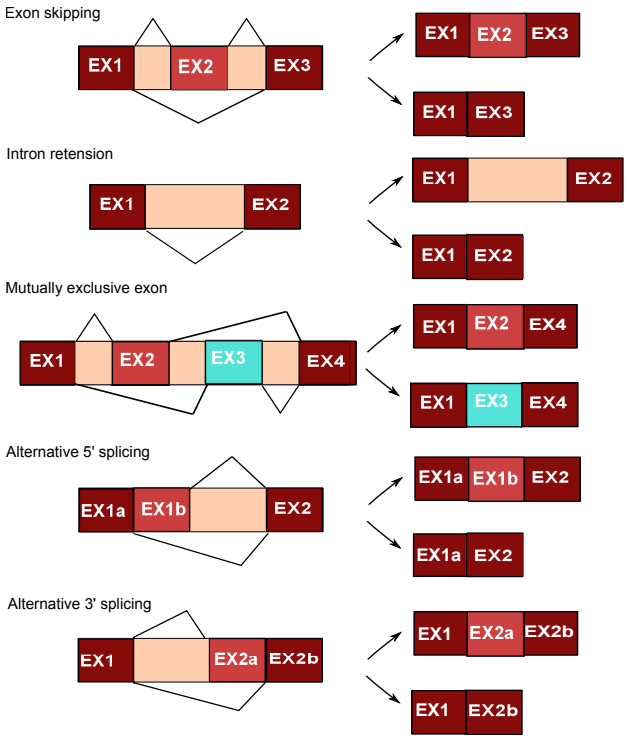


Figure 2.3: Different types of alternative splicing.

3. Mutually exclusive exons: One of two exons is retained in mRNAs after splicing, but not both.

4. Alternative donor site: An alternative 5' splice junction (donor site) is used, changing the 3' boundary of the upstream exon.

5. Alternative acceptor site: An alternative 3' splice junction (acceptor site) is used, changing the 5' boundary of the downstream exon.

In addition to the core splicing signals, other splicing regulatory elements (SREs) are pivotal to ensure that splicing events occur accurately and efficiently. A pair of strong splice sites is not sufficient to define an exon, as many pseudo-exons that are flanked by predicted splice sites are not spliced [66]. SREs are classified as exonic splicing enhancers (ESEs) or silencers (ESSs) if they promote or inhibit the inclusion of the exon where they reside, and as intronic splicing enhancers (ISEs) or silencers (ISSs) if they enhance or inhibit the inclusion of the exon adjacent to the intron where they reside [24, 99, 104].

Much attention has been given to ESEs. The first ESEs to be characterized were short, purine-rich motifs containing repeated GAA, or GAG trinucleotides, but after that many other sequences have been shown to have enhancer activities [72]. SREs are highly variable in sequence. Many ESEs contain binding sites for members of the SR family of proteins. SR

proteins have roles in several steps of spliceosome assembly, and function as both essential splicing factors and regulatory factors. Splicing silencers (ESSs and ISSs) are variable in sequence, but some of them bind to members of the extended family of heterogeneous nuclear ribonucleo-proteins (hnRNPs). A number of hnRNPs function as splicing repressors, such as hnRNP A1 [66, 99, 105].

As stated before, alternative splicing is a mechanism that generates different mRNA isoforms by variable use of splice sites. Therefore, the basis of splicing is the recognition of introns and exons by the splicing machinery [59]. The spliceosome, which is a molecular ribonucleoprotein complex, is the main machinery responsible for assembling the pre-mRNA [25]. It consists of five small nuclear ribonucleoprotein (snRNP) particles (U1, U2, U4, U5 and U6) as well as U2AF65 and serine-arginine rich (SR) proteins that are associated with a large number of additional proteins [59, 66, 77].

The spliceosome is considered a dynamic complex organelle as it is formed of about 100 core proteins in addition to the small nuclear RNAs [99, 105]. Splicing begins with a multi-step process of spliceosome assembly around the splice sites and the branch point. Splicing factors bound to nearby SREs can influence spliceosome assembly by facilitating or inhibiting the subunits of the spliceosome to recognize the splice sites [66, 103]. The first step in AS involves two complexes (U1,U2) that bind near the GU sequence. The RNA is then looped, and three other protein-RNA complexes (U4, U5, U6) bind. This final complex then undergoes a conformation change. The intron is then cleaved at the 5' GU sequence and forms a lariat at the A branch site. The 3' end of the intron is next cleaved at the AG sequence, and the two exons are ligated together. As the spliced mRNA is released from the spliceosome, the intron debranches (resolves) and is degraded [36, 77, 99]. In many eukaryotes, including most plant and animal species, there are two types of spliceosomes, the major and minor ones. The major spliceosome is responsible for splicing of the vast majority of introns in both plants and animals. The minor spliceosome participates in the splicing process of U12 introns and splices out 0.3% of introns in humans [88].

As stated earlier, there are specific proteins that bind to SREs, which are called splicing factors. One type of splicing factor is the SR proteins and they bind to enhancers. SR proteins have roles in several steps of spliceosome assembly. For example, as illustrated in Figure 2.4, the central exon has ESE and ESS elements in proximity. If an SR protein is present, it can bind to the ESE and stimulate exon inclusion by recruiting other splicesome proteins U1, and U2 to the core splicing signals. As a result, splicesome machineries are assembled, and the central exon is included. On the other hand, there may be another case if the other type of splicing factor is present, such as an hnRNP splicing factor that acts as a splicing repressor and inhibits the exon inclusion. This is performed by an hnRNP splicing factor binding to the silencer sequence and recruiting the binding of other inhibitory factors, which extends to the exon boundary and prohibits the binding of the SR factor. As a result the exon will be skipped.

Many studies have identified that splicing and alternative splicing regulate almost every bio-

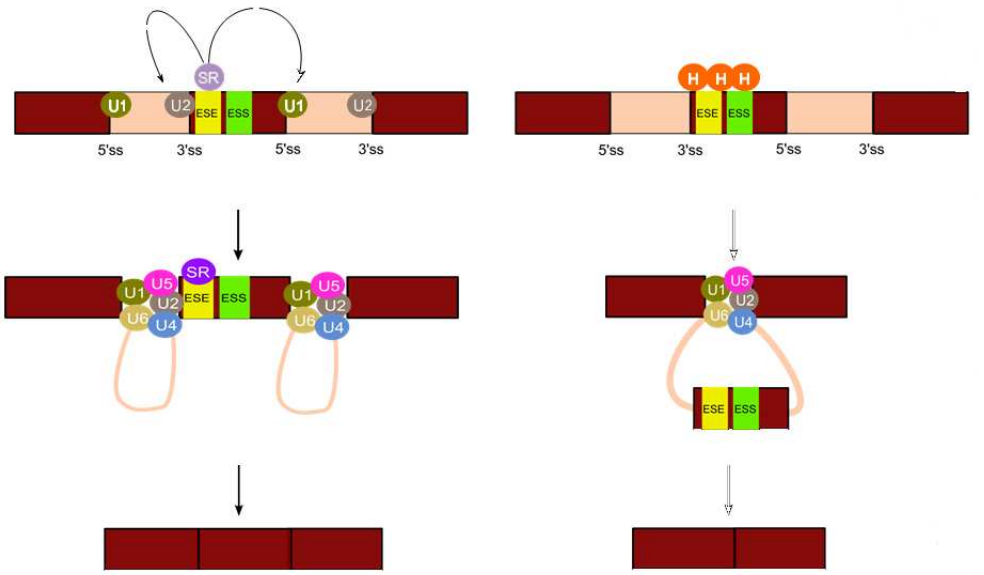


Figure 2.4: Effect of splicing factors on AS decisions.

logical process such as signal transduction and energy transfer in both multicellular animals and plants [36]. Accurate splicing is crucial. It is believed that up to 50% of human genetic diseases are the result of mutations either in consensus splice site sequences or in the exon and intron splicing enhancers and silencers [16, 41, 59, 65, 66]. For instance, alternative splicing is involved in familial isolated GH deficiency type II (IGHD II), Frasier syndrome, myotonic dystrophy diseases [60], neurodegenerative diseases (spinal muscular atrophy via SMN splicing), and frontotemporal dementia with parkinsonism-17 (FTFP-17) via tau splicing. Furthermore, it is involved in causing different types of cancers [62] such as breast and lung cancers via the Bcl-x gene [36]. For a detailed review, refer to [6, 45]. Another example that clarifies the importance of AS is the behavior of Sex-lethal (Sxl) and transformer (tra) genes in D. melanogaster. In each gene, sex-specific splicing gives rise to a functional protein product in females, but in males alternative splicing leads to the inclusion of stop codons, so that no functional protein is produced [66, 99]. As a result, AS can play an important role in generating tissue specificity [104].

The splicing process is influenced by several factors such as exon size, trans-factors (e.g. SR proteins), and cis-regulatory elements [59, 60]. For instance, there are various features that distinguish alternatively spliced exons from constitutive ones such as sequence conservation. The sequences of alternatively spliced cassette exons are more conserved than constitutive exons. The conservation is higher towards the exon boundaries. These sequences presumably direct the splicing machinery to the correct exon-intron junctions. Conservation extends from 80 to 100 bases into the introns flanking the alternatively spliced exons [59].

Another important factor is the regulatory elements. These sequences are also more con-

served in alternatively spliced exons than in constitutively spliced ones. Cassette exons also usually have weaker splice sites than constitutive exons and thus enhancers are needed to promote splicing in specific conditions [59]. The function of an SRE may depend on its relative location. For example, G triplets commonly enhance splicing from intronic locations, but they function as splicing silencers when located in exons. It is not only where SREs are located that promotes AS, but also the distance between the SRE and the adjacent splice site. In the same way, some SR proteins can promote splicing when bound to sites (ESEs) in exons, and also inhibit splicing when bound to intronic (ISS) sites [99, 105]. SRE behavior also changes according to its neighboring sequences. It has two categories: location-dependent activity, in which SRE behavior changes with relative positions in the pre-mRNA, and gene-dependent activity, where activity observed in one gene is lost when the SRE is moved to another gene [99].

In summary, despite the important role SREs play in alternative splicing, their nature imposes many challenges on accurate identification of SREs. Hence, there is a need for new approaches that can tackle the limitation of current approaches for identifying SREs.

Chapter 3

Literature Review

3.1 Experimental Approaches

There have been several experimental large-scale studies of AS. Several techniques were utilized to identify SREs such as systematic evolution of ligands by exponential enrichment (SELEX) [31], UV crosslinking and immunoprecipitation(CLIP) [92], and minigene-based systems [100].

SELEX experiments have been carried out with a number of SR proteins and hnRNPs [66, 34]. Optimal binding sites vary between individual SR proteins, where they often function as splicing enhancers [89, 90, 32, 64, 29, 81, 63, 27]. In these experiments, random oligomers were inserted into a poorly splicing exon and the effective sequences were then collected from spliced molecules after iterative selection. In several of these studies, splicing was assayed so as to depend on the presence of a single SR protein. In this way it was possible to identify not only the ESE sequences but also the SR proteins (SRp40, SRp55, SC35, ASF/SF2, 9G8) that targeted those sequences. Similarly, optimal binding sites for repressor hnRNPs are known as splicing silencers [1].

On the other hand, CLIP has allowed the identification of the binding sites *in vivo* of several splicing factors, such as NOVA [92], SRSF1 (ASF/SF2) [80], hnRNP A1 [50], and TDP-43 [91]. This is by allowing the isolation of RNA fragments that are directly bound by an RNA binding protein. There are 655 human splicing factor binding sites that are experimentally assessed currently [46].

Furthermore, there is the minigene-based technology that utilizes the natural transcriptional and splicing machinery of the original environment and concentrates only on the genomic segment from the gene of interest [101]. They are probably the most common methodology used in both basic science research and the clinical setting [25]. Several studies exploited minigene technology for identifying SREs [110, 109, 17, 57]. For example, Ke et al. [58]

placed all 4096 6-mers at five different sites in two model exons and the inclusion ratio of the central exon for each pre-mRNA molecule was measured. The authors identified 1182 6-mer ESEs and 1090 6-mer ESSs. I have used their data in my research.

Wang et al. [100] used the same approach to identify ESSs using a three-exon minigene. In this screen, which is called a fluorescence activated screen (FAS), GFP protein was utilized. A set of random 10-mers were inserted into the central exon, which was then transfected into cells. The cells were then sorted by the GFP signal level. One hundred and forty-one ESS 10-mers were isolated from this screen. By avoiding the iterative enrichment process of SELEX, any sequences that inhibit splicing were isolated, rather than only those that have the strongest ESS activity. Wang et al. [97] utilized the same approach to identify ISS, while they tried to identify all different types of SREs in [98].

3.2 Identifying Individual SREs

3.2.1 Word enrichment and statistical analysis of frequency

The word count enrichment-based approach is one of the the key computational techniques used. It identifies SREs as short nucleotide sequences (typically 6-mers or 8-mers) that are statistically enriched in a carefully selected set of exons and/or introns against a background or negative data set. However, this type of analysis heavily relies on the accurate selection of the background sequences.

For example, in the RESCUE-ESE (relative enhancer and silencer classification by unanimous enrichment) approach [39], 6-mers were identified in constitutive human exons by enrichment in exons versus introns and in exons with weak splice sites versus exons with strong splice sites. Determining whether the splice site is strong or weak was based on the summation of the log-odds score calculated for each position between -3 and +6 over the exon-intron junction. Using stringent cutoffs, 238 distinct 6-mers were identified as possible ESEs, then clustered into ten motifs.

Zhang et al. [110] utilized non-coding exons, exons that are not involved in protein synthesis although they exist in the pre-mRNA, instead of protein-coding exons. The authors argument was that protein coding exons are non-random as the distribution of the amino acids in these proteins is not random. However, this information is not related to splicing and hence can be considered as noise and affects finding SREs [112]. They compared the frequencies of 8-mers (allowing one mismatch) in constitutively spliced non-coding exons with those in pseudo-exons, intronic sequences that, although they are flanked by obvious consensus splice sites, they are not observed in spliced mRNAs, and the 5' untranslated regions (UTRs) of intronless genes. Their hypothesis was that the over-represented sequences in the non-coding exons can serve as ESEs, and the under-represented ones as putative ESSs. The authors computed z-scores of all the 8-mers of the non-coding exons versus the other two sets and

the 8-mers with a z-score above or below a pre-specified threshold were considered putative enhancers and silencers, respectively. Two thousand ninety-six putative enhancers and 974 putative silencers were identified and then grouped using hierarchical clustering. Some of the putative splicing elements were verified experimentally [109].

The same approach was employed in [104] to identify tissue-specific SREs in mouse genes. Mouse RNA-seq data for three tissues (brain, liver and skeletal muscle) were utilized to calculate the expression level of each isoform of genes for a set of predefined cassette exons. The goal was to determine whether each exon will be in the inclusion or exclusion sets. Then, for each tissue, the frequency of each 6-mer in the inclusion set was compared with the frequency of the same 6-mer in the exclusion set using a z-score. Those 6-mers that are over-presented in one tissue but not in the other two tissues can be identified as a tissue specific SRE. The same technique was applied on the intron regions flanking the incorporated cassette exons. They identified 456 putative enhancers and silencers. Among these, 45 were common to all tissues.

Fedorov et al. [40] compared the frequencies of 4-mers and 5-mers in exons to those in intronless genes, reasoning that, while both code for proteins the former require splicing signals, the latter do not. They identified 23 sequences that were significantly more abundant in exons.

Sironi et al. [84] utilized pseudo-exons. The authors collected a subset of pseudo-exons that was rich in predicted ESEs and then searched for over-represented 6-mers as candidates for ESSs. Their hypothesis was that some ESSs function prevent the splicing of pseudo-exons. They also chose the 6-mers that are over-represented in pseudo-exons compared to the sequences flanking the pseudo-exons to normalize the possible base compositional differences between pseudo-exon and exon regions. The identified motifs were clustered into families to generate three consensus sequences.

Pertea et al. [72] introduced another computational approach to identifying ESE motifs in the model plant *Arabidopsis thaliana*. They applied a similar approach to RESCUE-ESE to identify putative ESE 6-mers in the flanking ends of a set of known exons, and 84 potential ESE 6-mers were identified. Then, they applied the Estimated-Location-of-Pattern-Hits (ELPH) program on the 5' and 3' flanking ends of the internal exons. ELPH is a Gibbs sampling program that can identify the most common motifs in a set of sequences. The 84 putative ESEs were used as input seeds to identify motifs that contain the ESEs. Seventy-three of them were found to be significantly conserved by ELPH. Thirty-five SREs were verified experimentally. They also incorporated the SREs information into their splice predictor (GeneSplicer) as well as another program (SpliceMachine), and they found that incorporating this data enhances splice site prediction and reduces false positive recognition rates.

Studying the densities of known 6-mers and their association with specific splicing events, Galante et al. [3] focused on detecting intron retention events. They performed a large scale analysis on 21,106 known human genes and utilized both full insert cDNA and ESTs,

discovering 3127 genes with intron retention events. They argued that this data set can be used to study regulatory elements that affect retention events. Sakabe et al. [79] utilized the same procedure to detect intron retention events on human genome sequence data, mRNA, and EST sequences downloaded from the UCSC Genome Browser. Their data set includes two categories, one where the isoforms with the retained introns constitute the larger ratio for this gene (High RIF (Relative Isoform Frequency)). The other category is where the isoforms with the retained introns constitute the smaller ratio (Low RIF). They studied the density of known SREs from databases such as SELEX-ESEs and FAS-hex3 ESSs on retained introns, arguing that functional SREs will have larger densities. They found that in the high-RIF set, the density of SELEX- ESEs is higher than the ESSs density. On the other hand, in the low RIF set, the density of the ESSs was the higher one.

In [88], a large-scale analysis of plant introns and splice sites was performed. Seven plant species were studied including the model plant *Arabidopsis thaliana*. Splicing regulatory elements were identified by sliding a window of size m (input parameter) across all the intronic input sequences to identify all m-length substrings and count their occurrences in the input sequences. Then, all substrings are connected to each other by similarity. Using a position weight matrix (PWM), all connected substrings form a motif. The over-represented motifs are identified by calculating the ratio between the number of matches of corresponding PWMs in the input sequences and in the reference sequence. The reference sequence can be provided as an input parameter to the program or can be randomly generated with the probability of each residue taken from the input sequences. They have identified putative intronic cis-acting elements.

In [28], the authors designed microarrays monitoring 203,672 exons and 178,351 exon-exon junctions in 17,939 human genes. These microarrays covered all types of alternative splicing. A set of 48 human tissues was hybridized to the arrays. For the regulatory elements, they extracted nucleotide sequence in eight neighborhoods around regulated exons. The authors searched for over- and under-represented words between 4 and 7 nucleotides in size, using neighborhood-specific sequences adjacent to all monitored cassette exons as a background set. They identified 135 motifs where some of them represent *de novo* predictions and others are known.

3.2.2 Machine learning based approaches

Instead of employing statistical analyses, Zhang et al. [111] exploited a supervised machine learning technique. Specifically, support vector machines (SVM) classifiers were used to define specific sequence information that distinguish real exons from pseudo-exons. Pseudo-exons were utilized as a control set for evaluating the significance of potential splicing signal sequences. The features used to train SVMs were the existence or absence of k-mer elements in the exons. Sequences of 4-mers and 5-mers found at 50 nucleotides upstream and downstream in the exon proved to achieve the best results. Two hundred fifty-six 5-mers were

identified to be splicing elements.

Zhang et al. [106] utilized a hidden Markov model (HMM)-based algorithm for building the mCarts tool (motif-based predictor of clustered accessible RBP target sites) to predict functional RNA-binding proteins (RBP) binding sites. The features used to train the model include CLIP sequences that were identified as binding sites for the NOVA and Mbnl binding proteins, distance from neighboring sites, and conservation scores. The trained models were then used to predict clustered motif sites of Nova and Mbnl in all extended genic sequences.

3.2.3 Regression-based approaches

Regression-based approaches exploit both sequence information and gene or exon expression levels [33, 95]. However, the assumption of linearity might oversimplify this complex relationship [108]. Also, the current regression methods for AS were not developed systematically from a theoretical base, which may limit their performance [103].

In [108], a varying effect regression model on splicing elements (VERSE) was developed to predict the genome-wide intronic SREs. RNA-Seq data for 16 human tissues was used to extract alternative and constitutive exons based on exon inclusion rate in each tissue. A varying coefficient regression model was utilized to associate the inclusion rate of exons with 6-mer occurrences in intronic regions. The authors incorporated non-motif based biological features (the phyloP conservation scores) into the model as the baseline binding preference of splicing factors. Utilizing this feature enabled the model to predict the contribution of each 6-mer to the exon inclusion rate. Many motifs were uniquely discovered by VERSE but missed by linear regression. About half of the SREs (55.68%) were found to be significant only in one tissue.

3.2.4 Evolutionary based analysis and comparative genomics

A conservation-based approach utilizes comparative genomic methods to identify evolutionarily conserved motifs in introns and exons, which can also be combined with the enrichment-based approach to identify SREs. Goren et al. [49] argued that SREs would be both conserved and abundant in exons. Over-represented 6-mers were chosen, those di-codons that appeared more frequently than expected if codons were paired randomly. The 6-mers with high scores for both criteria were collected, resulting in a set of 285 splicing elements that represented the best combination of scores.

Ramalho et al. [73] employed a different approach. They investigated the evolution of 6-mers that are putatively involved with the constitutive to alternative transition in vertebrate evolution. The hypothesis was that skipped exons are accumulating slightly deleterious mutations and thus weakening the cis-splicing signals of constitutive exons. They identified 145 6-mers using rate-shift analysis on alternative exons and 198 significant motifs on constitutive exons.

The *Drosophila melanogaster* exon database (DEDB) was utilized in [24] to identify mutually exclusive (ME) exons and subsequently finds putative exonic splicing regulatory (ESR) motifs. DEBD was built from abundant expressed sequence tags (ESTs) and contains entire gene sequences of 15 insect species in splicing graph format. The ESTs-To-ESRs (E2E) method was introduced, where conserved fragments in a multiple alignment of mutually exclusive exons were searched for conserved words of length 5 to 9 bp using a sliding window algorithm to scan all columns of an exon multiple alignment with an initial window length of 5 bp. With a 100% conservation score, 482 words of length 5 to 22 bp were found.

Suyama et al. [87] analyzed multiple sequence alignments of genomic sequences and searched for conserved 5-mers in at least ten mammalian species. Eleven motifs were identified and a co-occurrence network of motif pairs was developed.

3.2.5 Other approaches

Kim et al. [60] utilized a distribution-based quantitative association rule mining to find individual exonic/intronic sequence motifs. Combinatorial cis-regulatory motifs were also discovered and the effect of motif pairs was investigated. A data set of exon skipping rates for cassette exons in 10 mouse tissues was employed. The goal of the association rule mining is to discover a relationship between a specific set of 7-mers and exon skipping rate in a specific gene in one or more tissues. Using the Apriori algorithm, the t-test, and Bonferroni's multiple testing correction, they identified several statistically significant associations between sequence motifs and tissue specific exon skipping rates. Ninety-seven interesting association rules were identified, of which 3 contain multiple 7-mers.

3.3 Identifying Combinatorial SREs

Recent methods have studied combinatorial SREs in AS regulation [57, 103], but some of them did not exploit transcript expression data and focused only on frequently co-occurring SREs. All the methods concentrated on SRE pairs only [103, 57, 43, 87].

Ke et al. [57] utilized a hypergeometric test to discover sequence pairs that are over-represented in intronic regions flanking human exons. They identified more than 60,000 5-mer sequence pairs with a p-value $\leq 10^{-4}$. Friedman et al. [43] employed a similar approach except they utilized a Poisson approximation instead of a hyper-geometric test. They identified SRE pairs at the two ends of introns in both human and mouse. A biophysical principals based regression model for the regulation of AS was developed in [103]. It captures both the main effects of individual SREs and the combinatorial effects of SRE pairs. The authors model the spliceosome assembling process with a simplified chemical reaction. All 6-mers for cassette exons were calculated in five different locations on and around each exon. The response variable was calculated for this set of exons and then the regression

model was utilized to conclude SREs and SREs interacting pairs. They employed different techniques to reduce the number of variables in the regression model and avoid over-fitting. They used lasso, adaptive lasso, refitted cross validation (RCV), and ordinary least squares (OLS). Overall, 619 different SREs and 196 SRE pairs were detected from different tissues. Their model was limited to the interaction of at most two SREs.

3.4 Tissue-specific Studies

Some studies have identified tissue-specific regulatory elements. As stated earlier, the authors in [104] utilized mouse RNA-Seq data for three tissues to identify tissue-specific SREs. The authors identified 456 putative enhancers and silencers. Among these, 45 were common to all tissues. Kim et al. [60] utilized a distribution-based quantitative association rule mining to discover a relationship between a specific set of 7-mers and exon skipping rate in a specific gene in one or more tissues. In [28], the authors designed microarrays monitoring 203,672 exons and 178,351 exon-exon junctions in 17,939 human genes. Samples from 48 human tissues were hybridized to the arrays. VERSE was introduced in [108] where RNA-Seq data for 16 human tissues was used. Approximately half of the identified SREs (55.68%) were found to be significant only in one tissue. The authors in [96] developed a linear regression model to estimate the effect of various splicing factors on exon inclusion between two tissues, and, hence, the binding sites of these splicing factors are predicted. They applied their model on data from liver and heart tissues and predicted 15 motifs that contribute to exon skipping events. The work was extended to 11 human tissues in [95]. In [23], the authors identified intronic regulatory elements that are brain-specific, while in [33] the identified regulatory elements were muscle-specific. In [16], a probabilistic approach was utilized and several intronic regulatory elements in different human tissues were identified. In [94], the authors analyzed 10 human tissues. A high frequency of tissue-specific regulation was observed for each of various alternative splicing event types, including over 60% of the analyzed skipped exons. Ke and Chasin [57] utilized a hypergeometric test to discover sequence pairs that are over-represented in intronic regions flanking human exons. They identified more than 60,000 5-mer sequence pairs with $p \leq 10^{-4}$. They showed that some pairs are associated with tissue-specific genes.

Chapter 4

Identifying Splicing Regulatory Elements with de Bruijn Graphs

4.1 Introduction

As illustrated in Chapter 3, there have been several large-scale experimental and computational studies of alternative splicing. In most of the previously stated studies, a predefined length for SREs is assumed, and the frequency of occurrence of these SREs is taken into account. What is known is that the motifs recognized by SR proteins are short and degenerate [72]. Their lengths range from 4 to 18 nucleotides [49], but most SRE studies have focused on 6-mers [3, 24, 72, 73, 79, 103]. Some utilized 7-mers [60, 88, 110] or 5-mers [111] instead. As experimental evidence indicates, SREs should not be restricted to be a fixed length. SpliceAid-F [46] is a recent database that includes all experimentally determined binding sites, from which it is clear that the experimentally verified SREs vary in length. Therefore, assuming a predefined size beforehand can lead to inaccurate results, especially when the SRE frequency is a key part in the analysis, as SRE length affects frequency.

Here, I propose a de Bruijn graph-based model to identify exonic splicing elements of variable length that entails word count enrichment analysis. The proposed model combines different data sources to accurately identify SREs. I utilize data from Ke et al. [58], who used a minigene approach to insert random 6-mers into the central exon. Based on their results, an enrichment index is calculated for all possible 6-mers, which is considered a measure of central exon inclusion ability. Utilizing these scores in my graph model, Longer k-mers can be identified. I apply my model on a data set of all known human coding exons and their flanking intronic regions to find exonic enhancers and silencers. The discovered ESEs and ESSs overlap with many of the experimentally verified splicing elements in the SpliceAid-F database [46], as well as several computationally predicted data sets.

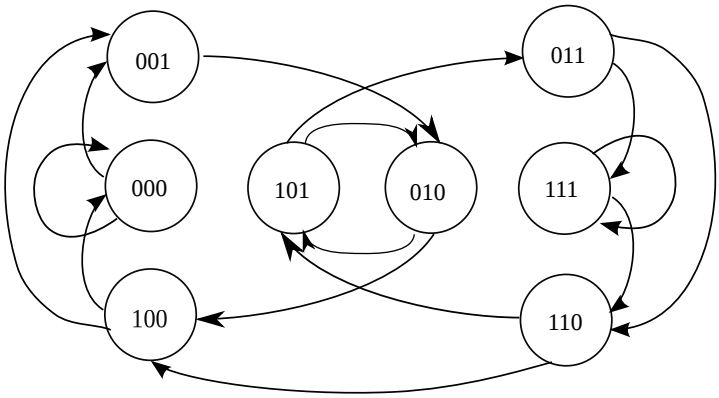


Figure 4.1: The 3-dimensional de Bruijn graph over the alphabet $\Sigma = \{0, 1\}$.

4.2 Preliminaries

I use terminology from formal language theory [55]. Let Σ be an alphabet, a finite set of symbols such as the DNA alphabet $\{A, C, G, T\}$. For $k \geq 1$, the *k-dimensional de Bruijn graph $G = (V, E)$ over* Σ is a directed graph with vertex set $V = \Sigma^k$, all length-k strings over Σ, and edge set

$$E \;=\; \{(\sigma w, w\tau) \mid w \in \Sigma^{k-1}, \sigma, \tau \in \Sigma\}.$$

In other words, an ordered pair of length-k strings $(u, v) \in E$ if the length-$(k-1)$ suffix of u equals the length-$(k-1)$ prefix of v [76]. Clearly, $|V| = |\Sigma|^k$, $|E| = |\Sigma|^{k+1}$, and the indegree and outdegree of each vertex is $|\Sigma|$.

For example, the 3-dimensional de Bruijn graph over the binary alphabet $\Sigma = \{0, 1\}$ has $2^3 = 8$ vertices, i.e., $V = \{000, 001, 010, 011, 100, 101, 110, 111\}$. This de Bruijn graph is depicted in Figure 4.1. Similarly, the 2-dimensional de Bruijn graph over the DNA alphabet $\Sigma = \{A, C, G, T\}$ has vertex set

$$V = \{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}.$$

as illustrated in Figure 4.2

Let $G = (V, E)$ be any de Bruijn graph, and let $U \subseteq V$. The *SRE graph $G_U = (U, E')$ for G and U* is the vertex-induced subgraph of G with edge set

$$E' \;=\; \{(u, v) \in E \mid u, v \in U\}.$$

A *weakly connected component* in a directed graph $G = (V, E)$ is a maximal, nonempty set of vertices $C \subseteq V$ such that, for every pair of vertices $u, v \in V$, there is path in the underlying undirected graph from u to v [71]. The set of weakly connected components of G clearly partition V.

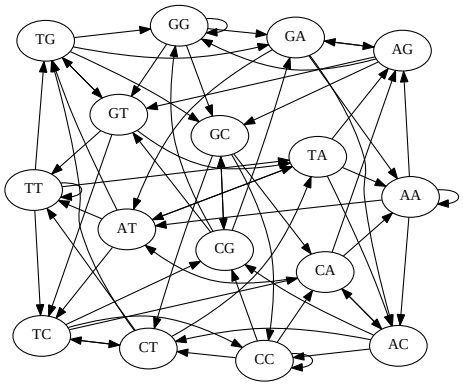


Figure 4.2: The 2-dimensional de Bruijn graph over the DNA alphabet $\{A, C, G, T\}$.

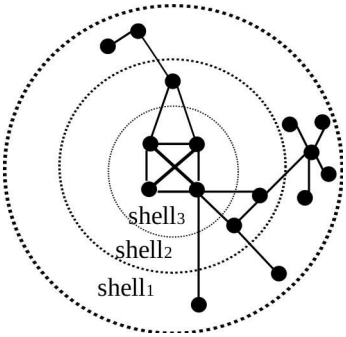


Figure 4.3: An example of j-core analysis where 3 shells are identified (adapted from [61]).

A j-core (or j-shell) decomposition analysis is a method to identify the most connected or important nodes in a graph [61]. Using j-core analysis, the graph is described in a layered structure as illustrated in Figure 4.3, where the innermost nodes are the most important ones and the other nodes will be positioned in the outer layers according to their importance, revealing a hierarchy for the graph. Therefore, finding the position of the node relative to the organization of the network can determine its influence better than utilizing a local property of nodes such as its degree [19]. The j-core of a graph is obtained by recursively removing all nodes with degree $< j$ and their incident edges; the remaining nodes and edges form the j-core graph.

4.3 Methods

4.3.1 Data sets

LEIsc (Log of the Enrichment Index, scaled) scores from Ke et al. [58] are used. Utilizing the minigene approach, they placed all 4096 6-mers at five different sites in two model exons. For each 6-mer, an LEIsc value was calculated. It represents a relative measure of central exon inclusion for each pre-mRNA molecule, with higher values representing greater inclusion.

A library of variant minigenes was constructed to include random 6-mers and then sequenced using an Illumina Genome Analyzer. A relative concentration was assigned to all 6-mers based on millions of high-confidence reads. The library was then transfected into human embryonic kidney cells (HEK293), and 24 hours after the transfection, the mRNA molecules that had successfully included the central exon were isolated and converted to cDNA. The output molecules were then similarly sequenced. For each 6-mer, an enrichment of output proportion over input proportion (enrichment index, EI) was calculated. The EI value represents the splicing efficiency of the central exon.

The spectra of activities of the 6-mers often differed among the five chosen sites. Much of this context effect was due to the creation of different overlapping sequences at each site. The 6-mer scores were identified based on the average of LEIsc values of a specific 6-mer in all five sites and all different places within a 16-nucleotide region of each site. In this way, LEIsc values can determine potential SREs that are generally used. Using a t-test to compare each LEIsc value of a specific 6-mer with the average of the LEIsc values of molecules that do not contain this 6-mer, Ke et al. identified 1182 potential ESEs and 1090 potential ESSs. The LEIsc scores range from 0.0534 to 1.034 in the ESE case. In the ESS case, they range from -0.0596 to -1.061. Figure 4.4 shows the distribution for both ESE values and ESS values.

Another data source is all the available unique coding exons for known human genes from the ENCODE project [56], which reports 205,163 exons from 29,179 genes. Data was acquired from the RefSeq Genes track, where known human protein-coding genes are recorded. The December, 2013, human genome assembly (GRCh38/hg38) is used. The 200 intronic nucleotides upstream and the 200 intronic nucleotides downstream of each exon are also retrieved.

For comparing my results with previously published results, several databases are utilized. SpliceAid-F [46] which contains 71 splicing factors and 655 binding sites for human. AEdb [86] was also used, which is a database for alternative exons and their properties from various species; it is the manually curated component of the Alternative Splicing Database (ASD). The exon data in AEdb have been experimentally verified.

In addition, I compared my ESE list with four other computational data sets. The RESCUE-ESE [39] data set contains 238 6-mers for human exons. Another data set is PESE [110], where 2096 8-mers were identified. The third data set is from [40] and contains 4- and 5-mers

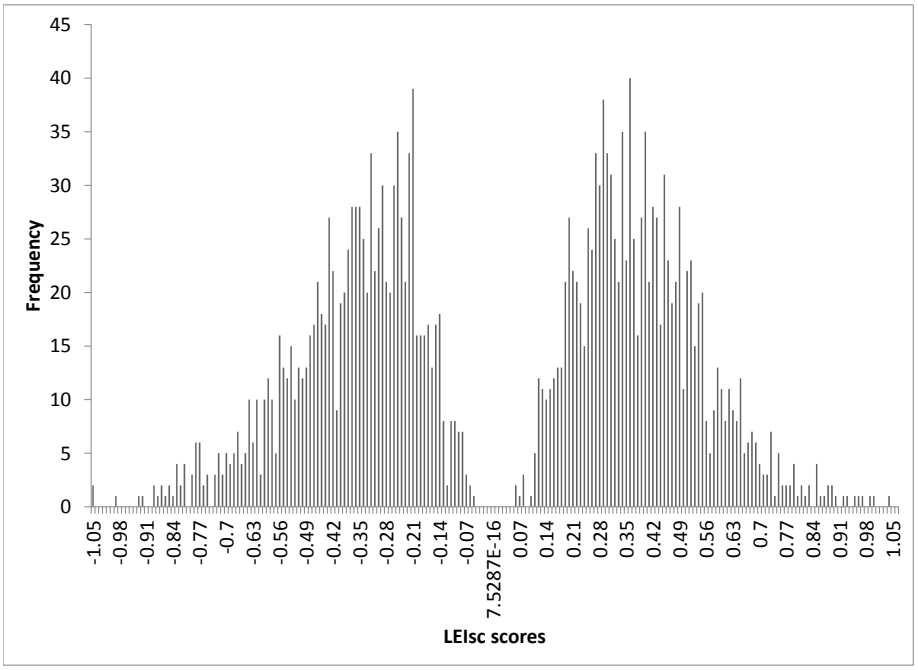


Figure 4.4: Distribution of the LEIsc scores. The x-axis represents LEIsc scores and the y-axis represents their frequencies. On the left, is ESS values range from -0.0596 to -1.061, while, on the right, the ESE values range from 0.0534 to 1.034.

as potential ESEs. Finally, in the data set from [111], the authors concentrated on 5-mer putative ESEs.

For ESSs, I compared my results with FAS-ESS [100], and PESS [110]. The FAS-ESS data set contains 130 10-mer sequences that were identified utilizing the mini-gene approach. PESS is another data set where the authors compared the frequencies of 8-mers (allowing one mismatch) in constitutively spliced non-coding exons with those in pseudo-exons and the 5' untranslated regions (UTRs) of intronless genes.

4.3.2 Outline of the computational strategy

A de Bruijn graph-based model followed by word count enrichment analysis is applied. My hypothesis is that SREs can be detected through both their effect on splicing (inclusion ratio and LEIsc scores) and their frequency in a specific data set (exons) with respect to a background data set (flanking introns). In particular, utilizing a de Bruijn graph allows us to detect potential SREs of different lengths based on the experimental data from Ke et al. [58]. The assumption that all SREs are of the same length can lead to inaccurate results as the actual length of an SRE is usually unknown. Therefore, developing a computational method to produce SREs that vary in length based on experimental data can achieve more

Algorithm 1 DBGBuild

Input: $k = 6$

Output: 6-dimensional de Bruijn graph G over $\{A, C, G, T\}$

1: G = GenerateVertices(0,"",k)
2: G = AddEdges(G)
3: **return** G

Figure 4.5: DBGBuild algorithm: Build k-dimensional de Bruijn graph.

accurate results.

If there are two 6-mers that overlap in five nucleotides and both of them have high LEIsc values, there is a greater probability that they form a potential 7-mer SRE. For example, if the two 6-mers $ACGTCA$ and $CGTCAT$ both have high LEIsc scores, there is a good chance of having one 7-mer SRE with the sequence $ACGTCAT$. The same applies with m consecutive 6-mers in the de Bruijn graph; if they all have high LEIsc values, then they can form one potential $(m + 5)$-mer SRE.

The processing in my model consists of six steps. First, the 6-dimensional de Bruijn graph $G = (V, E)$ is constructed over the DNA alphabet $\Sigma = \{A, C, G, T\}$ and associate each vertex with its rank based on LEIsc scores from Ke et al. [58]. Second, depending on whether I am searching for ESEs or ESSs, a subset $U \subseteq V$ is selected; for example, if I am looking for ESEs, then I might select U to be the 400 6-mers with the highest LEIsc values. Third, I construct the SRE graph G_U. Fourth, the weakly connected components is determined in G_U. Fifth, I apply the algorithm GenSRE to each weakly connected component to determine a set of potential SREs (see Section 4.3.4). Sixth, these sequences are submitted to word count enrichment analysis accompanied by all known human coding exons with their intronic flanks (see Section 4.3.5).

4.3.3 Constructing de Bruijn and SRE graphs

The 6-dimensional de Bruijn graph $G = (V, E)$ over the DNA alphabet $\Sigma = \{A, C, G, T\}$ is constructed. Each vertex v is a 6-mer. G represents all the possible one-character overlaps between pairs of 6-mers. It has 4096 vertices and 16,384 edges.

Algorithm 1 is used to build the graph where, I first generate the graph vertices with their corresponding 6-mers (Figure 4.6). Then, an edge between two vertices in the graph is added if their 6-mers satisfy the overlapping condition in a de Bruijn graph (Figure 4.7).

Theorem 1. *The algorithm DBGBuild has time complexity* $O(|V|)$

Algorithm 2 GenerateVertices

Input: *depth*,*base*, k
Output: G with all possible vertices

```
1: if depth == k then
2:    Add a vertex v in G such that v = base
3: else
4:    for newbase ∈ {A, C, G, T} do
5:       GenerateVertices(depth + 1, base + newbase, k)
6:    end for
7: end if
8: return G
```

Figure 4.6: GenerateVertices algorithm: Generate all possible k-mers

Algorithm 3 AddEdges

Input: G
Output: G connected graph

```
1: for v ∈ G do
2:    Let v = σ1σ2 ··· σk
3:    suffix = σ2σ3 ··· σk
4:    for base ∈ {A, C, G, T} do
5:       u = suffix + base
6:       Add edge (v, u) to G
7:    end for
8: end for
9: return G
```

Figure 4.7: AddEdges algorithm: a subroutine to connect de Bruijn graph vertices

Proof. Each call to the recursive algorithm GenerateVertices is performed in O(1) as all what required is to add a 6-mer vertex to the graph. The function is recursively called for concatenating letters to the previous generated sequence. It will continue until the sequence length is k where in my case $k = 6$. As the loop in line 5 will be repeated $|\Sigma|$ times, which in the DNA alphabet set case is 4, the recurrence relation of this function can be expressed as following:

$T[0] = O(1)$

$T[1] = 4T[0] + C_1$ for some constant $C_1 > 0$

$T[2] = 4T[1] + C_1 = 4(4T[0] + C_1) + C_1 = 4^2T[0] + 4C_1 + C_1$

$T[3] = 4T[2] + C_1 = 4(4T[1] + C_1) + C_1 = 4^3T[0] + 4^2C_1 + 4C_1 + C_1$

$T[n] = 4^kT[n - m] + \frac{4}{3}C_1(1 - 4^{m-1})$ for all m.

The function is called, in the worst case, k times. Hence, it is called (4^k) times which, in general , is $|\Sigma|^k$ times and that is the number of the vertices in the graph. Therefore, the recursive algorithm DBGBuild is of time complexity $O(|V|)$.

AddEdges subroutine time complexity is $O(k|V|)$ as the two loops in line 1 and 5 will take $kC_2|V|$ for some constant $C_2 > 0$. □

As stated before, each 6-mer has an LEIsc value that represents a relative splicing strength score for that 6-mer [58]. The higher the LEIsc value, the greater the potential enhancing effect of that 6-mer on splicing. Similarly, the lower the LEIsc value, the greater the potential silencing effect of that 6-mer (See Figure 4.4). I utilize the findings in Key et al. [58] of potential exonic enhancers and silencers. If a specific 6-mer was found to be an enhancer or silencer, I use its associated LEIsc score. If it is defined as neutral, its LEIsc value is considered to be zero. Then, I order all the scores in descending order and associate each vertex v in the G graph with its rank. The rank is suggestive of the strength of a 6-mer on splicing. As a result, the graph can capture hot spots where many connected vertices have high ranks (for enhancers) or low ranks (for silencers). Supplementary Table S1 from [14] contains all possible 6-mers in descending order according to their LEIsc scores.

Let R be a predefined number of ranks. A set U is constructed by choosing the top R vertices by rank in the case of searching for ESEs, and the lowest R vertices by rank in the case of ESSs. The SRE graph $G_U = (U, E')$ is constructed. Weakly connected components $C_i \subseteq U$, $i = 1, 2, .., w$, where w is the number of weakly connected components in G_U, are then extracted. Supplemental Figure S1 from [14] is an example of one of the weakly connected components for ESEs, where $R = 100$.

4.3.4 Identifying variable length SREs

I developed the GenSRE algorithm to generate all potential SREs. The pseudocode for GenSRE can be found in Figure 4.8. For each C_i, the SeqAssembly algorithm is applied as illustrated in Figure 4.9. Starting from each vertex $v \in C_i$, a modified depth-first traversal is performed. At each vertex x, a sequence s_x will be produced, representing the sequence going from v to x. Clearly, the sequence length $|s_x|$ depends on what level the traversal reaches. This process is repeated with each vertex in C_i as the starting vertex for the traversal, as given in the GenSRE algorithm. The result is all potential sequences with length six or more. As stated before, the idea is that, if there are two 6-mers that are overlapping in five nucleotides and both of them have high ranks, there is a greater probability that they form a 7-mer potential ESE. Figure 4.10 illustrates an example of the traversal and the output sequences. Consequently, the output sequences represent k-mers that can serve as potential SREs.

As illustrated in line 10 of the SeqAssembly algorithm, I mark the vertices I encounter in the traversal as visited. Therefore, these vertices will not be visited again. Such revisits are not allowed, because the existence of directed cycles will result in an infinite loop. The main drawback is that a vertex may be reached multiple times because two initially parallel paths from v intersect at some vertex x. Fortunately, these paths will rarely be shorter than length 6, so the algorithm does retrieve most SREs of length ≤ 12.

Algorithm 4 GenSRE

Input: $C_1, C_2, \ldots, C_w$
Output: $\{s_x\}$ generated by SeqAssembly
1: $S = \emptyset$
2: **for** $i \in \{1, 2, \ldots, w\}$ **do**
3: **for** $v \in C_i$ **do**
4: Mark all $v \in C_i$ as not visited
5: $S = S \cup \text{SeqAssembly}(C_i, v)$
6: **end for**
7: **end for**
8: **return** S

Figure 4.8: GenSRE algorithm: Generating all possible sequences from the weakly connected components.

Theorem 2. *The GenSRE algorithm has time complexity $O(|U|(|U| + |E'|))$, where $|U|$ is the number of nodes in the SRE graph and $|E'|$ is the number of edges.*

Algorithm 5 SeqAssembly

Input: C_i, v
Output: T

$T = \emptyset$
Initialize A to be a stack of strings
Let $v = \sigma_1\sigma_2 \cdots \sigma_m$
$s_v = \sigma_1\sigma_2 \cdots \sigma_m$
A.push(v)
while A is not empty **do**
  $x = A$.pop()
  $T = T \cup \{s_x\}$
  if x is not marked as visited **then**
    mark x as visited
    for $(x, y) \in E'$ **do**
      Let $y = \tau_1\tau_2 \cdots \tau_m$
      s_y = concatenate(s_x, τ_m)
      A.push(y)
    end for
  end if
end while
return T

Figure 4.9: SeqAssembly: Sequence assembling algorithm. A subroutine to traverse a weakly connected component starting from a specific vertex. Each vertex x is associated with a sequence s_x, which is extended as the traversal go deeper.

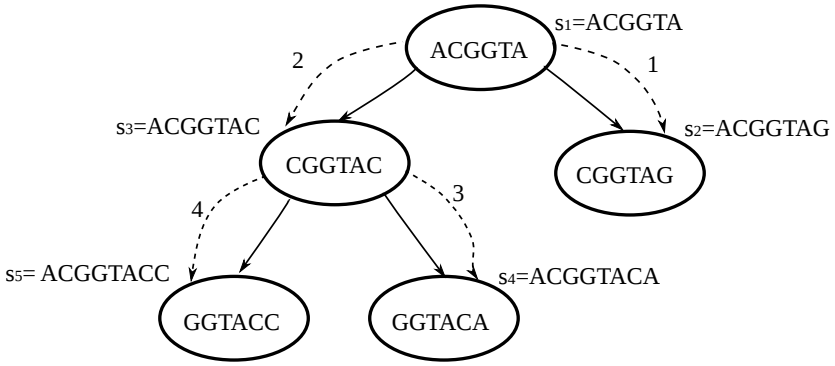


Figure 4.10: An illustration of GenSRE algorithm. The depth-first traversal starts at vertex $ACGGTA$ where the dotted lines with its associated number represent order of the traversal. The resulting sequences are labeled by the order they were produced. The output sequences in order are: $s_1 = ACGGTA, s_2 = ACGGTAG, s_3 = ACGGTAC, s_4 = ACGGTACA, s_5 = ACGGTACC$

Proof. The SeqAssembly algorithm will be repeated for each weakly connected component and each vertex in each component. Having $|U|$ nodes in all the components, this operation will be repeated $|U|$ times. The time complexity of the depth-first traversal is $O(|V_{C_i}|+|E_{C_i}|)$, where $|V_{C_i}|$ is the number of vertices in a weakly connected component C_i and $|E_{C_i}|$ is the number of its edges. Therefore the time complexity for the traversal on all components is $O(|U| + |E'|)$. Consequently, the time complexity of the algorithm is $O(|U|(|U| + |E'|))$. □

4.3.5 Word count enrichment analysis

As mentioned before, word count enrichment analysis is a computational technique that is widely used for identifying SREs. It searches for short nucleotide sequences that are statistically over-represented or under-represented through the comparison of foreground and background sequences [104, 108]. The same approach is followed here on the set S of sequences produced from the GenSRE algorithm. A data set consisting of the human coding exons and of their flanking intronic regions is utilized as well.

Consider any sequence $s \in S$; let $j = |s|$ be its length. Its frequency $f_E(s)$ in the first and last 50 nucleotides of all the exons is calculated. Its frequency $f_I(s)$ in the intronic flanking regions is calculated as well. Let N_E and N_I be the total number of j-mers in the exonic and intronic regions, respectively. Note that N_E and N_I change with each j-mer based on its length. The two-sample proportion z-score [39, 104, 110] of s is then given by

$$z_s = \frac{f_E(s) - f_I(s)}{\sqrt{(\frac{1}{N_I} + \frac{1}{N_E})p(1-p)}},$$

where

$$p = \frac{N_I f_I(s) + N_E f_E(s)}{N_I + N_E}.$$

I use pooled sample proportion p, as the null hypothesis states that $f_E = f_I$ [102]. Potential SREs are defined as over-represented j-mers in exonic regions but not in intronic regions. To test the statistical significance under the null hypothesis $f_E = f_I$, j-mers with $z \geq 1.64$ ($p < 0.05$, two-tail test) are identified as being over-represented. A false discovery rate (FDR) is calculated for each over-represented j-mers, and j-mer with FDR corrected p-value that is less than 0.05 are reported [22].

4.3.6 Analysis of the functional characteristics of predicted SREs

To assess the significance of the predicted SREs and whether they are good candidates for ESEs or ESSs, I utilized the command-line version of Ontologizer [20], with the goal of determining the enriched GO annotations for the experimentally verified SREs from SpliceAid-F [46] and checking whether my predicted SREs share the same enriched GO terms. This can be interpreted as both sets of SREs affect the regulation of similar pathways.

The genes that contain all the human coding exons that I use in my analysis are utilized as a background data set. For each exonic splicing element in SpliceAid-F, the exon data set is searched to allocate each splicing element, and the corresponding gene set is identified to form the study set. GO annotation files `gene_ontology_edit.obo` and `gene_association.goa_human` were downloaded. GO enrichment analysis is performed using the Topology-Elim algorithm. Westfall-Young Single Step multiple testing correction procedure is then applied. The same approach is applied on my predicted splicing elements.

I am interested in the biological process annotations. Therefore, for the previously known splicing elements, I choose the biological process category with the minimum adjusted p-value, where terms with $p < 0.05$ are only considered to be significant. Then, I categorize the known splicing elements according to their biological processes, and I did the same procedure for my set of putative splicing elements.

4.4 Results

For predicting potential ESEs, the highest 400 6-mers by LEIsc values were chosen. In other words, the SRE graph was extracted with $R = 400$. I chose the value of R to be 400 as most of the analysis done by Ke et al. [58] on their produced LEIsc scores, which I utilize, was on the highest or the lowest 400 LEIsc scores. However, R can be chosen to be any value based on the utilized data. Applying my model, 36 weakly connected components are produced with most of the 6-mers located in one large component. This component consists of 352

Table 4.1: Distribution of the 400 6-mers on the weakly connected components in case of extracting potential ESEs

Number of 6-mer ESEs	352	5	2	1
Number of weakly connected components	1	2	5	28

Table 4.2: Number of resulted ESEs using different exonic flank sizes

Exon flank size (n)	50	100	150	200
Number of utilized Exons	134596	34595	14970	10634
Number of putative ESEs	2001	1806	1595	1575

6-mers out of the 400. Table 4.1 provides the sizes of all the weakly connected components. Certainly, a weakly connected component of size 1 can produce only one ESE, a 6-mer, while most of the potential ESEs are harvested from the one of size 352.

The GenSRE algorithm recovered 53,984 potential ESEs. Their lengths range from 6 to 87 nucleotides, with an average length of 48 nucleotides. Having one large weakly connected component is the reason that there are many potential ESEs that are quite long. Applying word count enrichment analysis, about 1500 to 2000 ESEs is obtained based on how many nucleotides are taken into account from the start and the end of all the exons (exonic flanks), as shown in Table 4.2. I started with $n = 50$, where n is the size of the exonic flanks. Extending this to $n = 100$ nucleotides did not change the results significantly as many of the resulting ESEs are overlapping, as illustrated in Table 4.3. Different experiments were done utilizing different exonic flank lengths. These included 50, 100, 150, and 200 nucleotides. Supplementary Tables S2, S3, S4, and S5 from [14] contain the details of each experiment, including a list of predicted ESEs, the frequency of each ESE in the exonic and intronic regions, its z-score, its associated p-value, and its FDR corrected p-value. In case of 50 nucleotide exonic flanks, I identified 2001 potential ESEs, where their lengths range from 6 to 12 nucleotides. Figure 4.11 depicts the predicted ESE length distribution.

I compared my results, where the exonic flanks are 50 nucleotides, with exonic binding sites from SpliceAid-F [46]. Removing duplicate binding sites, SpliceAid-F includes 330 different sequences for human. Among those, 112 are exonic binding sites. I removed all sites that bind to members of the extended family of heterogeneous nuclear ribonucleoproteins (hnRNPs) and other splicing factors that are considered silencers according to the literature, The remaining 59 sequences are considered ESEs, as they bind to splicing factors that are involved in enhancing activities. Since my predicted ESEs are of variable length, as are SpliceAid-F binding sites, I calculated the overlap between the two sets by finding whether each sequence in the first list is totally contained in at least one sequence in the second list and vice versa. The total number of overlapped sequences is 105.

Another data set is AEdb [86]. It contains 294 splicing regulatory motifs. Among those, 124 are ESEs. I considered only the 64 ESEs that belong to human.

Table 4.3: Number of common ESEs between different experiments

n (Number of ESEs)	50 (2001)	100 (1806)	150 (1595)	200 (1575)
50 (2001)	2001	1704	1514	1467
100 (1806)	-	1806	1528	1475
150 (1595)	-	-	1595	1460
200 (1575)	-	-	-	1575

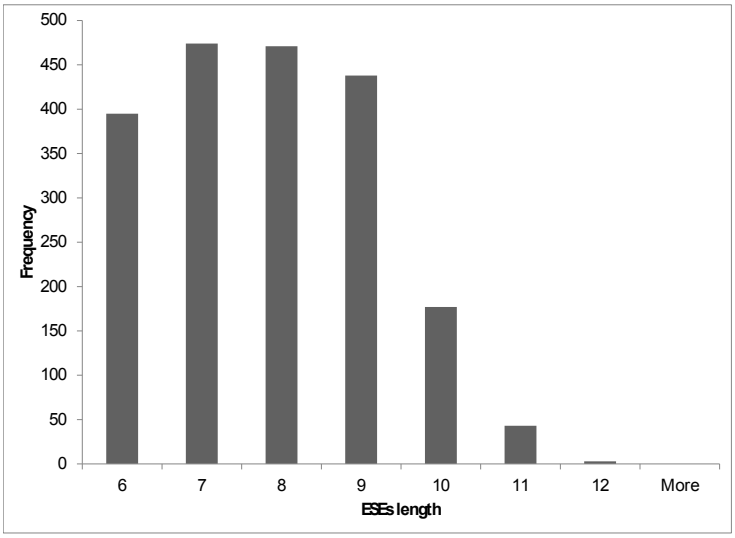


Figure 4.11: Distribution of the ESE lengths. The x-axis represents ESE length and the y-axis represents the frequency of occurrence.

Table 4.4: Number of overlapped ESEs with previously published data sets. Approximate match is calculated by calculating the overlapping between the two sets. This is performed by finding whether each sequence in the first list is totally contained in at least one sequence in the second list and vice versa. Exact match means the same ESE is found with the same length.

Data set	SpliceAid-F 2001/69	AEdb 2001/64	RESCUE-ESE 2001/238	PESE 2001/2060	Fedrove 2001/42	Zhang 2001/42
Approximate	103/9	62/6	54/54	447/51	-/16	-/12
Exact	7	5	54	44	-	-
Total	105	63	54	454	42	12

In addition, I compared my ESE list with four other computational data sets, such as the RESCUE-ESE [39] data set where the total overlap was 54 6-mers. The RESCUE-ESE approach is focused on exon skipping events [31], which may explain the low overlapping percentage. Another data set is PESE [110], where the overlap is 454 sequences. That includes 44 exact sequences (of length 8). The third data set is from [40]. As it contains only 4- and 5-mers as potential ESEs, I could only test if my data set includes any of these sequences. This also applies to the data set from [111]. Table 4.4 summarizes the overlapping results.

To verify the ability of word count enrichment analysis to filter the potential splicing elements, this analysis was applied to the 112 exonic and 87 intronic binding sites from SpliceAid-F [46]. Table 4.5 illustrates that 70.3% of the exonic binding sites were over-represented in the human coding exons and about 74% of the intronic binding sites were over-represented in the flanking intronic regions, which indicates the ability of this analysis to identify potential regulatory elements. The total number of exonic binding sits is 112 sequences. However, I am searching for over-represented sequences in the exonic flanks of length 50 nucleotides. Therefore, I limited the search for sequences with length less than or equal to 50 nucleotides (104 sequences). Many of the sequences were not found in my data set of all human coding exons (40 sequences). The remaining sequences (64 sequences) were tested for over-representation by calculating their z-scores. Using the same cutoffs, k-mers with $z \geq 1.64$ ($P < 0.05$, two-tail test) are identified as being over-represented. A false discovery rate (FDR) is calculated for each over-represented k-mers and k-mers with FDR corrected p-value that is less than 0.05 are reported. The same approach is applied on the intronic binding sites.

For the exonic splicing silencers, the lowest 400 6-mers in LEIsc values by rank is chosen. Applying my model, 18 weakly connected components are produced with most of the 6-mer silencers connected in one component, as in the case of 6-mer enhancers. The largest component consists of 369 6-mer ESSs out of the 400 silencers. Table 4.6 indicates the size of all the produced components.

The GenSRE algorithm resulted in 63,780 potential ESSs, with lengths ranging from 6 to

Table 4.5: Over-presented binding sites statistics from the SpliceAid-F data set utilizing word count enrichment analysis

Binding sites	Exonic	Intronic
Total number	112	87
Total number with length ≤ 50 nucleotides	104	77
Number of sequences found in my data set	64	50
Number of over-represented sequences	45 (70.3%)	37 (74%)

Table 4.6: Distribution of the 400 6-mers on the weakly connected components in case of extracting potential ESSs

Number of 6-mer ESSs	369	6	4	3	2	1
Number of weakly connected components	1	1	1	2	2	11

88 nucleotides and an average length of 47 nucleotides. For word count enrichment analysis, the exonic flank size was chosen to be 50 nucleotides as in the case of ESEs. This resulted in 3080 ESSs with lengths ranging from 6 to 15 nucleotides. Figure 4.12 illustrates the ESS length distribution. Supplementary Table S6 from [14] contains the ESS related information, including a list of predicted ESSs, the frequency of each ESS in the exonic and intronic regions, its z-score, associated p-value, and its FDR corrected p-value.

My ESSs are compared with other data sets as illustrated in Table 4.7 such as SpliceAid-F [46], AEdb [86], FAS-ESS [100], and PESS [110].

I used Ontologizer to analyze the functional similarities between the known splicing elements from SpliceAid-F data set and the predicted SREs. Table 4.8 depicts the different biological process categories for both ESEs from SpliceAid-F and my predicted ESEs. Out of 19 categories for the known ESEs, 14 are shared with my ESEs with the largest p-value is 0.00271. Supplementary Table S7 from [14] contains the complete list of the biological processes that the predicted ESEs are involved in. Supplementary Tables S8 and S9 from [14] depict the common biological process categories for ESSs and the biological processes for the predicted ESSs, respectively. My remaining ESEs have more functional categories that some of them are illustrated in Table 4.9.

4.5 Discussion

I introduce a de Bruijn graph formalism to identify exonic splicing elements of variable length. Utilizing this approach leads to the identification of new potential ESEs and ESSs. One of the advantages of my model is its scalability. This model allows building the de Bruijn graph from any k-mer (based on the available data). The number of k-mers that are taken into consideration (R) can be changed according to the available data as well. Here,

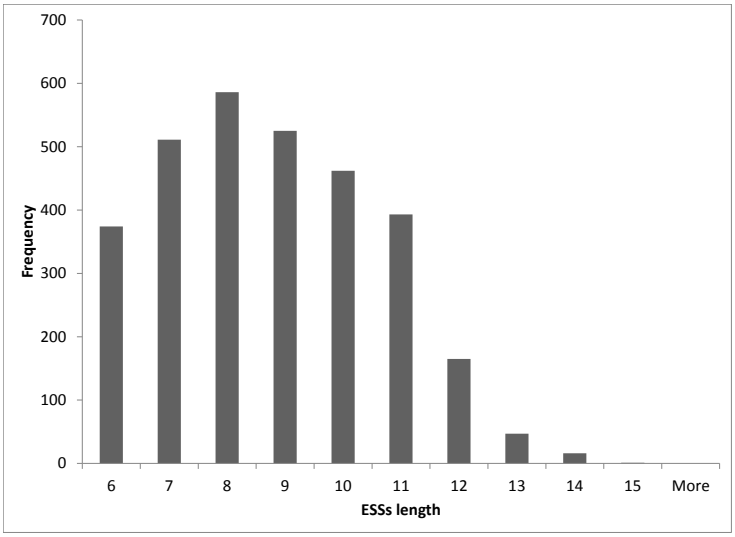


Figure 4.12: Distribution of the LEIsc scores. The x-axis represents ESS lengths and the y-axis represents their frequencies.

Table 4.7: Number of overlapped potential ESSs with previously published data sets. Approximate match is calculated by calculating the overlapping between the two sets. This is performed by finding whether each sequence in the first list is totally contained in at least one sequence in the second list and vice versa. Exact match means the same ESE is found with the same length.

Data set	SpliceAid-F 3080/53	AEdb 3080/24	FAS 3080/130	PESS 3080/1019
Approximate	88/10	22/3	190/-	338/35
Exact	3	3	-	34
Total	95	23	190	339

Table 4.8: Common biological process categories of my ESE list and ESE from SpliceAid-F based on GO term enrichment analysis

GO ID	Name	SpliceAid-F ESEs	Predicted ESEs
GO:0007250	actin filament capping	1	1
GO:0006200	ATP catabolic process	1	5
GO:0007411	axon guidance	11	178
GO:0030574	collagen catabolic process	1	39
GO:0032508	DNA duplex unwinding	1	6
GO:0022617	extracellular matrix disassembly	5	191
GO:0046037	GMP metabolic process	1	1
GO:0071044	histone mRNA catabolic process	1	4
GO:0086010	membrane depolarization during action potential	1	11
GO:0007018	microtubule-based movement	2	26
GO:0007528	neuromuscular junction development	1	7
GO:0090292	nuclear matrix anchoring at nuclear membrane	4	1
GO:0021860	pyramidal neuron development	1	1
GO:0060372	regulation of atrial cardiac muscle cell membrane repolarization	1	1

Table 4.9: Example of some biological process categories of the predicted ESEs based on GO term enrichment analysis

ID	Annotation	Number of Predicted ESEs
GO:0031532	actin cytoskeleton reorganization	3
GO:0008154	actin polymerization or depolymerization	2
GO:0070358	actin polymerization-dependent cell motility	2
GO:0007190	activation of adenylate cyclase activity	2
GO:0006919	activation of cysteine-type endopeptidase activity involved in apoptotic process	2
GO:0009060	aerobic respiration	3
GO:0097055	agmatine biosynthetic process	3
GO:0021960	anterior commissure morphogenesis	2
GO:0019885	antigen processing and presentation of endogenous peptide antigen via MHC class I	5
GO:0015991	ATP hydrolysis coupled proton transport	5
GO:0007409	axonogenesis	3
GO:0051016	barbed-end actin filament capping	4
GO:0006699	bile acid biosynthetic process	3
GO:0015878	biotin transport	11
GO:0007596	blood coagulation	3

the LEIsc scores were utilized as a measurement for rankings 6-mers. The rank can be based on other criteria such as conservation scores. Deciding on the significance of the produced k-mers may depend not only on the rank values and frequency but also on other data sources. For example, having a list of all protein binding sequences that are experimentally verified can increase the probability of having a certain k-mer as a putative SRE if a part of the sequence is in the verified list. Another possibility is utilizing the conservation score of the sequence of interest.

Another advantage of my model is its flexibility. I applied the model on a list of all known human coding exons and its flanking intronic regions to find potential ESEs. The same approach is also applied for finding ESSs. To do so, instead of selecting the highest 400 6-mers, I selected the lowest 400 6-mer in the LEIsc scores. Potentially, my model can also be utilized to find ISEs and ISSs by searching for the sequences of interest to be over-represented in the intronic regions and under-presented in the exonic flanks.

Using the parameter values of $R = 400$ and exonic flank size of 50 nucleotides, I identified 2001 potential ESEs. This includes some of the well-known ESEs such as $GAAGAA$, which is verified experimentally in the RESCUE-ESE data set [39]. It is noticed that this 6-mer is part of the consensus sequences $RGAAGAAC$ ($R = A$ or G) that have been verified as a SELEX binding motif to the ASF/SF2 splicing factor [89]. ASF/SF2 is one of the highly conserved proteins that affects alternative splicing [89]. My method could accurately identify this binding site as $GGAAGAAC$ with p-value 1.07×10^{-55}. Moreover, there are some other possibilities that contain the same sequence such as $GGAAGAACG$ and $GAAGAACG$ with p-values 2.01×10^{-9} and 2.43×10^{-41}, respectively.

Another consensus motif for the ASF/SF2 splicing factor is $GARGARGAR$ [82], which I have in my results as $GAAGAAGAG$ with p-value 9.58×10^{-23}, in addition to longer k-mers that contain this sequence (see Supplementary Table S3 [14]).

Utilizing the results obtained from Ontologizer. I investigated the effect of having similar binding sites on the biological processes they are involved in. In other words, I wanted to know if the genes that contain the binding site $GAAGAA$ are involved in the same biological process of the genes that contain the longer binding site $GGAAGAAC$. The answer is "no". For $GAAGAA$; the most enriched biological process is "axon guidance" while, for $GGAAGAAC$, it is "protein ubiquitination involved in ubiquitin-dependent protein catabolic process". It is obvious from Figure 4.13 that they are unrelated processes.

Axon guidance or axon path finding is a critical and complicated process for nervous system wiring, where there are certain tracts the axons should follow to reach specific targets [70]. Defects in this process can lead to various human disorders such as HGPPS [70, 38], congenital mirror movements, congenital fibrosis of the extraocular muscles [70], L1 Syndrome, and albinism [38]. Some of these disorders, such as HGPPS and L1 Syndrome, are caused from missense, splice site, and frameshift mutations of the ROBO3 and L1CAM genes, respectively.

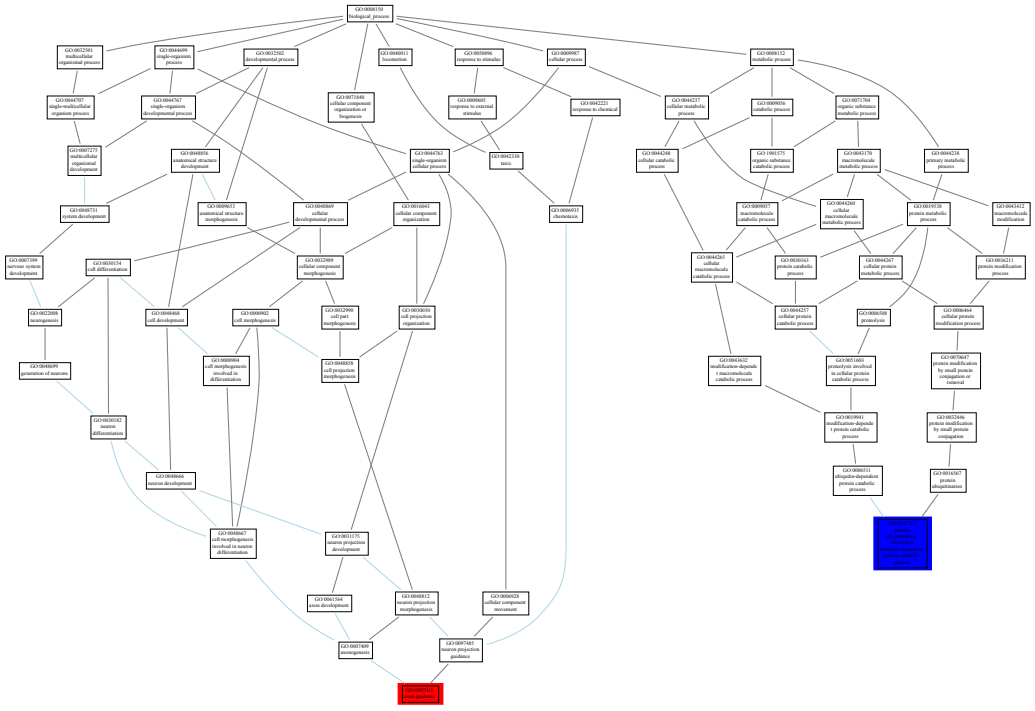


Figure 4.13: A tree map of Gene Ontology to illustrate the biological processes that the genes of the two binding sites *GAAGAA* and *GGAAGAAC* are involved in. *GAAGAA* biological process "axon guidance" is highlighted in red, while the other one "ubiquitin-dependent protein catabolic process" is highlighted in blue.

On the other hand, ubiquitin is a small regulatory protein that resides in eukaryotic cells and attaches to other proteins. This attachment can signal protein degradation [26]. It has been shown that ubiquitin has a controlling role in the splicing pathway and hence affects spliceosome assembly [21]. Moreover, according to [26], ubiquitin influences the stability and degradation of the SMN protein. In humans, SMN is encoded by two genes, SMN1 and SMN2. Mutations in SMN1 cause spinal muscular atrophy (SMA) disease. SMN stability is affected by its ability to oligomerize. Therefore, SMN mutations that prevent oligomerization lead to rapid degradation, and this may be the reason that it causes SMA [26]. It is also worth mentioning that the SMN protein is part of a large multiprotein complex (the SMN complex), which is essential for the biogenesis of small nuclear ribonucleoprotein particles (snRNPs). These snRNPs are major components of the spliceosome machinery.

It is clear that, although both of the these binding sites are overlapping on most of their sequences, the biological processes of the genes they reside in are highly different and alternative splicing is involved in both of the processes in different ways. The ability to determine a specific biological process can make it easier to investigate the actual effect the alternative splicing has in different contexts. Mutations in these binding sites can also affect the alternative splicing role. Therefore, having the ability to predict variable length SREs, instead of having a prefixed size before applying my analysis, gives the opportunity to discover new biological processes that alternative splicing may affect and gives an insight of how alternative splicing may work. Although I have a large number of biological processes in my analysis (947 categories), I see it as an opportunity for investigating specific contexts that alternative splicing may play a role in.

One of the ESSs in my results which is validated experimentally is $TAGTTAG$, a 7-mer ESS, which binds to the splicing repressor hnRNP A1 [69]. Another 7-mer exon silencer is $TTAAGGT$ [18], which is involved in optic atrophy disease.

Having one large weakly connected component that contains most of the SREs, whether for enhancers or silencers, indicates that there is much overlapping among the known SREs and confirms my hypothesis that longer k-mers can be a better and more accurate representation of SREs than shorter ones. As stated before, having an edge between two vertices means the they overlap in 5 nucleotides and perhaps they form one 7-mer SRE. Analyzing the largest component further using j-core analysis [19], Figure 4.14 illustrates the most influential nodes in the ESE case. In other words, these nodes are the most central and highly connected nodes. As a result, 6-mers that these nodes represent are the most repeated 6-mers in my ESE list.

These sequences are found to be GC enriched with GC content about 68% (Table 4.10, column 1), which is analogous to many data sets that are experimentally verified (Table 2 in [31]). This is also consistent with the fact that the regions around the splice sites are GC-enriched, which is considered one characteristic of having a stable pre-mRNA secondary structure [107]. Conserved and stable pre-mRNA secondary structures are thought to play an important role in splicing, as in [54], some of the experimentally verified SREs were found to be enriched near the splice sites in the regions of a single stranded local secondary

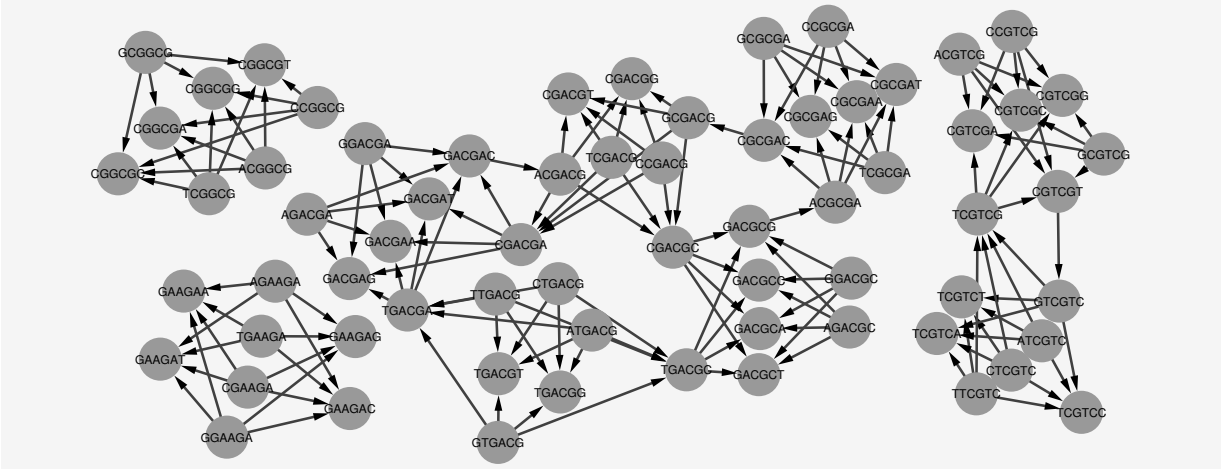


Figure 4.14: The most influential nodes in the largest weakly connected component for ESEs.

structure.

On the other hand, performing the same analysis on the silencers list, core sequences are found to be T-rich and C-poor just as in the PESS data set [110]. Figure S2 in the supplemental material [14] indicates the core nodes in the ESS case.

Table 4.10: Base compositions of core sequences in the case of exonic enhancers and silencers

Data	ESEs	ESSs
A%	21	23
C%	30	6
G%	38	33
T%	11	38

4.6 Conclusion

I have presented a new de Bruijn graph formalism to identify exonic splicing elements of variable length. Utilizing this approach leads to the identification of new potential ESEs and ESSs. Genomic structure, word count enrichment analysis, and experimental evidence were all utilized in my model to increase the accuracy of the results. I have developed GenSRE algorithm to produce potential variable length SREs. To demonstrate the usefulness of my approach, The produced results are compared with experimentally verified data sets and computational data sets as well. My results overlap with many of the experimental and computational results. I also analyzed the effect of having similar binding sites on the biological processes they are involved in. I indicated that although the binding sites may

overlap on most of their sequences, the biological processes of the genes they reside in can be highly different. Thus, the SREs length is a key part in the analysis where it cannot be assumed to be fixed. This approach can open new directions to study SREs and the roles they play in alternative splicing.

Chapter 5

CoSREM: A Graph Mining Algorithm for the Discovery of Combinatorial Splicing Regulatory Elements

5.1 Introduction

Most of the current approaches in identifying SREs have focused on individual motifs [60]. However, many AS events involve multiple regulators. Zhang et al. [109] showed experimentally that knocking out multiple ESEs affects splicing efficiency dramatically. Hence, AS is a complex process that involves cooperative or competitive interplay between splicing enhancers and silencers.

As stated earlier, if an exon has both ESE and ESS elements in proximity and in case of having an SR splicing factor with great affinity the SR protein will bind to the ESE and stimulate exon inclusion. However, if an inhibitory splicing factor such as hnRNP, which acts as a splicing repressor, is also present, it may inhibit the exon inclusion by binding to the silencer sequence and recruiting the binding of other inhibitory factors. These factors extend to the exon boundary and prohibit the binding of the SR protein. As a result, the exon will be skipped [66, 105].

Consequently, identifying individual SREs is not enough to explain tissue-specific or condition-specific AS. The challenge is that, because of the large number of possible SRE pairs that reside in different regions, experimental approaches for identifying SRE pairs will be prohibitively expensive [103]. Identifying larger SRE combinations, where multiple SREs are working together, will be even harder.

I have developed CoSREM (Combinatorial SRE Miner), an algorithm for discovering combinatorial SREs. CoSREM is a two-level graph mining algorithm that I apply to my SRE

graphs [14] to identify co-occurring sets of SREs. My focus is on identifying sets of exonic splicing regulatory elements whether they are enhancers or silencers. Experimental evidence is incorporated through the SRE graphs to increase the accuracy of the results. The identified SREs do not have a predefined length, and the algorithm is not limited to identifying only SRE pairs, as are current approaches. CoSREM is implemented as an open-source package (https://github.com/emanmostafabadr/CoSREM).

5.2 Preliminaries

I utilize definitions from Section 4.2. Let $G_{U_{ESE}}$ be an SRE graph where the chosen vertices U_{ESE} has an experimental evidence of enhancing activity. In analogy to $G_{U_{ESE}}$, let $G_{U_{ESS}}$ be an SRE graph where the chosen vertices U_{ESS} has an experimental evidence of silencing activity.

Let Y be a set of n 6-mers of interest $Y = \{y_1, y_2, \ldots, y_n\}$. For example, it can contain only 6-mers with evidence of enhancing activity. Let X be a set of m exons $X = \{x_1, x_2, \ldots, x_m\}$. The *SRE profile matrix* $P = (p_{i,j})$ is the $n \times m$ occurrence matrix, where $p_{i,j} = 1$, if 6-mer y_i is in exon x_j, and $p_{i,j} = 0$, if 6-mer y_i is not in exon x_j. Let P_{ESE} and P_{ESS} be two SRE profile matrices for enhancers and silencers, respectively.

For a 6-mer y_i, $T(y_i)$ is the set of exons containing y_i, that is, $T(y_i) = \{x_j \mid p_{i,j} = 1\}$.

Let $Y' \subseteq Y$ be a set of 6-mers. The set of shared exons for Y' is $T(Y') = \bigcap_{y_i \in Y'} T(y_i)$, the set of all common exons where the Y' 6-mers reside together.

Let $G_S = (S, E_S)$ be an induced connected subgraph of the SRE graph G_U. G_S is *α-cohesive* if $|T(S)| \geq \alpha$, where $\alpha \geq 1$. G_S is a *maximal α-cohesive subgraph (MCS)* if none of its supergraphs is α-cohesive.

Let $M = \{G_{S_1}, G_{S_2}, ..., G_{S_r}\}$ be a set of MCSs, where its shared exon set is $T(M) = \bigcap_{G_{S_i} \in M} T(S_i)$. M is called an *MCS collection* if it satisfies the following conditions: $|M| \geq \beta$ and $|T(M)| \geq \theta$, where β and θ are user defined thresholds.

5.3 Problem Definition

Let $C = \{M_1, M_2, ..., M_l\}$ be a set of all the MCS collections that can be identified given the two SRE graphs $G_{U_{ESE}}$ and $G_{U_{ESS}}$, SRE profile matrices P_{ESE} and P_{ESS}, and the parameters α, β, and θ. The problem of discovering combinatorial SREs is to find the set C such that $|M| \geq \beta$, $|T(M)| \geq \theta$, for any $M \in C$, and $|T(S)| \geq \alpha$, for any $G_S \in M$.

5.4 Methods

5.4.1 Data sets

I utilized the same data sets as explained in Section 4.3.1

5.4.2 Overview of the computational method

A de Bruijn graph based model is developed, and a two-level graph mining algorithm is applied to discover enhancers and silencers that occur in the same set of exons. Experimental evidence that a specific k-mer has enhancing or silencing behavior is incorporated through the graph model. My hypothesis is that combinatorial SREs can be discovered by their co-occurrence behavior in the same set of exons and the experimental evidence of their enhancing or silencing activities.

Utilizing a de Bruijn graph allows the detection of potential SREs of different lengths based on the experimental data from Ke et al. [58]. For example, if there are two 6-mers that overlap in five nucleotides and both of them have high LEIsc values, there is a greater probability that they form a potential 7-mer SRE. Suppose that the two 6-mers $GTCATC$ and $TCATCC$ have high LEIsc scores. Consequently, there is a good chance of having one 7-mer SRE with the sequence $GTCATCC$. The same applies with m consecutive 6-mers in the de Bruijn graph; if they all have high LEIsc values, then they can form one potential $(m+5)$-mer SRE [14].

The model starts with constructing the 6-dimensional de Bruijn graph $G = (V, E)$ over the DNA alphabet $\Sigma = \{A, C, G, T\}$ and associates each vertex with its rank based on LEIsc scores from Ke et al. [58] . The next step is **building the SRE graphs**. For example, if I am looking for ESEs, I select a subset $U_{ESE} \subset V$ that is associated with the highest LEIsc values. In the same manner, I select U_{ESS} to be the 6-mers with the lowest LEIsc values. As a result, I construct two SRE graphs, $G_{U_{ESE}}$ for enhancers and $G_{U_{ESS}}$ for silencers. The next step is **constructing the SRE profile matrices**, where I build profile matrices P_{ESE} and P_{ESS} for enhancers and silencers, respectively. I apply the first level of the CoSREM algorithm (GenMCS) for **discovering maximal α-cohesive subgraphs (MCSs)**. The goal in this level is to discover potential enhancer and silencer elements of different lengths where each element resides in a specific set of exons. With inputs $G_{U_{ESE}}$ and P_{ESE}, GenMCS generates several subgraphs, where each one represents a set of ESEs that resides in at least α exons. In addition, GenMCS is also applied with inputs $G_{U_{ESS}}$ and P_{ESS} to discover potential silencers as well. Combining the output from the two runs of GenMCS, the second level of CoSREM is then applied for **identifying MCS collections**. MCS collections are sets of cohesive subgraphs, whether they represent enhancers or silencers, that occur in at least θ exons. The output is sets of potential regulatory elements that are grouped together. The final step is **filtering the resulted MCS collections**. Each subgraph in an MCS

collection is mapped to the actual sequence in the associated exons. The resulted sequences are checked for overlapping. In case of overlapped sequences, they are replaced by one longer k-mer, which is evaluated to be included or eliminated in the final output.

5.4.3 Building the SRE graphs

The 6-dimensional de Bruijn graph $G = (V, E)$ over the DNA alphabet $\Sigma = \{A, C, G, T\}$ is constructed. The G graph has 4096 vertices and 16,384 edges. As mentioned earlier, the LEIsc scores (calculated in [58]) of potential exonic enhancers and silencers are utilized. If a specific 6-mer was found to be an enhancer or silencer, I use its associated LEIsc score. If it is defined as neutral, I consider its LEIsc value to be zero. All the scores are ordered in descending order and I associate each vertex v in the G graph with its rank. The rank suggests the strength of the effect of a specific 6-mer on splicing. Hence, the higher the rank, the greater the evidence of the enhancing activity for that specific 6-mer, and the lower the rank, the greater the evidence of the silencing activity. Let R be a predefined number of ranks. A set U_{ESE} is constructed by choosing the top R vertices by rank to create the SRE graph $G_{U_{ESE}} = (U_{ESE}, E')$, and the lowest R vertices by rank to create the SRE graph $G_{U_{ESS}} = (U_{ESS}, E'')$ as well.

5.4.4 Constructing the SRE profile matrices

Two SRE profile matrices (P_{ESE} and P_{ESS}) are then constructed based on the vertices in the SRE graphs $G_{U_{ESE}}$ and $G_{U_{ESS}}$ respectively. Utilizing the human coding exon database, I set $p_{i,j}$ equal to 1 or 0, according to the presence or absence of 6-mer y_i in exon x_j. The search for 6-mers in the exons is limited to the first 50 nucleotides as I showed that extending the exonic flanking length does not affect the results significantly [14].

5.4.5 Discovering maximal α-cohesive subgraphs (MCSs)

Given an SRE graph G_U and an SRE profile matrix P, the algorithm GenMCS from [4] is modified to find maximal α-cohesive subgraphs. GenMCS takes as an input, in case of ESEs, the SRE graph ($G_{U_{ESE}}$), the SRE profile matrix (P_{ESE}), and the user-defined threshold α. It starts by pruning all vertices that do not satisfy the threshold requirement. Then, starting from each vertex as an initial subgraph G, GenMCS extends the subgraphs in a depth first search manner. Each initial subgraph G will be extended with its neighboring vertices. GenMCS checks if the extended subgraph G' with one neighbor vertex will generate an α-cohesive subgraph (i.e a subgraph with its vertices sharing at least α exons, where $T(G') \geq \alpha$). If this is the case, GenMCS will proceed in a depth-first fashion to extend G'. If subgraph G cannot be extended without violating the α threshold, then G is a maximally

α-cohesive subgraph. Two pruning strategies are applied in the original algorithm to reduce the search space: if the extended subgraph has been seen before or if it is subsumed by any of the other discovered cohesive subgraphs. I modified GenMCS not to apply the second pruning strategy as, due to the nature of the data, it is allowed to have overlapping subgraphs with common vertices as long as the common exons are not the same. These overlapped subgraphs represent different SREs with some common nucleotides.

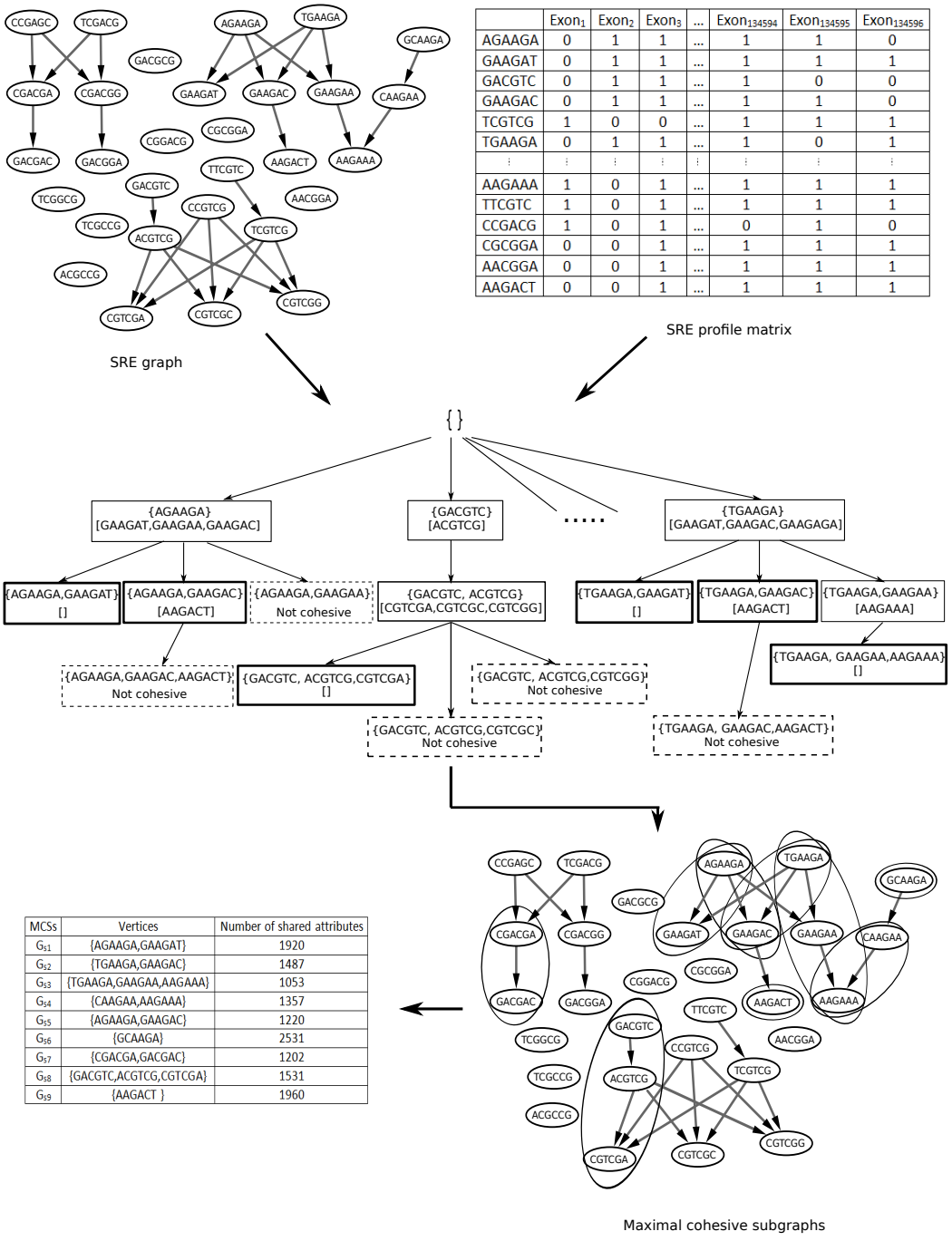
Figure 5.1 illustrates an example of the algorithm in case of ESEs. The output is a table called *MCStable*. It consists of maximal cohesive subgraphs and each subgraph is associated with a set of exons where the splicing enhancer, which this subgraph represents, resides. I apply GenMCS utilizing $G_{U_{ESS}}$ and P_{ESS} as inputs to get potential silencers as well.

5.4.6 Identifying MCS collections

The output from the first level of CoSREM is all the maximal α-cohesive subgraphs (MCSs), whether they represent enhancers or silencers, with their associated exons. The next step is to find collections of these already discovered subgraphs that share at least θ exons. To find such MCS collections, an *MCStree* is built. The *MCStree* is an ordered search tree, where each vertex contains an MCS collection M and its associated exons $T(M)$. The root of the *MCStree* is a vertex with $M = \emptyset$ and $T(M)$ containing all the exons.

The algorithm, given in Figure 5.2, uses a depth first search approach to build the *MCStree*. It takes the *MCStable* as an input. *MCStable* is a hash table where the MCS IDs are the keys and the exon set of each MCS is the value. Each vertex at the first level of the tree represents one of the already calculated MCSs as an initial M. Therefore, the exon set $T(M)$ is the exon set of the corresponding MCS (line 6). A child vertex u of vertex v is generated by extending M_v with one of the remaining MCSs and $T(M_u)$ is then calculated as depicted in Figure 5.3. As the *MCStree* is an ordered tree, M_v is extended by adding an MCS whose ID is only bigger than the largest MCS ID in the collection. Different pruning strategies are applied to reduce the search time and space. One pruning strategy is that the tree branches are extended in a depth-first manner as long as the generated M in the current vertex has shared exons with size $|T(M)| \geq \theta$. Once this constraint is violated, this branch is pruned (see Figure 5.3). Another strategy is to prune the branch if the generated M has been generated in a previous part of the tree with the same exon set. Figure 5.4 illustrates an example of an *MCStree*.

After building the *MCStree*, a breadth-first search (BFS) is applied to identify the path from the root to each vertex in the tree. Only vertices with distance $\geq \beta$ from the root are included in the results. Each vertex represents an MCS collection and its distance from the root represents the number of MCSs in that collection (Figure 5.4).



	$Exon_1$	$Exon_2$	$Exon_3$	...	$Exon_{134594}$	$Exon_{134595}$	$Exon_{134596}$
AGAAGA	0	1	1	...	1	1	0
GAAGAT	0	1	1	...	1	1	1
GACGTC	0	1	1	...	1	0	0
GAAGAC	0	1	1	...	1	1	0
TCGTCG	1	0	0	...	1	1	1
TGAAGA	0	1	1	...	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
AAGAAA	1	0	1	...	1	1	1
TTCGTC	1	0	1	...	1	1	1
CCGACG	1	0	1	...	0	1	0
CGCGGA	0	0	1	...	1	1	1
AACGGA	0	0	1	...	1	1	1
AAGACT	0	0	1	...	1	1	1

MCSs	Vertices	Number of shared attributes
G_{s1}	{AGAAGA,GAAGAT}	1920
G_{s2}	{TGAAGA,GAAGAC}	1487
G_{s3}	{TGAAGA,GAAGAA,AAGAAA}	1053
G_{s4}	{CAAGAA,AAGAAA}	1357
G_{s5}	{AGAAGA,GAAGAC}	1220
G_{s6}	{GCAAGA}	2531
G_{s7}	{CGACGA,GACGAC}	1202
G_{s8}	{GACGTC,ACGTCG,CGTCGA}	1531
G_{s9}	{AAGACT }	1960

Figure 5.1: An example of mining cohesive subgraphs. The graph at the top left corner represents the SRE graph $G_{U_{ESE}}$. I choose R = 30 which means the SRE graph contains the top 30 6-mers in rank. The matrix on the right is the SRE profile matrix P_{ESE}. Setting $\alpha = 1000$ means that the connected vertices should co-occur in at least 1000 exons to be considered a cohesive subgraph. The tree in the middle shows how GenMCS proceeds. The bold boxes represent cohesive subgraphs. The dotted boxes represent subgraphs that are not cohesive and the remaining branch will be pruned. The output is 9 subgraphs as illustrated in the bottom graph.

Algorithm 6 BuildMCStree

Input: $MCStable, Graphs = \{G_{S_1}, G_{S_2}, ..., G_{S_r}\}$ a list of MCSs ordered by their IDs
Output: $MCStree$

Add a root vertex t to the $MCStree$ such that $M = \emptyset$
for $i \in \{0, .., r\}$ **do**
  $G = Graphs[i]$
  $T(G) = MCStable[G]$
  if $T(G) \geq \theta$ **then**
    Add a vertex v to $MCStree$ such that $M_v = \{G\}$ and $T(M_v) = T(G)$
    Add edge (t, v) to $MCStree$
    ModDFS$(i, v, MCStree, Graphs, MCStable)$
  end if
end for
return $MCStree$

Figure 5.2: BuildMCStree algorithm: Build the $MCStree$.

Algorithm 7 ModDFS

Input: $i, v, MCStree, Graphs, MCStable$
for $j \in \{i + 1, ..., n\}$ **do**
  $G = Graphs[j]$
  $T(G) = MCStable[G]$
  $R = T(M_v) \cap T(G)$
  if $R \geq \theta$ **then**
    $M_u = M_v \cup G$
    Flag = TRUE
    if There exists $k \in MCStree : [M_u \subseteq M_k$ and $T(u) = T(k)]$ **then**
      Flag= FALSE
    end if
    if Flag **then**
      Add vertex u to $MCStree$ where $T(M_u) = R$
      Add edge (u, v) to $MCStree$
      ModDFS$(j, u, MCStree, Graphs, MCStable)$
    end if
  end if
end for

Figure 5.3: ModDFS: An algorithm to recursively extend the MCS collections.

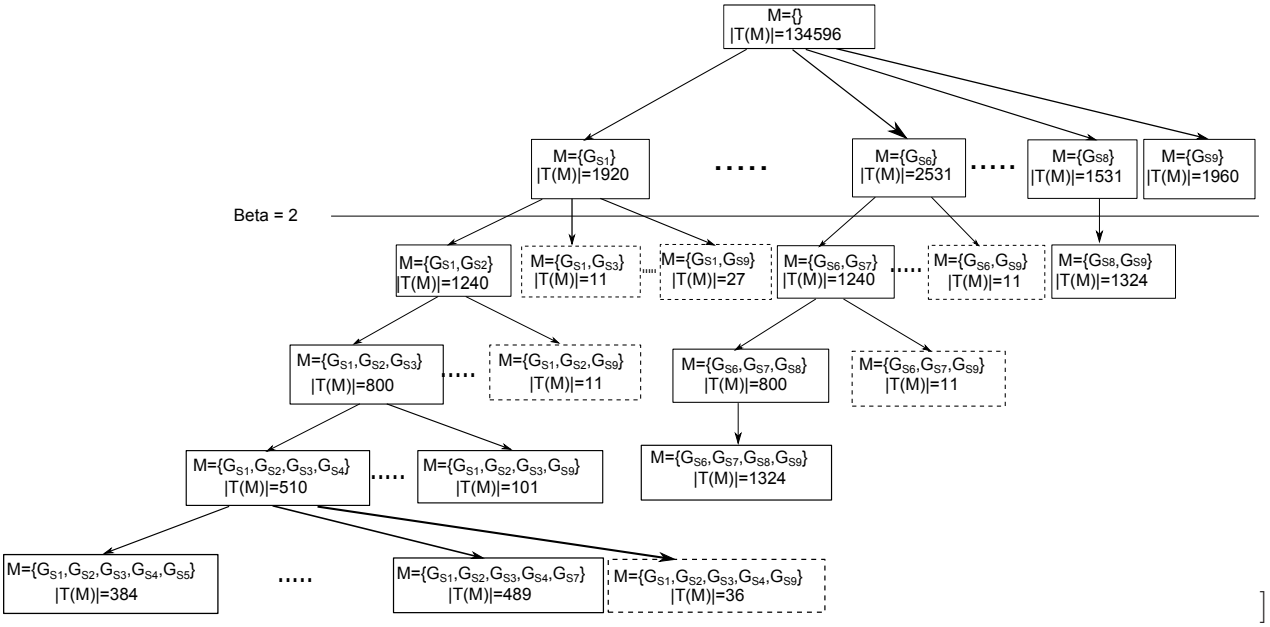


Figure 5.4: An example of an *MCStree*. The example shows a part of the tree where $\theta = 100$. The dotted boxes means that this MCS set does not satisfy the user threshold $T(M) \geq \theta$, where $T(M)$ is the number of shared exons between the MCSs, and this branch will be pruned. all vertices with distance from the root $\geq \beta$ threshold will be considered as potential MCS collection.

5.4.7 Filtering the MCS collections

The output of CoSREM is all MCS collections, which represent sets of potential enhancers and silencers that co-occur in specific sets of exons. The goal of the filtering step is to generate the corresponding sequences for each MCS collection. As overlapping is allowed between sequences in the first level of CoSREM, there is a possibility to have multiple regulatory elements that form a co-occurring MCS collection but they are actually overlapping sequences in the exons. As a result, they can be considered as one longer k-mer instead. Therefore, I replace the overlapping SREs of the same type (ESEs or ESSs) with one longer SRE. That may result in an MCS collection with only one long SRE, or still multiple SREs if not all of them are overlapping. In the former case, this MCS collection will be eliminated from the results. On the other hand, if the set contains both enhancers and silencers, overlapping between sequences is allowed as that is in accordance with the complex interplay between enhancers and silencers [66].

Therefore, for each MCS collection M, the corresponding sequences of each subgraph are generated. This is performed by applying a depth first traversal as in [14]. I eliminate the generated sequences that are subsumed by other sequences. Then, I check each exon in

$T(M)$ to locate these sequences and generate the new SREs if some of them are overlapped. For example, one of the MCS collections contains these four ESEs: $CCCGGA$, $CCGGAG$, $CGGAGC$, and $GGAGCC$. These sequences are overlapped in some of the exons in the associated exon set, forming one 9-mer element $CCCGGAGCC$. In this case, I consider it only one ESE, and I do not include it in the final results. In other cases, multiple sequences are generated such as ($CCCGGAGC$, $GGAGCC$), ($CCCGGAG$, $CGGAGCC$), and ($CCCGGA$, $CCGGAGCC$).

5.5 Results

5.5.1 Combinatorial SREs discovered in human exons

I applied CoSREM on all unique human coding exon data set as I indicate in Section 4.3.1. For predicting combinatorial SREs, I chose the highest and the lowest 400 6-mers by LEIsc values to construct the SRE graphs $G_{U_{ESE}}$ and $G_{U_{ESS}}$. These values were chosen since most of the analysis done by Ke et al. [58] on their produced LEIsc scores, which I utilize, was on the highest or the lowest 400 LEIsc scores. However, any number can be chosen based on the utilized data. I chose the user defined constraints α, θ, and β to be 1000, 100, and 2, respectively. Different values for α and θ have been tried as I will illustrate in Section 5.6. I chose β to be 2 to discover at least pairs of SREs.

GenMCS (the first level of CoSREM) produced 264 potential exonic regulatory elements as illustrated in Supplementary Table S1 [7]. That includes 175 enhancers and 89 silencers. Building the *MCStree* (the second level of CoSREM) generated 745 MCS collections as depicted in Supplementary Table S2 [7]. Filtering the results and generating the corresponding sequences, I generated 37 combinatorial SRE sets. That includes 30 sets of both enhancers and silencers and seven sets of co-occurring enhancers. The resulting regulatory element lengths are between 6-mers and 7-mers. The results are shown in Tables 5.1, and 5.2, where I also utilized SpliceAid-F [46] and the ESEfinder tool [27] to evaluate the resulting regulatory elements and whether they bind to known splicing factors.

Since the predicted SREs are of variable length, as are SpliceAid-F binding sites, I checked if my SREs are totally contained in at least one binding-site in the database or vice versa. Hence, I retrieved the associated splicing factor.

It should be noticed that although 37 combinatorial SRE sets were generated, the actual number of enhancers and silencers appeared in these sets are 25, 14, respectively. With total number of 39 SREs. This supports the known complex relationship between enhancer and silencer elements and that alternative splicing is a complex process that involves cooperative or competitive interplay between both types. These combinatorial SREs can be the basis to identify context-dependent regulation where the regulatory element behavior does not only depend on its sequence but also on its neighboring sequences [99].

Table 5.1: Combinatorial enhancers and silencers generated by CoSREM, the number of exons each set resides in, and the splicing factors that they may bind to according to SpliceAid-F [46] and ESEfinder tool [27].

Combinatorial SREs	Number of exons	Splicing factors
CGGGAG,GGGAGG	526	hnRNP A1
GAAGGC,AGGCAG	373	9G8,SC35,SF2/ASF
GCTGTC,TGTCAG	254	-
GAGGAC,GGGAGG	233	SF2/ASF*, hnRNP A1
CCGGGA,GGGAGG	229	hnRNP A1
AGAGAC,TAGAGA	218	-
GGAGTC,AGTCAG	217	-
GAAGTC,AGTCAG	213	-
TGAGGA,GGTGAG	200	SF2/ASF
CCGGGAG,GGGAGG	199	hnRNP A1
GCGGGA,GGGAGG	190	hnRNP A1
GATGTC,TGTCAG	171	-
GCGGGAG,GGGAGG	169	hnRNP A1
AGAGGA,AGGCAG	156	FMRP
GCAAGA,GTGCAA	154	-
GTGAAGA,AGGTGA	153	SF2/ASF
GAGGAT,GGGAGG	147	SF2/ASF**,hnRNP A1
AGAGGA,CAGCCA	133	FMRP ,hnRNP L
TGAGGA,AGGCAG	129	-
GATGCC,TGCCTA	127	SRp55*
GGAGCC,AGGTGG	114	-
GGAGCC,CCCACC	114	-
TGGACC,AGGTGG	112	-
TTCAAC,CTTTCA	112	SRp40*,hnRNP E1
TTCATC,CTTTCA	110	YB-1,SRp55*,hnRNP E1
GAACAA,AGGTGA	106	-
CAAGGA,CAGCCA	103	FMRP,hnRNP L
TGAGGA,AGGTGG	103	-
TGAGGA,AGGTGA	102	-
CAAGGA,TCCCAA	100	SRp40*,FMRP

[a]* identifies splicing factors identified by ESE finder.
** means the splicing factors is identified by both methods.

Table 5.2: Combinatorial enhancers generated by CoSREM, the number of exons each set resides in, and the splicing factors that they may bind to.

Combinatorial SREs	Number of exons	Splicing factors
AGAGGA,TGAGGA	185	FMRP
GAAGGC,TGAGGA	113	9G8,SC35,SF2/ASF
CAAGGA,TGAGGA	105	-
AGAGGA,GATGGA	104	FMRP
AGAGGA,CAAGGA	103	FMRP
AGAGGA,GAGGAC	101	FMRP
GGAGCC,TGAGGA	100	-

Figure 5.5 illustrates the the relationship between enhancer and silencer elements in my combinatorial SRE sets. It indicates the many-to-many relationship where, one enhancer element can co-occur with multiple silencers and vice versa. This many-to-many relationship does not only include regulatory elements of different types, it can also contains regulatory elements of the same type. For example the enhancer element $AGAGGA$ co-occur with other enhancers ($CAAGAA$,$GATGGA$,$TGAGGA$,$GAGGAC$).

5.5.2 Comparison with other data sets

Among the 39 SREs in the resulted combinatorial SRE sets, 35 were included in my previous results [14]. I also compared these results with previously published databases. I utilized exonic binding sites from SpliceAid-F [46]. I utilized the same approach I previously used in [14] to identify 59 exonic enhancers from SpliceAid-F. As stated earlier, since the predicted SREs are of variable length, as well as SpliceAid-F binding sites, the overlap between the two sets is calculated by finding whether each sequence in the first list is totally contained in the second list or vice versa. Another database is AEdb [86], which includes 294 splicing regulatory elements. I only utilized human enhancers (64 sequences) and silencers (24 sequences). I utilized PESE and PESS data sets as well [110]. Table 5.3 summarizes the overlapping results. Overall, 88% of the enhancers and 64% of the silencers I identified in my combinatorial SRE sets can be mapped to previous data sets.

Table 5.3: Number of overlapped enhancers and silencers from my combinatorial SRE sets with previously published data sets. The numbers between brackets are the number of enhancer and silencer elements in my SRE sets.

	SpliceAid-F	AEdb	PESE	PESS
Enhancers (25)	8	7	19	-
Silencers (14)	4	3	-	4

I also compared the results with results from [58]. Those authors identified 232 and 262 6-

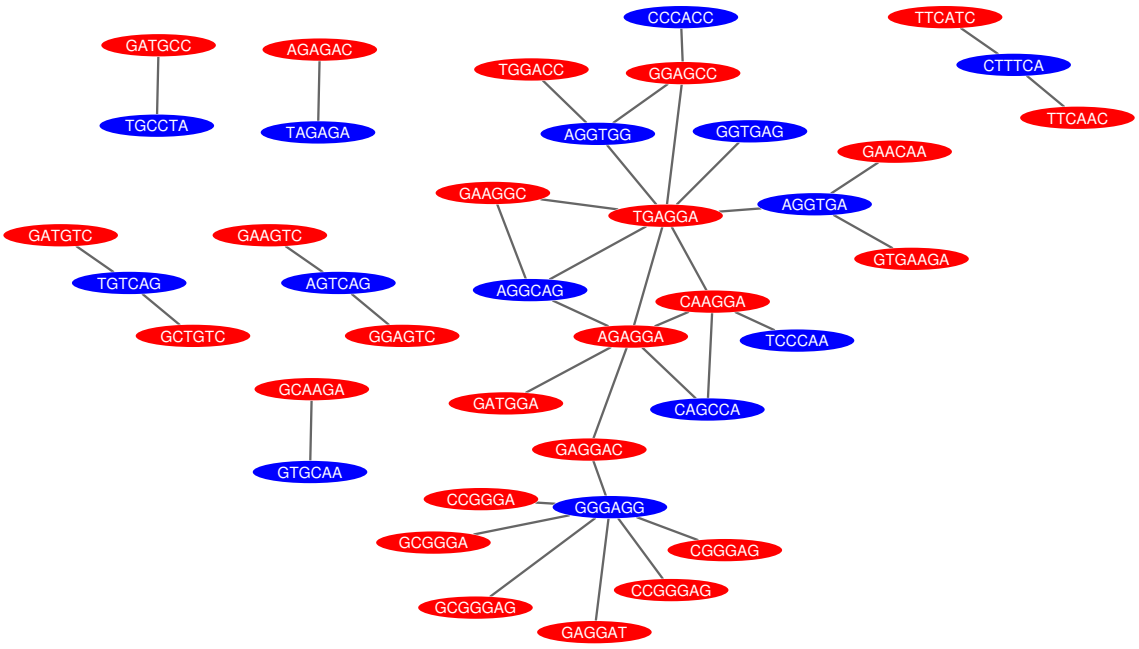


Figure 5.5: A regulatory network for enhancers and silencers. The red nodes represent enhancer elements, and the blue ones represent silencer elements. The network illustrates the many-to-many relationship between the enhancers and silencers.

mers that could have potential positive or negative synergy with other 6-mers. The authors did not identify an actual set of combinatorial 6-mers. From the 37 combinatorial SRE sets, 20 sets had at least one 6-mers from their list [58]. Most of the current approaches are applied on intronic regions [60, 103, 57, 43, 87]. Therefore, I was not able to utilize their results for verification.

I also wanted to verify whether the SRE sets I found are significant. To address this issue, the same number of exons that I have in my database was randomly generated and I applied CoSREM with the same threshold values. Although the number of generated MCS collections was considerably larger in the random case (4853), the filtering stage did not yield any results, as the generated groups did not pass the threshold $\theta = 100$.

5.5.3 SRE set (GAGGAC,GGGAGG) and the role it may play in cancer progression

I further investigated some of the combinatorial SREs. The SRE set ($GAGGAC$, $GGGAGG$) was chosen as it is one of the highest ranked sets, according to the number of exons it resides in, and they are potential binding sites to both types of splicing factors (SR proteins and hnRNP proteins) as illustrated in table 5.1. For example, when checking the exons that the SRE set ($GAGGAC$, $GGGAGG$) resides in, the two SREs were overlapped in most of

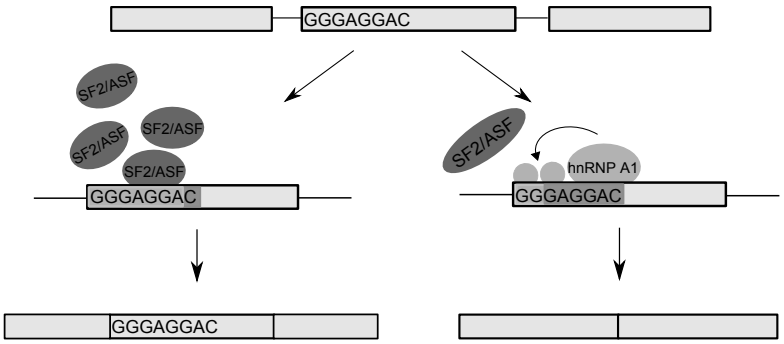


Figure 5.6: Possible combinatorial effect of the overlapped SREs ($GGGAGGA, GAGGAC$). One possible scenario is having SF2/ASF splicing factor with great affinity. It binds to the ESE and stimulate exon inclusion. Another possibility is if the splicing repressor hnRNP A1 exists, it may inhibit the exon inclusion by binding to the silencer sequence and recruit the binding of other inhibitory factors which extend to the exon boundary and prohibit the binding of the SF2/ASF protein. As a result, the exon will be skipped. The rectangles in this figure represent exons and lines represent introns.

the sequences, constituting the sequence $GGGAGGACA$. I utilized the Human Splicing Finder tool [51] to validate whether the sequence contains both an enhancer and a silencer as predicted. Human Splicing Finder is a tool to identify splicing motifs utilizing all the already known SRE experimentally and computationally. It also provides the splicing factors the sequence binds to if they are known. Utilizing Human Splicing Finder, the sequence $GGGAGGACA$ is found to have the ESE motif $GGGAGGA$, among other motifs, where the splicing factor SF2/ASF binds. It also contains the ESS motif $GAGGAC$ that binds to the splicing factor hnRNP A1.

This is one of the known classical examples of the combinatorial effect of having both an ESE and an ESS in adjacent positions. There are several studies that report the antagonistic behavior between the SF2/ASF and hnRNP A1 splicing factors [66, 105]. For example, in exon 3 of the HIV1 tat gene, the hnRNP A1 splicing factor may bind to an ESS and inhibit splicing by propagating hnRNP A1 molecules further towards the 3' splicing site. That propagation behavior can be inhibited by the SF2/ASF splicing factor when it binds to an ESE that resides upstream of the ESS, as in my sequence [68, 52, 113, 66, 105]. Furthermore, Mayeda et al. [67] showed *in vitro* that having different ratios of SF2/ASF to hnRNP A1 promotes exon skipping or inclusion by binding to different ESEs or ESSs. Therefore, that could provide an understanding of what might be the possible outcomes of combinatorial splicing regulation (Figure 5.6).

I further investigated the exons in the genes that have this SRE set and identified by CoSREM utilizing TCGA Spliceseq [78]. TCGA is an AS database that utilizes RNA-Seq samples from The Cancer Genome Atlas project to provide the splicing patterns differences between different tumor samples and between tumor and normal samples. Several of these exons

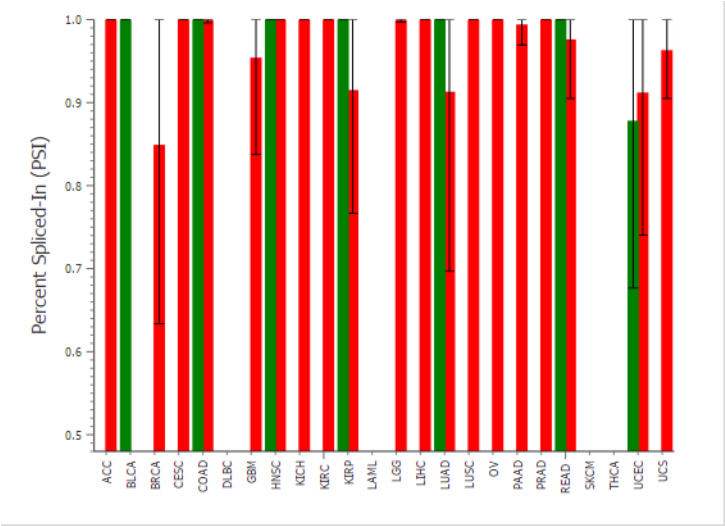


Figure 5.7: A bar plot of the PSI (Percent Spliced-In) values of exon 17 in PRKCG gene. It illustrates the difference in the PSI values between normal and tumor samples. The red bars represent the PSI of tumor samples while the green bars represent the normal samples. This figure is generated using TCGA Spliceseq [78].

were found to be included in several samples of different cancer types and skipped in the normal samples. For example, exon 17 in the PRKCG gene is included in 100% of all the transcripts of the samples for lung squamous cell carcinoma (LUSC), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), prostate adenocarcinoma (PRAD), and kidney chromophobe (KICH), while skipped in 100% of all the transcripts of the normal samples, as shown in Figure 5.7. The inclusion or exclusion of these exons may be related to the antagonistic behavior of their positive and negative regulators that I identify. PRKCG is known to be a major receptor for phorbol esters, a class of tumor promoters. As abnormal splicing events are a major contributor to cancer development [53], understanding the reasons behind specific exon inclusion or exclusion can play a role in understanding cancer. The complete list of exon skipping events is shown in Supplementary Table S3 from[7].

I also utilized Ontologizer [20] to identify the enriched GO terms for the same set of genes. GO enrichment analysis is performed utilizing the Topology-Elim algorithm. Then, the Westfall-Young Single Step multiple testing correction procedure is applied. The most enriched biological process was "activation of Ras GTPase activity" with adjusted p-value 0.00028, meaning any process that initiates the activity of Ras superfamily members. It is known that Ras family genes are oncogenes [75, 47, 35]. Several human tumors have constitutively active Ras proteins. The activation can be caused by mutations in the Ras genes or by modifications in the upstream or downstream signaling components in Ras pathways [35]. Supplementary Table S4 [7] contains the complete list of the biological processes that the

predicted SRE sets are involved in.

5.6 Discussion

I introduce CoSREM, a graph mining algorithm, to discover co-occurring groups of exonic enhancers and silencers. CoSREM utilizes experimental data to increase the accuracy of the results. Using a de Bruijn graph formalism allowed the identification of splicing regulatory elements with different lengths without any prior assumptions on SRE size.

One of the advantages of this algorithm is its generality. CoSREM is designed to discover multiple SREs not only pairs as with the current approaches. My current results do not include multiple SREs but the reason for that is the filtering step. In fact, the MCS collections that resulted from CoSREM include several larger sets of SREs, not only pairs (see Supplementary Table S2 [7]). As stated before, in the filtering step, I assume if SREs of the same type are overlapped, they constitute one longer SRE. This is one possibility to consider. Another possibility is that they are different regulatory elements that overlap and may have either cooperative or competitive behavior [48]. I chose to focus on the first possibility in my analysis. However, CoSREM provides the results for both possibilities. I provide both outputs in my open source package. So, the user can analyze both possibilities.

Another advantage is its flexibility. Utilizing a de Bruijn graph-based model allows building the main graph from any k-mer (based on the available data). The number of vertices chosen to build the SRE graphs can change according to the data as well. In my case, I utilized the LEIsc scores as a measurement for ranking 6-mers. The rank can be based on other criteria such as conservation scores or other data sources. For example, utilizing data from CLIP experiments where both the RNA binding protein and the location of its binding site is experimentally identified [92]. Having a list of all protein binding sequences that are experimentally verified can increase the probability of having a certain k-mer as a putative SRE if a part of the sequence is in that list. CoSREM can be applied on different parts of the genome as well to identify combinatorial SREs. For example, it can be applied to identify combinatorial SREs in both of the exonic flanking regions. It can be applied on intronic regions as well, depending on the provided data. I applied CoSREM on the first and last 50 nucleotides in the exons to discover SREs group that co-occur in both regions, and I found several co-occurring ESEs and, in some cases, the same ESE is repeated in these two different parts of the exon as shown in Supplementary Table S5 [7].

Another aspect of CoSREM flexibility is the ability to choose the user defined thresholds. I have tried several values for the thresholds α and θ. As illustrated in Figure 5.8, as α increases, the number of potential SREs decreases while the number of MCS collections increases and then decreases. This behavior can be explained, as α is the minimum number of exons that an SRE should reside in, and with increasing α, SREs that satisfy this constraint decreases and longer k-mer SREs are eliminated. However, as I set the θ threshold to a

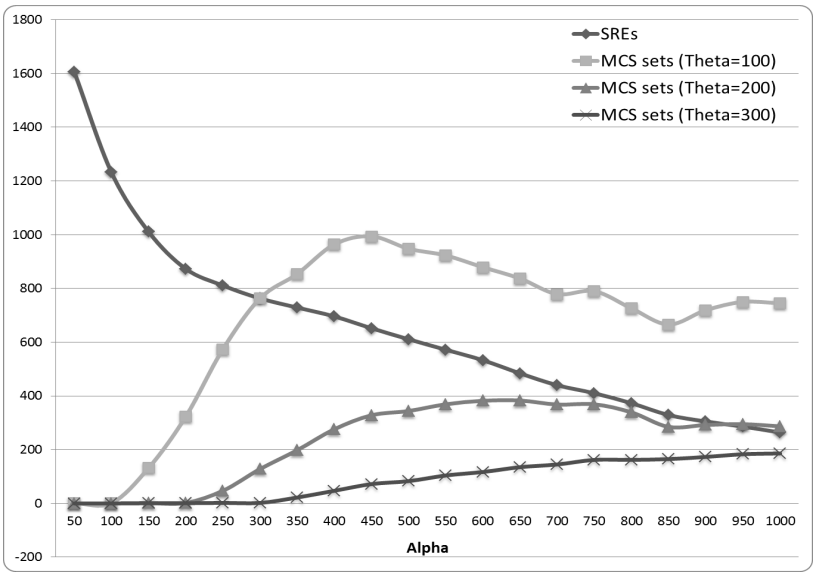


Figure 5.8: The number of generated MCSs and MCS sets using different values of α and θ.

relatively small number ($\theta = 100$), some of these longer k-mers are combined again as co-occurring groups and this is the reason for the increasing number of combinatorial SREs. Eventually with the constant decreasing number of the resulted SREs, the number of the resulting MCS collections are decreased. I chose α to be 1000 to have a reasonable number of common exons between 6-mers to start with. Another reason is the time performance as shown in Figure 5.9. The θ threshold eliminates only the groups with smaller exon sets. This is why I chose θ to be a small number relatively to have all the results for further filtering. I tried CoSREM with $\alpha = 500$ which resulted in 11 combinatorial SRE groups. These groups were a subset of the previous results with $\alpha = 1000$.

The ability to identify genes with different splicing events between normal and tumor samples, as in the case of the PRKCG gene, may shed further light on the important role that SREs may play in cancer progression and open the door for further experimental validation. Wan [93] introduces a protocol to manipulate the AS of exon 15 of the HER2 gene. Utilizing splice switching oligonucleotide (SSO), the splice site or an exonic enhancer is targeted to induce exon 15 skipping. That results in down-regulating the expression of HER2 mRNA and protein expression in HER2-overexpressing breast cancer cell line SK-BR-3. In fact, PRKCG has analogous behavior to HER2 where exon 17 is included in 100% of the transcripts in case of LUSC, KIRC, LIHC, PRAD, KICH cancer samples and skipped in the normal tissues as discussed earlier. That may open the way for further experimental validation.

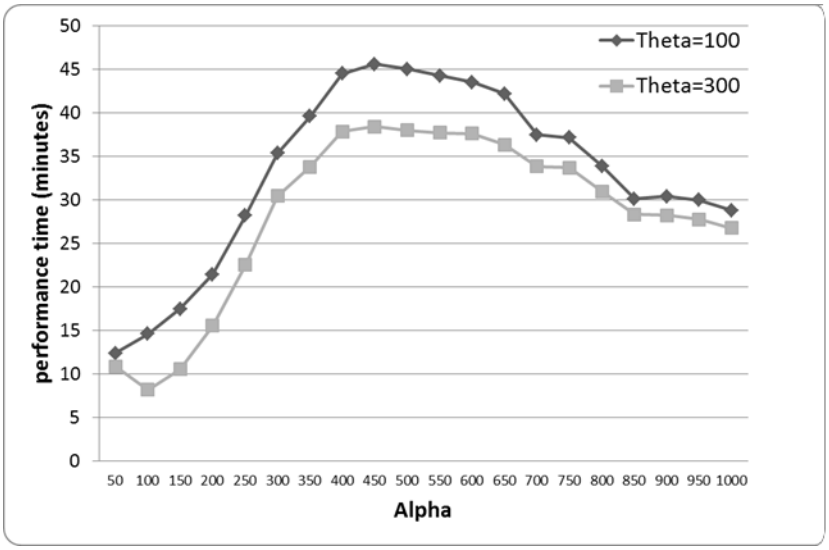


Figure 5.9: CoSREM time performance using different values for α in case of $\theta = 100$ and 300.

Conclusion

I have presented CoSREM, a graph mining algorithm to discover combinatorial SREs. Utilizing this approach allowed the identification of different combinations of splicing enhancers and silencers without assuming a predefined size or limiting the algorithm to find only pairs of SREs. This approach can open new directions to study SREs and the roles that AS may play in diseases.

Chapter 6

Computational Identification of Tissue-specific Splicing Regulatory Elements in Human Genes from RNA-Seq Data

6.1 Introduction

Alternative splicing plays a key role in tissue-specific expressed genes [104, 95]. Tissue-specific alternative splicing is regulated by a combination of tissue-specific and ubiquitously expressed RNA-binding factors [94]. They interact with the splicing regulatory elements to affect the spliceosome assembly (splicing machinery) and consequently the transcribed isoforms. There are several splicing factors that activate or repress splicing in different contexts [94]. Skipped exons are one of the notable alternative splicing events between different tissues. The authors in [94] suggested that these exons have more biological importance and that having that switch-like regulation between tissues requires additional splicing regulatory elements to be present. Most work in mammalian systems revealed that AS decisions are often made by a combinatorial action of general and tissue-specific regulators [105]. Even simple tissue-specific decisions can involve additional layers of complexity, where regulatory elements cooperate or compete with each other [85].

In this work, I performed genome-wide analysis to study alternative splicing on multiple tissues (brain, heart, liver, and muscle). The RNA-Seq data set from the Human BodyMap project [42] was utilized. I used DEXSeq [5] to identify tissue-specific exons. Then, I applied my algorithms, GenSRE [14] and CoSREM [15], to identify both individual and combinatorial regulatory elements responsible for exons that exist in one tissue but not in other tissues. Putative tissue-specific enhancers were discovered and a complicated enhancer regulatory

network was revealed. Multiple enhancers were found across multiple tissues, while some were found only in specific tissues. Putative combinatorial enhancers and silencers were discovered as well that may be responsible for exon inclusion or exclusion across tissues. This is, to my knowledge, the first analysis to focus only on discovering exonic regulatory elements (individual and combinatorial) across tissues.

6.2 Data Set

RNA-Seq data from the Human BodyMap 2.0 project is utilized [42]. This data originates from 16 different human tissues. It contains 50 bp paired-end reads, 75 bp single-end reads, and 100 bp single-end reads. I focus on four tissues, namely, brain, heart, liver, and muscle.

6.3 Overview of the Proposed Pipeline

In this pipeline, different tools are utilized to identify tissue-specific exonic regulatory elements. The first stage, as illustrated in Figure 6.1, is to identify tissue-specific exons. To do that, I utilized DEXSeq [5]. DEXSeq identifies exons that are differentially used between two tissues. I compare each tissue of interest with the other three tissues. The output of this stage is three sets of differentially used exons in the tissue of interest but not in the remaining tissues. GenSRE [14] is then applied separately on the exon sets to identify exonic enhancers. I identify tissue-specific enhancers by determining the common enhancers across the sets. CoSREM [15] is also applied to identify co-occurring exonic enhancers and silencers that may be responsible for exon inclusion or exclusion across tissues.

6.4 Results

6.4.1 Differentially used exons between tissue pairs

In this study, I analyzed four tissues from the RNA-Seq data of the Human BodyMap project [42]. DEXSeq [5] was utilized to identify differentially used exons between pairs of tissues. Table 6.1 illustrates a part of the DEXSeq output for each pairwise comparison.

For each tissue pair comparison, DEXSeq produces a list of exons that are differentially used in one tissue but not in the other tissue. These exons may be whole exons or parts of exons. Being conservative, I define *tissue-specific exons* as differentially used exons with $p \leq 0.05$, $\log_2$ fold change ≥ 2 or ≤ -2, and exons that have some reads in one tissue and zero reads in the other tissue. I preferred to have a strict set of exons that exist in one tissue against the other tissue. Table 6.2 illustrates the number of exons resulting from each pairwise

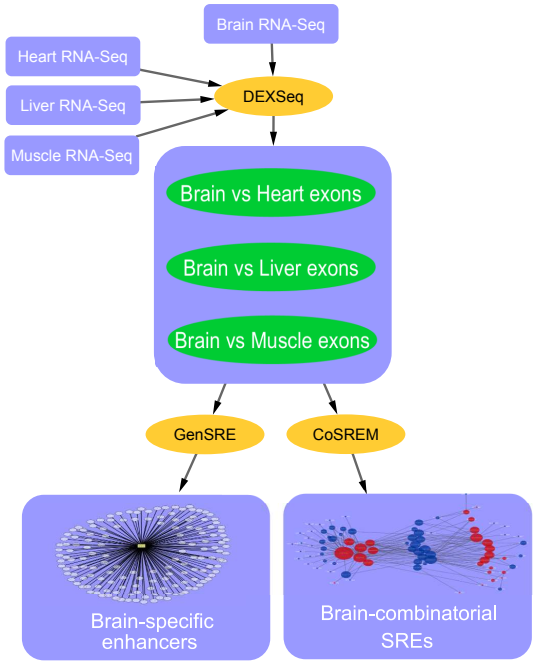


Figure 6.1: An example of the proposed pipeline applied on the brain tissue.

comparison. For example, applying DEXSeq on brain and heart tissues resulted in 1423 exons that are brain-specific. In the same manner, 857 exons were found to be heart-specific while having zero reads in brain tissue. It should be noticed that the brain tissue has the largest number of tissue-specific exons, which is consistent with results reported in [60] that the brain has a large number of tissue-specific alternative spliced exons. Six pairwise tissue comparisons were performed. Therefore, I have 3 different sets of exons for each tissue with a total of 12 tissue-specific exon sets for all 4 tissues.

6.4.2 Tissue-specific exonic enhancers

I used GenSRE algorithm to identify enhancer elements in each tissue [14]. Having three exon sets for each tissue, GenSRE was applied on each set separately. Table 6.3 indicates the number of enhancers identified in each set. The complete list of the identified enhancers is listed in Supplementary Tables S1–S6 [10]. I focused on identifying putative exonic enhancers that may play a role in the inclusion of these exons within these specific tissues. GenSRE identifies variable length SREs in the exonic flanking regions. Table 6.4 contains the number of exons that were utilized. Overall, I identified 1929 exonic splicing enhancers (ESEs) across the four tissues whose lengths range from 6 to 15 nucleotides.

Table 6.1: An example of DEXSeq output for brain and heart tissues. The ID column lists the gene name and the exon number. The stat column includes the likelihood ratio test (LRT) statistic value. Brain and heart columns contain the exon usage coefficients for both tissues. The count columns include the actual counts of the mapped reads.

ID	stat	padj	brain	heart	$\log_2$ fold	count1	count2
chr8_ANK1-:E006	124.56	4.08E-26	4.39	20.56	2.23	5	357
chr10_ABLIM1-:E015	121.81	1.58E-25	3.62	18.91	2.39	1	507
chr10_ACBD7-:E003	7.75	0.04	8.47	0.36	-4.57	415	0
chr10_ADD3+:E015	68.33	3.04E-14	12.16	1.59	-2.94	144	0
chr10_CCSER2+:E006	19.46	0.00027	3.59	15.95	2.15	7	161
chr10_NEBL-:E017	394.064	5.60E-84	3.88	18.06	2.22	1	1601

Table 6.2: Number of tissue-specific exons that are present in one tissue and excluded in the other tissue. The counts are tissue-specific to the rows, while the columns show the second tissue in a comparison.

	Brain	Heart	Liver	Muscle
Brain	-	1423	4592	7071
Heart	857	-	3952	6975
Liver	800	839	-	5002
Muscle	623	612	2495	-

To identify tissue-specific enhancers, I extracted the common set of enhancers that appeared in all the three exon sets for each tissue. This behavior suggests that these enhancers are tissue-specific as they repeatedly appeared in one tissue against all the other tissues. Table 6.3 illustrates the number of tissue-specific enhancers for brain, heart, liver, and muscle tissues. Supplementary Table S7 [9] includes the list of tissue-specific enhancers for all tissues.

Table 6.3: Number of identified putative enhancers in one tissue with respect to the other tissues using GenSRE algorithm.

	Brain	Heart	Liver	Muscle	Tissue-specific enhancers
Brain	-	449	793	923	205
Heart	250	-	695	877	85
Liver	277	398	-	752	50
Muscle	255	282	567	-	34

I also wanted to assess the accuracy of this pipeline in identifying tissue-specific enhancers. Therefore, for each tissue, I identified the set of exons that is present in one tissue but not in the other three tissues. This is done by finding the intersection between the different exon sets of each tissue. As stated earlier, each tissue has three sets of tissue-specific exons. These sets result from pairwise comparisons of one tissue with the other three tissues. For

Table 6.4: Number of utilized exons in GenSRE.

	Brain	Heart	Liver	Muscle
Brain	-	930	3132	4796
Heart	573	-	2638	4607
Liver	564	596	-	3404
Muscle	359	417	1634	-

example, in the case of the brain tissue, each exon set represents exons that are specific to the brain and not to the heart, liver, and muscle tissues, respectively. The common set of exons between these three sets represents brain-specific exons that are not present in all the other tissues, as illustrated in Figure 6.2.

Then, I searched for the tissue-specific enhancers that were previously identified in these unique sets of exons, to see if they are appearing in their exonic flanking regions or not. Table 6.5 illustrates the number of enhancers that were found for each tissue. For example, out of 205 exonic enhancers that I identify as brain-specific enhancers, I found 173 enhancers in the brain unique exon set, which represents about 85% of all the identified enhancer elements. Overall, I identified 46% of my tissue-specific enhancers in these exons. Although the percentage is not high, I believe that the reason is the small number of unique exons in some tissues. For example, as indicated in Table 6.5, for the heart tissue, although there are 85 enhancers identified as heart-specific, the number of heart unique exons is only 23 exons. In these 23 exons, 19 enhancers is found which is about 83% coverage for the exons. Overall, There is about 76% coverage of the unique exon sets, which suggests that this is a valuable approach to identify tissue-specific regulatory elements.

Table 6.5: A comparison table to identify the occurrence of tissue-specific enhancers in tissue unique exon sets (exons that are present in one tissue but not in all the other tissues).

	Unique exons	Tissue-specific enhancers	Enhancers found in unique exons	Overlapping	Coverage
Brain	173	205	174	85%	100.01%
Heart	23	85	190	22%	83%
Liver	32	50	190	38%	59%
Muscle	21	34	13	38%	62%
Overall %				45%	76%

6.4.3 Comparison with SRE databases

I compared these results with databases from [14, 39, 40, 46, 86, 110, 111]. I extracted only exonic enhancers and silencers to compare the results with. The same approach I used in [14] was utilized.

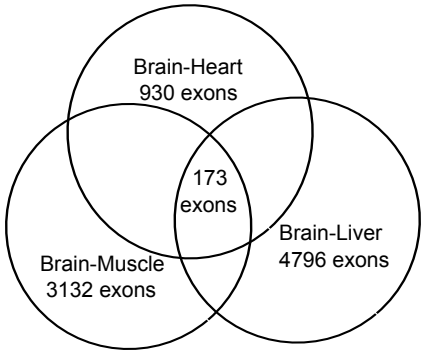


Figure 6.2: Each circle represents the number of brain-specific exons that resulted from brain pairwise comparisons with other tissues (heart, liver, and muscle). The intersection represents brain-specific exons against all other tissues.

I first compared my results with exonic binding sites from SpliceAid-F [46]. SpliceAid-F contains 330 different sequences for human. I utilized 59 binding sites of them as exonic enhancers. Since the predicted enhancer list are of variable length, as are SpliceAid-F binding sites, I calculated the overlap between the two sets by finding whether each sequence in the first list is entirely contained in at least one sequence in the second list or vice versa.

Another database is AEdb [86], where only the 64 human enhancers were considered.

In addition, I compared my enhancer list with five other computational data sets, such as the RESCUE-ESE [39] data set, the PESE [110] data set, and results from [14]. The fourth data set is from [40]. As it contains only 4- and 5-mers as potential enhancers, I could only test if sequences in my list include any of these sequences. This also applies to the data set from [111]. Table 6.6 summarizes the overlapping results. Twenty-two percent of predicted elements can be mapped to known binding sites of RNA binding proteins, and 99% can be matched to any of the previously published databases that predict enhancers using other bioinformatics approaches

Table 6.6: Number of overlapped exonic enhancers with previously published data sets. The numbers between brackets are the number of enhancers in each database. The first number in each cell represents an exact match and the second number is an approximate match.

Data set	SpliceAid-F	AEdb	RESCUE-ESE	PESE	Fedrove	Zhang	Badr
	69	64	238	2060	42	42	2001
Brain(205)	5/35	2/23	25/36	2/130	-/11	-/5	202/203
Heart(85)	1/11	1/8	7/12	1/47	-/12	-/1	85/85
Liver(50)	1/8	0/3	5/10	2/28	-/6	-/2	49/50
Muscle(34)	1/9	1/6	3/6	1/22	-/3	-/3	34/34

6.4.4 Tissue-specific enhancer regulatory network

To illustrate the complex relationship among the identified exonic enhancers found in multiple tissues, I constructed a regulatory network (Figure 6.3). It is a bipartite graph with two types of nodes: the circular nodes represent enhancers and the rectangular ones represent tissues. The size of any node type is proportional to its degree (number of incident edges). It is clear that there are enhancers that are tissue-specific and others that may regulate more than one tissue. That is in accordance with the suggestion in [105] that AS decisions are often made by a combinatorial action of general and tissue-specific regulators.

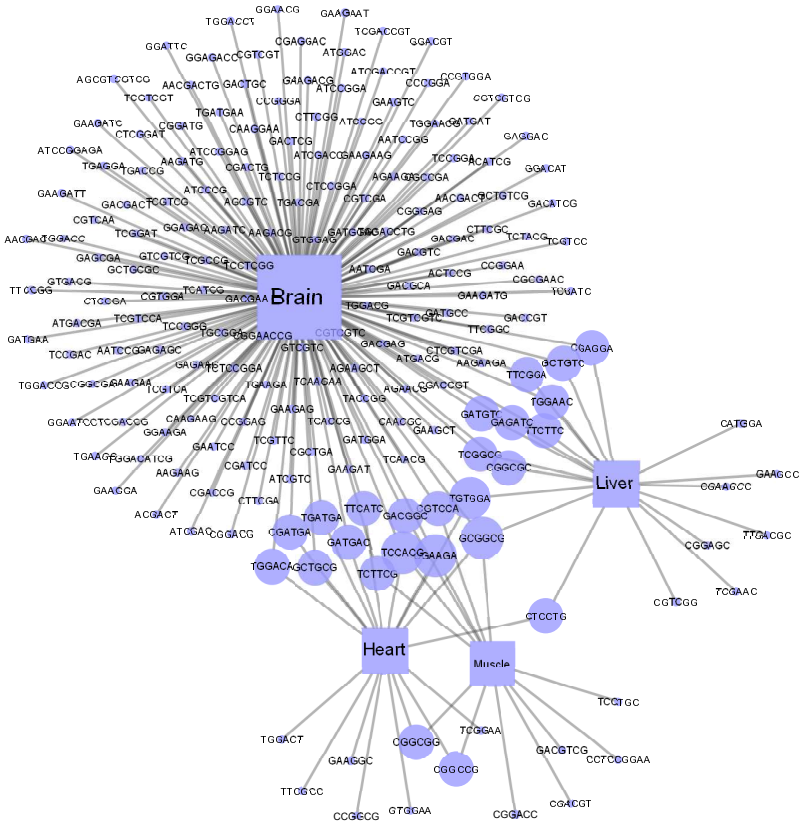


Figure 6.3: Tissue-specific enhancer regulatory network. The circular nodes represent enhancers, and the rectangular ones represent tissues. An edge indicates an enhancer contained in a tissue. The node size indicates the node degree.

Focusing on the enhancers that are involved in multiple tissues, I noticed a hierarchical relationship, where some enhancers regulate two tissues (10% of all enhancers). A smaller number regulate 3 tissues (1%), and only one enhancer element is found in all four tissues. The other 89% of the SREs are tissue-specific where brain-specific enhancers represent 64% of all the identified ESEs. These results are consistent with the conclusions from [60], that

brain tissue exhibits a very large number of tissue-specific SREs and a limited number of general ones. However, in [60], the focus was on intronic SREs.

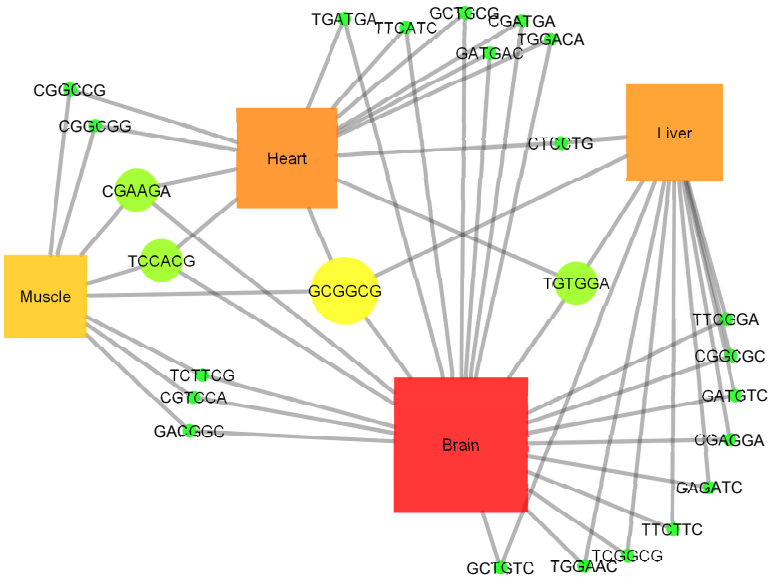


Figure 6.4: Enhancer regulatory network that focuses on enhancers that are involved in multiple tissues. The node size and color is proportional to its degree.

6.4.5 Tissue-specific combinatorial SREs

Identifying individual *cis*-regulatory elements does not suffice to explain tissue-specific or condition-specific AS.

I utilized my graph mining algorithm CoSREM [15] to identify co-occurring enhancers and silencers that may cause exon inclusion in one tissue and its exclusion in another tissue. To do that, for each tissue, I utilized all the identified tissue-specific exons against the other three tissues. Table 6.7 illustrates the number of exons used for each tissue and the number of identified combinatorial enhancers and silencers.

Table 6.7: Number of exons used in CoSREM and the resulted combinatorial SREs.

	No. of exons	Combinatorial SREs
Brain	8858	366
Heart	7818	283
Liver	4564	51
Muscle	2410	45

The actual combinatorial SRE sets are given in Supplementary Tables S8–S11 [12]. I notice that these SRE sets appear in most of the specified exons. For example, in the brain tissue, the 366 combinatorial SRE sets appeared in 8753 out of 8858 exons. The total number of exons that combinatorial SRE sets appeared in is given in Supplementary Tables S8–S11 [7] for all the tissues. Although the number of combinatorial enhancers and silencers in the brain tissue was quite large, the number of unique enhancers and silencers was surprisingly much smaller. There are 30 unique enhancers, of which 26 are previously identified by GenSRE, as individual enhancers and 63 as silencers. Table 6.8 illustrates the number of enhancers and silencers identified as a part of combinatorial SREs for all the tissues.

Table 6.8: Number of enhancers and silencers identified as a part of combinatorial SREs.

	No. of enhancers	No. of silencers
Brain	30	63
Heart	55	55
Liver	15	24
Muscle	26	28

To understand the complex relationship between these enhancers and silencers, I constructed a regulatory network for enhancers and silencers in the brain tissue as illustrated in Figure 6.5. The red nodes represent enhancer elements, and the blue ones represent silencer elements. The node size is proportional to node degree. The network illustrates the many-to-many relationship between the enhancers and silencers. In other words, one enhancer can co-occur with multiple silencers and vice versa.

I wanted to assess the accuracy of the discovered combinatorial SREs. Therefore, I investigated whether the discovered enhancers can be verified from previously published databases. Table 6.9 illustrates the number of enhancers that were verified. I utilized the same data sets from [14, 40, 39, 106, 46, 86, 110]

Table 6.9: The number of overlapped enhancers with previously published data sets. The numbers in parentheses are the numbers of enhancers discovered as a part of a combinatorial set.

	No. of verified enhancers
Brain(30)	26
Heart(14)	9
Liver(5)	4
Muscle(9)	1

I wanted to assess the accuracy of the discovered silencers as well. These silencers are compared with other data sets as illustrated in Table 6.10 such as SpliceAid-F [46], AEdb [86], FAS-ESS [100], PESS [110], and Badr and Heath [14].

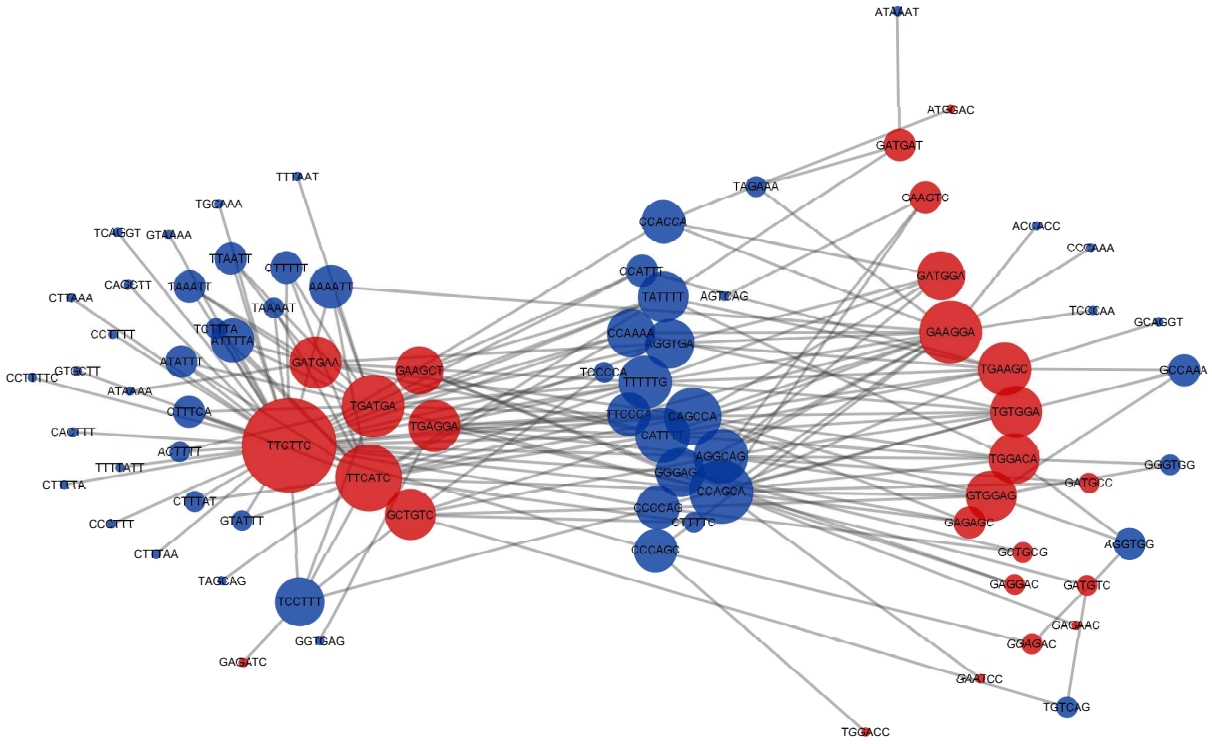


Figure 6.5: Each circle represents the number of brain-specific exons that resulted from brain pairwise comparison with the other tissues (heart, liver, and muscle). The intersection represents brain-specific exons against all other tissues.

Table 6.10: The number of overlapped ESSs with previously published data sets.

Data set	SpliceAid-F	AEdb	FAS	PESS	Badr
	59	24	130	1091	3080
Brain(63)	17	6	3	36	56
Heart(56)	16	5	3	37	50
Liver(24)	9	3	0	15	21
Muscle(28)	11	3	1	16	25

6.5 Discussion

I developed a novel pipeline to identify splicing regulatory elements across tissues. I utilized DEXSeq [5] to identify tissue-specific exons, and then I applied the GenSRE [14] and CoSREM [15] algorithms to identify both individual and combinatorial SREs that are tissue-specific. Overall, I identified 1929 exonic splicing enhancers across all the tissues. That includes either enhancers that appeared in only one tissue or in multiple tissues. I utilized Ontologizer [20] to assess the significance of the predicted tissue-specific enhancers and whether they are involved in tissue-specific biological processes. Therefore, for each tissue, I determined the enriched GO annotations of the genes that contain the identified enhancers. I focused on identifying significant biological processes with adjusted p-value ≤ 0.05. Several brain-related processes were identified, as illustrated in Table 6.11. I also identified some heart-related processes such as "regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion" with p-value 0.00295. Other biological processes were identified as liver and muscle-related such as "digestive tract morphogenesis", and "regulation of muscle system process" with p-values 0.00202, and 0.00253, respectively. Supplementary Tables S12-S15 [11] includes the complete list of the enriched GO annotations for the four tissues of interest.

Table 6.11: Examples of biological processes that are brain and nervous system-related, which resulted of GO enrichment analysis of gene sets that contain putative brain-specific enhancers identified by my approach.

Enhancer element	ID	Annotation	p-value
CGGAAGA	GO:0042428	serotonin metabolic process	0.00027
TCGGAT	GO:0021553	olfactory nerve development	0.00028
AATCGA	GO:0048708	astrocyte differentiation	0.00029
ATGACGA	GO:0060291	long-term synaptic potentiation	0.00031
CGTCGT	GO:0001505	regulation of neurotransmitter levels	0.00036
GGAGAC	GO:0046928	regulation of neurotransmitter secretion	0.00049
GAGAGC	GO:0021983	pituitary gland development	0.00070
CGTCGAC	GO:0090210	blood-brain barrier establishment regulation	0.00216
ATGACG	GO:0007212	dopamine receptor signaling pathway	0.00248
TTCGGAT	GO:0007269	neurotransmitter secretion	0.00384
ACCGGGA	GO:0007269	neurotransmitter secretion	0.00718

Tissue-specific combinatorial SREs were also identified. These are sets of co-occurring enhancer and silencer elements in each tissue. As my focus here is exons that are differentially used in one tissue with respect to other tissues, identified combinatorial SREs may play a role in tissue-specific exon inclusion in one tissue and its exclusion in another tissue. Mayeda et al. [67] showed *in vitro* that having different ratios of SF2/ASF to hnRNP A1 splicing factors promotes exon skipping or inclusion by binding to different enhancers or silencers. In other words, it was shown that the ratio of SF2/ASF to hnRNP Al can affect whether the internal

exon is included or excluded. An excess of SF2/ASF promotes exon inclusion, while hnRNP Al excess promotes exon exclusion. Therefore, this could provide us with an understanding of what might be the possible outcomes of combinatorial splicing regulation. Therefore, I investigated this hypothesis by incorporating splicing factor information. Splicing factor proteins (enhancers or repressors) were identified from SpliceAid-F [46] that bind to SREs in my combinatorial SRE sets. Then, the expression levels for these splicing factors were calculated from the RNA-Seq data as illustrated in Figure 6.6. The actual expression levels in FPKM are in supplementary Table S16 [8]. For each combinatorial SRE set, I identified the splicing factors that binds to its elements, if any. I then calculated the ratio between the expression levels of the splicing factors involved (enhancer factor and inhibitor factor) across the tissues. Table 6.12 provides an example from the combinatorial SREs identified in the brain tissue. I focused on results where the ratio ≥ 1 in one tissue and ≤ 1 in at least another tissue, which suggests that these splicing factors and their binding sites may play a role in regulating exon inclusion or exclusion between the tissues.

One interesting result involved the splicing factors FMRP and hnRnpLL where the FMRP to hnRNPLL expression level ratio was (≈ 1.83) in brain tissue while it was (≈ 0.52) in muscle tissue as illustrated in Table 6.12. More interestingly, multiple co-occurring enhancer and silencer elements were identified that all are potential binding sites to these splicing factors in the brain tissue. For example, one silencer element identified as a potential binding site to hnRNPLL protein is $CAGCCA$. It co-occurs with different enhancer elements ($AAGAGA, AAGGAA, CAAGGA, GAGAGC$, $GTGGAG, TCCTGC, TGTGGA$). All the enhancers elements were identified as binding sites to FMRP splicing factor. Other silencer elements were identified as well as binding sites to hnRNPLL such as $CCACCA, CCAGCA$. All the identified binding sites include CA dinucleotide repeats which is known to be preferentially recognized by HNRNPLL [30]. This suggests the hypothesis that these two splicing factors may have an antagonistic behavior that results in some exons being included in the brain tissue and excluded in the muscle tissue. Although this hypothesis needs further experimental validation, my pipeline can highlight interesting results for more experimental testing.

6.6 Methods

6.6.1 Identifying tissue-specific exons

The main goal is to identify splicing regulatory elements that are tissue-specific. To do that, DEXSeq [5] was utilized to identify exons that are differentially used between two tissues. Differential exon usage analysis aims at identifying the changes in relative usage of exons caused by a certain condition [74].

Let $R_g = \{t_1, t_2, ..., t_n\}$ be the set of all the transcripts from a specific gene g. Let $X_g =$

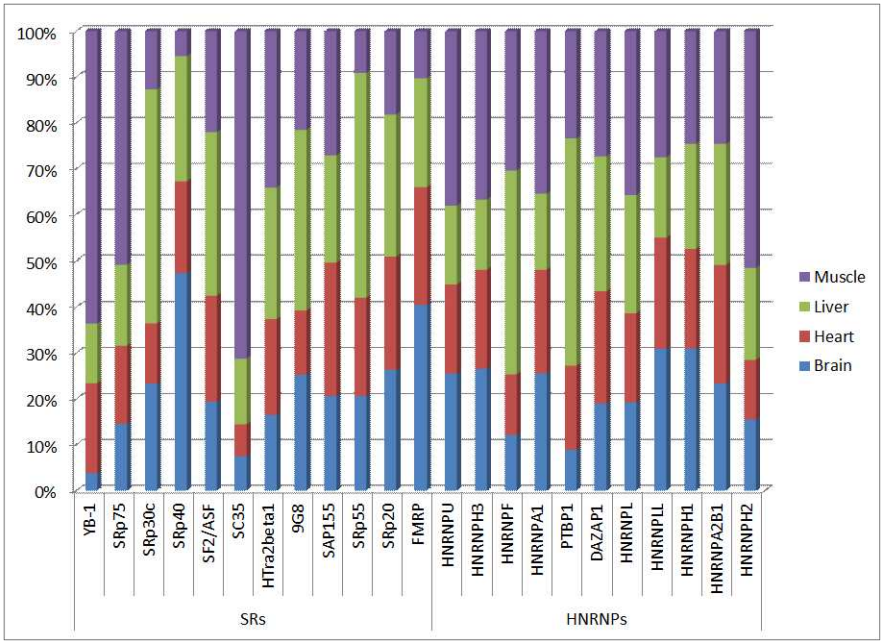


Figure 6.6: Relative expression levels for several splicing factors across the tissues from the RNA-Seq data.

Table 6.12: Identifying splicing factors that binds to combinatorial SREs in the brain tissue. The ratio columns contain the ratio between the expression level of the enhancer factor and the inhibitor factor in the specified tissue. The cells with 'non' indicates that I could not identify an associated splicing factor and hence no ratios are provided.

Combinatorial SRE set	Enhancer factor	Inhibitor factor	Brain ratio	Heart ratio	Liver ratio	Muscle ratio
CAAGGA,CAGCCA	FMRP	hnRNPLL	1.831	1.484	1.899	0.522
TGTGGA,CCAGCA	SRp55	hnRNPLL	1.706	2.279	7.115	0.847
TGTGGA,CAGCCA	SRp55	hnRNPLL	1.706	2.279	7.115	0.847
TGTGGA,TCCTTT	SRp55	DAZAP1	1.162	0.935	1.786	0.356
GAAGGC,AGGCAG	9G8	non	-	-	-	-
AGAAGAT,TTAGAA	9G8	non	-	-	-	-
GAAGGA,GGGAGG	HTra2beta1	hnRNPF	0.113	0.131	0.054	0.0933
TTCTTC,TCCTTT	non	hnRNPA2B1	-	-	-	-
GAGGAT,GGGAGG	SF2/ASF	hnRNPA1	0.0642	0.086	0.183	0.0525
GAGGAT,GGGAGG	SF2/ASF	hnRNPF	0.697	0.761	0.350	0.316

$\{e_1, e_2, ..., e_m\}$ be the set of all exons that constitute the transcripts of gene g. Let $T(e, g)$ be the set of all the transcripts from R_g that contain an exon e. Exon usage $U(e, g)$ of exon e is defined to be

$$U(e, g) = \frac{T(e, g)}{n}.$$

In DEXSeq, generalized linear models are utilized to model read counts, and the χ^2 likelihood ratio test is then used to get an analysis of deviance p-value. The null hypothesis in this test is that none of the conditions influence exon usage. Rejecting the null hypothesis indicates that the count of sequencing reads that map to the exon under the test differs significantly between the different conditions. One of the advantages of this model is accounting for biological variability when the data has replicates for different conditions in contrast to other methods [2].

Before conducting differential exon usage analysis, flattening gene models and counting the reads steps were first performed. Flattening gene models means aligning the sequencing reads to a reference genome and accumulating all the reads for each exon in each tissue from all the transcripts that contain this exon. As some of the transcripts may contain only a part of an exon, the *exon counting bins* term is used to refer to an exon or a part of an exon [5, 74]. It is notable that having exons with differential usage does not mean that their corresponding transcripts are differentially expressed. The output of DEXSeq is a table that contains the differential counting bins, their genes, their read counts in both tissues, and the p-values to determine significance.

For each pair of tissues, DEXSeq was applied. Differentially used exons were identified by choosing the exons with p-value ≤ 0.05. I further filtered these exons by choosing the exons with $\log_2$ fold change ≥ 2 or ≤ -2. To focus on exons that exist in one tissue but not in the other, I only used exons that have reads in one tissue and no reads at all in the other tissue.

The output of this stage is a set of exons in each tissue that are differentially used and present in one tissue but not the other.

These exons were then retrieved from the ENCODE project [56]. The February, 2009, human genome assembly (GRCh37/hg19) was used. The 200 intronic nucleotides upstream and the 200 intronic nucleotides downstream of each exon were also retrieved.

6.6.2 GO enrichment analysis

The command-line version of Ontologizer [20] was utilized, with the goal of determining the enriched GO annotations of the genes that contain the identified enhancers that appears only in one tissue.

For each tissue, the genes, whose exons were identified as tissue specific exons, are utilized as a background data set. For each exonic splicing enhancer in specific tissue, the

exon data set is searched to allocate each splicing enhancer, and the corresponding gene set is identified to form the study set. GO annotation files gene_ontology_edit.obo and gene_association.goa_human were downloaded. GO enrichment analysis is performed using the Topology-Elim algorithm. The Westfall-Young Single Step multiple testing correction procedure is then applied.

I am interested in the biological process annotations. Therefore, the biological process category was chosen with the minimum adjusted p-value, where I consider only terms with $p \leq 0.05$ to be significant.

Chapter 7

Conclusions

Alternative splicing is a vital process for regulating gene expression and promoting proteomic diversity. A key factor in the process is the splicing regulatory elements (SREs). Accurate identification of SREs is crucial in understanding alternative splicing and the role it plays in different biological contexts.

I developed two different methods to tackle some of the limitations that exist in the current approaches for identifying SREs. First, I built a de Bruijn graph based model and developed GenSRE algorithm to identify SREs of different lengths. My model utilizes genomic structure, word count enrichment analysis, and experimental evidence to increase the accuracy of the results. The results overlap with many of the experimental and computational results. This method can open new directions to study SREs and the roles they play in alternative splicing.

Second, I introduced CoSREM, a graph mining algorithm to discover combinatorial SREs. This method is capable of identifying different combinations of splicing enhancer and silencer elements without assuming a predefined size on the identified elements or limiting the algorithm to find only pairs of SREs. This method allows studying and analyzing the complex relationship between different groups of regulators and how splicing decisions are affected by it.

CoSREM, as a graph mining approach, can be applied in different biological contexts. It can be utilized to find other types of motifs such as transcription factors. It is not limited only to find splicing regulatory elements. It can also be extended to other application areas, such as text mining. One interesting application is to find co-occurring sentences in a specific set of texts. These texts can be the abstracts of literature papers in a specific discipline. The way CoSREM works can identify common keywords between different sentences. Moreover, it is able to discover sentences that co-occur in the same set of abstracts. Consequently, it reduces the time taken to search for specific keywords in the literature. The variability inherited in CoSREM design allows discovering even new relationship between key terms

that may have not discovered before.

Lastly, I applied both algorithms on RNA-Seq data for multiple human tissues to identify tissue-specific SREs where a complicated enhancer regulatory network was identified. Combinatorial exonic enhancers and silencers across the tissues were discovered as well, which may be responsible for exon inclusion or exclusion across tissues.

Both algorithms, GenSRE and CoSREM, utilize a de Bruijn graph-based model, where certain vertices are only used (SRE graph) based on their ranks. I utilized the data from Ke et al. [58] to rank the graph vertices. However, it will be interesting to use other measures in ranking, such as utilizing data from CLIP experiments where both the RNA binding protein and the location of its binding site are experimentally identified [92]. This information can greatly help in identifying combinatorial SREs with their associated splicing factors. It can help understand the competitive or cooperative behavior between different types of splicing factors. Moreover, if there is a similar data to the one I utilized from [58] but for intronic SREs, it will allow applying my algorithms on intronic regions as well. Both algorithms can be applied in either exonic or intronic regions or even a combination of both as long as experimental data is provided.

I reported a genome wide analysis on human exons across four tissues (brain, heart, liver, and muscle). The next step is to apply the same pipeline on a larger number of human tissues. This will allow us to understand the complexity in alternative splicing regulation across different tissues and may help us to understand how splicing decisions are taken. Another point is to apply this pipeline on different species such as *Arabidopsis thaliana*. Applying the pipeline on a model plant will help us better understand alternative splicing in plants. All the developed algorithms are general enough to be used with any species as long as suitable data are available. Another potential use is to study alternative splicing and its behavior in diseases, for example, by applying the pipeline on RNA-Seq data for normal and tumor samples. In summary, there are several potential applications where both algorithms GenSRE and CoSREM can be utilized.

Bibliography

[1] N. Abdul-Manan and K. R. Williams. hnRNP A1 binds promiscuously to oligoribonucleotides: Utilization of random and homo-oligonucleotides to discriminate sequence from base-specific binding. *Nucleic Acids Research*, 24(20):4063–70, October 1996.

[2] Gael P Alamancos, Eneritz Agirre, and Eduardo Eyras. Methods to study splicing from high-throughput RNA sequencing data. *Methods in Molecular Biology*, 1126:357–97, January 2014.

[3] Pedro Alexandre, Favoretto Galante, Noboru Jo Sakabe, Natanja Kirschbaum-Slager, and Sandro José D. E. Souza. Detection and evaluation of intron retention events in the human transcriptome. *RNA*, 10(5):757–765, 2004.

[4] Rami Alroobi, Syed Ahmed, and Saeed Salem. Mining maximal cohesive induced subnetworks and patterns by integrating biological networks with gene profile data. *Interdisciplinary Sciences, Computational Life Sciences*, 5(3):211–24, October 2013.

[5] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–17, October 2012.

[6] Faustino Andre and Thomas A. Cooper. Pre-mRNA splicing and human disease. *Genes Development*, 17(4):419–437, 2003.

[7] Eman Badr. Chapter 5, Supplementary Tables S1–S5. `http://bioinformatics.cs.vt.edu/~ebadr/Chapters/ch5/SupplementaryTables.xlsx`, 2015.

[8] Eman Badr. Chapter 6, Supplementary Table S16. `http://bioinformatics.cs.vt.edu/~ebadr/Chapters/ch6/SupplementaryTableS16.xlsx`, 2015.

[9] Eman Badr. Chapter 6, Supplementary Table S7. `http://bioinformatics.cs.vt.edu/~ebadr/Chapters/ch6/SupplementaryTableS7.xlsx`, 2015.

[10] Eman Badr. Chapter 6, Supplementary Tables S1–S6. `http://bioinformatics.cs.vt.edu/~ebadr/Chapters/ch6/SupplementaryTablesS1-S6.xlsx`, 2015.

[11] Eman Badr. Chapter 6, Supplementary Tables S12–S15. `http://bioinformatics.cs.vt.edu/~ebadr/Chapters/ch6/SupplementaryTablesS12-S15.xlsx`, 2015.

[12] Eman Badr. Chapter 6, Supplementary Tables S8–S11. `http://bioinformatics.cs.vt.edu/~ebadr/Chapters/ch6/SupplementaryTablesS8-S11.xlsx`, 2015.

[13] Eman Badr, Mahmoud M. ElHefnawi, and Lenwood S. Heath. Computational identification of tissue-specific splicing regulatory elements in human genes from RNA-Seq data. 2015(submitted).

[14] Eman Badr and Lenwood S. Heath. Identifying splicing regulatory elements with de bruijn graphs. *Journal of Computational Biology*, 21(12):880–97, December 2014.

[15] Eman Badr and Lenwood S. Heath. CoSREM: A graph mining algorithm for the discovery of combinatorial splicing regulatory elements. 2015(submitted).

[16] Yoseph Barash, Benjamin J. Blencowe, and Brendan J. Frey. Model-based detection of alternative splicing signals. *Bioinformatics*, 26(12):i325–i333, June 2010.

[17] Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[18] Olivier Baris, Cécile Delettre, Patrizia Amati-Bonneau, Marie-Odile Surget, Jean-François Charlin, Antoine Catier, Laurence Derieux, Jean-Laurent Guyomard, Hélène Dollfus, Philippe Jonveaux, Carmen Ayuso, Irene Maumenee, Birgit Lorenz, Shehla Mohammed, Yves Tourmen, Dominique Bonneau, Yves Malthièry, Christian Hamel, and Pascal Reynier. Fourteen novel OPA1 mutations in autosomal dominant optic atrophy including two de novo mutations in sporadic optic atrophy. *Human Mutation*, 21(6):656–656, June 2003.

[19] Vladimir Batagelj and M. Zaversnik. An O(m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*, pages 1–10, 2003.

[20] Sebastian Bauer, Steffen Grossmann, Martin Vingron, and Peter N. Robinson. Ontologizer 2.0– A multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, July 2008.

[21] Priya Bellare, Eliza C. Small, Xinhua Huang, James A. Wohlschlegel, Jonathan P. Staley, and Erik J. Sontheimer. A role for ubiquitin in the spliceosome assembly pathway. *Nature Structural & Molecular Biology*, 15(5):444–451, May 2008.

[22] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300, 1995.

[23] M. Brudno, M. S. Gelfand, S. Spengler, M. Zorn, I. Dubchak, and J. G. Conboy. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Research*, 29(11):2338–2348, June 2001.

[24] Patricia Buendia, John Tyree, Robert Loredo, and Shu-Ning Hsu. Identification of conserved splicing motifs in mutually exclusive exons of 15 insect species. *BMC Genomics*, 13(Suppl 2):S1, 2012.

[25] Emanuele Buratti, Marco Baralle, and Francisco E. Baralle. From single splicing events to thousands: The ambiguous step forward in splicing research. *Briefings in Functional Genomics*, 12(1):3–12, January 2013.

[26] B. G. Burnett, E. Munoz, A. Tandon, D. Y. Kwon, C. J. Sumner, and K. H. Fischbeck. Regulation of SMN protein stability. *Molecular and Cellular Biology*, 29(5):1107–1115, December 2008.

[27] L. Cartegni. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, 31(13):3568–3571, July 2003.

[28] John C. Castle, Chaolin Zhang, Jyoti K. Shah, Amit V. Kulkarni, Auinash Kalsotra, Thomas A. Cooper, and Jason M. Johnson. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics*, 40(12):1416–25, December 2008.

[29] Yvon Cavaloc, C. F. Bourgeois, Liliane Kister, and James Stévenin. The splicing factors 9G8 and SRp20 transactivate. *RNA*, pages 468–483, 1999.

[30] Xing Chang, Bin Li, and Anjana Rao. Rna-binding protein hnrnpll regulates mrna splicing and stability during b-cell to plasma-cell differentiation. *Proceedings of the National Academy of Sciences*, 112(15):E1888–E1897, 2015.

[31] Lawrence A. Chasin. Searching for splicing motifs. *Advances in Experimental Medicine and Biology*, 623(Cv):85–106, January 2007.

[32] L. R. Coulter, M. A. Landree, and T. A. Cooper. Identification of a new class of exonic splicing enhancers by *in vivo* selection. *Molecular and Cellular Biology*, 17(4), 1997.

[33] Debopriya Das, Tyson A. Clark, Anthony Schweitzer, Miki Yamamoto, Henry Marr, Josh Arribere, Simon Minovitsky, Alexander Poliakov, Inna Dubchak, John E. Blume, and John G. Conboy. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Research*, 35(14):4845–57, January 2007.

[34] Marko Djordjevic. SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering*, 24(2):179–189, June 2007.

[35] Julian Downward. Targeting RAS signalling pathways in cancer therapy. *Nature Reviews. Cancer*, 3(1):11–22, January 2003.

[36] Zhiguo E, Lei Wang, and Jianhua Zhou. Splicing and alternative splicing in rice and humans. *BMB Reports*, 46(9):439–447, September 2013.

[37] Johannes Eichner, Georg Zeller, Sascha Laubinger, and Gunnar Rätsch. Support vector machines-based identification of alternative splicing in *Arabidopsis thaliana* from whole-genome tiling arrays. *BMC Bionformatics*, 12(1):55–55, January 2011.

[38] Elizabeth C. Engle. Human genetic disorders of axon guidance. *Cold Spring Harbor Perspectives in Biology*, 2(3):a001784, March 2010.

[39] William G. Fairbrother, Ru Fang Yeh, Phillip A. Sharp, and Christopher B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, August 2002.

[40] A. Fedorov, S. Saxonov, L. Fedorova, and I. Daizadeh. Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Research*, 29(7):1464–1469, April 2001.

[41] Elisa Napolitano Ferreira, Pedro A. F. Galante, Dirce Maria Carraro, and Sandro José de Souza. Alternative splicing: A bioinformatics perspective. *Molecular BioSystems*, 3(7):473–477, July 2007.

[42] Paul Flicek, Ikhlak Ahmed, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Carlos García-Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Monika Komorowska, Eugene Kulesha, Ian Longden, Thomas Maurel, William M. McLaren, Matthieu Muffato, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet Singh Riat, Graham R. S. Ritchie, Magali Ruffier, Michael Schuster, Daniel Sheppard, Daniel Sobral, Kieron Taylor, Anja Thormann, Stephen Trevanion, Simon White, Steven P. Wilder, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Jennifer Harrow, Javier Herrero, Tim J. P. Hubbard, Nathan Johnson, Rhoda Kinsella, Anne Parker, Giulietta Spudich, Andy Yates, Amonida Zadissa, and Stephen M. J. Searle. Ensembl 2013. *Nucleic Acids Research*, 41(Database issue):D48–55, January 2013.

[43] Brad A. Friedman, Michael B. Stadler, Noam Shomron, Ye Ding, and Christopher B. Burge. *Ab initio* identification of functionally interacting pairs of cis-regulatory elements. *Genome Research*, 18(10):1643–51, October 2008.

[44] Kaiping Gao, Akio Masuda, Tohru Matsuura, and Kinji Ohno. Human branch point consensus sequence is yUnAy. *Nucleic Acids Research*, 36(7):2257–67, April 2008.

[45] Mariano A. Garcia-Blanco, Andrew P. Baraniak, and Erika L. Lasda. Alternative splicing in disease and therapy. *Nature Biotechnology*, 22(5):535–546, May 2004.

[46] Matteo Giulietti, Francesco Piva, Mattia D'Antonio, Paolo D'Onorio De Meo, Daniele Paoletti, Tiziana Castrignanò, Anna Maria D'Erchia, Ernesto Picardi, Federico Zambelli, Giovanni Principato, Giulio Pavesi, and Graziano Pesole. SpliceAid-F: A

database of human splicing factors and their RNA-binding sites. *Nucleic Acids Research*, 41(Database issue):D125–D131, January 2013.

[47] DS Goodsell. The molecular perspective: the ras oncogene. *The oncologist*, pages 263–264, 1999.

[48] Amir Goren, Eddo Kim, Maayan Amit, Keren Vaknin, Nir Kfir, Oren Ram, and Gil Ast. Overlapping splicing regulatory motifs–combinatorial effects on splicing. *Nucleic Acids Research*, 38(10):3318–27, June 2010.

[49] Amir Goren, Oren Ram, Maayan Amit, Hadas Keren, Galit Lev-Maor, Ida Vig, Tal Pupko, and Gil Ast. Comparative analysis identifies exonic splicing regulatory sequences–The complex definition of enhancers and silencers. *Molecular Cell*, 22(6):769–781, June 2006.

[50] Sonia Guil and Javier F. Cáceres. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nature Structural & Molecular Biology*, 14(7):591–596, July 2007.

[51] Dalil Hamroun and Marine Lalande. Human Splicing Finder : an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, 37(9):1–14, 2009.

[52] Akemi Hanamura, Javier F. Cáceres, Akila Mayeda, B Robert Franza Jr, and Adrian R Krainer. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA*, 4:430–444, 1998.

[53] Chunjiang He, Fang Zhou, Zhixiang Zuo, Hanhua Cheng, and Rongjia Zhou. A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. *PLOS ONE*, 4(3), 2009.

[54] Michael Hiller, Zhaiyi Zhang, Rolf Backofen, and Stefan Stamm. Pre-mRNA secondary structures influence exon recognition. *PLoS Genetics*, 3(11):e204, November 2007.

[55] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Publishing Company, Massachusetts, 1979.

[56] Donna Karolchik, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The UCSC table browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue):D493–D496, January 2004.

[57] Shengdong Ke and Lawrence A. Chasin. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biology*, 11(8):R84, January 2010.

[58] Shengdong Ke, Shulian Shang, Sergey M. Kalachikov, Irina Morozova, Lin Yu, James J. Russo, Jingyue Ju, and Lawrence A. Chasin. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Research*, 21(8):1360–1374, August 2011.

[59] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: Diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, May 2010.

[60] Jihye Kim, Sihui Zhao, Brian E. Howard, and Steffen Heber. Mining of cis -regulatory motifs associated with tissue-specific alternative splicing. *Springer*, 5542:260–271, 2009.

[61] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, August 2010.

[62] José Eduardo Kroll, Pedro A. F. Galante, Daniel T. Ohara, Fábio C. P. Navarro, Lucila Ohno-Machado, and Sandro J. de Souza. SPLOOCE: A new portal for the analysis of human splicing variants. *RNA Biology*, 9(11):1339–43, November 2012.

[63] H. X. Liu, S. L. Chew, and Luca Cartegni. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Molecular and Cellular Biology*, 20(3), 2000.

[64] H. X. Liu, M. Zhang, and A. R. Krainer. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes & Development*, 12(13):1998–2012, July 1998.

[65] Yi Lv, Zhixiang Zuo, and Xiao Xu. Global detection and identification of developmental stage specific transcripts in mouse brain using subtractive cross-screening algorithm. *Genomics*, 102(4):229–236, May 2013.

[66] Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, May 2005.

[67] Akila Mayeda, David M. Helfman, and Adrian R Krainer. Modulation of Exon Skipping and Inclusion by Heterogeneous Nuclear Ribonucleoprotein Al and Pre-mRNA Splicing Factor SF2 / ASF. *Molecular and Cellular Biology*, 13(5), 1993.

[68] Akila Mayeda and A. R. Krainer. Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell*, 68:365–375, 1992.

[69] Stefania Millevoi, Sandra Bernat, Dominique Telly, Françoise Fouque, Laurence Gladieff, Gilles Favre, Stéphan Vagner, and Christine Toulas. The c.5242C>A BRCA1 missense variant induces exon skipping by increasing splicing repressors binding. *Breast Cancer Research and Treatment*, 120(2):391–399, April 2010.

[70] Alicia A. Nugent, Adrianne L. Kolpak, and Elizabeth C. Engle. Human disorders of axon guidance. *Current Opinion in Neurobiology*, 22(5):837–843, October 2012.

[71] Sriram Pemmaraju and Steve Skiena. *Computational Discrete Mathmatics Combinatorics and Graph Theory with Mathematica*. University of Cambridge, 2003.

[72] Mihaela Pertea, Stephen M. Mount, and Steven L. Salzberg. A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. *BMC Bioinformatics*, 8:159–159, January 2007.

[73] Rodrigo F. Ramalho, Sahar Gelfman, Jorge E. De Souza, Gil Ast, Sandro J. De Souza, and Diogo Meyer. Testing for natural selection in human exonic splicing regulators associated with evolutionary rate shifts. *Journal of Molecular Evolution*, 76(4):228–239, April 2013.

[74] Alejandro Reyes, Simon Anders, and Wolfgang Huber. Inferring differential exon usage in RNA-Seq data with the DEXSeq package. *R package version 1.10.8*, 2014.

[75] Sjoerd Rodenhuis and R. J. Slebos. Clinical significance of ras oncogene activation in human lung cancer. *Cancer Research*, 9, 1992.

[76] Arnold L. Rosenberg and Lenwood S. Heath. *Graph Separators, With Applications.* Kluwer Academic/Plenum Publishers, 2000.

[77] Scott William Roy and Manuel Irimia. Splicing in the eukaryotic ancestor: Form, function and dysfunction. *Trends in Ecology & Evolution*, 24(8):447–55, August 2009.

[78] Michael C. Ryan, James Cleland, Ryangguk Kim, Wing Chung Wong, and N. Weinstein. SpliceSeq : a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, 28(18):2385–2387, 2012.

[79] Noboru Jo Sakabe and Sandro José De Souza. Sequence features responsible for intron retention in human. *BMC Genomics*, 8:59–59, January 2007.

[80] Jeremy R. Sanford, Pedro Coutinho, Jamie A. Hackett, Xin Wang, William Ranahan, and Javier F. Caceres. Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. *PLOS ONE*, 3(10):e3369, January 2008.

[81] T. D. Schaal and T. Maniatis. Selection and characterization of pre-mRNA splicing enhancers: Identification of novel SR protein-specific enhancer sequences. *Molecular and Cellular Biology*, 19(3):1705–1719, 1999.

[82] M. Selvakumar and D. M. Helfman. Exonic splicing enhancers contribute to the use of both 3' and 5' splice site usage of rat beta-tropomyosin pre-mRNA. *RNA*, 5(3):378–394, March 1999.

[83] Phillip A. Sharp and Christopher B. Burge. Classification of Introns : U2-Type or U12-Type. *Cell*, 91(Table 2):875–879, 1997.

[84] Manuela Sironi, Giorgia Menozzi, Laura Riva, Rachele Cagliani, Giacomo P. Comi, Nereo Bresolin, Roberto Giorda, and Uberto Pozzoli. Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Research*, 32(5):1783–91, January 2004.

[85] C. W. Smith and J. Valcárcel. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends in Biochemical Sciences*, 25(8):381–8, August 2000.

[86] Stefan Stamm, Jean-Jack Riethoven, Vincent Le Texier, Chellappa Gopalakrishnan, Vasudev Kumanduri, Yesheng Tang, Nuno L. Barbosa-Morais, and Thangavel Alphonse Thanaraj. ASD: A bioinformatics resource on alternative splicing. *Nucleic Acids Research*, 34(Database issue):D46–D55, January 2006.

[87] Mikita Suyama, Eoghan D. Harrington, Svetlana Vinokourova, Magnus von Knebel Doeberitz, Osamu Ohara, and Peer Bork. A network of conserved co-occurring motifs for the regulation of alternative splicing. *Nucleic Acids Research*, 38(22):7916–26, December 2010.

[88] Micha Wojciech Szcześniak, Micha Kabza, Rafa Pokrzywa, Adam Gudyś, and Izabela Makaowska. ERISdb: A database of plant splice sites and splicing signals. *Plant & Cell Physiology*, 54(2):e10, February 2013.

[89] R. Tacke and J. L. Manley. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *The EMBO Journal*, 14(14):3540–3551, July 1995.

[90] H. Tian and R. Kole. Selection of novel exon recognition elements from a pool of random sequences. *Molecular and Cellular Biology*, 15(11), 1995.

[91] James R. Tollervey, Tomaž Curk, Boris Rogelj, Michael Briese, Matteo Cereda, Melis Kayikci, Julian König, Tibor Hortobágyi, Agnes L. Nishimura, Vera Zupunski, Rickie Patani, Siddharthan Chandran, Gregor Rot, Blaž Zupan, Christopher E. Shaw, and Jernej Ule. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience*, 14(4):452–458, April 2011.

[92] Jernej Ule, Kirk B. Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B. Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215, November 2003.

[93] Jing Wan. Antisense-Mediated Exon Skipping to Shift Alternative Splicing to Treat Cancer. *Humana Press*, 867:201–208, 2012.

[94] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, November 2008.

[95] Xin Wang, Kejun Wang, Milan Radovich, Yue Wang, Guohua Wang, Weixing Feng, Jeremy R. Sanford, and Yunlong Liu. Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. *BMC Genomics*, 10 Suppl 1:S4, January 2009.

[96] Xin Wang, Kejun Wang, Guohua Wang, Jeremy R. Sanford, and Yunlong Liu. Model-based prediction of cis-acting RNA elements regulating tissue-specifc alternative splicing. In *8th IEEE International Conference on BioInformatics and BioEngineering*, 2008.

[97] Yang Wang, Meng Ma, Xinshu Xiao, and Zefeng Wang. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nature Structural & Molecular Biology*, 19(10):1044–52, October 2012.

[98] Yang Wang and Zefeng Wang. Systematical identification of splicing regulatory cis-elements and cognate trans-factors. *Methods*, August 2013.

[99] Zefeng Wang and Christopher B. Burge. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, 2008.

[100] Zefeng Wang, Michael E. Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–845, December 2004.

[101] Laura E. Warner and Jeffrey S. Chamberlain. Minigenes. *Encyclopedia of Life Sciences*, pages 1–6, 2006.

[102] Neil A. Weiss. *Introductory Statistics*. Pearson Education Inc., 7th edition, 2005.

[103] Ji Wen, Zhibin Chen, and Xiaodong Cai. A biophysical model for identifying splicing regulatory elements and their interactions. *PLOS ONE*, 8(1):e54885, January 2013.

[104] Ji Wen, Akira Chiba, and Xiaodong Cai. Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. *Nucleic Acids Research*, 38(22):7895–7907, December 2010.

[105] Xinshu Xiao and Jae-Hyung Lee. Systems analysis of alternative splicing and its regulation. *Wiley interdisciplinary reviews. Systems biology and medicine*, 2(5):550–65, 2010.

[106] Chaolin Zhang, Kuang-Yung Lee, Maurice S Swanson, and Robert B Darnell. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Research*, 41(14):6793–807, August 2013.

[107] Jing Zhang, C. C. Jay Kuo, and Liang Chen. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics*, 12(1):90–90, January 2011.

[108] Jing Zhang, C. C. Jay Kuo, and Liang Chen. VERSE: A varying effect regression for splicing elements discovery. *Journal of Computational Biology*, 19(6):855–865, June 2012.

[109] Xiang H. Zhang, Thaned Kangsamaksin, S. P. Mann, Joydeep K. Banerjee, Lawrence A. Chasin, and Mann S. P. Chao. Exon inclusion is dependent on predictable exonic splicing enhancers. *Molecular and Cellular Biology*, 25(16):7323–7332, 2005.

[110] Xiang H. F. Zhang and Lawrence A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Development*, 18(11):1241–1250, June 2004.

[111] Xiang H. F. Zhang, Katherine A. Heller, Ilana Hefter, Christina S. Leslie, and Lawrence A. Chasin. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Research*, 13(12):2637–2650, December 2003.

[112] Xiang H. F. Zhang, Christina S. Leslie, and Lawrence A. Chasin. Computational searches for splicing signals. *Methods*, 37(4):292–305, December 2005.

[113] Jun Zhu, Akila Mayeda, and A. R. Krainer. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Molecular Cell*, 8:1351–1361, 2001.