UNIVERSITY OF CALIFORNIA

Los Angeles

Unveiling Hidden Interactions:
Local Inference of Splicing Regulatory Networks in RNA Binding Protein Knockdown studies by the Coordinated Multi-Neighborhood Learning Algorithm

A thesis submitted in partial satisfaction
of the requirements for the degree
Masters of Science In Statistics

by

Raymond Alexander Benitez

2024

ABSTRACT OF THE THESIS

Unveiling Hidden Interactions:
Local Inference of Splicing Regulatory Networks in RNA Binding Protein Knockdown
studies by the Coordinated Multi-Neighborhood Learning Algorithm

by

Raymond Alexander Benitez
Masters of Science in Statistics
University of California, Los Angeles, 2024
Professor Qing Zhou, Chair

RNA-binding proteins (RBPs) are integral to RNA metabolism and their dysregulation is linked to cancer and neurodegenerative diseases. Understanding RBP-RNA interactions is therefore crucial. CML, a local constraint-based structure learning algorithm, utilizes conditional independence tests and deterministic rules to infer a graph from observed data. Traditional structure learning algorithms face challenges in high-dimensional settings, common in genomics, due to the rapid expansion of the search space as the number of variables increases. CML mitigates this by coordinating learning across multiple neighborhoods, reducing computational costs, and focusing on local graph structures around target variables. In this work, we implement the CML algorithm on an augmented dataset derived from RNA-seq and rMATS data obtained from RBP knockdown experiments in HepG2 and K562 cell lines. We investigate causal interactions between transcripts and genes within five selected RBP knockdown experiments for each alternative splicing (AS) event in each cell line. This resulted in 50 datasets that combined differential AS patterns and gene expression changes. Our findings revealed numerous causal relationships between transcripts and genes in the context of RBP knockdown experiments, highlighting the efficacy of CML in uncovering intricate molecular interactions in high-dimensional genomics data while dramatically improving computation time.

The thesis of Raymond Alexander Benitez is approved.

Xinshu Xiao

Jingyi Li

Qing Zhou, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Directed Acyclic Graphs

A graph $G = (V, E)$ is formed by a set of nodes, or vertices, $V = V_1, \cdots, V_p$ and a set of edges $E = \{(i, j) \in V \times V\}$. Vertices represent a set of random variables $X = (X_1, X_2, \cdots X_p)$. For nodes $i, j \in V$, an undirected edge is an edge such that $\{(i, j), (j, i) \in E\}$, denoted as $i - j$. A directed edge is an edge where $\{(i, j) \in E\}$ and $\{(j, i) \notin E\}$, denoted as $i \to j$, where i is a parent node of j, and j is a child node of i. For any node i, its parent and child set is denoted as $pa_G(i), ch_G(i)$ respectively. A path in G is a sequence of nodes $< V_1, V_2, \cdots, V_j >$, for $1 \leq n \leq j - 1$, V_i precedes V_{i+1}. If every edge on the path is oriented as $V_i \to V_{i+1}$, then this is a directed path from V_1 to V_j. A node V_k is said to d-separate V_i from V_j if and only if V_k blocks every path from node V_i to V_j. V_i is an ancestor of V_j if there is a directed path from V_1 to V_j. If $i \to j$ and j is an ancestor of i, then there is a directed cycle in G. A Directed Acyclic Graph (DAG) is a graph that consists of only directed edges and contains no directed cycles. A DAG G is often used to model the causal relationships among the random variables in X by the following structural equation model.

$$X_i = f_i(X_{pa_G(i)}, \epsilon_i), \quad i = 1, \cdots, p.$$

Any random variable X_i is determined by a function of its parent set $X_{pa_G(i)}$ and an independent error term ϵ_i. The joint probability distribution $P(X)$ is Markov with respect to graph G and has the following factorization.

$$P(X_1, X_2, \cdots, X_p) = \prod_{i \in V} P(X_i | X_{\mathsf{pa}_G(i)}).$$

The probability distribution P is faithful to a DAG G if there is a one-to-one correspondence between the conditional independence relations in P and the d-separations in G. The adjacency set of node i refers to the set of nodes that are directly connected to i by an edge, denoted as $adj_G(i)$. A v-structure in a DAG is a triplet of nodes (i, j, k) where E contains directed edges $i \to k$ and $j \to k$ and i, j are not adjacent. The set of spouses for node i, denoted as $sp_G(i)$ is the set of non-adjacent nodes that share a child node with i. The Markov blanket of node i, denoted as $mb_G(i)$, is the minimum set of nodes such that X_i is conditionally independent of all other nodes in the graph. If a probability distribution is faithful to its DAG, then the Markov blanket of a node is the union of its parents, children, and spouses.

1.2 RNA-Binding Proteins

RNA-binding proteins (RBPs) are a diverse class of proteins involved in the process of regulating gene expression. These proteins play crucial roles in virtually every step of RNA metabolism including transcription, splicing, polyadenylation, RNA localization, translation, and degradation. RBPs can be involved in one or multiple of these processes. They can have binding specificity to one or more multiple categories of RNA, such as messenger RNA (mRNA), transfer RNA (tRNA), and many more. They are critical for maintaining homeostasis for gene expression and thus normal human physiology. Disruption or dysregulation with RBPs is known to be associated with various neurodegenerative diseases, and cancer (Lukong et al. 2008). Given the vital role RBPs play in RNA metabolism, it is important to elucidate the mechanisms underlying RBP-RNA interactions. By unraveling the interactions within the RBP-RNA landscape, researchers can gain valuable insights into disease mechanisms and identify potential therapeutic targets for intervention.

1.3 Motivation

The study (Mukherjee et al. 2019) suggests that there are at least 1542 RBP-encoding genes in the human genome. Given the number of genes dedicated to producing RBPs, and the multiple modalities RBPs are involved with in the maturation of RNA, it necessitates uncovering the exact mechanism by which RBPs interact with RNA. In their seminal study, Van Nostrand et al. 2020 made significant efforts to elucidate the functions of 356 RBPs in K562 and HepG2 cells. They employed a range of assays to provide different perspectives and lenses to view the functions and behaviors of these RBPs. These include eCLIP, RNA Bind-N-seq, Immunofluorescence, Knockdown RNA-seq, and RBP ChIP-seq. Secondary analysis included the implementation of DEseq, rMATS, MISO, and CUFFDIFF. The results of these experiments provided valuable insights into the diversity and complexity of RBP-RNA interactions, revealing binding patterns across the transcriptome.

Schadt 2009 Highlights the importance of properly constructing a network so that causal relationships for biological phenomena can be unearthed. Networks are graphical models with nodes and edges that are used to visualize causal relationships between variables within the network. The creation of networks in which the relationship between genes and other biological events can be understood is a key goal for life sciences and biological research. Correlation methods that only examine gene expression data is insufficient by themselves to achieve this aim. Constructions of more complex and informative gene networks involve integrating multiple data types. We will construct a transcript-gene network for selected RBP knockdowns. We will use the normalized RBP knockdown RNA-seq and rMATS data generated from Van Nostrand et al. 2020 to infer causal relationships between differentially expressed transcripts and genes.

One method of identifying the interactions within this transcript-gene network is to use directed acyclic graphs (DAG) to represent causal relationships between genes and transcripts. Structure learning algorithms aim to learn a DAG that best represents causal relationships inferred from the given data. Global structure learning algorithms estimate a DAG over the entire variable space. However as the number of variables in a network increases, we quickly observe a rapid deterioration in performance from these structure learning algorithms (Gu and Zhou 2020). When working with data in the field of genomics, this is frequently an issue since datasets tend to have many features and fewer observations. In addition, only a select few variables and their causal effects tend to be of interest to researchers. The Coordinated Multi-Learning Neighborhood Learning (CML) algorithm (Smith and Zhou 2024) is a structure learning algorithm designed to address both of these concerns. The CML algorithm learns local DAG structures around predetermined target variables of interest. This is achieved by implementing Markov Blanket estimation methods around the target variables, and then implementing a three-stage constraint-based algorithm to prune and orient edges. CML is a local structure learning algorithm and greatly improves runtime compared to global structure learning algorithms in high dimensional settings. Thus, making it a well-suited structure learning algorithm to implement in the field of genomics, and in particular on the constructed transcript-gene network.

1.3.1 Objectives

In this thesis, we aim to address the aforementioned research gap in applying local DAG learning to research problems in genomics. We will apply the CML algorithm to the data from the RBP Map experiments to infer causal relationships between transcripts and genes. Specifically, our objectives are to: construct a suitable network from the various data modalities from the RBP Map experiments to infer causal relationships and apply the CML algorithm to assess its utility in identifying meaningful causal interactions within the RBP-RNA network. By achieving these objectives, we aim to deepen our understanding of how certain RBPs can affect the expression levels of genes and transcripts.

Chapter 2

Methods

2.1 CML Algorithm

2.1.1 Background

In contrast to other global structure algorithms that assume causal sufficiency, that is all common causes of variables have been observed, CML does not make this assumption and treats other variables in the network as latent. This feature necessitates the use of a different class of graphs to accommodate latent variables while still being able to infer causal relationships. Ancestral graphs will be used since they are well-suited for representing causal information from observed data in the presence of latent variables (Richardson and Spirtes 2002). Ancestral graphs belong to the class of mixed graphs and only directed $(\rightarrow)$ and bi-directed edges $(\leftrightarrow)$ are considered. In this algorithm, selection bias is not considered and thus there will be no undirected edges in an ancestral graph.

2.1.2 Definitions

Two nodes V_i and V_j are siblings if $V_i \leftrightarrow V_j$. A bi-directed edge between two nodes implies that they share a common latent cause. Similar to how a directed cycle is defined, an almost directed cycle exists if $i \leftrightarrow j$ and $j \in an_G(i)$. A mixed graph is ancestral if there is no directed or almost directed cycle.

A node i is called a collider on a path p if any two non-adjacent nodes on path p have edges directed into node i. In an ancestral graph, a path p between V_i and V_j is m-connecting relative to a set S with $X, Y, \notin S$ if every non-collider on p is not a member of S and every collider on p is an ancestor of some node in S. If there are no m-connecting paths from V_i to V_j given S, then they are m-separated by S. The m-separations in an ancestral graph imply conditional independence among the observed variables by the global Markov property (Richardson and Spirtes 2002).

An inducing path l is a path in which every node except for the endpoints is a collider on the path and every collider is an ancestor of some endpoint on path l. An ancestral graph is maximal if there is no inducing path between any two non-adjacent nodes. The graph is maximal (MAG) in the sense that no additional edges can be added without changing the conditional independence relations. A MAG represents conditional independence relations among the observed nodes with latent variables. Multiple MAGs may also encode the same conditional independencies, and thus will belong to the same Markov equivalence class. We denote the use of a partial ancestral graph (PAG) to represent the set of all MAGs that belong to the same equivalence class. A PAG introduces a distinctive edge mark (o) in addition to the conventional tail and arrowhead marks typically observed in graphical models. Circle marks denote ambiguity in the edge direction and are variant within the equivalence class. Conversely, each non-circle mark remains invariant across the equivalence class of a Maximal Ancestral Graph (MAG).

The algorithm defines the set of first-order neighbors of a node i to be its Markov blanket, denoted N_i^1. We call $NB_i := N_i^1 \cup i$ the neighborhood of node i. The set of second-order neighbors of node i is defined to be the union of the Markov blankets for each node in the first-order neighborhood, excluding nodes in NB_i, denoted as $N_i^2 = \cup_{j \in N_i^1} N_j^1 \backslash NB_i^1$. For a set of nodes T, the union of their neighborhoods is denoted as $NB_T = \cup_{t \in T} NB_t$

2.1.3 Algorithm Details

It is assumed that there is sufficient background knowledge to identify a set of target nodes T. It is also assumed that estimated first- and second-order neighborhoods,N_t^1 and N_t^2, are provided for each target node $t \in T$. This estimation is performed using existing Markov blanket learning algorithms. Algorithm 1 (Smith and Zhou 2024) shows the steps of the CML algorithm.

Algorithm 1 Coordinated Multi-Neighborhood Learning Algorithm

```
 1: $E \leftarrow$ edge set of complete, undirected graph on $V \leftarrow NB_T$ .
 2: for $(i, j) \in E$ do
 3:         Search for separating set $S_{ij} \subset V \setminus \{i, j\}$ such that $X_i \perp\!\!\!\perp X_j \mid S_{ij}$.
 4:         if $S_{ij}$ is found then update $E \leftarrow E \setminus \{(i, j)\}$.
 5: end for
 6: $E_t \leftarrow \{(i, j) \in E : i, j \in NB_t\}$ for all $t \in T$.
 7: for $t \in T$ do
 8:     for $(i, j) \in E_t$ do
 9:             Search for $S_{ij} \subset N_i^1 \cup N_j^1$ such that $X_i \perp\!\!\!\perp X_j \mid S_{ij}$.
10:             if $S_{ij}$ is found then update $E \leftarrow E \setminus \{(i, j)\}$.
11:     end for
12: end for
13: Apply $R_0$ of the FCI algorithm to identify v-structures based on $E$ and $S_{ij}$.
14: Apply FCI rules $R_1$ to $R_4$ and $R_8$ to $R_{10}$ until none of them apply.
15: Modify edge marks within each single neighborhood with rule $R_N$.
```

The algorithm first begins with learning the skeleton of a graph, which is the undirected graph corresponding to ignoring edge orientations in the true underlying graph. Beginning with the complete graph over NB_T, the skeleton recovery corresponds to lines 1-12 and edges are deleted from the complete graph based on conditional independence tests. The skeleton recovery is broken into two phases to facilitate the retention of between neighborhood edges for future edge orientations. Lines 2-5 correspond to the first phase and is equivalent to the FCI algorithm (Spirtes 2001) with $V = NB_T$ being the observed nodes. As a result, only subsets of NB_T can be candidate separation sets, and between neighborhood edges will be preserved. The second phase of the skeleton recovery involves pruning superfluous edges that may be present within each target neighborhood. Lines 6-12 perform conditional independence tests within a single target neighborhood at a time. The second-order neighbors are utilized to search for separating sets for nodes within a single target neighborhood.

After the skeleton recovery stage, the algorithm utilizes the separation sets to identify v-structures and then applies the appropriate FCI rules from line 14 to orient edges further. In the estimated PAG, there may exist four types of edges $(\leftrightarrow, \rightarrow, \circ\!\!\rightarrow, \circ\!\!-\!\!\circ)$. Rule R_N is then applied to simplify the edge markings from the resulting PAG.

R_N : For nodes i, j in the same target neighborhood, convert $i \leftrightarrow j$, and $i \circ\!\!-\!\!\circ j$ to an undirected edge $i - j$ and convert $i \circ\!\!\rightarrow j$ to a directed edge $i \rightarrow j$

2.2 rMATS

Alternative splicing is a fundamental mechanism in eukaryotic gene expression that allows a single gene to produce multiple mRNA and protein isoforms. Thereby greatly contributing to the complexity of the transcriptome. Differential alternative splicing, where the splicing pattern of a gene differs between conditions, plays a critical role in development and disease. Replicate multivariate analysis of transcript splicing (rMATS) Shen et al. 2012 is a statistical method designed to identify differential alternative splicing events between two different sample groups with multiple replicates from RNA seq-data.

rMATS requires aligned RNA-seq reads in BAM format and a gene annotation file in GTF format. The RNA-seq reads are mapped to various exon isoforms and the read counts are used to estimate isoform proportions. Specifically, the counts of reads that map to an exon inclusion and exclusion isoform are used to calculate the exon inclusion level, denoted as ψ. More formally, ψ is defined as the proportion of exon transcripts that splice from the upstream exon into the alternative exon and then into the downstream exon, relative to the total number of such transcripts plus the transcripts that skip the alternative exon by splicing directly from the upstream exon to the downstream exon. Figure 2.1 from (Shen et al. 2012) is an example of a skipped exon alternative splicing event. The estimation for the exon inclusion level ψ is calculated using the count of reads that map to the exon inclusion isoform(I) and the count of reads that map to the exon skipping isoforms (S). Since the length of the isoform-specific segments (junction sites and alternative exons) can vary between alternative isoforms, it is necessary to normalize the read counts by the length of the exons. For any isoform segment with length l, and read length r, the normalized effective length is given by the number of unique reads in this region, $l - r + 1$. Given the exon inclusion and skipping effective lengths l_I, l_S and the number of exon inclusion and skipping isoform reads $(I), (S)$, the exon inclusion level ψ for any exon event can be estimated as,

$$\hat{\psi} = \frac{\frac{I}{l_I}}{\frac{I}{l_S} + \frac{S}{l_S}}.$$

In each replicate, there exists estimation uncertainty for ϕ for an AS event influenced by the sequencing depth. With greater sequencing depths leading to more reliable estimates of ψ. Furthermore, there can be variation between estimates of ψ between replicates in a sample group due to biological or technical reasons. rMATS accounts for the estimation uncertainty in individual replicates by assuming the read count I follows the following

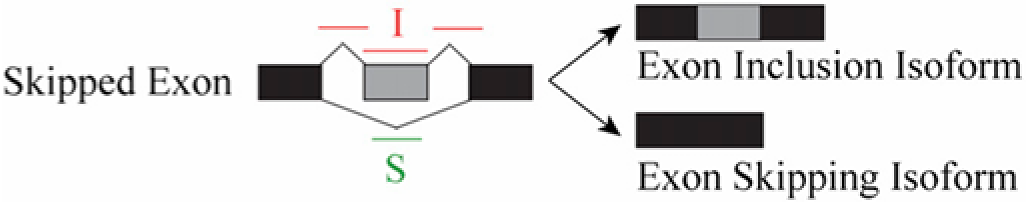


Figure 2.1: Toy visual of exon skipping event

binomial distribution.

$$I|\psi \sim \text{Binomial}\left(n = I + S, p = f(\psi) = \frac{l_I\psi}{l_I\psi + l_S(1-\psi)}\right). \qquad (1)$$

The binomial distribution models the estimation uncertainty of ψ influenced by the total number of reads n, the sequencing depth. Furthermore, the proportion of reads from the exon inclusion isoform is represented by $f(\psi)$ that normalizes ψ by the length of the isoforms. A hierarchical estimation framework to account for the estimation uncertainty in individual replicates and variability among replicates. The variation between replicates within a sample group is modeled by random effects in a mixed model. Within each sample group $j = 1, 2$ with $k = 1, \cdots, M_1$ and $k = 1, \cdots, M_2$ replicates respectively, for each exon i, the group mean of exon inclusion levels (ψ_{i1}, ψ_{i2}) are measured as fixed effects. Then the algorithm assumes the logit transformation of exon inclusion levels ψ_{ijk} follows a normal distribution with parameters $(\mu, \sigma^2) = (logit(\psi_{ij}), \sigma^2_{ij})$.

$$\text{logit}(\psi_{ijk}) \sim \text{Normal}(\mu = \text{logit}(\psi_{ij}), \sigma^2 = \sigma^2_{ij}). \qquad (2)$$

Equations (1) and (2) show how rMATS accounts for within-replicate and between-replicate variability for estimating ψ. A likelihood ratio test is then used to test whether the difference of the group means between the two groups exceeds a user-defined threshold c against the null hypothesis $|\Delta\psi_i| = |\psi_{i1} - \psi_{i2}| \leq c$.

2.3 Data Preparation

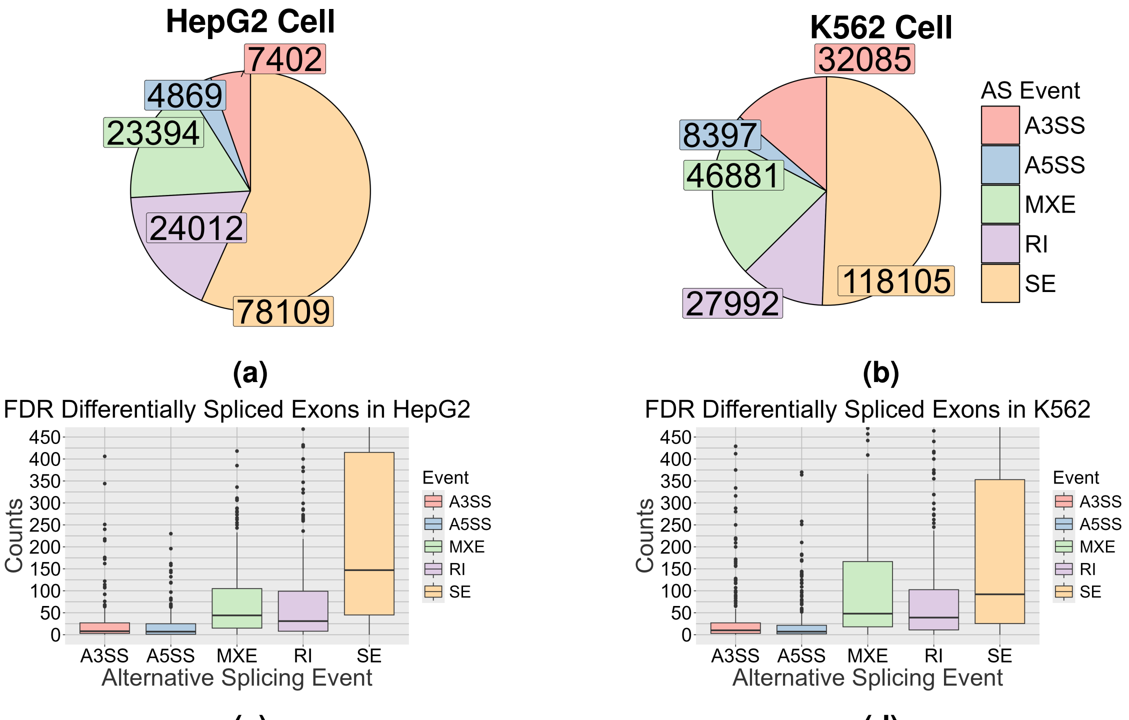
2.3.1 Introduce Data

Two biological replicates were conducted for each knockdown experiment targeting 237 and 235 RNA-binding proteins (RBPs) in the HepG2 and K562 cell lines, respectively. Additionally, two biological replicates of control experiments involving 29 and 28 non-target knockdowns were performed in the HepG2 and K562 cell lines, respectively. Reads were aligned to GRCH37 using the GENCODE v19 annotations. The reads from HepG2 and K562 cell line experiments were mapped to 15,046 and 14,942 genes respectively.

Differential alternative splicing events were analyzed using rMATS. It was implemented on the knockdown and control replicate bam files with the GENCODE v19 annotation for both cell lines. This reports five types of differential alternative splicing events: A3SS (alternative 3' splice site), A5SS (alternative 5' splice site), MXE (mutually exclusive exons), RI (retained introns), and SE (skipped exons).Figures 2.2a and 2.2b represent the number of reported alternatively spliced significant transcripts at FDR $< .05$ in HepG2 and K562 cells. Figures 2.2c and 2.2d show the distribution of FDR significant transcripts across all RBP knockdown experiments.

2.3.2 Processing Steps

Two requirements had to be met to prepare the data for applying the CML algorithm. The first is to transform the rMATS and RNA-seq data to make it compatible with the CML algorithm. The second is to augment the two mentioned data types to create a transcript-gene network when CML is applied to the augmented dataset. The log ratio fold change transformation will be used on the exon inclusion isoform levels from the rMATS output,



(a) (b) (c) (d)

Figure 2.2: Summary Statistics for number of significant transcripts

and the read counts from the RNA-seq data.

For the rMATS data processing, we consider transcripts significant if their FDR value is < 0.05. All non-significant transcripts were filtered out for each type of alternative splicing event in both cell lines. Across all RBP knockdown experiments, these FDR significant transcripts were grouped into different datasets by the type of AS events. Significant transcripts were labeled according to which RBP knockdown experiment they were detected in. A pseudo-count of 1 was then added to the inclusion and exclusion read counts for both knockdown and control replicates. Subsequently, exon inclusion levels were recalculated and averaged across replicates. Log ratio fold changes were calculated for each exon to quantify the change in its inclusion isoform compared to the control. Regarding the RNA-seq data, a pseudo-count of 2 was added to all count entries. Next, read counts across biological cell replicates were averaged, and log ratio fold changes were calculated to assess the change in gene expression relative to the control condition.

The log ratio fold change values from the rMATS and RNA-seq data will be used as a metric to determine causal relationships between genes and significant transcripts within an RBP knockdown experiment. The dataset we construct follows a structured format: each row represents an individual RBP knockdown experiment, while each column contains the log ratio fold change values pertaining to gene expression levels and exon inclusion levels for selected transcripts. Transcripts were selected according to the following criteria. We restricted exon selection belonging to the five RBP knockdown experiments having the highest number of significant transcripts, up to the top five. Shown in Figures 2.3a and 2.3b are the 5 RBPs whose knockdown experiments yielded the highest number of FDR significant A3SS transcripts, along with their respective quantities. Within each RBP knockdown experiment, we identify the most significant transcripts, with a minimum absolute value inclusion fold change of 1.5, ensuring that all selected transcripts originate from unique genes. Figure 2.1 shows an example of ten selected A3SS transcripts from the U2AF2 knockdown experiment in HepG2 along with log ratio fold change values for the inclusion isoform. The gene from which the exon was spliced and the event coordinates are presented. ES represents the exon start base, and EE represents the exon end base. FDR 0 are generated when the actual value is smaller than the numerical accuracy cutoff. Zero FDR values can be interpreted as $\leq 2.2e^{-16}$. We generate five different datasets according to these five RBPs, encompassing each AS event across both cell lines, resulting in a total of 50 datasets.

Subsequently, within each dataset, we search the aforementioned ten transcripts across all other RBP knockdown experiments. It was observed that any one of these ten transcripts was not always observed in the rMATS data for other experiments. If any of these ten transcripts are absent in at least 25% of the other experiments, we proceed to select the next most significant exon based on the same criteria. To address missing values,

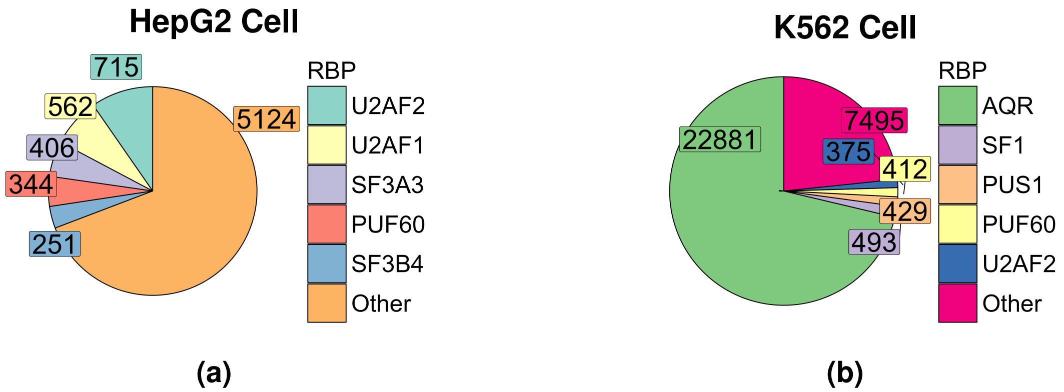


Figure 2.3: Distribution of significant A3SS exon events across RNA binding protein knockdown experiments

Gene	longES	longEE	shortES	shortEE	flankingES	flankingEE	FDR	LR
CTNNAL1	111,704,850	111,705,151	111,704,850	111,705,114	111,705,240	111,705,324	0	2.27
C17orf76-AS1	16,342,894	16,343,567	16,343,498	16,343,567	16,342,640	16,342,728	0	−1.54
PDLIM7	176,916,771	176,917,032	176,916,771	176,916,833	176,917,870	176,917,907	0	2.07
HSP90AA1	102,552,553	102,552,755	102,552,553	102,552,715	102,553,322	102,553,381	0	3.28
TSTA3	144,698,275	144,698,610	144,698,275	144,698,390	144,698,736	144,698,893	0	2.03
MICAL1	109,768,559	109,768,780	109,768,559	109,768,643	109,768,883	109,768,925	0	2.52
USP34	61,522,302	61,522,461	61,522,302	61,522,425	61,523,934	61,524,060	0	3.9
SLC35C2	44,983,711	44,983,918	44,983,711	44,983,880	44,984,440	44,984,513	0	2.66
AP3D1	2,123,828	2,123,963	2,123,828	2,123,878	2,127,150	2,127,200	0	4.41
NOTUM	79,916,810	79,916,993	79,916,810	79,916,871	79,917,346	79,917,442	0	2.33

Table 2.1: Selected A3SS transcripts from U2AF2 knockdown in HepG2 cell line

we uniformly simulate them based on the 5th quantile of the distribution of log ratio fold changes for all transcripts from rMATS within the respective RBP experiment, AS event, and cell line. Furthermore, we exclude gene expression columns from the dataset if the average read count between replicates for the target RBP falls below 10, or if the log ratio fold change in gene expression is less than 1.75 in absolute value. These columns are deemed less significant for analysis. Furthermore, the log-transformed gene expression values will be appended to each of the 50 processed rMATS datasets. The application of the Coordinated Multi-Neighborhood Learning (CML) algorithm to these datasets yields 50 distinct transcript-gene networks. These networks effectively capture causal relationships within each RBP knockdown experiment across various AS events in both cell lines.

Chapter 3

Results

The CML algorithm was run with the following parameters on each dataset. All transcripts were selected as target nodes. The MMPC algorithm was used to estimate Markov blankets and obtain first and second-order neighborhoods for each target Tsamardinos, Aliferis, and Statnikov 2003. Significance levels $\alpha_{Mb} = .01$ and $\alpha_{skel} = .01$ were used for the Markov blanket estimation and skeleton recovery steps. To assess confidence in the learned graphs from CML, we utilized bootstrapping on each dataset across experimental configurations. Running 200 bootstrap iterations, we applied CML to each resampled dataset. For each iteration, we compared the original graph $G = (V, E)$ with the bootstrapped graph $G_B = (V_B, E_B)$. An edge $(i, j) \in E$ is considered confident if it appears in at least 50% of the bootstrap iterations.

We provide tables and graphs for specific experimental configurations, summarizing the CML algorithm's results. Orange-highlighted nodes denote selected target transcripts, while white nodes represent genes. We use the notation `T_GeneSymbol` to denote a transcript that has been spliced from the given gene symbol. The exact coordinates for a transcript can be found in the supplementary. Dashed lines indicate edges between neighborhoods. Notably, the displayed graphs omit the application of rule R_N (line 15 of the CML algorithm). This omission acknowledges the potential presence of confounding variables, given the complex effects of RBPs on transcript-gene interactions. The table's ”Transcript” column lists the transcripts selected as target nodes. The ”Children” column lists the nodes that are directly affected by the transcript in the same row, indicating a causal relationship where the transcript acts as a parent node. The ”Parents” column lists the nodes that directly affect the transcript in the same row, indicating a causal relationship where the transcript acts as a child node. The ”Neighbors” column lists the nodes that are neither parents nor children but are connected to the transcript through other types of relationships. The data summarizing our results can be accessed here.

Table 3.1: HepG2 A5SS U2AF2 Causal Relationships

Transcript	Children	Parents	Neighbors
(5) T_ABHD3	KRT15	NA	NA
(6) T_ACSS2	(7) T_FDPS	NA	NA
(2) T_CASP8	SLC10A3	NA	NA
(3) T_NUP160	OAZ3 CREB5	NA	NA
(1) T_RPL10A	PIDD ADAT3	NA	NA
(4) T_SRSF7	NA	NA	MED30
(7) T_FDPS	NA	(6) T_ACSS2 CCDC17	NA

Figure 3.1: HepG2 A5SS U2AF2 Graph

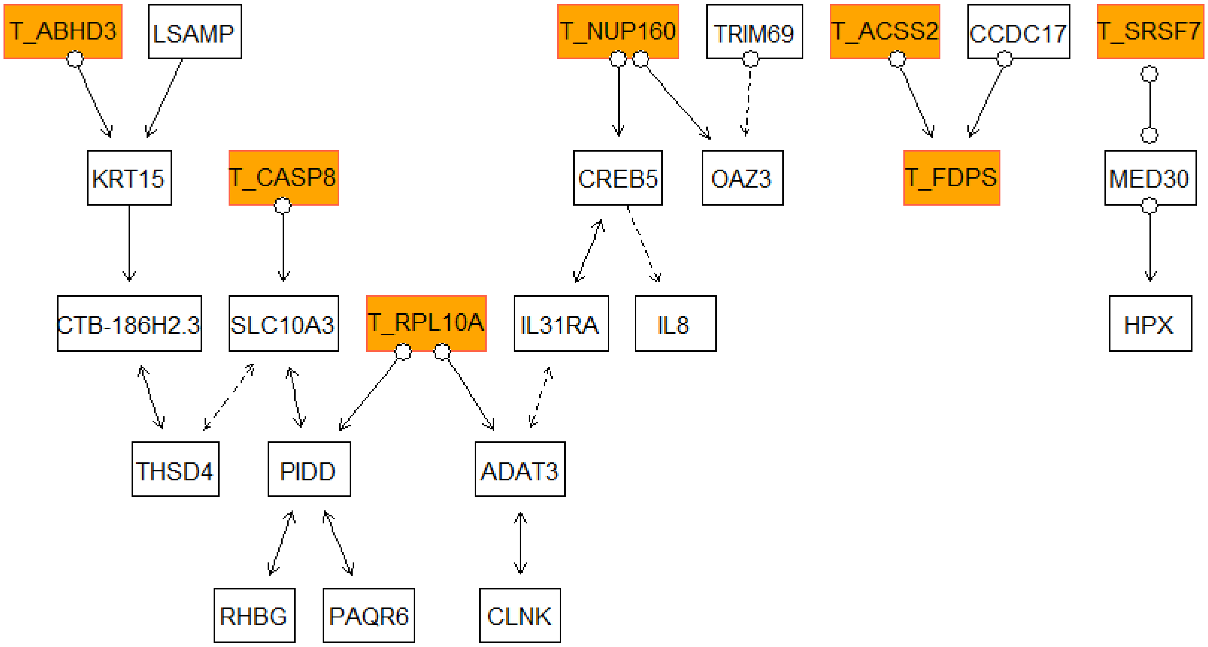


Table 3.2: HepG2 RI SF3A3 Causal Relationships

Transcript	Children	Parents	Neighbors
(9) T_ARAP1	NA	NA	(14) T_SSB
(14) T_SSB	NA	NA	(9) T_ARAP1
(12) T_MARS	(11) T_RPS18 AC024592.12	NA	NA
(15) T_PILRB	NA	SOD3	NOTCH1 SYNPO
(10) T_PTBP1	NA	(13) T_SYVN1	RGS17
(11) T_RPS18	(8) T_RPL4	(12) T_MARS ARTN	RP11-345J4.3
(13) T_SYVN1	(10) T_PTBP1	NA	NA
(8) T_RPL4	NA	(11) T_RPS18	NA

Figure 3.2: HepG2 RI SF3A3 Graph

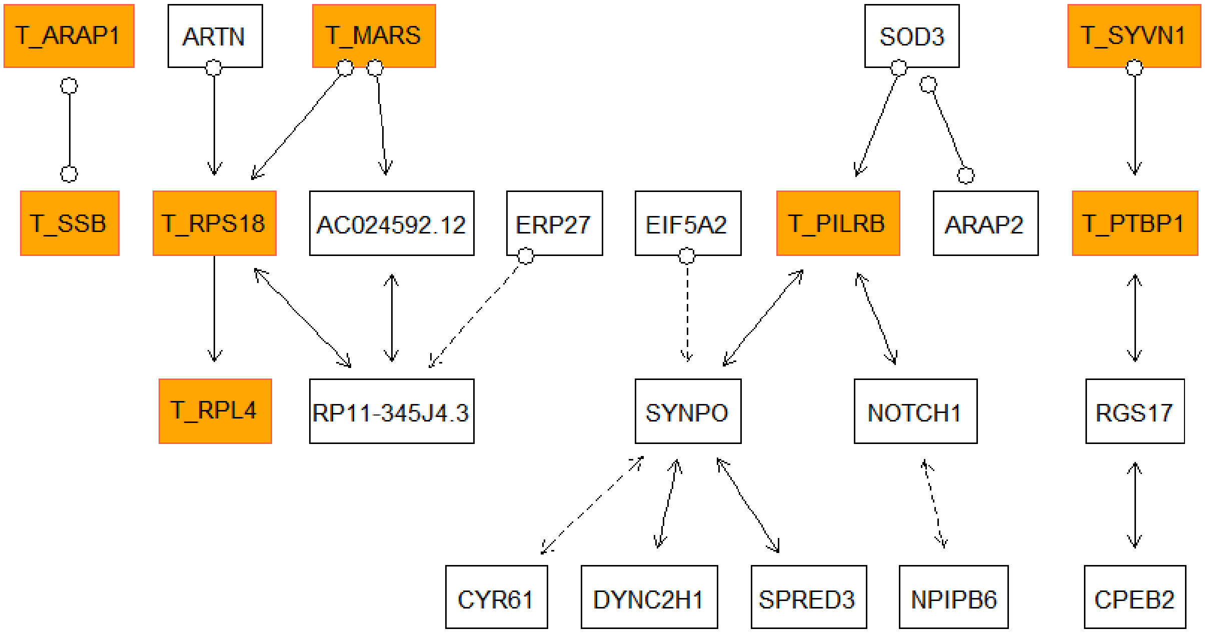


Table 3.3: HepG2 SE PUF60 Causal Relationships

Transcript	Children	Parents	Neighbors
(16) T_ACSL4	(23) T_HNMT	NA	NA
(20) T_BAG6	HR	NA	NA
(18) T_EIF4H	SECTM1	NA	NA
(19) T_SREK1	(21) T_STX16	NA	NA
(21) T_STX16	NA	(19) T_SREK1 ESRRB	CES4A
(17) T_TPM1	C1orf222	NA	NA
(22) T_WIPI2	(23) T_HNMT	NA	NA
(23) T_HNMT	NA	(16) T_ACSL4 (22) T_WIPI2	NA

Figure 3.3: HepG2 SE PUF60 Graph

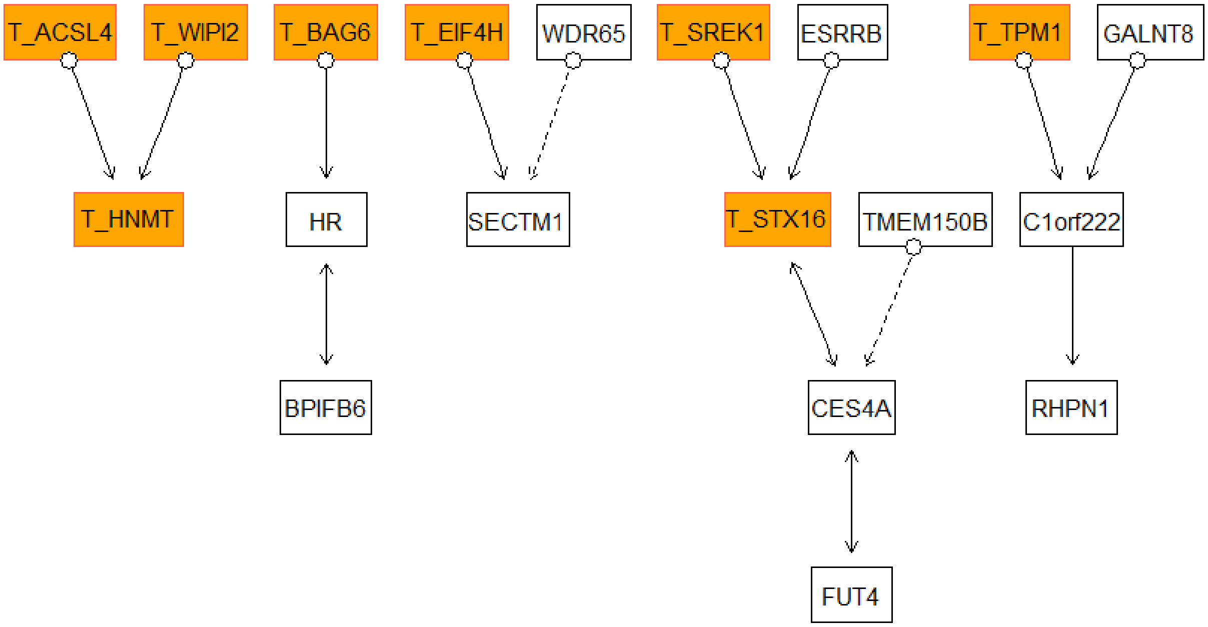


Table 3.4: K562 A3SS AQR Causal Relationships

Transcript	Children	Parents	Neighbors
(24) T_AKT2	NA	(28) T_BAZ2A (30) T_VRK1	ITGA9 PLCD3
(28) T_BAZ2A	(24) T_AKT2 GOLGA6L6	NA	NA
(27) T_CARS	(25) T_DDIT3	NA	NA
(26) T_CCNB1	NA	NA	(30) T_VRK1
(30) T_VRK1	(24) T_AKT2	NA	(26) T_CCNB1
(25) T_DDIT3	SARM1	(27) T_CARS	GAGE12I
(29) T_RPS10	CNIH2	NA	NA

Figure 3.4: K562 A3SS AQR Graph

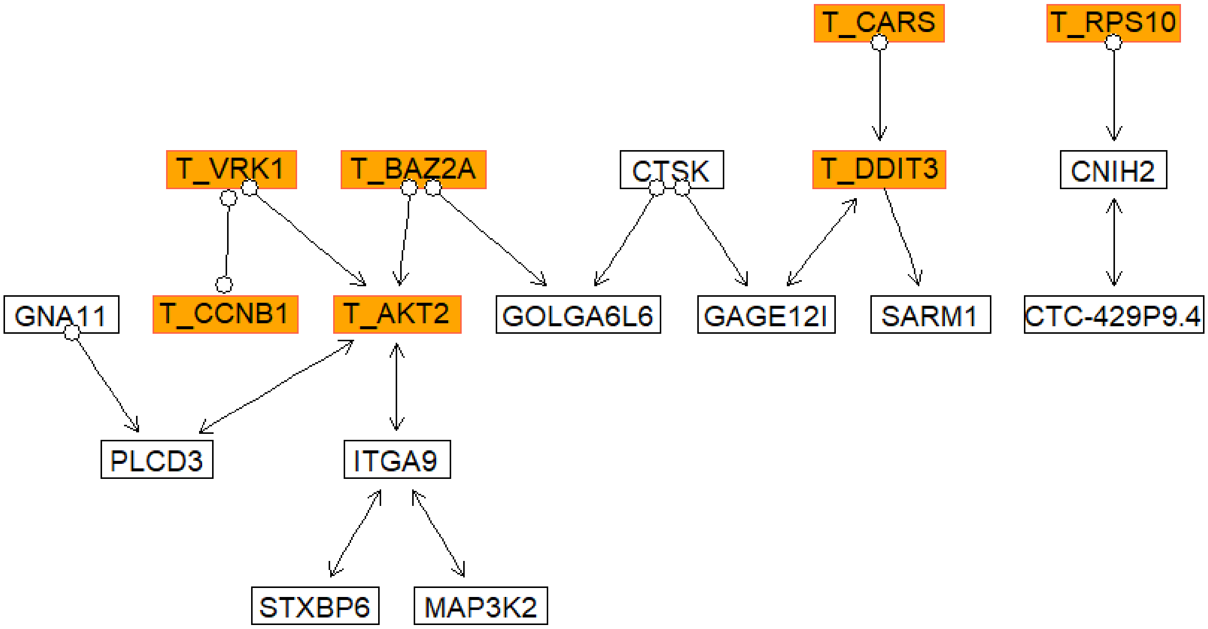
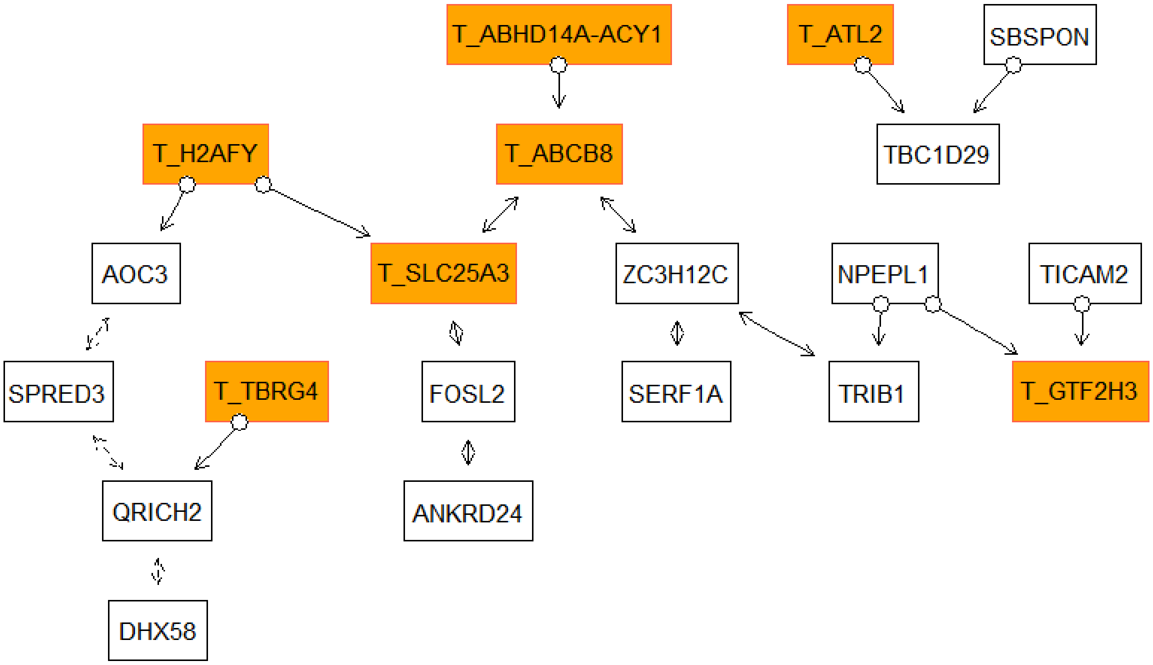


Table 3.5: K562 MXE EIF4AS Causal Relationships

Transcript	Children	Parents	Neighbors
(34) T_ABCB8	NA	(33) T_ABHD14A-ACY1	(31) T_SLC25A3 ZC3H12C
(31) T_SLC25A3	NA	T_H2AFY	(34) T_ABCB8 FOSL2
(33) T_ABHD14A-ACY1	(34) T_ABCB8	NA	NA
(35) T_ATL2	TBC1D29	NA	NA
(32) T_H2AFY	(31) T_SLC25A3 AOC3	NA	NA
(37) T_TBRG4	QRICH2	NA	NA
(36) T_GTF2H3	NA	NPEPL1 TICAM2	NA

Figure 3.5: K562 MXE EIF4A3 Graph



Chapter 4

Discussion

In this thesis, we implement the CML algorithm on an augmented dataset utilizing RNA-seq and rMATS data acquired from RBP knockdown experiments in the HepG2 and K562 cell lines. RBPs are a class of proteins with many diverse functions, including the regulation of RNA metabolism. Disruption of the functions of RBPs is known to be associated with cancer and neurodegenerative diseases. It is crucial to uncover the mechanisms governing RBP-RNA interactions. CML is a local constraint-based structure learning algorithm akin to the PC and FCI algorithms. It implements conditional independence tests and deterministic rules to recover a graph from observed data. Traditionally, structure learning algorithms struggle to infer the underlying graphical structure for observed data in high-dimensional settings. This is, in part due to the fact that as the number of variables increases, the number of possible edges in a graph grows quadratically, leading to an exponential increase in the number of possible graph structures. The search space becomes large quickly, making it computationally infeasible to explore all possible graph structures. This is commonly an issue in the field of genomics where datasets have many features and fewer observations.

A novel feature of CML is that it implements coordinated learning across multiple neighborhoods, and dramatically lowers computation cost by forgoing the estimation of an entire graph structure. The reduced computation cost allowed CML to be applied to perform structure learning to recover a graph local to target variables of interest. Within each cell line, for each of the 5 types of AS events, 5 RBPs were selected to infer causal relationships between transcripts and genes from their knockdown experiments. A total of 50 datasets were constructed using rMATS and RNA-seq data corresponding to each experimental configuration. Within each configuration of the datasets, AS exons were selected according to our filtering criteria. The datasets' columns corresponding to exons encompassed log ratio fold change values, which gauged alterations in exon inclusion levels, while the columns pertaining to genes encompassed log ratio fold change values, reflecting changes in gene expression. With our augmented dataset containing information about differential AS patterns and gene expression, several causal relationships between transcripts and genes were elucidated within an RBP knockdown experiment.

Chapter 5

Supplementary

The tables below summarize the coordinates of the selected transcripts for the selected experimental configurations shown in the results section. The ES and EE labels correspond to an exon's start and end bases respectively. The AS event coordinates are defined as follows:

- **SE (Skipped Exon)**:
  - Coordinates: ES, EE, upStrES, upStrEE, downStrES, downStrEE
  - Inclusion form includes the target exon: (ES, EE)
- **MXE (Mutually Exclusive Exons)**:
  - Coordinates: 1stESe, 1stEE, 2ndES, 2ndEE, upStrES, upStrEE, downStrES, downStrEE
  - If the strand is +, then the inclusion form includes the 1st exon: (1stES, 1stEE) and skips the 2nd exon
  - If the strand is -, then the inclusion form includes the 2nd exon: (2ndES, 2ndEE) and skips the 1st exon
- **A3SS (Alternative 3' Splice Site)**:

  - Coordinates: longES, longEE, shortES, shortEE, flankingES, flankingEE
  - Inclusion form includes the long exon: (longES, longEE) instead of the short exon: (shortES, shortEE)

- **A5SS (Alternative 5' Splice Site)**:
  - Coordinates: longES, longEE, shortES, shortEE, flankingES, flankingEE
  - Inclusion form includes the long exon: (longES, longEE) instead of the short exon: (shortES, shortEE)

- **RI (Retained Intron)**:
  - Coordinates: riES, riEE, upStrES, upStrEE, downStrES, downStrEE
  - Inclusion form includes (retains) the intron: (upStrEE, downStrES)

Table 5.1: Selected A5SS Exons from HepG2 Cell

RBP	Gene	longES	longEE	shortES	shortEE	flankingES	flankingEE	FDR	LR
U2AF2	RPL10A	35436723	35437062	35436723	35436804	35437157	35437306	0	1.88
U2AF2	RP1-283E3.8	1654026	1654270	1654146	1654270	1653034	1653150	1.19E-12	-2.32
U2AF2	CASP8	202141549	202141827	202141549	202141691	202149538	202150040	5.35E-08	-2.10
U2AF2	NUP160	47834418	47834599	47834433	47834599	47833853	47833981	1.00E-07	1.81
U2AF2	SRSF7	38976039	38976488	38976381	38976488	38975720	38975795	1.03E-07	-2.10
U2AF2	ABHD3	19243638	19244191	19244078	19244191	19239130	19239304	2.28E-07	2.19
U2AF2	IVD	40699836	40700011	40699836	40699926	40700137	40700189	1.57E-06	1.98
U2AF2	EIF3K	39116667	39116902	39116667	39116742	39123069	39123136	1.10E-05	3.23
U2AF2	ACSS2	33509132	33509276	33509132	33509265	33509346	33509403	3.95E-05	2.41
U2AF2	FDPS	155282045	155282195	155282045	155282186	155287731	155287812	3.95E-05	1.54

Table 5.2: Selected RI Exons from HepG2 Cell

RBP	Gene	riES	riEE	upStrES	upStrEE	downStrES	downStrEE	FDR	LR
SF3A3	RPL4	66794949	66795502	66794949	66795088	66795395	66795502	0	1.90
SF3A3	ARAP1	72407593	72408240	72407593	72407699	72408027	72408240	0	1.70
SF3A3	PTBP1	804035	804438	804035	804208	804291	804438	0	2.42
SF3A3	RPS18	33243741	33244044	33243741	33243843	33243952	33244044	0	1.77
SF3A3	SRRM2	2808988	2809173	2808988	2809048	2809140	2809173	0	2.83
SF3A3	MARS	57910027	57910438	57910027	57910120	57910217	57910438	0	1.51
SF3A3	CYTH2	48978094	48981402	48978094	48978206	48981322	48981402	0	1.99
SF3A3	SYVN1	64898744	64899067	64898744	64898844	64898940	64899067	2.49E-12	1.68
SF3A3	SSB	170664986	170665412	170664986	170665063	170665369	170665412	3.13E-11	2.32
SF3A3	PILRB	99950186	99950746	99950186	99950537	99950619	99950746	3.70E-11	-2.14

Table 5.3: Selected SE Exons from HepG2 Cell

RBP	Gene	ES	EE	upStrES	upStrEE	downStrES	downStrEE	FDR	LR
PUF60	ACSL4	108934231	108934360	108926364	108926601	108939372	108939425	0	2.47
PUF60	TPM1	63335904	63336030	63334956	63335142	63336225	63336351	0	3.53
PUF60	EIF4H	73604576	73604636	73604151	73604248	73609070	73609208	0	-2.68
PUF60	SREK1	65451892	65454760	65449290	65449424	65455046	65455162	0	-3.66
PUF60	BAG6	31612083	31612129	31611858	31611971	31612301	31612379	0	2.10
PUF60	UQCRC2	21990386	21990572	21987487	21987564	21991867	21992021	0	2.62
PUF60	SAT1	23802410	23802520	23801916	23802000	23803444	23803546	0	-1.57
PUF60	STX16	57234678	57234690	57226921	57227143	57242545	57242653	0	-1.65
PUF60	WIPI2	5232748	5232802	5229818	5230124	5239206	5239289	0	-2.56
PUF60	HNMT	138724666	138724956	138722048	138722198	138727734	138727787	0	-1.53

Table 5.4: Selected A3SS Exons from K562 Cell

RBP	Gene	longES	longEE	shortES	shortEE	flankingES	flankingEE	FDR	LR
AQR	AKT2	40761064	40761206	40761064	40761176	40762832	40762961	0	3.34

Continued on next page

5.4 – continued from previous page

RBP	Gene	longES	longEE	shortES	shortEE	flankingES	flankingEE	FDR	LR
AQR	INSIG1	155094438	155094556	155094456	155094556	155093960	155094127	0	4.56
AQR	DDIT3	57911051	57911242	57911051	57911221	57911488	57911536	0	2.04
AQR	CCNB1	68467078	68467279	68467096	68467279	68463999	68464170	0	4.25
AQR	CARS	3023199	3023404	3023199	3023283	3023770	3023830	0	1.52
AQR	BAZ2A	56993747	56993901	56993747	56993880	56993984	56994274	0	2.12
AQR	RPS10	34392445	34392632	34392445	34392617	34392848	34392998	0	3.76
AQR	CRLF3	29124328	29124437	29124328	29124416	29130918	29131126	0	3.82
AQR	DDX23	49230676	49230886	49230676	49230820	49231063	49231440	0	3.35
AQR	VRK1	97347492	97347728	97347513	97347728	97342366	97342457	0	3.57

Table 5.5: Selected MXE Exons from K562 Cell

RBP	Gene	Strand	1stES	1stEE	2ndES	2ndEE	upStrES	upStrEE	downStrES	downStrEE	FDR	LR
EIF4A3	SLC25A3	+	98989210	98989335	98989504	98989626	98987756	98987913	98991633	98991813	0	2.90
EIF4A3	H2AFY	-	134688635	134688735	134696186	134696470	134681657	134681747	134705095	134705293	1.56E-13	1.83
EIF4A3	TTLL3	+	9862229	9862425	9867483	9867632	9854931	9855029	9868680	9868924	9.43E-06	1.84
EIF4A3	ABHD14A-ACY1	+	52018062	52018174	52019222	52019287	52012274	52012390	52019376	52019481	4.46E-05	1.87
EIF4A3	ABCB8	+	150729930	150730148	150731359	150731515	150725536	150725697	150731591	150731686	1.26E-04	2.21
EIF4A3	ATL2	-	38570409	38570654	38581208	38581319	38546026	38546161	38604284	38604404	2.05E-04	2.27
EIF4A3	GTF2H3	+	124143976	124144022	124144062	124144131	124140317	124140371	124144341	124144477	2.64E-03	1.83
EIF4A3	TBRG4	-	45143697	45143855	45144136	45144345	45142931	45143042	45145039	45147063	3.60E-03	1.57
EIF4A3	PPOX	+	161137144	161137276	161138782	161138973	161136889	161137024	161139449	161139510	4.16E-03	1.86
EIF4A3	FANCA	-	89877114	89877210	89877336	89877479	89874701	89874775	89882944	89883065	4.19E-03	2.18

Bibliography

Gu, Jiaying and Qing Zhou (2020). “Learning big Gaussian Bayesian networks: Partition, estimation and fusion”. In: *Journal of machine learning research* 21.158, pp. 1–31.

Lukong, Kiven E et al. (2008). “RNA-binding proteins in human genetic disease”. In: *Trends in Genetics* 24.8, pp. 416–425.

Mukherjee, Neelanjan et al. (2019). “Deciphering human ribonucleoprotein regulatory networks”. In: *Nucleic acids research* 47.2, pp. 570–581.

Richardson, Thomas and Peter Spirtes (2002). “Ancestral graph Markov models”. In: *The Annals of Statistics* 30.4, pp. 962–1030.

Schadt, Eric E (2009). “Molecular networks as sensors and drivers of common human diseases”. In: *Nature* 461.7261, pp. 218–223.

Shen, Shihao et al. (2012). “MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data”. In: *Nucleic acids research* 40.8, e61–e61.

Smith, Stephen and Qing Zhou (2024). “Coordinated Multi-Neighborhood Learning on a Directed Acyclic Graph”. In: *arXiv preprint arXiv:2405.15358*.

Spirtes, Peter (2001). “An anytime algorithm for causal inference”. In: *International Workshop on Artificial Intelligence and Statistics*. PMLR, pp. 278–285.

Tsamardinos, Ioannis, Constantin F Aliferis, and Alexander Statnikov (2003). “Time and sample efficient discovery of Markov blankets and direct causal relations”. In: *Proceed-*

ings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 673–678.

Van Nostrand, Eric L et al. (2020). “A large-scale binding and functional map of human RNA-binding proteins”. In: *Nature* 583.7818, pp. 711–719.