Abstract

KIM, JIHYE. Mining of C*is*-Regulatory Motifs Associated with Tissue-Specific Alternative Splicing. (Under the direction of Steffen Heber).

Alternative splicing (AS) is an important post-transcriptional mechanism that increases protein diversity and may affect mRNA stability and translation efficiency. Despite its importance, our knowledge about its mechanism and regulation is very limited. Although it is known that the regulation of AS is influenced by multiple factors, most previous studies have focused on analyzing an individual regulator. In this dissertation, we apply three types of association rule mining techniques to discover *cis*-regulatory motifs or motif groups that are associated with specific AS patterns in mouse. General association rule mining for categorical attributes is used to find "motif=>motif" rules in gene groups that show similar exon skipping patterns. This method provides candidates for interacting motifs. Discretization-based and distribution-based quantitative association rule mining techniques are used to find "motif => exon skipping profile" rules. Many of the discovered motif candidates coincide with known splicing factor binding sites. Our ultimate goal is to find motifs and motif combinations that are involved in the dynamic regulation of AS. Based on our observations we hypothesize that some *cis*-regulatory elements affect AS only in combination with other elements. Interacting motifs show interesting differences to motifs that act individually. For example, interacting motif pairs are more conserved, they occur on average closer to the splice sites, motif pairs derived from distribution-based association rule

mining, occur also in higher multiplicity. Based on these observations, we hypothesize that interacting *cis*-regulatory motifs might often correspond to weaker binding sites that occur in clusters close to the regulated splice sites.

Mining of *Cis*-Regulatory Motifs Associated with Tissue-Specific Alternative Splicing

by
Jihye Kim

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2009

APPROVED BY:

Zhao-Bang Zeng	Barbara Sherry
Eric A. Stone	Steffen Heber Chair of Advisory Committee

Dedication

To my parents

Biography

Jihye Kim was born in Yeosu, Republic of Korea on August 18 of 1977, to Yoojo Kim and Youngsook Ahn. She spent her childhood in very small islands due to her father's mission as a teacher. She spent childhood and youth with wide experience from her parent's love and support. She received her B.S. degree in computer science from Chonnam Nat'l university in 1999. After a short break, she went to Gwangju Institute of Science and Technology (GIST). She completed M.S. degree in computer science there in 2002. In the last semester, she became interested in biological data and bioinformatics. After obtaining her M.S. degree she worked for several months in National Genome Information Center. In 2003, just after her marriage, she stared her doctoral work in the Ph.D. program in bioinformatics at North Carolina State University. During her study at NCSU, under the direction of Dr. Steffen Heber, she focused on her research on alternative splicing regulation.

Acknowledgments

I would like to thank my advisor Dr. Steffen Heber. He has guided and advised me throughout all years at BRC. His enthusiasm has been impetus for me and his unlimited patience has encouraged me all the time. I would also like to thank my committees, Dr. Zhao-Bang Zeng, Dr. Barbara Sherry and Dr. Eric A. Stone. Their advice and support have helped me to step forward. In addition, I would like to thank my good friend and co-worker, Sihui Zhao, for many good conversations.

I would like to express my appreciation to all the faculty members, staffs, and friends at BRC. I cannot thank everybody by name, but I would like to record my great debts of gratitude to the following: Dr. Jeff Thorne, Dr. Spencer Muse, Dr. Ben Redelings, Dr. Reed Cartwright, Dr. Dahlia Nielsen, Chris Smith, Stan Martin, Sunil Suchindran, Li Zhang, ClarLynda Williams-Devane, Xiaohua Gong, Youfang Liu, Jessica Maia, Li Li, Monnat Pongpanich, Alex Griffing, Brian Howard, Sang Chul Choi, Dr. Tae-Kun Seo, Paveena Lertampaiporn, Haojun Ouyanag, Noffisat Oki. They all have helped me a lot with good comments, suggestions, wide-ranging questions, and discussions over the years.

Finally, I would like to thank my family. My parents have given me unfathomable support, love, and sacrifice. I always can regain big energy and confidence from them. My big mentor, Father Minsu Timothy Kim has shown me his big support and love just like my

parents. My only sister, Jihee as my best friend has encouraged me all the time whenever I spent hard time. Aunt Sooknam Ha also encouraged me with good stories. My parents-in-law have prayed for me all the time. My husband, Hyunmin as my best co-worker has helped me very much at every moment. Without my family's endless love and support, I can never grasp a chance to study bioinformatics.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction: Basic facts about alternative splicing and its regulation

From the recent discovery that the estimate of the number of human genes has been revised down to 20,000 to 25,000 (Claverie 2001; Stein 2004) and that there are estimated 100,000 gene products, alternative splicing could be a important mechanism for producing diversity of gene products (Brett, Pospisil et al. 2002; Black 2003). Also, Previous estimates of the ratio of alternative splicing have been revised to 70% (Ladd and Cooper 2002) and 74% of genes show at least one alternative splicing form (Mironov, Fickett et al. 1999; Brett, Hanke et al. 2000; Kan, Rouchka et al. 2001; Modrek, Resch et al. 2001; Modrek and Lee 2002; Johnson, Castle et al. 2003; Leipzig, Pevzner et al. 2004). Bioinformatical analysis became essential for discover mechanism of alternative splicing with growing importance and the number of alternative splicing events.

This dissertation contributes to alternative splicing research by developing computational algorithms for the identification of alternative splicing regulatory elements. The computational methods and algorithms developed have applicability to other areas of bioinformatics research and the results produced are directly relevant to alternative splicing research.

1.1 Alternative Splicing

In higher eukaryotes, genes often contain intervening sequences (introns). In the central dogma from gene to protein eukaryote genes take an additional step to mature RNA, RNA splicing, before going out of the nucleus. During splicing the introns are removed and the remaining sequences (exons) are concatenated. In the general splicing process, the

spliceosome, which contains five small nuclear ribonucleoproteins (snRNPs), cleaves the 5' splice site (5' ss), joins the 5' end of the intron to the branch point forming a loop, cleaves the 3'splice site (3' ss), and finally ligates the exons (Fig 1) (Cooper). This process results in excision of the introns and ligation of exons. Often, a gene might be spliced in various ways, resulting in several splice variants and the corresponding protein isoforms (Fig 2). This process is known as alternative splicing (AS). There are examples of hundreds and even thousands of functionally divergent mRNAs and proteins being generated from a single gene (Black 2000).

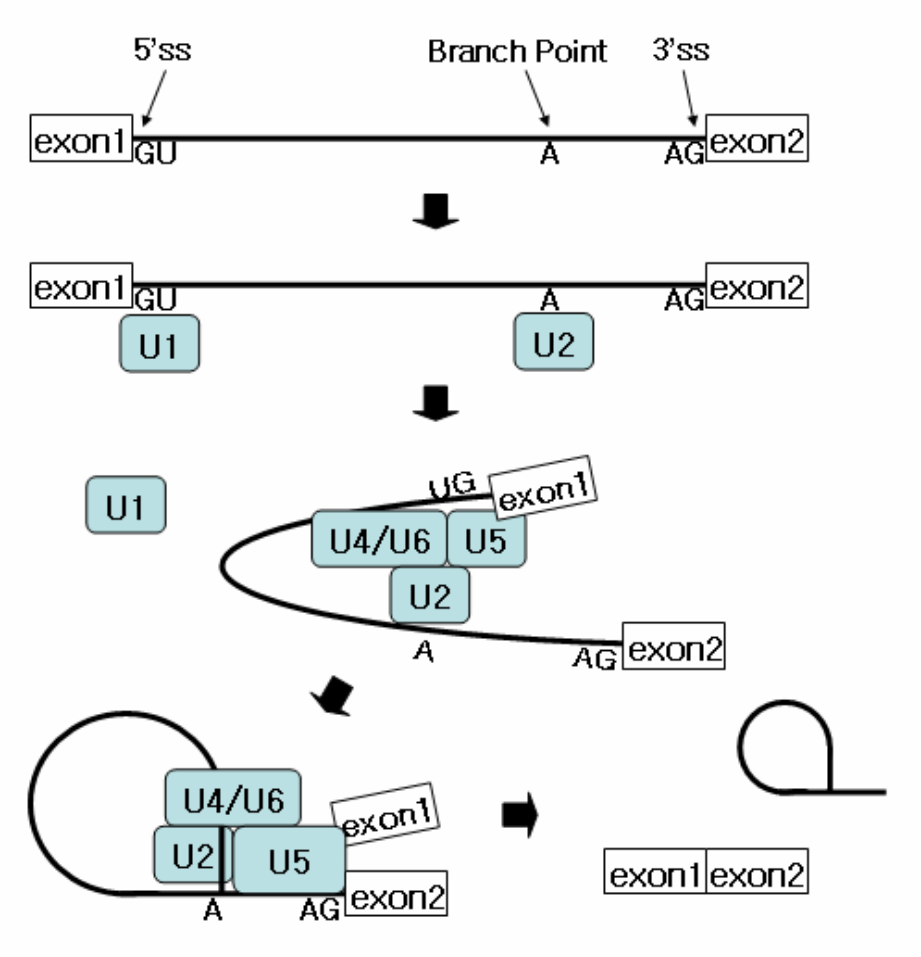


Fig 1 Splicing of pre-mRNA by spliceosome. U1 snRNP binds to the 5'ss and U2 snRNP binds to the branch point. U4/U6 and U5 snRNP complex enters the spliceosome. U5 binds to the upstream of the 5'ss, U6 displaces U1 and then U5 binds to the 3'ss, followed by removal of intron and concatenation of exons.

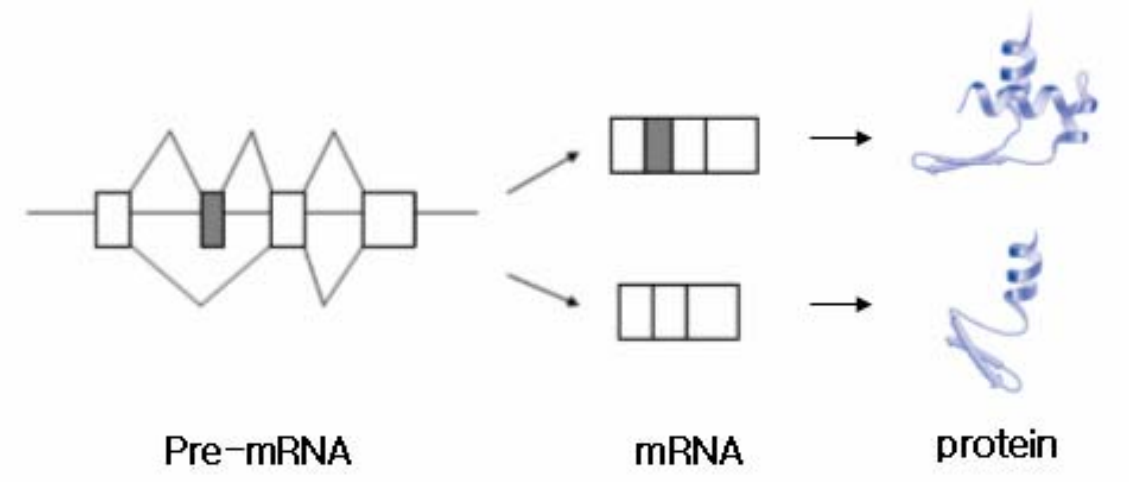


Fig 2 Alternative splicing (exon skipping). Rectangles represent exons in the gene sequence. Two mRNA isoforms and corresponding proteins are generated. Exon 2 (marked in grey) can be either included or skipped.

AS plays an important role in the generation of protein diversity, subcellular localization, as well as processes such as transcription and signal transduction (Brudno, Gelfand et al. 2001). AS is also involved in diseases such as familial isolated GH deficiency type II (IGHD II), Frasier syndrome, and myotonic dystrophy (Faustino and Cooper 2003; Garcia-Blanco, Baraniak et al. 2004).

It is estimated that up to 70% of human genes are alternatively spliced (Ladd and Cooper 2002), and this percentage might even increase if one takes into account that often AS events occur only in specific tissues, and at specific developmental stages (Yeo, Holste et al. 2004).

There are several patterns of AS (Fig 3)(Cartegni, Wang et al. 2003). Exons that are always included in the mRNA are called constitutive. A cassette exon is an exon that is

sometimes included and sometimes skipped. In animals, exon skipping is the most frequent AS pattern. A famous example for exon skipping is sex determination in drosophila. In male flies the gene *Sxl* includes exon 3, while in female flies this exon is skipped (Baker and Rubin 1989; Black 2003).

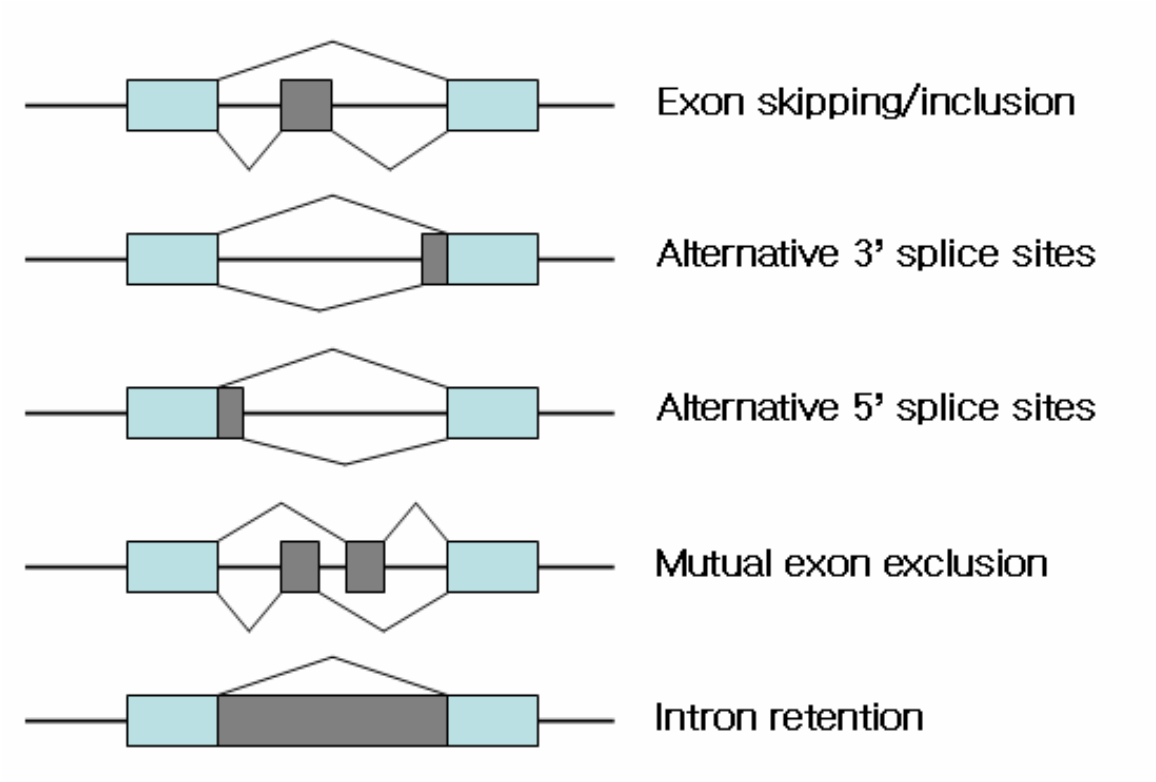


Fig 3 Alternative Splicing Patterns. In each case, one splicing path is indicated in the upper line and the other AS path is indicated in the lower line. The exon/intron part which makes a difference is marked as a gray box. In the intron retention type, the alternative pattern path represents no splicing. The whole intron is included in the final mRNA.

1.1.1 Analysis of Alternative Splicing

There are few bioinformatics tools for de novo AS prediction, and due to our limited knowledge of the splicing process, their predictions typically represent only a fraction of the true transcripts (Stanke, Keller et al. 2006; Stanke, Tzvetkova et al. 2006).

The most common way to identify AS involves aligning and comparing EST/cDNA sequences. Examples include UniGene (Schuler 1997), TIGR Gene Indices (Lee, Tsai et al. 2005), and GeneNest (Haas, Beissbarth et al. 2000). This method has significant limitations due to biases in transcript coverage, non-uniformity of EST/cDNA libraries, and transcript sampling (Lee and Roy 2004). Some of the problems of EST/cDNA analyses have been overcome by the development of AS oligonucleotide microarrays (Hu, Madore et al. 2001; Pan, Shai et al. 2004). Microarray experiments can measure whether a specific splice form constitutes an important fraction of a gene's transcripts, and investigate its regulation across different tissues.

Very recently, high-throughput sequencing technology became available to investigate AS (Pan, Shai et al. 2008). Second-generation DNA sequencing (including 454 pyrosequencing, Illumina, and SOLiD platforms) is capable to detect and quantify alternative mRNA isoforms (Margulies, Egholm et al. 2005; Shendure, Porreca et al. 2005; Turcatti, Romieu et al. 2008) is capable to detect and quantify alternative mRNA isoforms. This sequencing approach is expected to cover many of the limitations of previous technologies used to investigate AS

1.1.2 Alternative Splicing Resources

High throughput techniques such as sequencing and microarray experiments allow us to generate a huge amount of data. Information about AS is spread among many databases. Some data is available with direct or indirect annotation in various sequence databases such as GenBank and EBI. Parallel to the increasing data volume, programs and web servers for analysis of AS are developed. A summary of currently available AS resources is given in Table 1.

PALSdb (Huang, Chen et al. 2002) and EASED (Pospisil, Herrmann et al. 2004) collected AS events by comparisons of cDNA and protein sequences, ASAP (Lee, Atanelov et al. 2003), AltExtron and AltSplice (Thanaraj, Stamm et al. 2004), and SpliceInfo (Huang, Horng et al. 2005) collected AS events from genomic exon-intron structures. Besides sequence and splicing information, databases also contain annotations such as tissue specificity, developmental stage, expression level, GC content, repeat information, conservation information and the biological function. AEDB database (Stamm, Zhu et al. 2000), as part of ASD (Thanaraj, Stamm et al. 2004), contains experimentally identified splicing regulatory signals extracted from the literature.

In addition to databases, several programs and web servers are developed to annotate alternatively spliced transcripts based on alignment of cDNAs and protein sequences. ASG (Leipzig, Pevzner et al. 2004) provides splice graphs for several eukaryotic genomes and ASmodeler (Kim, Shin et al. 2004) help users to create their own splice graph annotations. TIGR Gene Indices (Schuler 1997) and NCBI UniGene (Liang, Holt et al. 2000) provide

comprising gene index collections.

Table 1 Resources of alternative splicing on the web.

Resource	Description	Address
ASAP (Lee, Atanelov et al. 2003)	Database of alternative splice events by comparison of cDNA and genome alignments	http://bioinfo.mbi.ucla.edu/ASAP/
ASD, AltExtron, AltSplice (Stamm, Riethoven et al. 2006)	Database of alternative splice events by comparison of cDNA and genome alignments	http://www.ebi.ac.uk/asd/altextron/ http://www.ebi.ac.uk/asd/altsplice/
ASD, AEDB (Stamm, Riethoven et al. 2006)	Database of alternative splice events by experimental data from the literature	http://www.ebi.ac.uk/asd/aedb/
PALSdb (Huang, Chen et al. 2002)	Database of alternative splice events by comparison of mRNA and EST sequences	http://ymbc.ym.edu.tw/palsdb/
SpliceInfo (Huang, Horng et al. 2005)	Database of alternative splice events by comparison of cDNA, protein and genome alignments	http://spliceinfo.mbc.nctu.edu.tw/
ASG (Leipzig, Pevzner et al. 2004)	Genome-based splice graphs by a collection of transcripts	http://statgen.ncsu.edu/asg/
Asmodeler (Kim, Shin et al. 2004)	Genome-based splice graphs for transcript prediction	http://genome.ewha.ac.kr/ECgene/ASmodeler/
TIGR (Liang, Holt et al. 2000)	Gene indices by cDNA clustering and assembly	http://compbio.dfci.harvard.edu/tgi/
UniGene (Schuler 1997)	Gene indices by cDNA clustering and assembly	http://www.ncbi.nlm.nih.gov/unigene

1.2 Alternative Splicing Regulation

There is growing evidence about the importance of regulated AS for novel therapeutics and diagnostic markers (Lyddy 2002). However, the splicing code, a set of rules for AS regulation is poorly understood. The first layer of the splicing code consists of consensus splice site sequences located at exon/intron boundaries. A second layer consists of proteins (AS factors) that tend to recognize short sequences (*cis*-elements, see Fig 4) that are located close to regulated splice sites, and that selectively control splice site choice (see (Cooper 2001) for a comprehensive review). For example, a study that used a custom microarray to profile AS in mice showed that deletion of the neural specific AS factor Nova-2 primarily affects AS events in genes related to synaptic proteins, or axon guidance (Ule, Ule et al. 2005). It has been shown that tissue-specific AS and tissue-specific transcription primarily regulate independent sets of genes, and that often splicing is not only regulated by a single protein binding to a single *cis*-element, but by a combination of multiple proteins (Burge, Padgett et al. 1998; Frilander and Steitz 1999; Pilpel, Sudarsanam et al. 2001; Hannenhalli and Levy 2002; Kato, Hata et al. 2004; Chan, Elemento et al. 2005; Vardhanabhuti, Wang et al. 2007; Sinha, Adler et al. 2008).

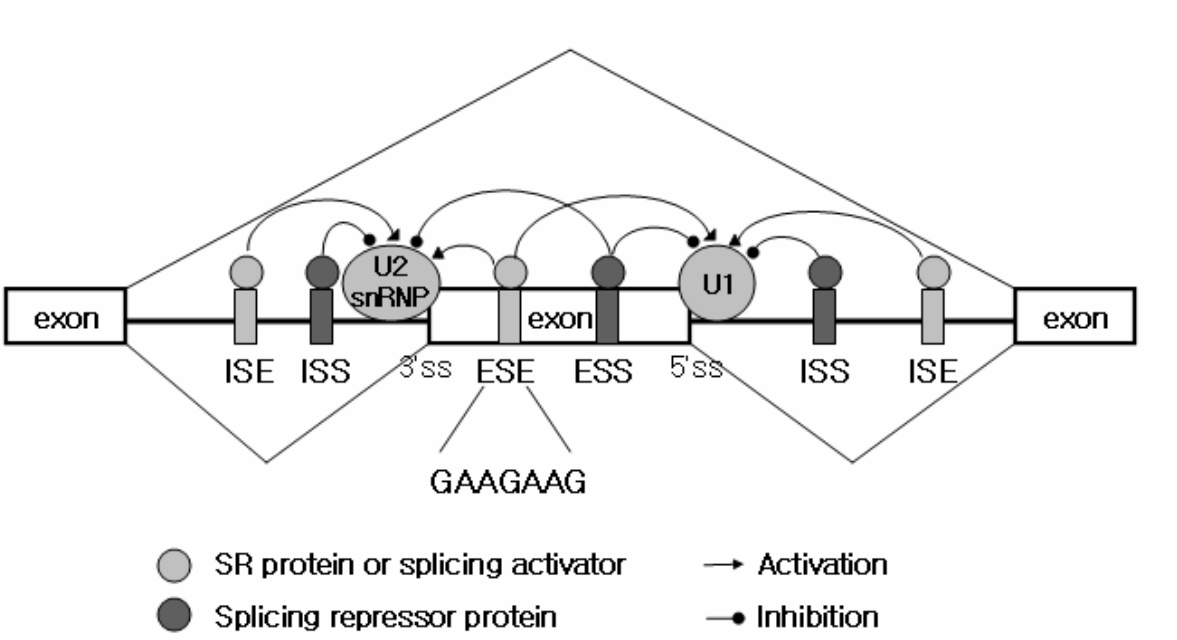


Fig 4 *Cis*-elements on a pre-mRNA sequence. *Cis*-elements can be located on both exon and intron with a function as an enhancer or silencer – ISE (intronic splicing enhancer), ISS (intronic splicing silencer), ESE (exonic splicing enhancer), and ESS (exonic splicing silencer). They are known to be generally short (5 to 10-mer). For example a sequence GAAGAAG is an exonic splicing enhancer for a rat gene, COT (Caudevilla, Codony et al. 2001).

Recently, Florea (Florea 2006) classified patterns of AS mechanisms (Fig 5). AS enhancement (or activation) is preceded by recruiting splicing activator proteins such as SR (Ser/Arg) proteins at RRM (RNA recognition motif) domains, RS domain or exon splicing enhancers (ESEs). AS repression (or silencing) is induced by splicing silencers (splicing repressor proteins) such as hnRNPs or polypyrimidine tract binding proteins (PTBs).

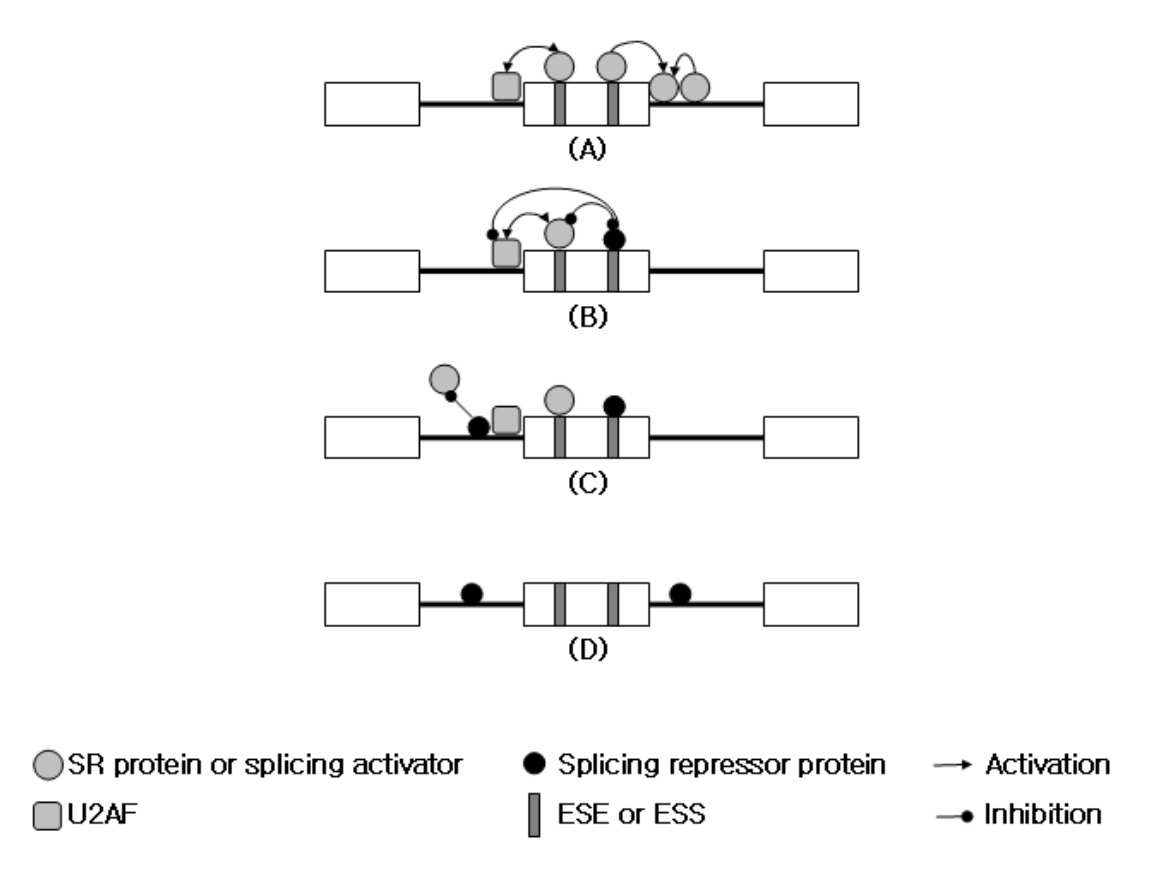


Fig 5 Mechanisms of AS. (A) Splicing activator binds ESE RS-domain to enhance AS. Splicing activators activates U2AF or other splicing activators. (B) Splicing activator and Splicing repressor compete against each other. Winner protein activates its assistant proteins to activate or inhibit AS. (C) Sometimes combinatorial splicing factors (repressors in this picture) blocks splicing activator protein. (D) Combinatorial splicing factors (repressors in this picture) bind to the motifs on the pre-mRNA sequences to silence the exon.

1.2.1 Methods to Investigate the Regulation of Alternative Splicing

Because of a growing interest in AS, many experimental and computational methods have been developed to discover AS regulatory elements.

1.2.1.1 Experimental methods

Early detection of AS regulatory elements resulted in identifying individual genes with certain features. A number of publications describe specific experimental studies for special cases of individual genes in a specific environment. Large-scale experiments are also being pursued to find AS regulatory elements.

1.2.1.1.1 SELEX (Systematic Evolution of Ligands by Exponential Enrichment)

SELEX (Tuerk and Gold 1990) allows users to discover AS regulatory elements in vivo. Randomly generated oligonucleotides are exposed to gene sequences or part of gene sequences, which are supposed as pre-mRNA. Those that do not bind the gene sequences are removed and spliced sequences are amplified by RT-PCR to go through the next cycle. After several cycles, sequences with high binding affinity for specific splicing profiles are obtained. Often, ESE (exon skipping enhancer) elements are found such as ASF/SF2, SC35, SRp40 or SRp55 (Liu, Zhang et al. 1998; Cartegni, Wang et al. 2003). A web tool, ESEfinder(Cartegni, Wang et al. 2003) provides ESE searching with query sequences by using weight matrices of four different human SR proteins found by SELEX method. Many other computational methods also use the SELEX method to validate their motifs.

1.2.1.1.2 In vivo splicing reporter system

C. Burge and his colleagues used a green-fluorescent protein (GFP) reporter to detect ESS (exon skipping silencer) elements (Wang, Rolish et al. 2004). They used three-exon minigenes whose middle exons are suspected to be skipped. The first and last exons together form the complete green fluorescent protein (GFP), and the middle exon contains a sequence that interrupts the reporter. The middle exon is used for testing ESS properties. A putative ESS oligonucleotide is inserted into the second exon. If the oligonucleotide is indeed an ESS sequence, the middle exon is skipped and the transcribed mRNA encodes a functional GFP protein. Minigenes are transfected into cultured cells, the cells with GFP are detected by fluorescent-activated cell sorting (FACS), and their inserts are sequenced to identify the oligonucleotides responsible for exon skipping.

1.2.1.2 Computational methods

As genome-wide methods, such as microarrays, have been applied to detect AS events, computational and statistical approaches methods were also applied to investigate the regulation of AS. Many methods have been developed and they can be categorized into several groups.

1.2.1.2.1 Methods for finding exonic elements

RESCUE-ESE (Fairbrother, Yeh et al. 2002) identifies ESEs by a statistical analysis of exons, introns, and strong, weak splice site signals. They collected all possible hexamers in exons and flanking intron regions. They also selected hexamers that were enriched in exons (as

compared to introns) and genes with weak splice sites. The idea assumes that exonic enhancing motifs are more frequent in exons with weak splice sites. They found 238 hexamers overrepresented in exons with weak splice site signals. After clustering similar heptamers by CLUSTALW(Thompson, Higgins et al. 1994), 10 motifs were discovered and validated by in-vivo experiments using SELEX. Later, they applied the VERIFY (variant elimination reinforces functionality) method to assess the natural selection acting on hexamers they found (Fairbrother, Holste et al. 2004). From aligning human SNPs to the chimpanzee genome, they analyzed overlapping mutations for SNPs and their hexamers. They concluded that one-fifth of the mutations that break their prediction have vanished. This suggested valuable factors to identify variants of splicing as well as phenotypes.

Similarly, Chasin et al.(Zhang and Chasin 2004) identified ESEs by comparing frequent oligonucleotides in non-coding exons with pseudo-exons and 5'UTRs of intronless genes to avoid protein coding information.

Also, Itoh and his co-workers(Itoh, Washio et al. 2004) reported the comparative analyses indicating that AS exons have weaker splice sites and more regulatory motifs than constitutive exons. From *M.musculus*, they extracted 62 motifs including GAAGAAG, which overlaps with RESCUE-ESE results from humans.

Regulation activity of many predicted ESEs and ESSs is measured by neighborhood inference (NI) that predicts sequences with activity in regulating a biochemical process (Stadler, Shomron et al. 2006). Hexamers that are candidates for splicing elements are identified by their predictive power, measured by cross-validation, and their degree of sequence conservation, and validated by their effects shown through in-vivo splicing reporter

assays. The experiments revealed that orthologous exons in mammals are highly conserved over background than ESE primary sequence and that ESE sequences are frequently interchangeable in the exon of mammals.

1.2.1.2.2 Methods for finding intronic elements

Lim et al.(Lim and Burge 2001) analyzed short intron sequences. They measured information content using five eukaryote genomes and used Monte Carlo simulations to determine the necessary information for detecting reliably short introns in each organism. They discovered the fact that additional pentamers as motifs can improve splicing prediction, while 5'ss, 3'ss, and branch signals are not enough for the prediction of splicing.

Yeo et al.(Yeo, Van Nostrand et al. 2007) recently analyzed conserved oligonucleotides in 4 different species. They focused on the flanking 400 bp long introns. Statistically overrepresented oligonucleotides are extracted and grouped by conservation information. A similar approach with nematodes is presented by Kabat et al.(Kabat, Barberan-Soler et al. 2006)

1.2.1.2.3 Methods for finding tissue specific AS regulatory elements

Brudno and his co-workers(Brudno, Gelfand et al. 2001) applied computational approaches to identify tissue-specific AS elements. They started by retrieving 25 brain-specific alternative-splicing- cassette exons from the literature and assumed that splicing is regulated by short sequences near introns. They compared the introns near the 25 brain- specific alternative- splicing- cassette exons with a corresponding set derived from constitutive exons.

From this, they found divergent 5'ss, highly pyrimidine-rich upstream introns, a paucity of GGG motifs in the downstream intron, and enriched UAGAUG in the downstream intron.

A study by Stamm et al.(Stamm, Zhu et al. 2000) reported tissue-specific expressed cassette exons. Applying a Gibb's algorithm on the database from the literature, they identified several motifs in exons surrounded by weak splice sites and in tissue-specific exons. They also showed some features of alternative exons. For example, they showed that they are significantly skewed towards small lengths while lengths of constitutive exons are normally distributed; their splice site are more variant than the consensus; their 3' splice sites contains many purines; their 5' splice sites are more variant at +4 and +5 positions; adenosine is more frequently used at -3 position of the 3' splice site for a single tissue specific expressed exons. They suggested that there is a combinatorial effect of weak splice sites, atypical nucleotide usage at certain positions, and functional enhancers for alternative exon regulation.

Similarly, Zavolan et al (Zavolan, Kondo et al. 2003) also compared constitutive exons and cassette exons. They analyzed full-length cDNA sequences and public mRNA sequences. They identified a significant length difference between cassette and constitutive exons. To identify sequence motifs, they collected overrepresented and underrepresented motifs in cassette exons relative to constitutive exons in mRNA. Interestingly, they found that TGAAG and AAGAA containing motifs reported as ESE in RESCUE-ESE are overrepresented in both cassette and constitutive exons while TGGA-containing motifs are overrepresented only in constitutive exons. They found additional motif features, e.g. constitutive exons have CG-containing motifs while cassette exons have many pyrimidine-

rich motifs similar to SRp20 motifs reported by Schall and Maniatis(Schaal and Maniatis 1999). Many of their motifs are hRNP binding sites, which are G-rich (frequently AGGG containing motifs).

1.2.1.2.4 Combinatorial AS regulatory elements

Han et al.(Han, Yeo et al. 2005) discovered tissue-specific combinatorial motifs by an experimental approach. They identified that UAGG and GGGG motifs functions together to silence the brain-specific cassette exon of the glutamate NMDA R1 receptor (GRIN1) transcript. Their results indicate that combinatorial signals may strongly influence tissue-specific regulation of the cassette exon.

Recently, Burge's group identified interacting pairs of *cis*-regulatory elements by finding statistically co-conserved and co-occurring oligonucleotides (Friedman, Stadler et al. 2008). Compositionally orthogonalized co-occurrence analysis (coCOA) identified three clusters of oligonucleotide pairs that frequently co-occur at 5' and 3' boundaries of human and mouse introns. They describe GC-rich sequences at the 5' ends of introns that co-occur and are co-conserved with specific AU-rich sequences near intron 3' ends. The motif pair is expected to silence the intervening exons. This was verified by a splicing reporter assay.

1.2.2 Databases of AS Regulatory Elements

Currently, there are only a few available databases of AS regulatory elements. AEDB (Alternative Exon Database)(Stamm, Zhu et al. 2000) is a database of AS regulatory elements covering several species including humans. AEDB is a sub-menu of ASD (Alternative

Splicing Database)(Stamm, Riethoven et al. 2006) Project in EBI (http://www.ebi.ac.uk/). AEDB is manually generated from the literature with experimental verification. It provides alternatively spliced exon sequences, their function, regulatory motifs, minigenes and associated diseases. About 300 regulatory motifs with various lengths (3 to several tens bp long) are stored in AEDB.

Hollywood (Holste, Huo et al. 2006) is a database made by Burge's group. It is based on the genomic annotation of splicing patterns of known genes from alignment of cDNAs and ESTs. Hollywood also provides splicing features such as splice site, strength, type of splicing factors (enhancer, silencer), and conserved/non-conserved patterns of splicing as well as splicing regulatory elements of human and mouse. Hollywood is a collection of Burge's group's knowledge from all AS regulation projects including RESCUE-ESE (Fairbrother, Yeh et al. 2002), ACEs (Yeo, Van Nostrand et al. 2005), and FAS-ESSs (Wang, Rolish et al. 2004).

1.3 Outline of Dissertation

This dissertation describes new methods, motif association rule mining algorithms for finding *cis*-regulatory elements or motifs which are involved in tissue-specific alternative splicing. Through computational experiments with association rule mining we predicted individual and combinatorial motifs in alternatively spliced mouse genes. For representing *cis*-regulatory elements, we restricted AS motifs with short sequences (5-9mers) and we treated them as items or attributes in association rule mining techniques. The hypothesis of the

dissertation is that computational methods taking the biological contexts of alternative splicing are able to provide accurate predictions and are able to discover novel AS motifs. This hypothesis was evaluated by developing methods for discovery of motifs in alternatively spliced mouse genes and examining predictions. The subsequent chapters of this dissertation describe the introduction of association rule mining technique and following research projects

1. Introduction of association rule mining (ARM) and its algorithms and concepts
2. Design and evaluation of algorithm of motif association rule mining in alternatively spliced mouse gene datasets
3. Design and evaluation of algorithm of discretization-based quantitative association rule mining in alternatively spliced mouse gene datasets
4. Design and evaluation of algorithm of distribution-based quantitative association rule mining in alternatively spliced mouse gene datasets

Chapter 2

Association Rule Mining

2.1 Association Rule Mining (ARM)

Association rule mining is a tool for finding unexpected relationships or associations among a set of items (Agrawal and Srikant 1994). The association relationships are described in association rules. Each rule has two measurements, support and confidence. Confidence is a measure of the rules' strength, while support corresponds to statistical significance indicating how frequently the items present. Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence respectively. Formal description is followed.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its *TID*. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. X is called antecedent (or left hand side, lhs) while Y is called consequence (or right hand side, rhs) of the rule. The rule $X \Rightarrow Y$ holds in the transactions in D with support s, where s is the percentage of transactions in D that contain $X \cap Y$. This is taken as the probability, $P(X \cap Y)$. The rule $X \Rightarrow Y$ has confidence c in the transaction set D if c is the percentage of transactions in D containing X that also contain Y. This is taken to be the conditional probability, $P(Y|X)$. That is

$$\text{Support } (X \Rightarrow Y) = P(X \cup Y)$$
$$\text{Confidence}(X \Rightarrow Y) = P(Y|X) \qquad (1)$$

2.1.1 Algorithm

Given a user specified minimum support and minimum confidence, the problem of mining association rules is to find all the association rules whose support and confidence are larger than the respective thresholds. Thus, it can be decomposed into two sub-problems:

1. Finding all frequent itemsets whose support values are above the user-determined minimum support.
2. Deriving all rules, based on each frequent itemset, which have more than the user-determined minimum confidence.

Apriori is the basic algorithm for association rule mining. Many ARM algorithms such as DHP (Park, Chen et al. 1995), FDM (Cheungt, Hans et al.), CD (Agrawal and Shafer 1996), DD (Agrawal and Shafer 1996), IDD (Han and Karypis), HD (Han and Karypis), and CCPD (Zaki, Ogihara et al. 1996) are based on the Apriori and hash tree concept. Fig 5 describes the Apriori algorithm. Starting by finding all frequent 1-itemsets (1-sized itemsets), we then consider 2-itemsets, and so forth. During the each iteration only candidates found to be frequent in the previous iteration are used to generate a new candidate set during the next iteration. The algorithm terminates when there are no frequent *k*-itemset.

The apriori-gen function takes the argument L_{k-1} and returns a superset of the set of all frequent *k*-itemsets. It consists of a join step and prune step. In the join step, generate C_k from joining L_{k-1} with itself. In the prune step, delete all itemsets $c \in C_k$ such that some (k-1)-subset of c is not in L_{k-1} since any $(k-1)$-itemset that is not frequent cannot be a subset of a

frequent *k*-itemset. A subset function is used to find all the candidate *k*-itemsets in a transaction database using a hash tree.

In the rule generation step, we make all possible candidate rules with all frequent itemsets. Each frequent itemset can be antecedent and consequent, but not both simultaneously. After generating candidate rules, we discard rules that cannot satisfy a minimum confidence threshold. We call only rules satisfying minimum confidence association rules. Here, we note that a rule A=>B is not same as B=>A. Also, an association rule A=>B does not guarantee that B=>A also satisfies the minimum confidence.

Step 1 : Finding Frequent Itemset

C_k : {Candidate k-sized itemset}

L_k : {frequent k-sized itemset}

L_1 = {frequent 1-sized itemsets};

For (k = 2; L_{k-1} != 0; k++} do begin

    C_k = apriori-gen (L_{k-1}); // New Candidates

    foreall transaction $t \in D$ do begin

        C_t = substet (C_k, t); // Candidates contained in t

        forall candidates $c \in C_t$ do

            $c.count$ ++;

        end

        L_k = {$c \in C_k$ | $c.count$ > minsupp}

    end

    $FrequentItems$ = $\cup_k L_k$;

Step 2: Rule Generation

Rules = rule-generate ($FrequentItemsets$)

Fig 6 Apriori algorithm

2.1.2 Example of Apriori

We have a sample database in Fig 7 (A). Suppose that we want to find association rules with the minimum support 40%. We assume that the transactions in the database are lexicographically ordered. Fig 8 shows how to generate frequent itemsets. From all frequent itemsets we generate candidate rules and discard ones which are not satisfying the minimum confidence.

TID	Items
1	A B C D E G
2	A B E
3	B D E F
4	B C F G
5	A B E F G

(A) Sample DB

1-sized itemsets (support)

A (0.6), B(1.0), C(0.4), D(0.4), E(0.8),

F(0.6), G(0.6)

2-sized itemsets (support)

{A,B}(0.6), {A,C}(0.2), {A,D}(0.2), {A,E}(0.6), {A,F}(0.2), {A,G}(0.4), {B,C}(0.4), {B,D}(0.4), {B,E}(0.8), {B,F}(0.6), {B,G}(0.6), {C,D}(0.2), {C,E}(0.2), {C,F}(0.2), {C,G}(0.4), {D,E}(0.4), {D,F}(0.2), {D,G}(0.2), {E,F}(0.4),{E,G}(0.4), {F,G}(0.4)

3-sized itemsets (support)

{A,B,E}(0.6), {A,B,G}(0.4), {A,E,G}(0.4), {B,C,G}(0.4), {B,D,E}(0.4), {B,E,F}(0.4), {B,E,G}(0.4), {B,F,G}(0.4), {E,F,G}(0.2)

4-sized itemsets (support)

{A,B,E,G}(0.4)

(B) Finding frequent itemsets

Fig 7 An example of finding frequent itemsets. From a sample database, we find frequent itemsets with 40% of minimum support. Underlined itemsets are dropped in choosing frequent itemsets.

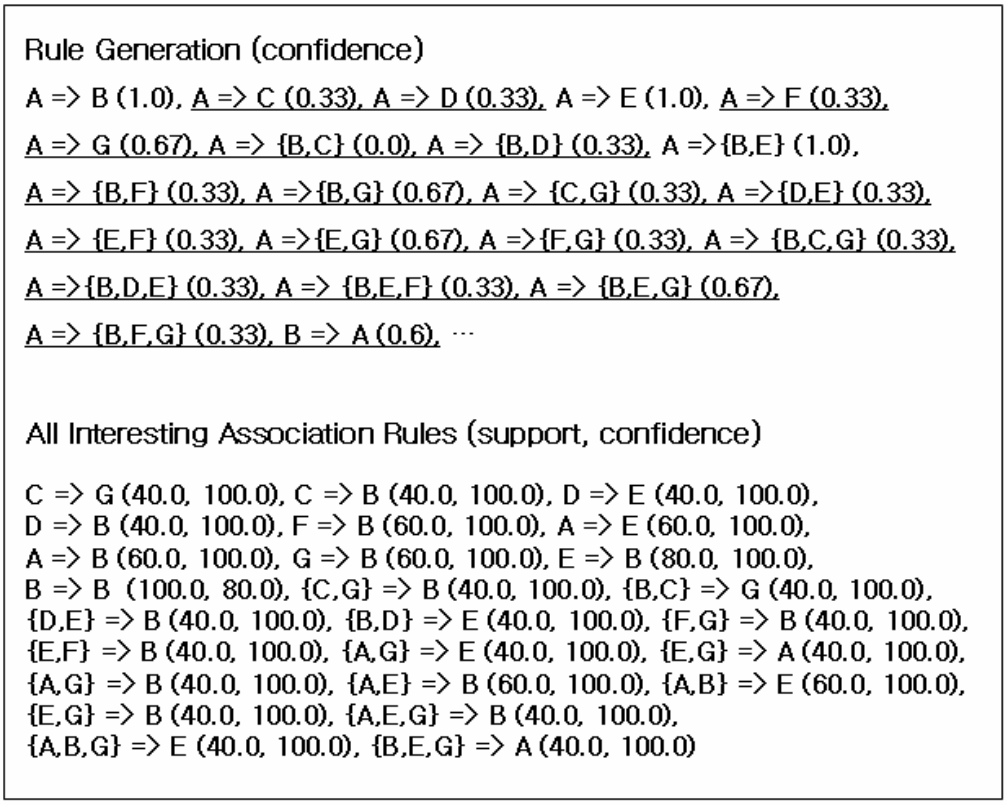


Fig 8 An example of finding interesting association rules. After the rule generation step with all frequent itemset of sample DB (Fig 7), we calculate the confidence of each candidate association rule. From frequent itemsets in Fig 7, we find interesting association rules with 80% of minimum confidence. Underlined rules are dropped in choosing interesting association rules.

2.1.3 Concepts in ARM

To find frequent itemsets efficiently, several types of data structures are suggested. The hash tree is the most popular concept in ARM. Many algorithms of association rule mining applied a hash table and a hash tree. In mining association rules from the sample database in Fig 7, a

hash tree (bucket size = 3) is created in every step from the sample database (Fig 7).

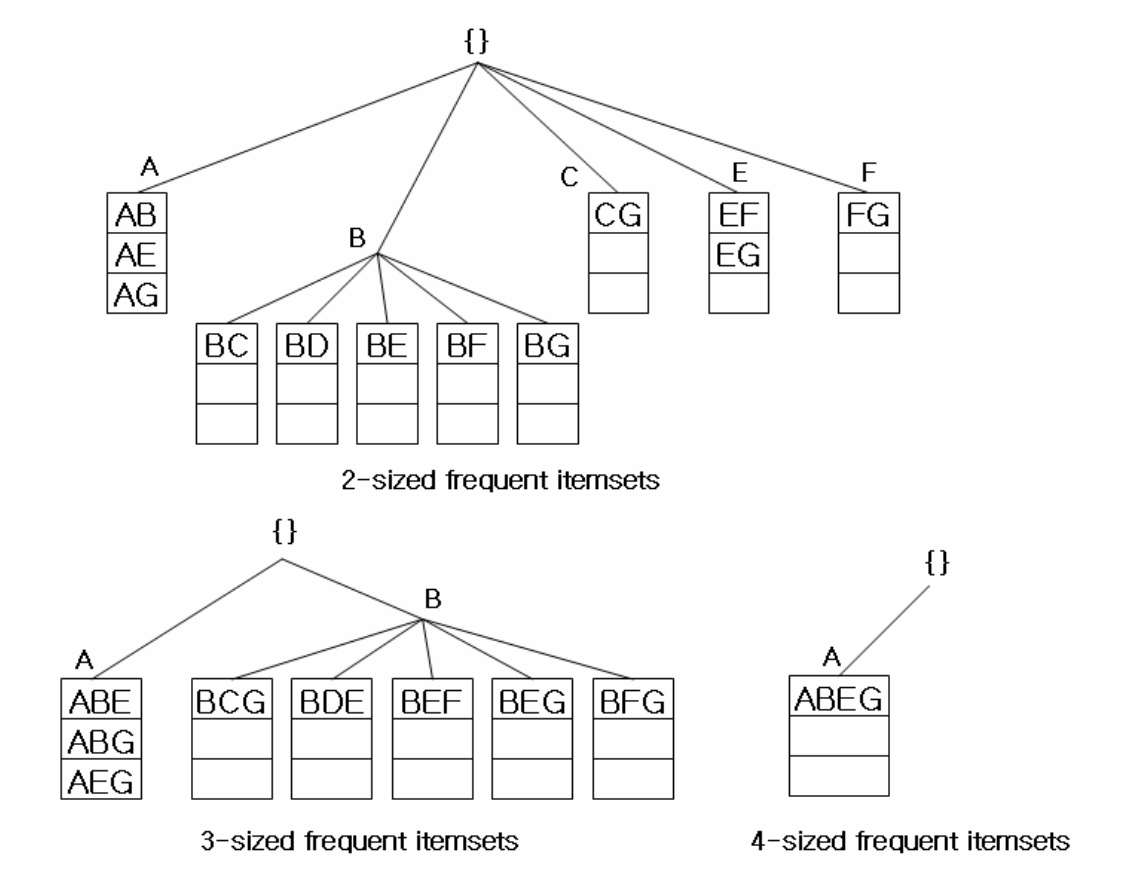


Fig 9 A sample database and hash trees for 2-sized (a), 3-sized (b), and 4-sized (c) candidate itemsets

Lattice as another concept for ARM is created once during a step of finding frequent itemsets. It does not contain unnecessary nodes compared with the Apriori algorithm. The lattice-based algorithms are PARTITION (Savasere, Omiecinski et al. 1995), Eclat-based (Zaki, Parthasarathy et al. 1997), DIC (Brin, Motwani et al. 1997), CHARM (Zaki and Hsiao 2005). Fig 10 is the lattice from the sample database shown in Fig 7.

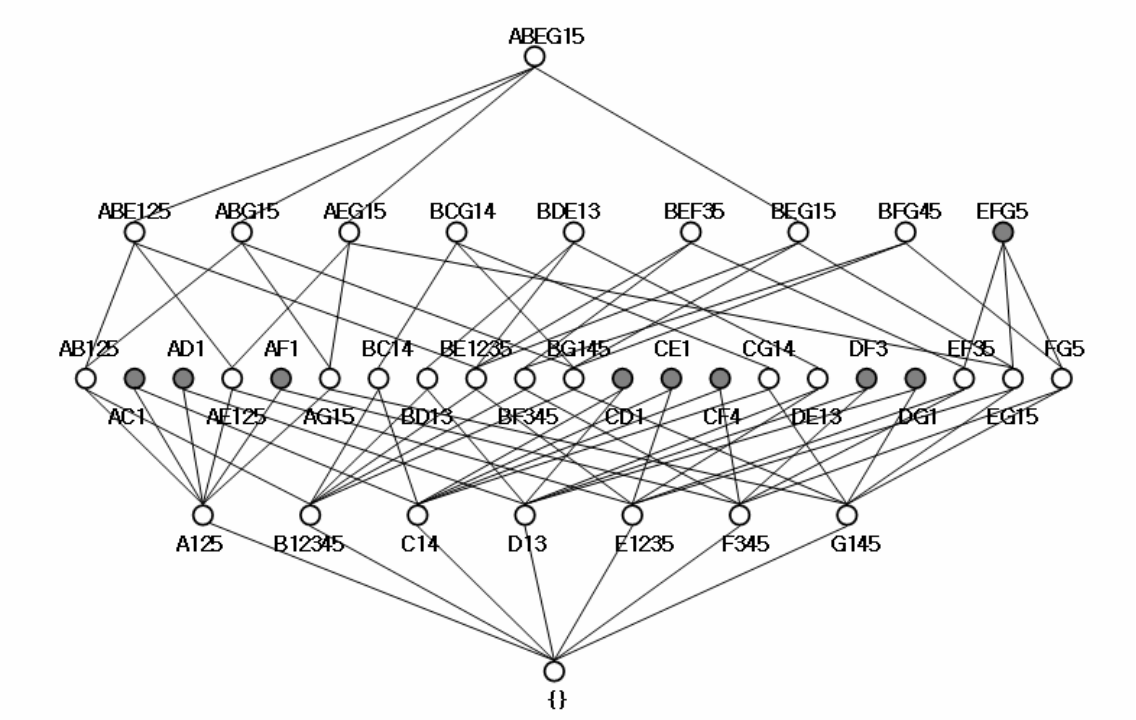


Fig 10 Lattice of a sample database. Grey nodes are not frequent.

FP-tree (frequent pattern tree) (Han, Pei et al. 2000) is an extended prefix tree structure for storing compressed, crucial information about frequent patterns. FP-growth is an efficient FP-tree-based mining method for mining the complete set of frequent patterns by pattern fragment growth. FP-tree scans a database only twice like the CHARM algorithm (Zaki and Hsiao 2005). It, however, does not produce candidate itemsets or select frequent itemsets from the database itself. Thus, FP-tree saves the space for the candidate itemsets. Fig 11 shows an FP-tree from the sample database shown in Fig 7.

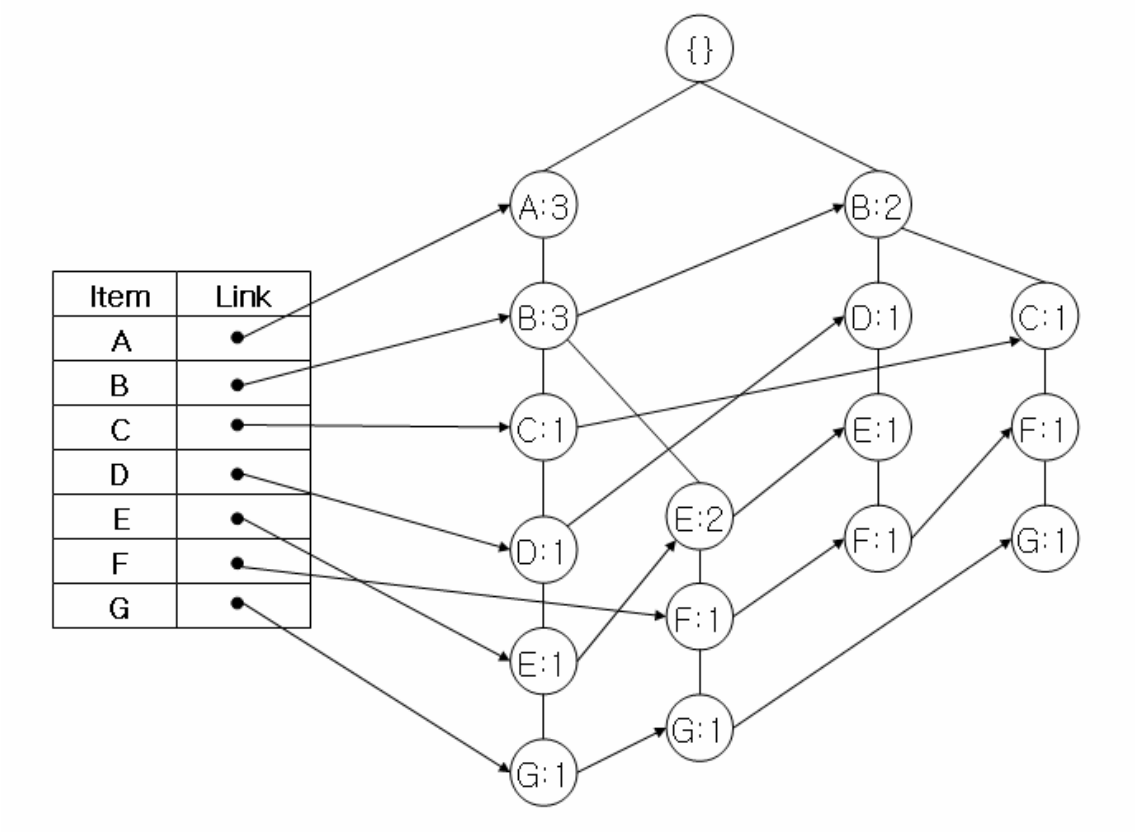


Fig 11 FP-tree with a sample database

Table 1 presents construction time and supporting counts time of concepts discussed previously. n is the number of 1-sized frequent items and $|D|$ is the number of transactions in database. Table 1 shows that each concept depends on the characteristics of data including n and $|D|$. When $|D|$ is large and sparse, we expect that support counting in lattice is faster than the one in FP-tree because support counting in lattice is done by intersecting of tid-lists (vertical formatted database). However, when n is large, the intersection in lattice is proportional to n^2 and FP-tree shows better performance.

Table 2 Concept Comparison (n = 1-sized frequent item).

Concept	Construction time	Support counting time
Hash Tree	$O(n^2)$	$O(n^2 \mid D \mid)$
Lattice	$O(n^2)$	
FP-Tree	$O(n)$	$O(n \times \mid D \mid)$

2.2 Quantitative Association Rule Mining

In many practical problems, there are not only categorical, but also quantitative attributes that are measured on a numerical scale. Since quantitative attributes in general cannot be treated as categorical ones, it is necessary to define quantitative association rules and corresponding rule mining algorithms. Several recent papers have addressed this problem (Aumann and Lindell 1999; Fukuda 1999; Brin, Rastogi et al. 2003). To solve the problem and to find association rules from quantitative attributed databases many methods have been developed. We can divide them into groups by their basic ideas.

2.2.1 Discretization-Based Methods

One of the most popular approaches is based on discretization (also called binning) of quantitative attributes. An example of a rule according to a discretization-based method is X

$\in$[20, 30] => Y$\in$[5, 10]. The discretization-based approach has an additional preprocessing step, discretization before performing mining work. Many methods have been developed based on either equi-depth or equi-width (Wang, Hock et al.; Ramakrishnan and Rakesh 1996; Lent, Swami et al. 1997; Miller and Yang 1997). The bin that original values are assigned in is too sensitive to the bin size. To reduce the influence of bin definition, distance-based methods such as clustering and interval merging have been suggested (Wang, Yang et al.; Yager 1995; Ester, Kriege et al. 1996; Guha, Rastogi et al. 1998; Chun-Hung, Ada Waichee et al. 1999). However, this approach is sensitive to outliers. Table 3 shows examples of discretization with these three methods.

Table 3 Example of discretization of quantitative attributes.

value	Equi-depth (depth = 3)	Equi-width (width = 10)	Distance-based
3	[3, 7]	[1, 10]	[3, 7]
4	[3, 7]	[1, 10]	[3, 7]
7	[3, 7]	[1, 10]	[3, 7]
19	[19, 21]	[11, 20]	[19, 33]
20	[19, 21]	[11, 20]	[19, 33]
21	[19, 21]	[21, 30]	[19, 33]
22	[22, 33]	[31, 40]	[19, 33]
24	[22, 33]	[31, 40]	[19, 33]
33	[22, 33]	[31, 40]	[19, 33]

2.2.2 Distribution-Based Methods

An alternative approach that overcomes the challenge to choose the "correct" bin size was proposed by Aumann and Lindell (Aumann and Lindell 1999). Their method directly considers the distribution of continuous data via standard statistical measures, such as the mean and the variance. A quantitative association rule is an association between a subset of a database (left-hand side of a rule) and its extraordinary behavior (right-hand side of rule). An example of a quantitative rule is $\{A, B\}$ => {mean (X) = 68.7}, where A and B are categorical items and X is a quantitative attribute. This rule is interesting if it reveals that a group containing A and B shows a significantly different average of X from the rest of the data. Webb (Webb 2001) extended the measures to standard deviation, minimum, count etc.

2.2.3 Optimization-Based Methods

Fukuda and his colleagues defined a new optimization parameter that is called *Gain* to get a trade-off between support and confidence. Based on their work, several extended ideas have been suggested (Fukuda 1999). Although this optimization-based approach produces optimized association rules from the image segmentation technique, it has a limitation of one or two numeric attributes (Salleb-Aouissi, Vrain et al. 2007). Another optimization-based algorithm, GAR (Mata 2002) is performed by a genetic algorithm to optimize the support of itemsets. QuantMiner (Salleb-Aouissi, Vrain et al. 2007) also used genetic algorithm to define intervals and optimized support and confidence thresholds. Ruchkert et al. used linear

inequation on the left and right hand sides of an association rule (Ruckert, Richter et al. 2004).

In this dissertation, we suggest motif association rule mining applying discretization-based and distribution-based quantitative association rule mining approaches.

Chapter 3

Association Rule Mining – based Motif Association Rules

Alternative splicing (AS) is a major mechanism to generate protein diversity. A single gene might generate hundreds or even thousands of different proteins. AS plays an important role in cell proliferation, differentiation and death. Human diseases caused by AS have been shown in many studies, examples are familial isolated growth hormone deficiency type II, Frasier syndrome, and myotonic dystrophy (see (Faustino and Cooper 2003; Garcia-Blanco, Baraniak et al. 2004) for a detailed review).

It is assumed that splice sites, exonic splicing enhancers and silencers, intronic splicing enhancers and silencers, and gene-specific splicing regulators contribute to the regulation of splicing during development or in different tissues (Grabowski 2002). SR proteins are required for constitutive pre-mRNA splicing, and often regulate alternative splice-site selection. They have a modular structure that consists of one or two RNA-recognition motifs (RRMs) and a C-terminal motif which is rich in arginine and serine residues (RS domain). Their activity in alternative splicing is antagonized by members of the hnRNP A/B protein family (Caceres and Kornblihtt 2002). Several splicing factor binding sites (*cis*-regulatory elements) which influence the amount and type of alternative splicing have been identified (Yeo, Holste et al. 2004). Often, these sequence motifs can be found in close proximity to the corresponding splicing sites (Akerman and Mandel-Gutfreund 2006).

Considerable effort has been made to discover regulatory elements in experimental and computational analyses (Famulok and Szostak 1993; Fairbrother, Yeh et al. 2002; Zhang, Leslie et al. 2005). Several studies investigate *k*-mer frequencies (k usually ranges from 5-10) in spliced sequences and compare them against a control (Brudno, Gelfand et al. 2001; Fairbrother, Yeh et al. 2002; Yeo, Holste et al. 2004). Another study uses a support vector

machine to identify regulatory elements (Zhang, Leslie et al. 2005). All of these studies focus on finding single motifs, and do not investigate systematically the co-occurrence of motif combinations.

Recently, powerful large-scale AS profiling microarrays have been developed, but computational methods which investigate the regulation of AS are still lagging behind. Researchers have focused on finding *cis*-regulatory motifs in pre-mRNA sequences. However, most studies are searching for single motifs, while many splicing events seem to be regulated by a combination of splicing factors.

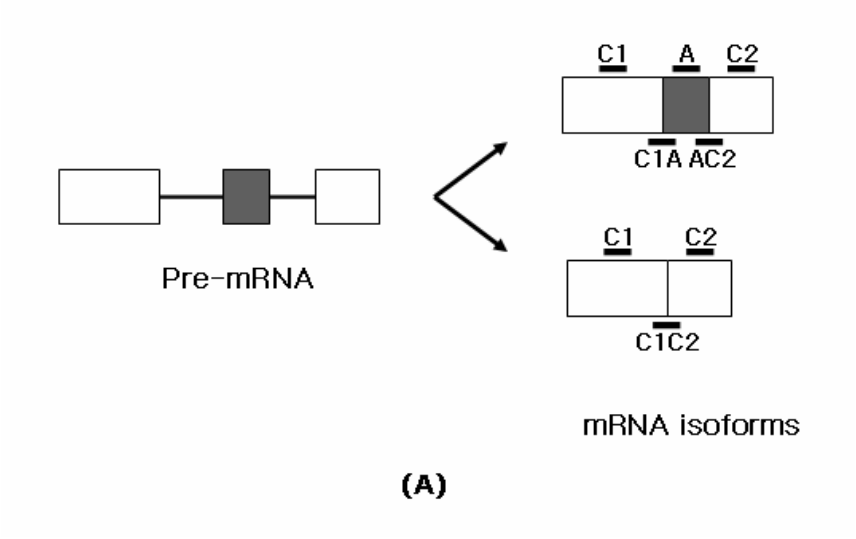
In this chapter, we use association rule mining to discover *cis*-regulatory motifs that are responsible for distinct alternative splicing patterns in 10 mouse tissues. Our approach generates motifs and motif association rules in different alternative splicing pattern groups in mouse. We search for exonic and intronic regulatory elements and their association rules in the exon/intron sequences flanking an exon skipping event. The inferred association rules indicate that alternative splicing pattern in different tissues might be explained by different motif combinations. Many of our discovered *cis*-regulatory motif candidates coincide with known splicing factor binding sites.

3.1 Datasets

Pan and colleagues (Pan, Shai et al. 2004) measured AS patterns of mouse genes using a custom splice array. They published exon skipping rates for 3126 alternatively spliced exons from 2647 genes in 10 tissues. To estimate the relative exon skipping rate, they developed a

generative model for the AS Array Platform (GenASAP, (Shai, Morris et al. 2006)). The exon skipping rate is defined as the ratio of the expression value of the transcript isoform skipping the cassette exon divided by the total expression of both isoforms. For example, Fig 12 shows exon skipping rates of the *BG046833* gene in 10 tissues.

We retrieved 3126 whole-length transcripts from NCBI using GeneBank (Benson, Karsch-Mizrachi et al. 2006) identifiers provided by Pan and colleagues. We trimmed polyA tails using the trimest program from the EMBOSS package (Rice, Longden et al. 2000) and mapped the transcripts onto the mouse genome (Build 36 v.1 released in May 2006) via BLAT (Kent 2002). Only transcripts that aligned with more than 95% identity over whole transcripts were used for our experiments. Each transcript may contain more than one partial match of the sequence (called blocks in BLAT), indicating potential exons. Blocks separated by less than 5bp were merged. We compared the original cassette exons that Pan's group provided and their neighboring constitutive exons with the corresponding set of blocks. Only genes where the exon borders differed by less than 5bp from the corresponding block borders were used for our study, resulting in a total of 2565 alternatively spliced pre-mRNA sequences.



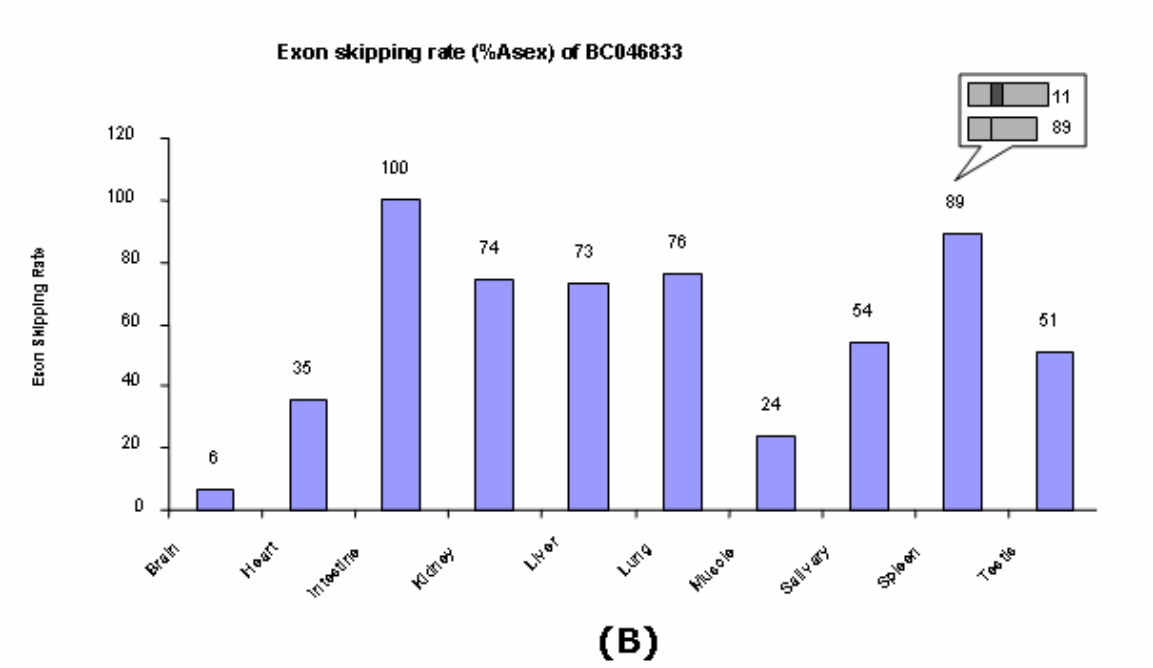


Fig 12 (A) Probe design of Pan's quantitative microarray platform. The dark rectangle represents an alternatively spliced exon; grey rectangles correspond to constitutive up and downstream exons. Six probes (C1, A, C2, C1-A, A-C2, C1-C2) are chosen from exons, introns, and splice junctions. (B) Exon skipping rates of the *BG046833* gene in 10

mouse tissues. The value of 89 in spleen is calculated by the ratio of mRNA without a cassette exon/the ratio of total mRNA.

3.2 Clustering Alternative Spliced Genes

To find motif candidates for different AS patterns we clustered the gene set based on their exon skipping profile. We used the Pearson correlation and Euclidean distance and applied three different clustering methods (complete linkage, average linkage, and Ward). Our clustering resulted in 50, 70, 100 clusters with each method. Clusters range from 5 to 174 genes. Fig 13 shows an example cluster from complete linkage method. The genes in this cluster show a sharp exon skipping rate peak in salivary tissue and a small variance and small exon skipping rates in other tissues. All clusters of genes are stored in http://statgen.ncsu.edu/~jihye/ASCluster.html

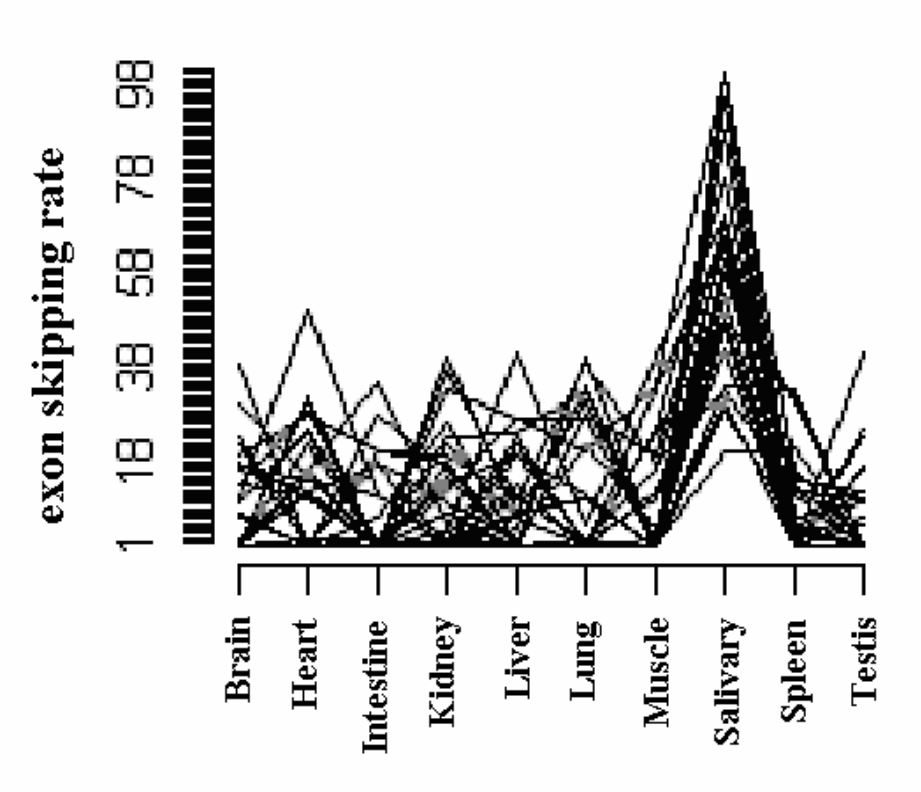
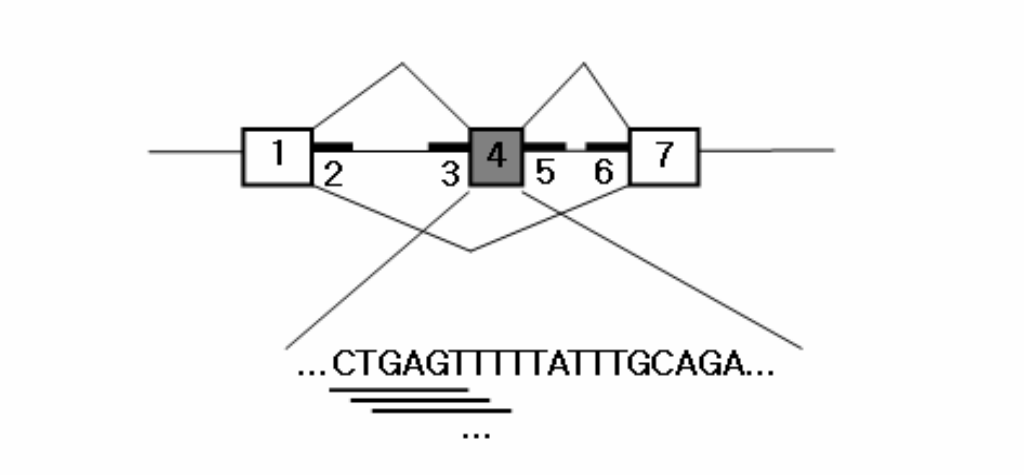


Fig 13 Exon skipping profiles of the genes contained in a cluster from complete linkage method based on Euclidean distance.

3.3 Algorithm

We define seven regions around each alternatively spliced exon. Since it is assumed that the majority of *cis*-regulatory elements involved in splicing are found close to splice sites (Cooper 2001; Akerman and Mandel-Gutfreund 2006; An and Grabowski 2007), we restrict our analysis to 200 base pairs flanking the splice sites. Each gene corresponds to a transaction. We count the occurrence of all possible hexamers and use them as items. We consider hexamers from different regions of a gene as different items (see Fig. 14), and

combine them into a transaction.



Gene ID	...	4_CTGAGT	4_CTGAGG	4_TGAGTT	...
1	...	1	0	1	...
2	...	0	0	1	...
...	...	...	...	...	...

Fig 14 For each alternatively spliced exon (grey box) we define seven regions (1-7) in the corresponding genomic sequence. The hexamer composition of each region is analyzed separately, and the corresponding hexamer counts are stored in an occurrence table.

Our goal is to apply association rule mining to find sequence motifs rules associated with different exon skipping rate patterns. We are searching for interesting rules of the form "hexamer set => hexamer set", where the hexamers derived from seven exon/intron regions are treated as categorical attributes. A rule indicates that genes that include a specific set of hexamers are likely to show a similar pattern of exon skipping rates in 10 tissues.

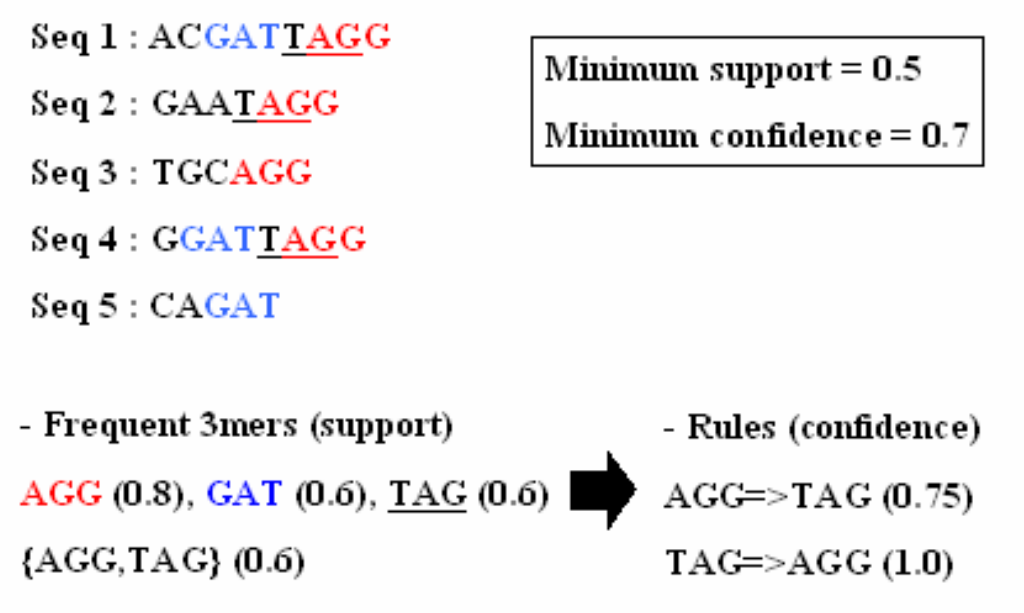


Fig 15 A simple example of finding association motif rules. From five sequences, we have four frequent 3-mer sets with 0.5 minimum support threshold. From these 4 frequent 3-mer sets, we finally extract two rules satisfying 0.7 minimum confidence threshold.

Fig 15 shows how we apply ARM to find frequent hexamers and association rules. In this figure, we only show one of the seven investigated splice site flanking regions, however we used all possible hexamers in all 7 regions as independent items. The ”Eclat” and “apriori” algorithms in the R package “arules” are used to extract frequent itemsets and association rules (Hahsler, Grün et al. 2007).

3.3.1 Overlap Handling

Since we slide a hexamer window over sequences, hexamers, as items, might overlap. We say that a rule A=>B shows overlap iff A and B overlap in a gene sequence. We discard

hexamers with insufficient support during the ARM process, and we perform an additional pass through the extracted rules in order to select the rules with non-overlapping motifs.

In the example of Fig 16, although the rule TAG=>AGG satisfies both minimum support and minimum confidence, it shows overlap since the suffix of TAG, AG is exactly the same as the prefix of AGG in sequence 1 to 4. To avoid overlapping association rules, the maximum distance of two motifs in gene sequences should be considered, in case we have repeats of motifs.

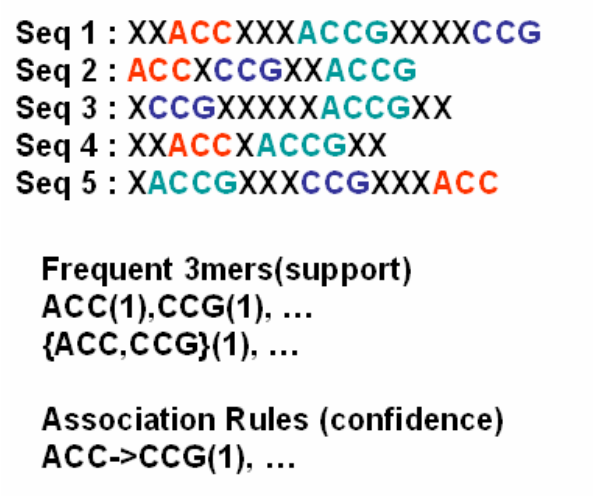


Fig 16 An example of motif repeats in gene sequences. Since two motifs ACC and CCG are repeated in gene sequences, we still have a rule ACC→CCG although they overlap.

3.3.2 Significance of Motif Association Rules

To measure importance of a rule, a lift value (also called interest) is frequently used (McNicholas, Murphy et al. 2008). The lift value of a rule is defined as:

$$\begin{aligned} \text{Lift}(X \rightarrow Y) &= \text{confidence} / \text{expected confidence} \\ &= \text{confidence}(X \rightarrow Y) / (\text{support}(X) * \text{support}(Y) / \text{support}(X)) \\ &= \text{confidence}(X \rightarrow Y) / \text{support}(X) \end{aligned} \quad (2)$$

Generally, a lift value greater than 1 indicates that the antecedent and the consequent appear more often together than expected. This can be interpreted as the occurrence of the antecedent has a positive effect on the occurrence of the consequent. We find all interest motif association rules with a lift value of 2 or higher.

3.4 Results

We tried various minimum support thresholds to find frequent hexamer sets. From the highest 0.2 minimum support, we decreased the threshold by 0.05. To find association rules, we needed lower minimum supports. In total, we found 4 frequent hexamers (0.2 minimum support or 430 genes) and 1 association rule (0.05 minimum support and 0.5 minimum confidence). Table 4 shows the frequent motifs with different minimum support thresholds from all exon skipping genes in mouse.

Table 4 Frequent hexamer sets from all AS genes in mouse. Hexamer sets are merged when they have only one nucleotide difference. Also they are extended when they overlap at the sequence level. Numbers after motifs indicate the different gene regions (see Fig 14) the motifs are derived from. For example, a frequent motif TGAAGA and

GAAGAA are from the downstream exon. Four frequent 6-mers were combined to two 7-mers, TGAAGAA and TTTTCTT. With 0.15 minimum support (or 323 genes), 37 frequent 6-mers are found and combined to 11 longer motifs when they are overlapped in gene sequences.

Minimum support = 0.2	Minimum support = 0.15
TGAAGA(7),	CTGAAGAAGA(7),
GAAGAA(7),	CTGC{A/T}G (7),
TTTCTT(6),	CAGC{A/T}G(7),
TTTTCT(6)	CCTGGAGA(7),
	AAAGAAAA(7),
	AGAGAAG(7),
	AGGAAGA(7),
	GAGGAGA(7),
	TTTTT{C/G}TTT(6),
	TTTTTCTTTT(3),
	GT{A/G}AGT(2)

We found A/G rich hexamers in the frequent items derived from exon sequences, and A/T rich hexamers in frequent items derived from intron sequences. This is in good concordance with previous studies that show that the major exonic enhancers such as SR protein binding sites are often A/G rich sequences (Zheng, Huynen et al. 1998). A/G rich motifs in exons and A/T rich motifs in introns are also found by the RESCUE algorithm (Fairbrother, Yeh et al. 2002). We also found CACC-containing *cis*-regulatory motif candidates, as predicted from SELEX experiments (Famulok and Szostak 1993). Some G/C rich motifs from exons may be a sign of coding regions.

Table 5 and table 6 show association motif rules from all AS genes and 50 clusters with Ward's method and correlation, respectively.

Table 5 Association motif rules from all AS genes. All motifs in rules are from region 4 (alternatively skipped exon) in Fig 14.

Minimum support = 0.05, Minimum confidence = 0.5	Minimum support = 0.05, Minimum confidence = 0.4
AAAAAT→TGAAGA	AAAAAT→TGAAGA, AAAGGA→AGAAGA, GAAAAA→AAGAAG, CTGCCT→CTGGAG, AGGAAA→AAGAAG, AATAAA→AAGAAG

We also compared the support of rules inside and outside clusters using a χ^2-test, but we did not find a significant difference. In addition, we computed the lift values as a parameter to measure the importance of a rule. All motif association rules we found have a lift value greater than 2.

Table 6 Example of frequent hexamers and their rules from 50 clusters with Ward's method. We applied 0.2 minimum support and 1.0 minimum confidence. The numbers in parentheses indicate exon/intron regions in Fig 14. We show only hexamers and rules for the selected two clusters.

Cluster	Association Rules, minsupp, minconf, lift
3	{7_AGGAAG} => {6_TCTTTT} 0.226 0.875 3.01 {6_TCTTTT} => {7_AGGAAG} 0.226 0.778 3.014
30	{7_TTATCT} => {7_GAGAAA} 0.220 1.000 3.154 {7_GAGAAA} => {7_TTATCT} 0.220 0.692 3.154 {7_GAGAAA} => {7_CTTCAG} 0.220 0.692 2.580 {7_CTTCAG} => {7_GAGAAA} 0.220 0.818 2.580 {7_CTCACT} => {7_TCCTGT} 0.220 1.000 3.727 {7_TCCTGT} => {7_CTCACT} 0.220 0.818 3.727 {7_TGGCAC} => {7_CCTGCT} 0.220 1.000 4.100 {7_CCTGCT} => {7_TGGCAC} 0.220 0.900 4.100 {7_TGGCAC} => {7_TCCTGT} 0.220 1.000 3.727 {7_TCCTGT} => {7_TGGCAC} 0.220 0.818 3.727 {7_GATCTC} => {7_CCTGCT} 0.220 0.900 3.690 {7_CCTGCT} => {7_GATCTC} 0.220 0.900 3.690 {7_CTCACC} => {7_TCCTGT} 0.220 0.900 3.355 {7_TCCTGT} => {7_CTCACC} 0.220 0.818 3.355

We also computed frequent motif sets and their motif rules from gene clusters. Since the SR protein binding site motifs GAAGAA and TGAAGA are known to occur in about 20% of all splice sites, we chose a minimum support threshold of 0.2 for our clusters. In the case of small clusters, we also set a minimum number of occurrences for frequent motifs. Table 7

shows our results. We have 16 association rules from 14 clusters (among a total of 50 clusters) when we use Ward's method based on correlation information. Many of the rules in each cluster contain A/G rich motifs in exons and A/T and C/T rich motifs in introns.

We also searched for the AS pattern specific association rules. Given an association rule A=>B we counted the number of occurrences of A, B, A and B, and non A and non B in and outside a cluster. We used the Cochran-Mantel-Haenszel chi-square test (Agresti 2002) to test for homogeneity between inside cluster and outside cluster frequencies. All association rules found in clusters show highly significant p-values (less than 2.2e-16). Also, most rules inside a cluster show 20% higher confidence than outside the cluster. Therefore, we hypothesize that these association rules are cluster specific- or AS pattern specific association rules. Table 7 shows the cluster specific association rules that have at least a 20% higher confidence inside the cluster than compared to outside the cluster.

Table 7 Association rules from clusters based on the Pearson correlation coefficient with Ward's method. The rules satisfy 0.05 minimum support thresholds, and they are supported by at least 7 genes. Confidence difference = (confidence inside cluster – confidence outside cluster). The number in the parentheses indicates exon/intron regions in Fig 14.

Cluster	Non-overlapped association rules from clusters (number after motifs indicates region of sequence)	Confidence difference
2	AGCAGC (1) → GCAGCC (1)	0.38
3	TGAAGA (7) → GAAGAA (7), AGGAAG (7) → TCTTTT (6) TCTTTT (6) →AGGAAG (7)	0.25 0.60 0.70
6	TTCCTT (3) → TTTCCT (3)	0.26
13	TCCAAA (7) → CCAAAG (7)	0.30
16	TTTCTT (6) → TTTTCT (6)	0.25
21	AAGCAG (7) → GAAGCA (7)	0.27
30	GAGAAA (7) → TTATCT (7)	0.53
35	CTTTTC (3) → TTTTCT (3)	0.18
36	GGAAGA (7) → GAAGAA (7)	0.26
42	TTTTTA (3) →TTTTAT (3)	0.38
43	GCTCCA (7) → CTCCAG (7)	0.27
44	TATTTT (3) → ATTTTT (3)	0.31
47	CTGTTT (6) → TGTTTT (6)	0.29
49	GTGTTT (6) → TGTTTT (6)	0.30

3.5 Conclusion and Discussion

We applied association rule mining to discover co-occurrence of potential *cis*-regulatory motifs in alternatively spliced genes. We used the "eclat" and "apriori" algorithm to find frequent sequence motifs and their association rules in exonic/intronic sequences flanking an exon skipping event. To guarantee rules with non-overlapping motifs, we extracted only rules with more than a 6 bp distance between antecedent and consequent. However, significantly larger sequence motifs might still be misunderstood with independent motif combinations. We plan to address this problem together with a more flexible motif description in future work. Altogether, we found 37 and 2471 frequent hexamers in all AS genes and clusters, respectively. Among these frequent hexamers 1799 have been already described in the transcript regulatory motif section of the Alternative Exon DataBase (AEDB) (Stamm, Riethoven et al. 2006), and 672 are new candidates for splicing regulating sequence motifs. We also computed association rules in clusters of genes with similar AS profiles. All inferred associations show a highly significant AS pattern specificity, and a large difference between rule confidence inside and outside the cluster. They are very promising candidates for cooperation of *cis*-regulatory elements involved in the regulation of tissue- and condition-specific AS. We hypothesize that they could be used to predict the exon skipping profile of mouse genes that are not included in our data set.

Chapter 4

Discretization – based Motif Association Rules

In general association rule mining, items or attributes should be categorical such as shopping items from market basket data, so that our we read sequences by sliding window and counted frequency of every hexamers. To obtain frequent hexamer motifs in genes with similar AS patterns, we grouped genes with similar patterns of exon skipping rates by clustering techniques. However, gene groups are dependent on clustering techniques and can be changed by applying different clustering methods, therefore, different motif association rules may be produced.

In this chapter, to solve this problem, we use exon skipping rate value itself as items instead of clustering. We applied a discretization-based motif association rule mining method to find candidates of or part of *cis*-regulatory motifs that may influence tissue-specific exon skipping rate in mouse. Based on our observations we hypothesize that some *cis*-regulatory elements only affect AS in combination with other elements. Also, combinational motifs are close to the splice site while individual motifs are located at some distance from the splice site. From this observation, we expect that individual motifs have a stronger signal of binding, locating far from the splice site, and also attracting other splicing factor binding as well as spliceosome.

4.1 Algorithm

In this project, we also use mouse genes by Pan and colleagues (Pan, Shai et al. 2004) that are introduced in chapter 3. They are composed of 3126 alternatively spliced cassette exons, as well as accompanying measures of their exclusion levels in ten tissues. As in chapter 3, we

define seven regions around each alternatively spliced exon reported in Pan's data set (Pan, Shai et al. 2004), and counted the occurrence of all hexamers in these regions (see Fig. 17). Since it is assumed that the majority of *cis*-regulatory elements involved in splicing are found close to splice sites (Cooper 2001; Akerman and Mandel-Gutfreund 2006; An and Grabowski 2007), we restrict our analysis to 200 base pairs flanking the splice sites. We consider hexamers from different regions of a gene as different items, and combine them into a transaction.

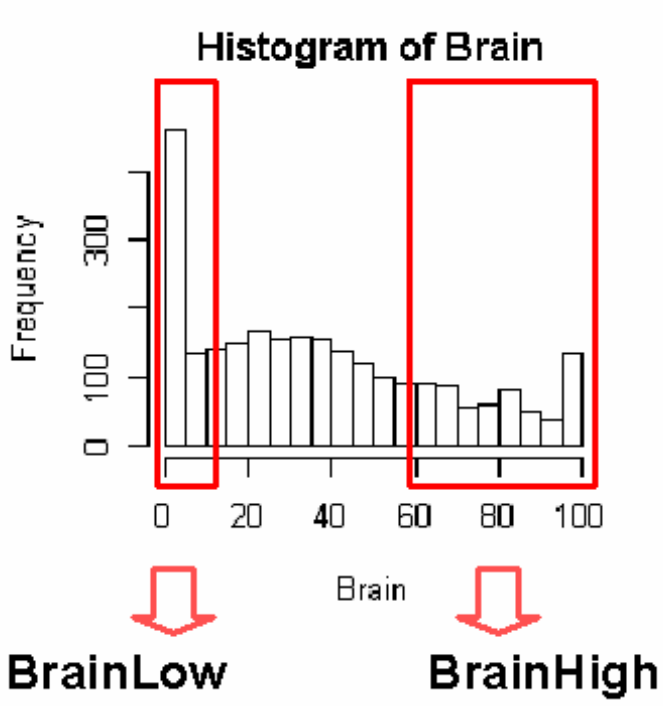
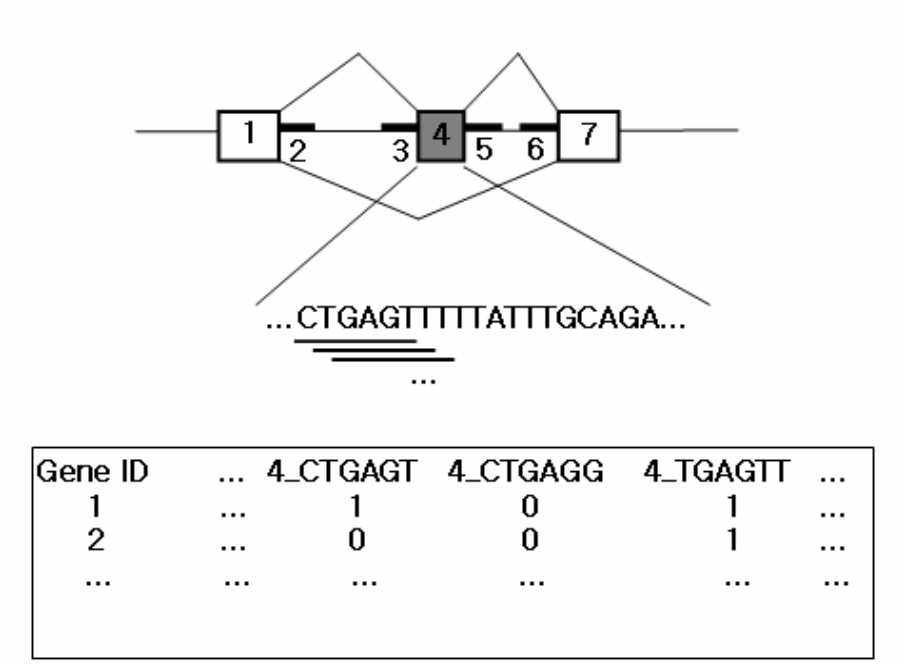


Fig 17 Discretization of quantitative exon skipping rates. We apply quartiles to convert numeric exon skipping rates to character items. *BrainLow* describes the first exon skipping rate quartile and *BrainHigh* describes the last exon skipping rate quartile in the brain.

We group the numeric %ASex values into four categories: *TissueHigh, TissueMedium1, TissueMedium2*, and *TissueLow* based on equi-depth bins (Han and Kamber 2000) that represent the exon skipping rate in the different tissues. In this study we only focus on rules involving the extreme skipping rates *TissueHigh*, and *TissueLow* (Fig 17).

Fig. 18 shows how we apply ARM to find frequent hexamers, frequent AS profiles, and their association rules. Although this figure shows only one of the seven investigated regions, our algorithm uses hexamers from each region as independent items. We use the "apriori" algorithm of the R package "arules" to extract frequent itemsets and association rules (Hahsler, Grün et al. 2007). To find sequence motif combinations, we set the maximum itemset length to two. To find clusters of sequence motifs associated with tissue-specific AS, we restrict the occurrence of hexamers to the antecedent, and the %ASex value intervals to the consequent.



Gene ID	...	4_CTGAGT	4_CTGAGG	4_TGAGTT	...
1	...	1	0	1	...
2	...	0	0	1	...
...	...	...	...	...	...

Fig 18 For each alternatively spliced exon (grey box) we define seven regions (1-7) in the corresponding genomic sequence. The hexamer composition of each region is analyzed separately, and the corresponding hexamer counts are stored in an occurrence table.

To find all frequent hexamers and their association rules Borgelt's C-version apriori program (Borgelt 2003) is used, which carries out a breadth first search on the lattice and uses a prefix tree to organize the counters for the itemsets.

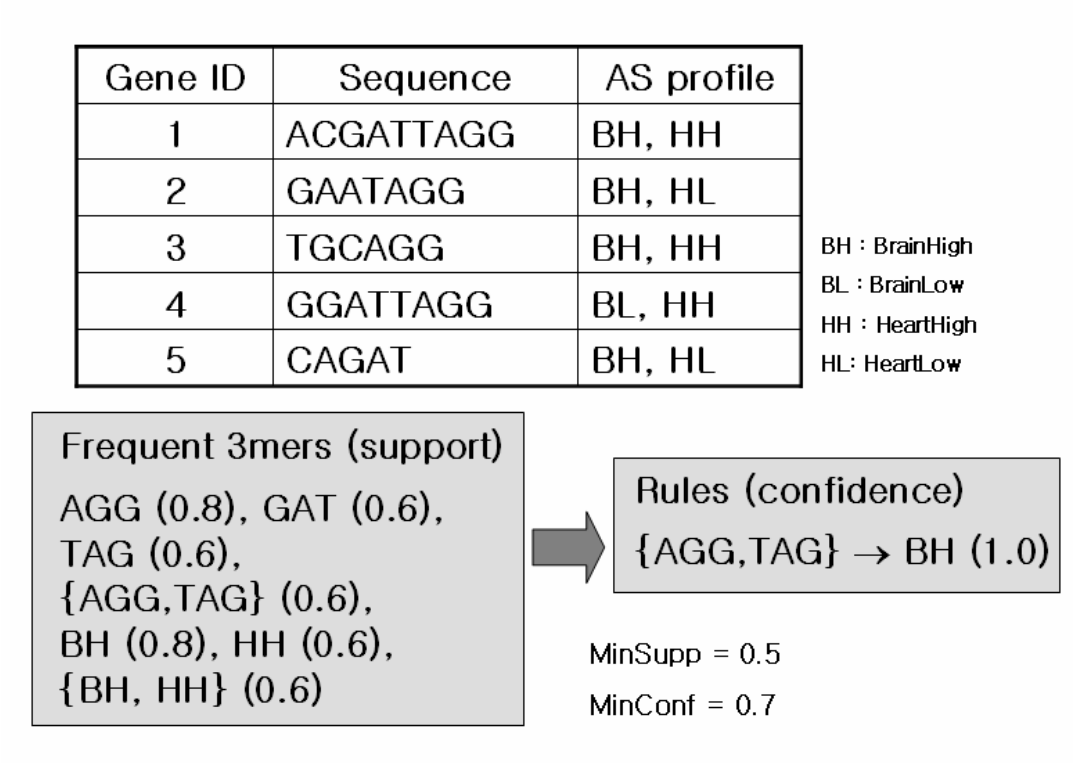


Fig 19 A simple example of finding motif association rules. From five sequences, we have four frequent 3-mer sets and three frequent AS profile sets with 0.5 minimum support threshold. From these 4 frequent 3-mer sets and 4 AS profile sets, we finally extract one rule satisfying 0.7 minimum confidence threshold. Association rule appearance is defined so that only an AS profile item can be located in consequent.

4.1.1 Significance of Motif Association Rules

To compute statistically significant association rules we use a chi-square analysis (Brin, Motwani et al. 1997). As suggested by Brin and colleagues (Brin, Motwani et al. 1997), we use the *lift* value to define the dependence between antecedent and consequent of an association rule.

Sergio Alvarez gave the following relationship between the chi-square statistic, and the values for support, confidence, and lift of a rule (see Alvarez 2003 and Appendix A for detail):

$$\chi^2 = n(lift - 1)^2 \frac{supp \cdot conf}{(conf - supp)(lift - conf)} \quad (1)$$

We use equation (1) to compute the relationship between support and confidence for fixed and lift. For α=0.05, n=2565 mouse genes and lift=1.2, we computed the maximized support (=0.032) and the corresponding minimum confidence (=0.195). To accommodate for multiple comparisons, we compute the p-value of each rule, and report significant rules after Bonferroni adjustment (Bonferroni 1936).

4.1.2 Overlap Handling

We noticed that the hexamer items of complex rules often overlap, and could be replaced by a single, longer sequence item. To identify such cases, we analyze the overlap and distance pattern of hexamer items involved in complex rules. If for an overlapping hexamer pair the thresholds for support and confidence are exceeded, we replace the hexamer pair by a single, larger sequence item and update the rule correspondingly.

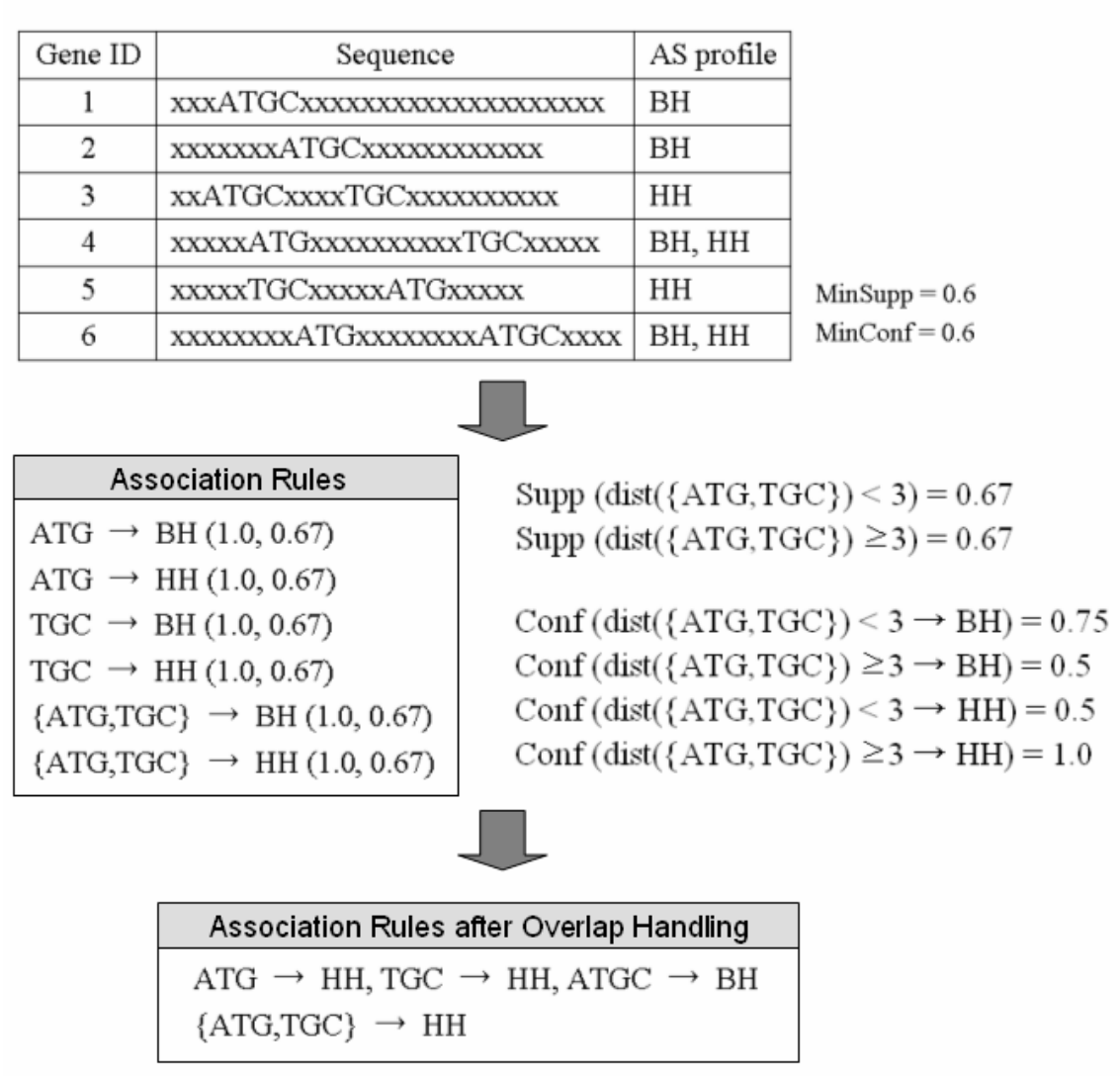


Fig 20 From the gene sequence and AS profile dataset, 6 association rules are generated by 0.6 minimum support and 0.6 minimum confidence. To check if a hexamer pair rule is a longer simple motif rule or a complex rule, supports of overlapping and non-overlapping hexamer pairs are computed. Both supports of hexamer pairs exceed the minimum support as well as minimum confidence. Overlapping hexamer pair rule {ATG, TGC} → BH, and non-overlapping hexamer pair rule {ATG, TGC} → HH exceed that threshold. Finally, we produce three simple rules and one complex association rule.

4.2 Results

We compute all association rules for minimum support = 0.032, minimum confidence = 0.195, and minimum lift = 1.2, and report all significant rules after Bonferroni correction. After overlap handling we obtained a total of 1260 single-hexamer association rules and 204 hexamer pair association rules. The entire set of association rules is available in Appendix A (also, http://statgen.ncsu.edu/~jihye/MotifARM.html).

Complex rules with two or more hexamers in the antecedent suggest a complex regulation of AS where multiple factors cooperate. We found 204 complex association rules, of which 117 rules contain hexamer pairs from different regions of the pre-mRNA sequence. For example, the rule {CCTGGG(2), TGTTTT(6)} → HeartHighQuan indicates that the occurrence of TGTTTT in the downstream intron of the cassette exon, and the occurrence of CCTGGG in the upstream intron seem to be associated with an increased exon skipping rate in the heart (Table 8).

Table 8 Examples of motif association rules. The number after the hexamer indicates a region on pre-mRNA sequence (Fig 18). P-values are Bonferroni adjusted.

Association rule	Supp.	Conf.	Lift	p-value
TTTCTC(6),TTCTTT(3)→BrainLowQuan	0.034	0.391	1.545	1.93e-4
TTTCTG(6)→HeartHighQuan	0.169	0.309	1.246	3.24e-9
TTTCTG(6),CTTTCT(3)→HeartHighQuan	0.033	0.417	1.678	5.31e-7
CCTGGG(2),TGTTTT(6)→HeartHighQuan	0.036	0.380	1.532	1.48e-4
GGTGGG(2),TTTCTT(3)→HeartHighQuan	0.034	0.384	1.545	1.70e-4
TTGTTT(5)→IntestineHighQuan	0.172,	0.304	1.220	3.53e-7
TTTTAT(5),TTGTTT(5)→IntestineHighQuan	0.044,	0.360	1.444	2.92e-3
GGTGGG(2),TTTTCT(3)→IntestineHighQuan	0.034	0.379	1.523	8.03e-4
CTTCCC(2),TTTTCT(6)→KidneyLowQuan	0.033	0.400	1.531	1.07e-3
TTTCCT(6),TGTTTT(3)→LiverLowQuan	0.039	0.434	1.688	1.22e-11
CCTGGG(2),TGTTTT(6)→LiverHighQuan	0.036	0.370	1.500	1.40e-3
GTAAGT(2),TTTTGT(6)→LungHighQuan	0.033	0.405	1.625	6.30e-7
TGTCTT(6)→MuscleHighQuan	0.159	0.292	1.204	1.02e-4
TTTGTT(3),TTCTTT(6)→MuscleLowQuan	0.039	0.374	1.491	7.65e-4
GTGAGT(2),TGTTTT(6)→SalivaryLowQuan	0.053	0.378	1.488	1.86e-6
CTTTTT(6),TTCTTT (3)→SalivaryLowQuan	0.036	0.391	1.542	8.29e-5
GTGAGT(5),TCTTTT(6)→SpleenHighQuan	0.033	0.393	1.587	1.29e-5
CTTCCC(2),TTTTCT(6)→TestisLowQuan	0.033	0.388	1.518	2.17e-3

4.2.1 Motif Combinations

For a given consequent Y there are three different types of antecedents.

- A complex antecedent {A, B}, where neither hexamer A nor hexamer B occurs in a simple rule. One might speculate that the motif pair {A, B} influences exon skipping only in combinatorial way. e.g., CCTGGG in region 2 and TGTTTT in region 6 alone do not appear in simple rules, but they appear together in a complex rule with consequent *HeartHighQuan* (Fig 21. (A)). We found 138 complex association rules with such a feature.
- A complex antecedent {A, B}, where hexamer A and/or hexamer B also occur in a simple rule. For example, TTTCTG in region 6 appears in simple rules and a combination with CTTTCT in region 3 also appears in a complex rule with consequent *HeartHighQuan* (Fig 21 (B)). We found 66 complex association rules with this feature.
- A simple antecedent {A}, where hexamer A occurs only in a simple rule. For example, TTGTTT in region 5 appears in a simple rule with consequent *IntestineHighQuan* but it is not shown any complex rule whose consequent is *IntestineHighQuan* (Fig 21 (C)). We found 1194 simple association rules with this feature.

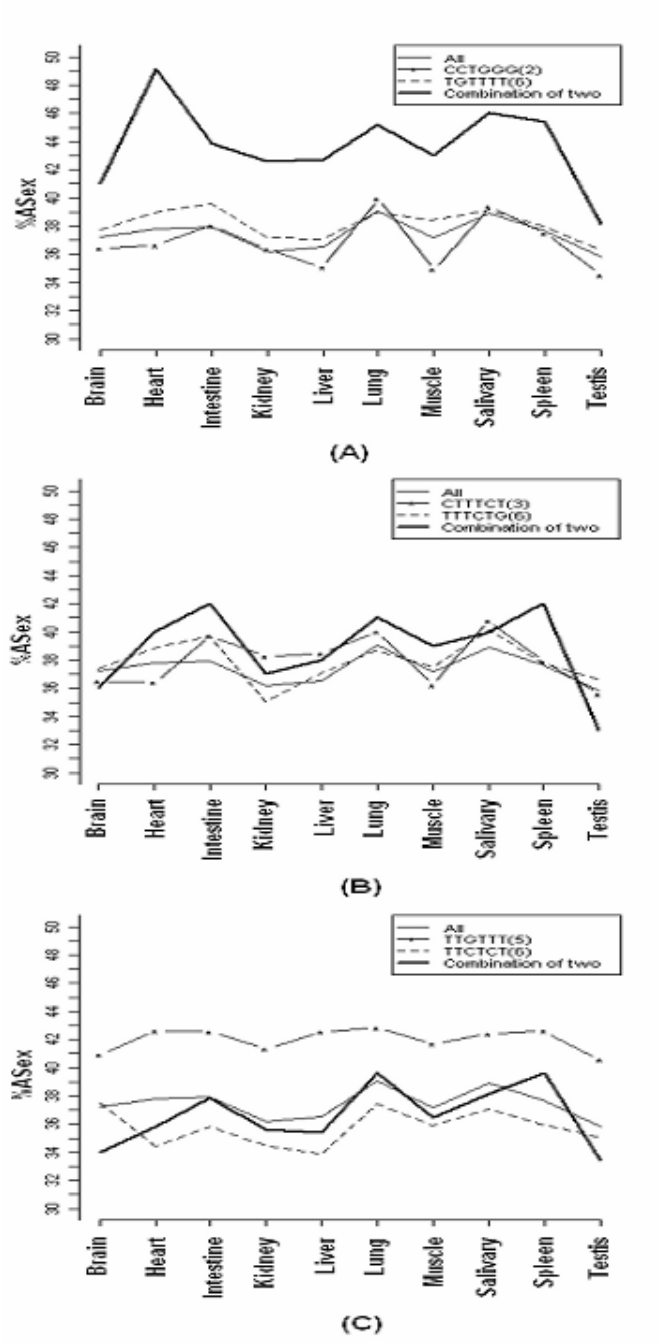


Fig 21 Three types of motif or motif combination effects on exon skipping. (A) Only genes with both hexamers, CCTGGG(2), and TGTTTT(6) show big different exon skipping value from all or genes with either hexamer. (B) A hexamer TTTCTG(6)

increases %ASex in the heart and cooperation with CTTTCT(3) increases more. (C) A hexamer TTGTTT(5) increases exon skipping in the intestine, however, cooperation with TTCTCT(6) does not show difference from the average %ASex even though they together are a frequent hexamer set.

4.2.2 Motif Position Distribution

To assess if there is a difference between simple and complex rules, we analyzed the distribution of hexamer occurrences with respect to the location of splice sites. For each hexamer involved in an association rule we counted the number of occurrences that are less/more than 100 bps away from the adjacent splice site. We also prepared a background set of 200 randomly selected hexamers not involved in any rule. While hexamers from simple rules do not significantly differ from the background (p-value = 0.564 from a t-test), hexamers from complex rules show a significant enrichment of occurrences close to the splice site (p-value <2.2e-16). On average, 64% of the hexamer occurrences from complex rules are less than 100 bps away from the adjacent splice site, versus 49.5% for the background set. Fig 22 shows distributions of hexamer positions for complex and simple association rules as well as the background.

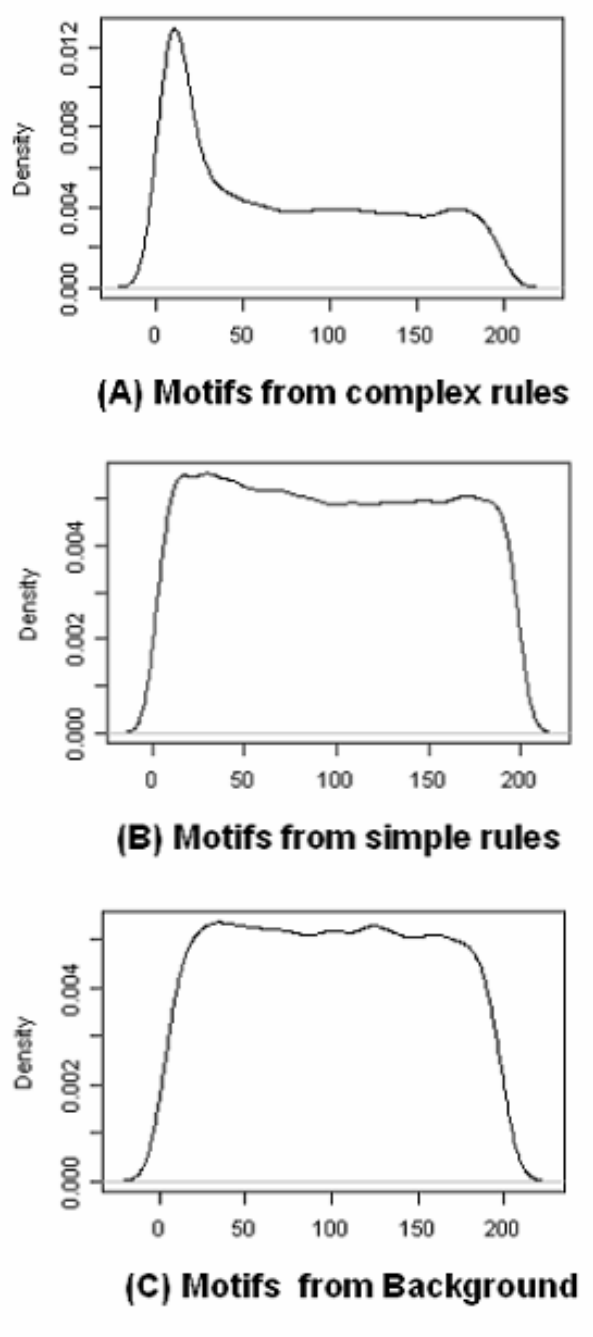


Fig 22 Distance distributions of (A) motifs from complex rules, (B) motifs from simple rules, and (C) random motifs. Motifs from complex rules are dense near the splice site while motifs from simple rules and random motifs are evenly distributed.

4.3 Conclusion and Discussion

We applied association rule mining to discover putative *cis*-regulatory motifs and motif pairs in alternatively spliced genes. We used the "apriori" algorithm to identify statistically significant association rules between frequent sequence motifs in exonic/intronic sequences flanking exon skipping events, and exon skipping levels. Association Rule Mining provides a convenient framework for the systematic investigation of sequence motifs involved in the regulation of AS. We found 1260 simple and 204 complex rules with statistically significant associations to tissue specific AS events. Among the complex rules, 117 rules contain hexamer pairs from different regions of the pre-mRNA sequence. Among the complex rules, 66 rules involve hexamers that also occur in simple rules, while 138 rules involve hexamer items not contained in any simple rule. An approach that targets only individual motif candidates would have overlooked these motifs. Surprisingly, we found a strong positional bias for sequence motifs involved in complex association rules, but not for motifs derived from simple rules. We hypothesize that different biological mechanisms might be involved in combinatorial regulation of AS.

We assessed the overlap of our predictions with known AS regulatory sequence motifs stored in AEDB (Stamm 2000). Among all hexamers involved in simple and complex rules, 42% of the hexamers located in exonic regions, and 63% of the hexamers located in intronic regions overlap with enhancer/silencer sequences from AEDB. This is significantly higher (p-value<=2.18e-13) than a similar value computed for randomly selected hexamers (19% for exonic regions, 18% for intronic regions). We hypothesize that our results

correspond to AS regulating factors. Our motif catalog provides a promising list of candidates for subsequent validation experiments.

Chapter 5

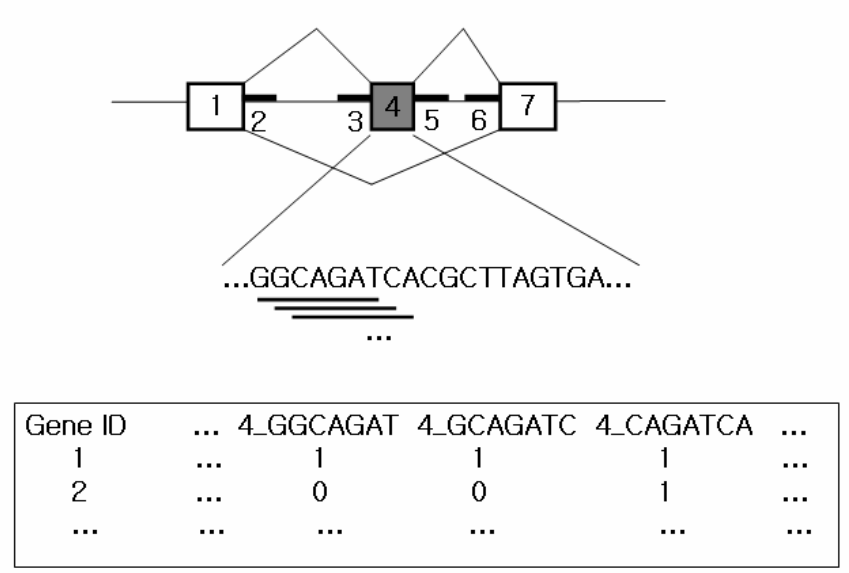
Distribution – based Motif Association Rules

In the previous chapter, we discretized numeric exon skipping rates to categorical items. Many interesting rules are discovered and supported by known and evaluated splicing factor binding sites. However, discretization-based rule mining is dependent on binning method and motif rules can be changed by applying different binning method and different bin size. Also, we met a problem of decision for proper thresholds to define interesting rules. To avoid these problems, in this chapter, we adapted distribution-based association rule mining idea and used numeric exon skipping rate itself instead of converting to categorical items. We found that motifs from cooperating motif pairs typically occur multiple times per gene, and that they are more conserved than motifs which act individually.

5.1 Algorithm

The goal of our study is to apply quantitative association rule mining to find sequence motifs associated with tissue-specific exon skipping rate changes. We searched for interesting rules of the form "a set of heptamer(s) => exon skipping rate", where a set of heptamer(s) from seven exon/intron regions are categorical attributes, and the exon skipping rate is a quantitative attribute. An "interesting" rule indicates that genes that include a specific set of heptamer(s) are likely to show an extraordinary exon skipping rate in one or several tissue(s) as compared to the remaining genes. After testing *k*-mers with k ranging from 5 to 9, we chose heptamers because of their superior performance, and because they are capable of detecting binding sites of splicing factors such as SR proteins (Fairbrother, Yeh et al. 2002; Yeo, Holste et al. 2004; Voelker and Berglund 2007).

We define seven regions around each alternatively spliced exon. Since it is assumed that the majority of *cis*-regulatory elements involved in splicing are found close to splice sites (Cooper 2001; Akerman and Mandel-Gutfreund 2006; An and Grabowski 2007), we restrict our analysis to 200 base pairs flanking the splice sites. Under this framework, each gene corresponds to a transaction. Each transaction contains as items, the counts for all occurrences of all possible heptamers in each of the 7 different gene regions. Fig 23 shows how the transaction database can be represented in tabular form. Each row corresponds to the transaction for a single gene. The generated table contains columns for each possible heptamer/region combination, for a total of 4^7 x 7 = 114,688 columns. Also included in the table are 10 additional columns containing the exon skipping rates for the various tissues. In Fig 23 (and throughout the text), the heptamers from a given region are fixed with the region number; for example, the heptamer GGCAGAT from region 4 is designated by 4_GGCAGAT.



Gene ID	...	4_GGCAGAT	4_GCAGATC	4_CAGATCA	...
1	...	1	1	1	...
2	...	0	0	1	...
...	...	...	...	...	...

Fig 23 For each alternatively spliced exon (grey box) we define seven regions (1-7) in the corresponding genomic sequence. The heptamer composition of each region is analyzed separately, and the corresponding heptamer counts are stored in an occurrence table.

To identify sequence motifs associated with changes in exon skipping rates, we used an adaptation of Aumann and Lindell's method. The algorithm for finding heptamer association rules follows three steps, outlined below:

1. Find all "frequent" heptamer sets, where a heptamer set is called frequent if its support is greater than a user-defined minimum support threshold.
2. For each frequent heptamer set and tissue type, compute the mean exon skipping rates for genes *having* the heptamer set, and genes *lacking* the heptamer set.

3. Identify and report “interesting” association rules using a t-test of the skipping rates computed in step 2. Association rules are considered interesting if the exon skipping rate is significantly different depending on whether the heptamer set on the left-hand side of the rule is found in the gene.

5.1.1 Data Structure

We computed frequent heptamer sets (which include both location and sequence information) based on an the Apriori algorithm (Agrawal and Srikant 1994). To efficiently compute frequent heptamer sets containing multiple heptamers, we used an itemset inclusion lattice, as described in (Zaki and Hsiao 2005). The lattice, G = (V, E), is composed of nodes of frequent heptamer sets with edges showing parent/children relationships (Fig. 24). An item superset cannot be frequent if any of its subsets is not frequent. For example, in Fig 24, a frequent heptamer set, {A, B} is frequent and all of its subsets, {A}, {B} are frequent. Also, {A, B, C} cannot be a frequent set because one of its subsets, {A, C} is not frequent.

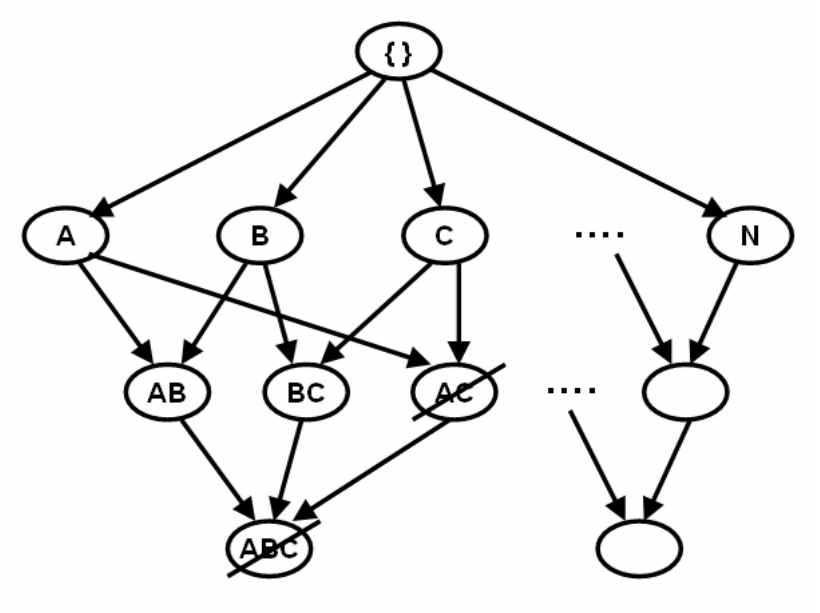


Fig 24 Lattice of frequent heptamer sets. Each node stores (for every tissue) the mean exon skipping rate of the genes containing the corresponding heptamer set. If {A, C} is not a frequent heptamer set, a superset {A, B, C} cannot be a frequent heptamer set either. We compare each node with the root node.

To identify interesting rules, we use a standard t-test. Let D denote the full gene set; let T_A denote the set of genes that include a given frequent heptamer set, A which occurs on the left-hand side of a rule for some tissue, t; and, let $D - T_A$ denote the remaining genes. Given the heptamer set, (Agrawal and Srikant 1994), we first compared $\mu_{T_A,t}$, the mean exon skipping rate in tissue t for genes having this heptamer with $\mu_{D-T_A,t}$, the mean exon skipping rate of genes lacking the heptamer. The rule was reported if the corresponding null hypothesis is rejected at an alpha level of 0.05, after applying Bonferroni's method to adjust for multiple testing.

5.1.2 Overlap Handling

We noticed that the heptamer items for rules often overlapped, and could be simplified by substitution with a single, longer sequence motif on the left-hand side of the rule. To identify such cases, we analyzed the overlap and distance patterns of heptamer items involved in complex rules. If an overlapping heptamer pair exceeded the support threshold we replaced the heptamer pair by a single, larger sequence item and updated the rule correspondingly. Fig 25 describes how motif overlapping is defined in association rules. In finding frequent 3-mer sets, two or more sized frequent sets are counted on two assumptions that frequent 3-mers are separated on sequence and that they are from one longer sequence.

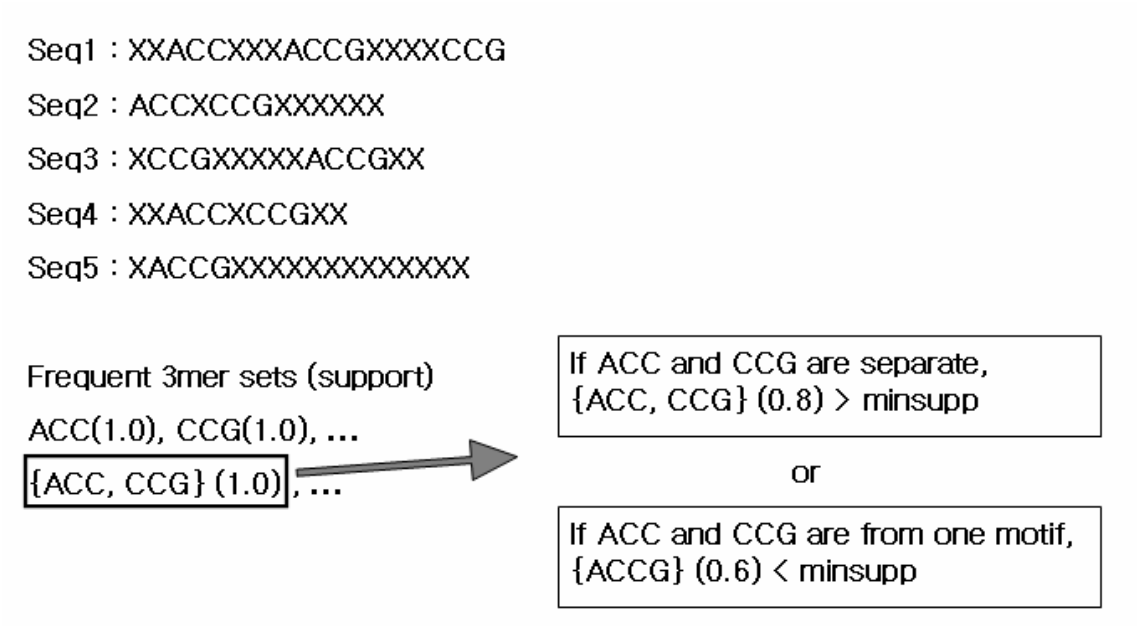


Fig 25 Overlap handling. We suppose that we want to find frequent 3-mers with 70% of minimum support. For a 2-sized frequent 3-mer set, {ACC, CCG}, we assume two cases

that they are separate on sequence and that they are from one motif. We count both cases and select the case that exceeds the minimum support.

5.2 Results

We computed all association rules for minimum support values ranging from 20 to 70 in steps of 5, corresponding to 2.72% to 0.77% of the whole dataset. Based on previous experience, we assumed that 20 genes is the smallest number to safely support sequence motifs as candidates for binding sites; we then increased the minimum support threshold and extracted the corresponding interesting rules until we could no longer find any interesting rules. In total, we mined 97 interesting rules, of which 3 contain multiple heptamers. There are 59 different heptamer sets and 71 individual frequent heptamers in the left hand sides of the rules. Table 9 shows heptamer association rules with absolute minimum support 20 genes in each tissue. The rules found for exon skipping rates in all tissues are listed in Table 15 of the Appendix B. All rules extracted are statistically significant after correcting for multiple testing.

Table 9 Heptamer association rules with 20 minimum support (0.77%). P-values are Bonferroni adjusted.

Tissue	Heptamer set , p-value, mean difference
Brain	3_TGACTAG , 0.026, -23.094 3_TTGGTTC 3_TGGTTCT , 0.009, -23.613 4_GCTGGAG , 0.001, -13.545

Table 9 Continued.

	4_TGCTGGA , 0.004, -16.373 4_TGCTGGA 4_GCTGGAG , 0.018, -19.440 4_TGGGCTG , 0.015, -19.357 6_TTTAAAA 3_TTATTTT , 0.004, -20.216 7_ACCTCAC , 0.018, -18.713
Heart	4_TGCTGGA , 0.010, -15.039 4_TGTGGAG , 0.003, -14.416
Intestine	4_TGCTGGA , 0.000, -19.731 4_GTGCTGG 4_TGCTGGA , 0.001,- 25.057 4_TGCTGGA 4_GCTGGAG , 0.001,-24.696 4_GCTGGAG , 0.024, -13.480 4_CTGCTGG 4_GCTGCTG , 0.003, -21.049 4_GCTGCTG , 0.032, -13.389 2_GAAGTCC , 0.042, -20.071 4_GACATCA , 0.035, -17.672
Kidney	7_TTGCTAA , 0.004, -19.492 4_TGCTGGA , 0.000, -19.273 4_TGCAGAA , 0.003, -15.417 6_AACAGGA , 0.005, -16.567 4_GAGAAGA 4_GGAGAAG , 0.003, -19.899 4_GACATTG , 0.022, -20.627 4_GGAGGTG , 0.002, -17.227
Liver	4_TGCTGGA , 0.000, -19.649 4_TGGGCTG 4_CTGGGCT , 0.027, -20.160 6_GGTCCAG , 0.004, -21.932 3_GACCTCT 3_TGACCTC , 0.003, -22.357

Tale 9 Continued.

	4_GTGCTGG 4_TGCTGGA , 0.026,- 21.740 2_TCACTCC , 0.029, -19.859 4_TGCTGGA 4_GCTGGAG , 0.007, 22.490 4_GCTGGAG , 0.000, 16.229 6_CTCCTTC 6_CCTCCTT , 0.003, -21.575 4_GAGAAGA 4_GGAGAAG , 0.025, -18.454 4_GACATTG , 0.008, -20.388 4_GGAGGTG , 0.002, -18.561 2_AGGCCTG 2_GGCCTGG , 0.000, -19.685
Lung	4_TGCTGGA , 0.000, -20.773 6_CCTAGTC , 0.002, -22.784 4_TGCTGGA 4_GCTGGAG , 0.000, -25.354 4_GCTGGAG , 0.000, -15.564 3_GAAGAGC 3_AAGAGCA , 0.001, -23.591 2_AGGCCTG 2_GGCCTGG , 0.010, -17.604 6_GTTTTTG 6_TTGTTTT 6_TTTGTTT , 0.024, -21.830
Muscle	2_GCCGGGC , 0.034, +15.141 4_GCTGGAG , 0.000, -13.921 2_GGAGCGG , 0.037, +18.662 4_TGTGGAG , 0.007, -14.381
Salivary	4_TGGGCTG , 0.004, -21.590 7_AGGGAGC , 0.021, +29.095 4_TGCTGGA , 0.001, -17.907 4_TGCTGGA 4_GCTGGAG , 0.003, -24.457 4_GCTGGAG , 0.015, -13.987 4_GGAGGTG , 0.002, -19.290

Table 9 Continued.

Spleen	4_TGCTGGA , 0.000, -18.461 3_AAAATAT 3_TTTGTTT , 0.002, -24.253 2_TTTCTCT 3_TTTCTCT , 0.023, +32.536 3_AAAATAT 3_TTTGTTT 3_TTGTTTT , 0.002, -24.879
Testis	6_ATAAAAT 6_TAAAATG , 0.021, -21.401 3_TTTTTCA 3_TTCATTT , 0.034, -22.472 4_TGCTGGA , 0.016, -15.240 3_ACCCACC 3_CACCCAC , 0.002, -25.001 3_TTGGTCT , 0.045,- 20.042

We performed a permutation experiment to estimate the number of rules obtained from a randomized data set. To do so, we shuffled gene sequences and exon skipping rates and then re-ran our algorithm. This procedure was repeated 100 times. Using the same minimum supports we found that the mean number of simple rules obtained from the randomized data sets was 14.7, compared to the 97 rules we found in the original database. Furthermore, we were unable to extract any complex rules using the randomized data sets.

In general, the number of reported rules decreased with increasing minimum support, but some rules were especially robust (Fig 26). Several heptamer sets in region 4 (cassette exon) are commonly found for a wide range of minimum support values. For example a rule with left hand side GCTGGAG was reported for all tested support values in association rules describing exon skipping in brain, intestine, kidney, liver, lung, muscle and salivary tissue. This heptamer overlaps with the 5' end of a potential SC35 binding site. It has been shown

that this binding site is crucial for the correct splicing of exon 5 of muscle-specific cardiac troponin T transcripts (Hodges, Cripps et al. 1999).

The complex rules we uncovered included two complex rules having heptamers from different regions. In brain tissue, the rule {6_TTTAAAA, 3_TTATTTT} => {meandiff(Brain) = -20.216} indicates that genes with both TTTAAAA in a downstream intron and TTATTTT in an upstream intron show, on average, a 20.216% lower exon skipping rate in brain compared to the exon skipping rate of brain of other genes (Fig 27 (A)). Interestingly, neither of these heptamers is included in a simple rule in any of the tissues. The other complex rule with two heptamers from different regions was found in the spleen: {2_TTTCTCT, 3_TTTCTCT} => {meandiff(Spleen) = 32.536}. This rule indicates that genes with two TTTCTCTs in the upstream intron show, on average, a 32.536% higher exon skipping rate in the spleen compared to the rest of the genes (Fig 27 (B)).

The third complex rule also occurred in spleen, and contained two heptamers from the same regulatory region: {3_AAAATAT, 3_TTTGTTT} => {meandiff(spleen) = -24.253}.

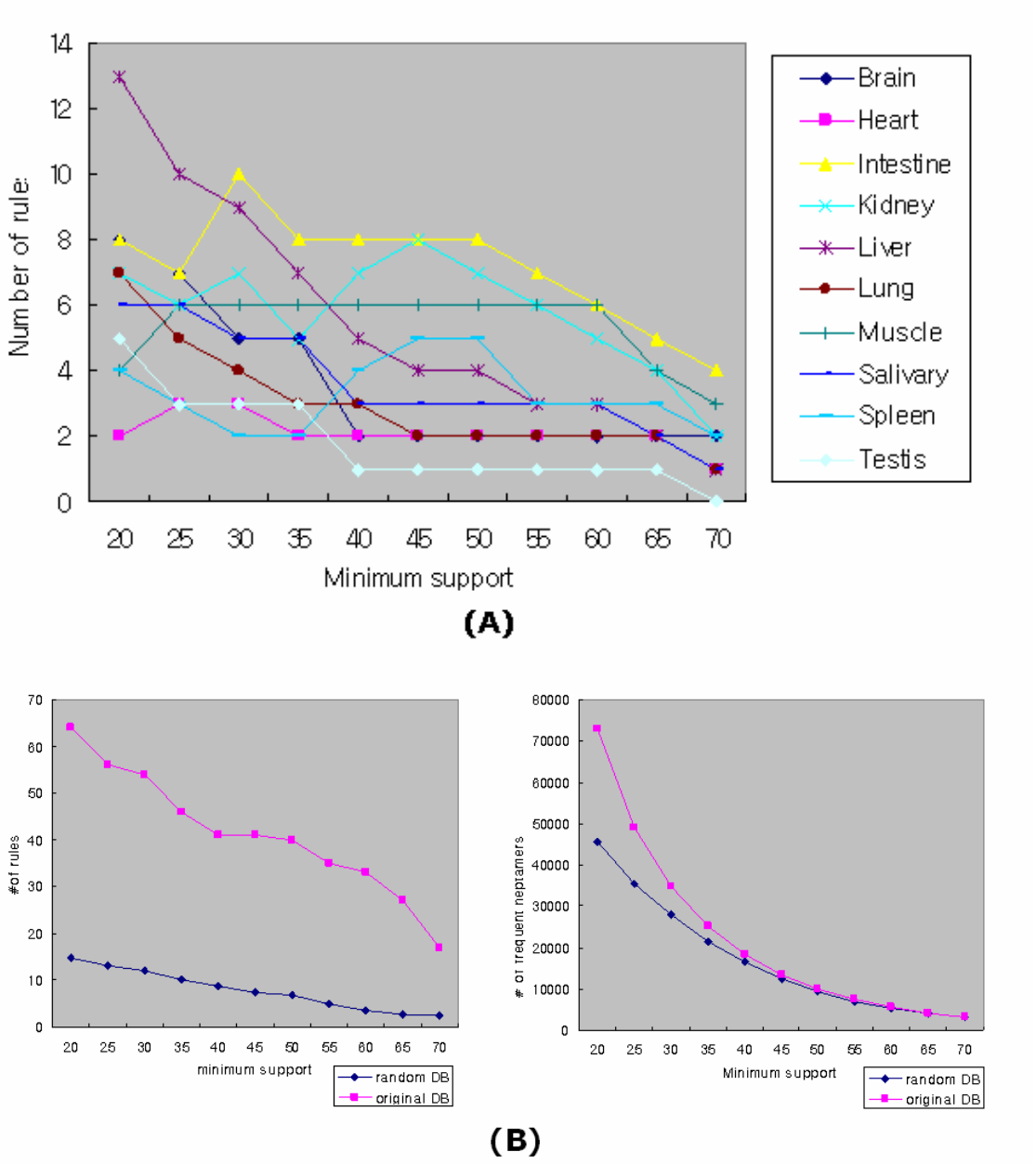


Fig 26 Number of Rules according to a minimum support threshold. (A) the number of rules reported by tissue decreases with increasing minimum support but exceptions exists. (B) As the minimum support decreases, the number of frequent heptamers increases exponentially while the number of rules increases linearly.

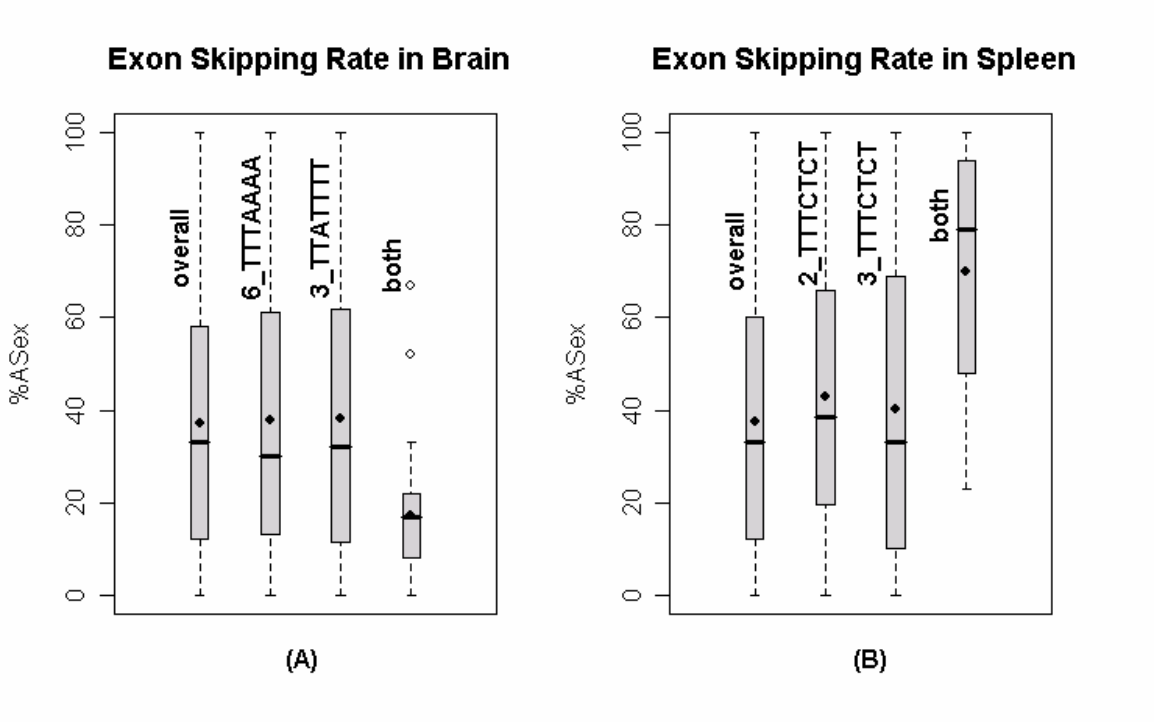


Fig 27 Exon skipping rates of complex rules. (A) {6_TTTAAAA, 3_TTATTTT} => {meandiff(Brain) = -20.216} (B) {2_TTTCTCT, 3_TTTCTCT} => {meandiff(Spleen) = 32.536}. Genes with only one heptamer do not show a statistically significant difference in the mean exon skipping rate while genes with both heptamers show statistically significant lower exon skipping rates in both cases.

5.2.1 Repeats of Motifs

The heptamers corresponding to complex rules were, on average, repeated higher multiplicity within their genes than heptamers from simple rules (Fig 28). In genes with two or more heptamer occurrences, heptamers from complex rules occurred in greater numbers than heptamers from simple rules regardless of whether the heptamers were from the same region (p-value of 0.067) or from all regions (p-value of 0.009).

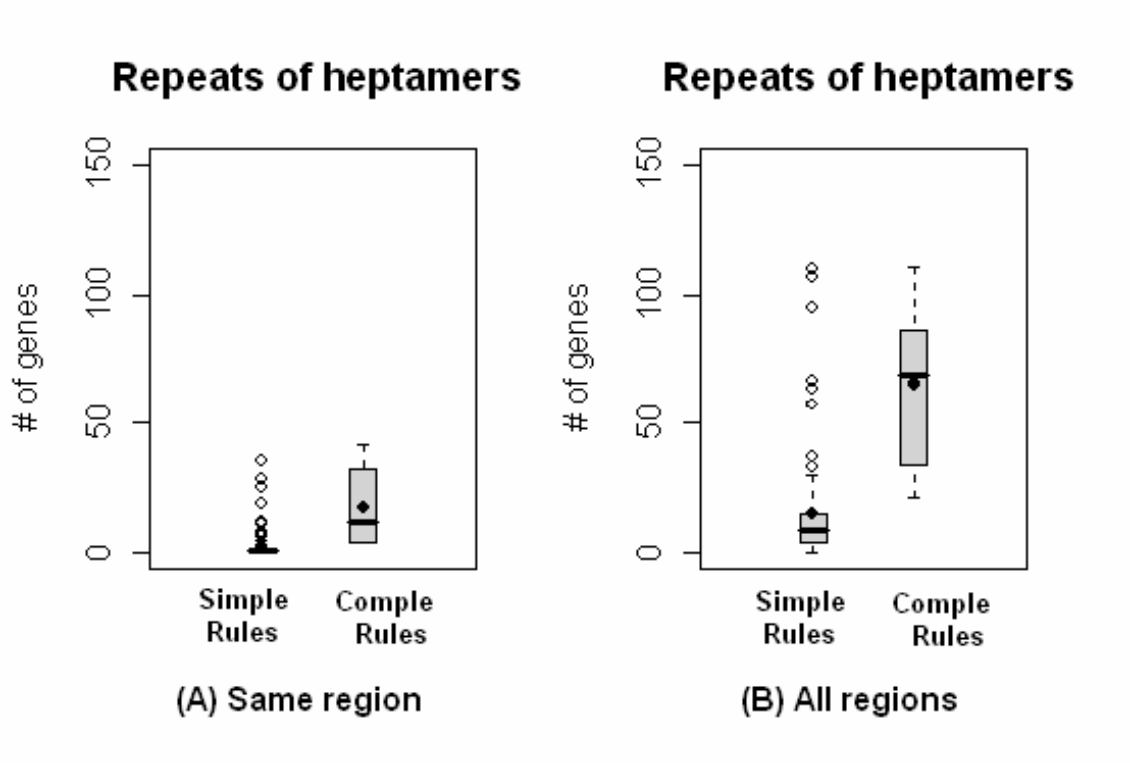


Fig 28 Number of genes with two or more heptamer repeats from simple and complex rules.

5.2.2 Motif Conservation Score

We also compared the motif conservation score of heptamers from simple and complex rules using PhastCons (Siepel, Bejerano et al. 2005) scores stored in UCSC. PhastCons fits a phylo –HMM to the data using maximum likelihood, and then predicts conserved elements based on this model (Siepel, Bejerano et al. 2005). Half of the heptamers from complex rules are significantly more conserved than random heptamers (p-values < 0.05) and a third of the heptamers from simple rules are significantly more conserved than random heptamers (p-

values < 0.05, data not shown). The motif conservation score of each heptamer is shown in Table 10.

Table 10 Motif conservation score of heptamers in association rules. Heptamer* is complex rules

Heptamer	Motif Conservation Score	p-value (description)
1_CCGGAGC	0.4759371	0.875 (not different)
1_GCCAAAG	0.7630357	0.005543 (more conserved)
2_AGGCCTG	0.1576149	0.3902 (not different)
2_CGCGCGG	0.1275714	0.1068 (not different)
2_GAAGTCC	0.1125476	0.4533 (not different)
2_GCCGGGC	0.1603598	0.4606 (not different)
2_GCGCGGG	0.1990612	0.01302 (more conserved)
2_GGAGCGG	0.09390476	0.006311 (less conserved)
2_GGCCTGG	0.1386000	0.05597 (not different)
2_TCACTCC	0.009918367	8.206e-14 (less conserved)
2_TTTCTCT*	0.002371429	3.195e-08 (less conserved)
3_AAAATAT*	0.2795306	0.004029 (more conserved)
3_AAGAGCA	0.1821746	0.461 (not different)
3_ACCCACC	0.1191429	0.7986 (not different)
3_CACCCAC	0.2122262	0.5112 (not different)
3_GAAGAGC	0.3526286	0.4211 (not different)
3_GACCTCT	0.05195122	0.0001790 (less conserved)
3_TGACCTC	N/A	N/A
3_TGACTAG	0.004457143	< 2.2e-16 (less conserved)
3_TGGTTCT	0.1700071	0.05903 (not different)
3_TTATTTT*	0.17487500	0.01479 (more conserved)
3_TTCATTT	0.2955678	0.2106 (not different)
3_TTGGTCT	0.4114898	0.2871 (not different)

Table 10 Continued.

3_TTGGTTC	0.1498831	0.9735 (not different)
3_TTGTTTT	0.2702045	0.2607 (not different)
3_TTTCTCT*	0.1828286	0.9383 (not different)
3_TTTGGTC	0.2199464	0.08885 (not different)
3_TTTGTTT*	0.04705357	4.715e-09 (less conserved)
3_TTTTCTG	0.3057817	0.4841 (not different)
3_TTTTTCA	0.2803077	0.2038 (not different)
4_AGGTGGT	0.7570238	0.04783 (more conserved)
4_CAACAGC	0.9246364	6.003e-06 (more conserved)
4_CTGCTGG	0.8603352	< 2.2e-16 (more conserved)
4_CTGGGCT	0.8892347	4.773e-06 (more conserved)
4_CTGGTGG	0.8295238	9.75e-12 (more conserved)
4_GACATCA	0.8726032	1.181e-07 (more conserved)
4_GACATTG	0.9939388	< 2.2e-16 (more conserved)
4_GAGAAGA	0.7370714	1.278e-08 (more conserved)
4_GCTGCTG	0.7899732	5.773e-11 (more conserved)
4_GCTGGAG	0.7013886	2.742e-09 (more conserved)
4_GGAGAAG	0.8079598	< 2.2e-16 (more conserved)
4_GGAGGTG	0.7520159	0.004809 (more conserved)
4_GGCTGTG	0.8156807	4.174e-11 (more conserved)
4_GTGCTGG	0.6434286	0.008951 (more conserved)
4_GTGGAGT	0.6524935	0.01505 (more conserved)
4_TGAGCTT	0.7767347	0.04894 (more conserved)
4_TGCAGAA	0.7505179	1.540e-07 (more conserved)
4_TGCTGGA	0.7824000	5.937e-06 (more conserved)
4_TGGCTGT	0.8259925	2.06e-12 (more conserved)
4_TGGGCTG	0.6890317	0.2228 (not different)
4_TGTGAAG	0.7787857	0.0003654 (more conserved)

Table 10 Continued.

4_TGTGGAG	0.8583442	< 2.2e-16 (more conserved)
4_TTGTGGA	0.7950084	2.207e-09 (more conserved)
6_AACAGGA	0.06032653	0.06032653 (more conserved)
6_ATAAAAT	0.2040905	0.8862 (not different)
6_CCTAGTC	0.006171429	1.729e-14 (less conserved)
6_CCTCCTT	0.02410714	7.784e-07 (less conserved)
6_CTCCTTC	0.04338961	4.78e-14 (less conserved)
6_CTTTCCT	N/A	N/A
6_GCAGCTG	0.2440143	0.2602 (not different)
6_GGTCCAG	0.0260000	2.518e-11 (less conserved)
6_GTTTTTG	0.1706286	0.5052 (not different)
6_TAAAATG	0.04618797	8.066e-06 (less conserved)
6_TTGTTTT	0.1538424	0.9539 (not different)
6_TTTAAAA*	0.4209580	0.0003128 (more conserved)
6_TTTCCTT	0.1540514	0.5167 (not different)
6_TTTGTTT	0.2385350	0.0001040 (more conserved)
7_ACCTCAC	0.02132143	< 2.2e-16 (less conserved)
7_AGGGAGC	0.7358810	0.6214 (not different)
7_ATGAAAA	0.7263109	6.505e-10 (more conserved)
7_TTGCTAA	0.9074935	1.228e-13 (more conserved)

Finally, to further validate our motif predictions, we assessed the overlap of our predictions with known AS regulatory sequence motifs stored in AEDB (Stamm, Riethoven et al. 2006). Among all heptamers involved in simple and complex rules, 43% occur within enhancer/silencer sequences from AEDB. This is a significantly (p-value = 0.017) higher percentage than we observed for a randomly selected set of heptamers of equal size.

5.2.3 Various Sizes of *k* of *k*-mer

We also found association rules with 6- to 9-mer with various minimum supports. Generally, *k*-mers proven after motif association rule mining are similar. For example, a heptamer rule, {4_GCTGGAG} => {meandiff(salivary) = -13.987} is extracted with all minimum support intervals we tried. It is found in hexamer rules in a rule with overlapping hexamers, {4_GCTGGA, 4_CTGGAG} => {meandiff(salivary) = -14.251}. It is also found in an octamer rule in a shape of superset rule, {4_TGCTGGAG} => {meandiff(salivary) = -24.457}. All association rules with different size of *k* in *k*-mer are stored at http://statgen.ncsu.edu/~jihye/KmerRule.html.

5.3 Conclusion and Discussion

We have applied distribution-based quantitative association rule mining to discover putative *cis*-regulatory motifs and motif combinations in alternatively spliced genes. Quantitative association rule mining provides a convenient framework for the systematic investigation of sequence motifs involved in the regulation of AS. Using t-tests and Bonferroni's multiple testing correction, we identified several statistically significant associations between sequence motifs, and tissue specific exon skipping rates. We found 94 simple rules containing 1 sequence motif in the antecedent, and 3 complex rules that contain 2 sequence motifs in the antecedent. Among the complex rules, 2 rules contain heptamer pairs from different regions of the pre-mRNA sequence. None of the heptamers from a complex rule is

also found in a simple rule. We hypothesize that these heptamer pairs correspond to factors that have to co-occur in order to influence AS. An approach that targets only individual motif candidates would have overlooked these motifs.

Many heptamer sets are found in multiple tissues, even when using high support thresholds. For example, the heptamer TGTGGAG in the cassette exon appears in rules describing heart, intestine, and muscle expression. Genes including this heptamer show lower exon skipping rates in all tissues (Fig 29 (A)). In addition, two very similar heptamers 4_GCTGGAG and 4_TGTGAAG, appear in rules that also correspond to a reduction in exon skipping rates. We speculate that the heptamers TGTGGAG, GCTGGAG and TGTGAAG might correspond to a single degenerate *cis*-regulatory element associated with a reduction of exon skipping. Among all 59 heptamer sets, 16 heptamer sets are found in two or more rules affecting exon skipping in different tissues.

On the other hand, some heptamers affect only exon skipping rates in a single tissue. For example, the rule {1_GCCAAAG} => {meandiff(spleen) = -18.186} occurs only in the spleen, with a support of 29 genes. The genes with this heptamer show significantly (p-value = 0.040) lower exon skipping in the spleen (Fig 29 (B)). We hypothesize that this heptamer motif increases exon inclusion specifically in the spleen.

This work has demonstrated that distribution-based quantitative association rule mining is a viable approach for discovering putative complex regulatory motifs for AS. In addition, comparison with known regulatory motifs stored in AEDB (Stamm, Riethoven et al. 2006) shows a significant enrichment of our heptamer set. Thus, we hypothesize that our motif catalog provides a promising list of candidates for subsequent experimental validation.

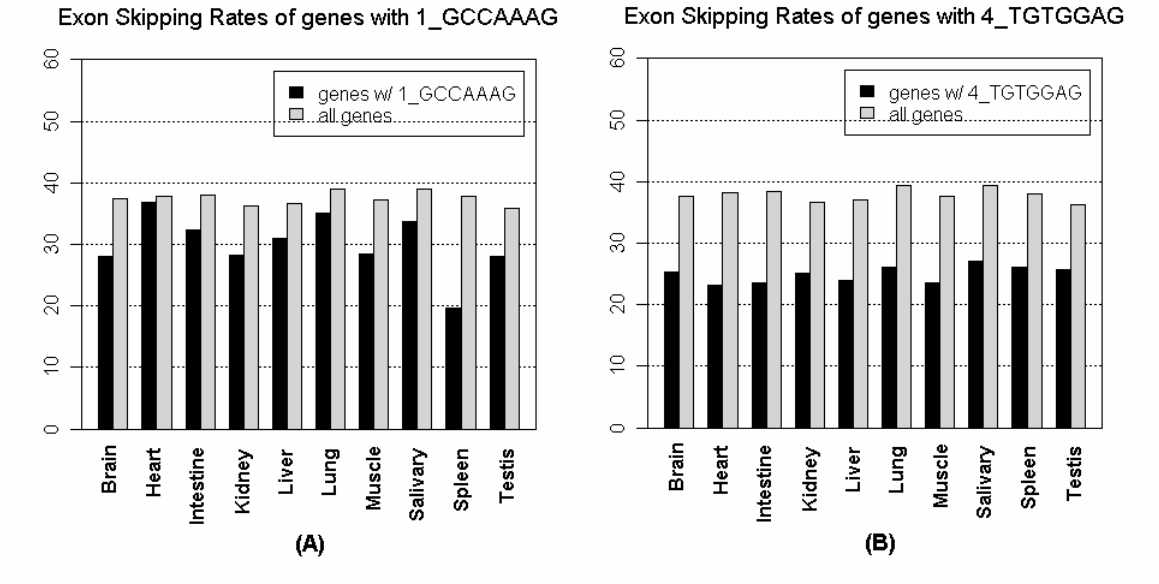


Fig 29 Exon skipping rates in 10 tissues. Gray bars represent the mean of exon skipping of genes with overall genes. Black bars represent the mean of exon skipping of genes with a frequent heptamer, GCCAAAG in upstream exon (A) and TGTGGAG in cassette exon (B), respectively.

Chapter 6
Summary

This dissertation has addressed the problem of identifying *cis*-regulatory elements involved in tissue-specific AS. The research focuses especially on interacting motifs. We have adapted three types of association rule mining to discover putative *cis*-regulatory motifs and motif combinations. General association rule mining with categorical attributes is used for finding motifs in genes with a similar exon skipping profile. The results of this method suggest that some motifs might influence exon skipping rates only if they appear in groups together with other motifs. A potential shortcoming of this approach is that the produced rules depend on user-specified thresholds and gene clusters.

To avoid these drawbacks, we used numeric exon skipping rates as items in quantitative association rule mining. In the discretization-based association rule mining method, we chose the equal–depth method to categorize numeric exon skipping rates in tissues. This method does not need to define clusters by exon skipping profiles. Using the results from this method we discovered individual motifs and also motif combinations which are involved in one or more tissue-specific AS. Many combinatorial motifs are from different exon and introns, indicating that they might be binding sites of different splicing factors which may work together in splicing process. This method is good to look at big patterns, but we cannot preclude the concern of missing important motifs and extracting trivial short sequences. It still contains problems in defining numeric values because it is sensitive to the bin size.

Lastly, distribution-based association rule mining methods free us from defining clusters or categorizing numeric values. Instead, we use the mean of the exon skipping rate as a measurement. It extracts association rules between motifs and splicing patterns more safely.

This method delivered us an interesting finding; some combinatorial motifs are not shown in a simple rules indicating that they work only together to influence AS. We also expect some interesting rules from different measurements such as variance.

From all methods, we extracted several typical splicing factor binding sites such as SR binding sites, GAAGAA, in common. In the validation with known motifs, they showed many overlapping sequences even though we still face problems of sensitive thresholds for a general association rule mining and categorization for discretization-based association rule mining. Association rule mining is a modern and promising framework for motif discovery.

The most important contribution of this dissertation is the method we developed to mine tissue-specific regulatory element sets. It is the first such method we know of to predict both individual and combinatorial motifs simultaneously. Over 40% of predictions are found to be known elements with support of a validated database. The findings are very useful for biologist interested in investigating tissue-specific AS events by helping to direct and prioritize their efforts and resources. In addition, they can make valuable contributions toward the creation of a catalogue of all splice regulatory elements and their respective condition distribution.

In the dissertation, we generate motif items with short sequences and see the exact matching of them during reading sequences. As a future direction, we can use more flexible motif representation by defining motif items with allowing gaps or mis-match or defining motifs with matrices. Also, we can include additional potential features of AS as association rule mining items such as trans-factors, exon length, splice site strength and RNA folds. Then, building a predictive model of AS with input of features or items of a gene and output of

tissue specific AS profile. Also, we can compare other approaches like tree-based methods including decision trees, regression trees, multivariate adaptive regression splines, and so on.

Association rule mining is also suitable for other areas. For example, with a similar approach, relationships between transcription factor motifs and conditions of transcribed genes can be estimated. As biological data grows faster, manually finding interesting features and relationships of features becomes impossible. Since, the nature of association rule mining is efficient discovery of unexpected relationships among a huge database, we expect association rule mining is a useful bioinformatical method which is helping to understand biological mechanism.

References

Agrawal, R. and J. Shafer (1996). "Parallel Mining of Association Rules." IEEE Transactions on Knowledge and Data Engineering **8**: 962-969.

Agrawal, R. and R. Srikant (1994). "Fast algorithms for mining association rules." 487-499.

Agresti, A. (2002). Categorical data analysis. New York, Wiley.

Akerman, M. and Y. Mandel-Gutfreund (2006). "Alternative splicing regulation at tandem 3' splice sites." Nucleic Acids Res **34**(1): 23-31.

Alvarez, S. A. (2003). Chi-squared computation for association rules: preliminary results. Boston, Boston College.

An, P. and P. J. Grabowski (2007). "Exon silencing by UAGG motifs in response to neuronal excitation." PLoS Biol **5**(2): e36.

Aumann, Y. and Y. Lindell (1999). A statistical theory for quantitative association rules. In Knowledge Discovery and Data Mining.

Baker, N. E. and G. M. Rubin (1989). "Effect on eye development of dominant mutations in Drosophila homologue of the EGF receptor." Nature **340**(6229): 150-3.

Benson, D. A., I. Karsch-Mizrachi, et al. (2006). "GenBank." Nucleic Acids Res **34**(Database issue): D16-20.

Black, D. L. (2000). "Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology." Cell **103**(3): 367-70.

Black, D. L. (2003). "Mechanisms of alternative pre-messenger RNA splicing." Annu Rev Biochem **72**: 291-336.

Bland, J. M. and D. G. Altman (1995). "Multiple significance tests: the Bonferroni method." BMJ **310**(6973): 170.

Borgelt, C. (2003). Efficient implementations of apriori and eclat. Workshop of Frequent Item Set Mining Implementations Melbourne, FL, USA.

Brett, D., J. Hanke, et al. (2000). "EST comparison indicates 38% of human mRNAs contain possible alternative splice forms." FEBS Lett **474**(1): 83-6.

Brett, D., H. Pospisil, et al. (2002). "Alternative splicing and genome complexity." Nat Genet **30**(1): 29-30.

Brin, S., R. Motwani, et al. (1997). "Dynamic itemset counting and implication rules for market basket data." SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data: 255-264.

Brin, S., R. Rastogi, et al. (2003). "Mining optimized gain rules for numeric attributes." Knowledge and Data Engineering, IEEE Transactions on **15**(2): 324-338.

Brudno, M., M. S. Gelfand, et al. (2001). "Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing." Nucleic Acids Res **29**(11): 2338-48.

Burge, C. B., R. A. Padgett, et al. (1998). "Evolutionary fates and origins of U12-type introns." Mol Cell **2**(6): 773-85.

Caceres, J. F. and A. R. Kornblihtt (2002). "Alternative splicing: multiple control mechanisms and involvement in human disease." Trends Genet **18**(4): 186-93.

Cartegni, L., J. Wang, et al. (2003). "ESEfinder: A web resource to identify exonic splicing enhancers." Nucleic Acids Res **31**(13): 3568-71.

Caudevilla, C., C. Codony, et al. (2001). "Localization of an exonic splicing enhancer responsible for mammalian natural trans-splicing." Nucleic Acids Res **29**(14): 3108-15.

Chan, C. S., O. Elemento, et al. (2005). "Revealing posttranscriptional regulatory elements through network-level conservation." PLoS Comput Biol **1**(7): e69.

Cheungt, D., J. Hans, et al. "A Fast Distributed Algorithm for Mining Association Rules *."

Chun-Hung, C., F. Ada Waichee, et al. (1999). Entropy-based subspace clustering for mining numerical data. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, United States, ACM.

Claverie, J. M. (2001). "Gene number. What if there are only 30,000 human genes?" Science **291**(5507): 1255-7.

Cooper, G. M. (2001). the Cell, a molecular approach, ASM Press.

Ester, M., H.-p. K. Kriege, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. KDD, AAAI Press 226-231.

Fairbrother, W. G., D. Holste, et al. (2004). "Single nucleotide polymorphism-based validation of exonic splicing enhancers." PLoS Biol **2**(9): E268.

Fairbrother, W. G., R. F. Yeh, et al. (2002). "Predictive identification of exonic splicing enhancers in human genes." Science **297**(5583): 1007-13.

Famulok, M. and J. W. Szostak (1993). "Selection of Functional RNA and DNA Molecules from Randomized Sequences." Nucleic Acids and Molecular Biology **7**.

Faustino, N. A. and T. A. Cooper (2003). "Pre-mRNA splicing and human disease." Genes Dev **17**(4): 419-37.

Florea, L. (2006). "Bioinformatics of alternative splicing and its regulation." Brief Bioinform **7**(1): 55-69.

Friedman, B. A., M. B. Stadler, et al. (2008). "Ab initio identification of functionally interacting pairs of cis-regulatory elements." Genome Res **18**(10): 1643-51.

Frilander, M. J. and J. A. Steitz (1999). "Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions." Genes Dev **13**(7): 851-63.

Fukuda, T. (1999). "Mining Optimized Association Rules for Numeric Attributes." Journal of Computer and System Sciences **58**(1): 1-12.

Garcia-Blanco, M. A., A. P. Baraniak, et al. (2004). "Alternative splicing in disease and therapy." Nat Biotechnol **22**(5): 535-46.

Grabowski, P. (2002). "Alternative splicing in parallel." Nat Biotechnol **20**(4): 346-7.

Guha, S., R. Rastogi, et al. (1998). CURE: an efficient clustering algorithm for large databases. SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data, ACM.

Haas, S. A., T. Beissbarth, et al. (2000). "GeneNest: automated generation and visualization of gene indices." Trends Genet **16**(11): 521-3.

Hahsler, M., B. Grün, et al. (2007). "arules: Mining Association Rules and Frequent Itemsets, 2006, URL http://cran.r-project.org/, R package version." SIGKDD Explorations **2**: 0-4.

Han, E.-H. and G. Karypis "Vipin Kumar."

Han, J. and M. Kamber (2000). Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems), {Morgan Kaufmann}.

Han, J., J. Pei, et al. (2000). "Mining frequent patterns without candidate generation." 1-12.

Han, K., G. Yeo, et al. (2005). "A combinatorial code for splicing silencing: UAGG and GGGG motifs." PLoS Biol **3**(5): e158.

Hannenhalli, S. and S. Levy (2002). "Predicting transcription factor synergism." Nucleic Acids Res **30**(19): 4278-84.

Hodges, D., R. M. Cripps, et al. (1999). "The role of evolutionarily conserved sequences in alternative splicing at the 3' end of Drosophila melanogaster myosin heavy chain RNA." Genetics **151**(1): 263-76.

Holste, D., G. Huo, et al. (2006). "HOLLYWOOD: a comparative relational database of alternative splicing." Nucleic Acids Res **34**(Database issue): D56-62.

Hu, G. K., S. J. Madore, et al. (2001). "Predicting splice variant from DNA chip expression data." Genome Res **11**(7): 1237-45.

Huang, H. D., J. T. Horng, et al. (2005). "SpliceInfo: an information repository for mRNA alternative splicing in human genome." Nucleic Acids Res **33**(Database issue): D80-5.

Huang, Y. H., Y. T. Chen, et al. (2002). "PALS db: Putative Alternative Splicing database." Nucleic Acids Res **30**(1): 186-90.

Itoh, H., T. Washio, et al. (2004). "Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes." RNA **10**(7): 1005-18.

Johnson, J. M., J. Castle, et al. (2003). "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays." Science **302**(5653): 2141-4.

Kabat, J. L., S. Barberan-Soler, et al. (2006). "Intronic alternative splicing regulators identified by comparative genomics in nematodes." PLoS Comput Biol **2**(7): e86.

Kan, Z., E. C. Rouchka, et al. (2001). "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs." Genome Res **11**(5): 889-900.

Kato, M., N. Hata, et al. (2004). "Identifying combinatorial regulation of transcription factors and binding motifs." Genome Biol **5**(8): R56.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.

Kim, N., S. Shin, et al. (2004). "ASmodeler: gene modeling of alternative splicing from genomic alignment of mRNA, EST and protein sequences." Nucleic Acids Res **32**(Web Server issue): W181-6.

Ladd, A. N. and T. A. Cooper (2002). "Finding signals that regulate alternative splicing in the post-genomic era." Genome Biol **3**(11): reviews0008.

Lee, C., L. Atanelov, et al. (2003). "ASAP: the Alternative Splicing Annotation Project." Nucleic Acids Res **31**(1): 101-5.

Lee, C. and M. Roy (2004). "Analysis of alternative splicing with microarrays: successes and challenges." Genome Biol **5**(7): 231.

Lee, Y., J. Tsai, et al. (2005). "The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes." Nucleic Acids Res **33**(Database issue): D71-4.

Leipzig, J., P. Pevzner, et al. (2004). "The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome." Nucleic Acids Res **32**(13): 3977-83.

Lent, B., A. Swami, et al. (1997). Clustering Association Rules. ICDE '97: Proceedings of the Thirteenth International Conference on Data Engineering, IEEE Computer Society.

Liang, F., I. Holt, et al. (2000). "An optimized protocol for analysis of EST sequences." Nucleic Acids Res **28**(18): 3657-65.

Lim, L. P. and C. B. Burge (2001). "A computational analysis of sequence features involved in recognition of short introns." Proc Natl Acad Sci U S A **98**(20): 11193-8.

Liu, H. X., M. Zhang, et al. (1998). "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins." Genes Dev **12**(13): 1998-2012.

Lyddy, J. (2002). "ExonHit Therapeutics." Pharmacogenomics **3**(6): 843-6.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-80.

Mata, J. (2002). An evolutionary algorithm to discover numeric association rules. In Proceedings of the ACM symposium on Applied computing SAC??002.

McNicholas, P. D., T. B. Murphy, et al. (2008). "Standardising the lift of an association rule." Comput. Stat. Data Anal. **52**(10): 4712-4721.

Miller, R. J. and Y. Yang (1997). "Association rules over interval data." SIGMOD Rec. **26**(2): 452-461.

Mironov, A. A., J. W. Fickett, et al. (1999). "Frequent alternative splicing of human genes." Genome Res **9**(12): 1288-93.

Modrek, B. and C. Lee (2002). "A genomic view of alternative splicing." Nat Genet **30**(1): 13-9.

Modrek, B., A. Resch, et al. (2001). "Genome-wide detection of alternative splicing in expressed sequences of human genes." Nucleic Acids Res **29**(13): 2850-9.

Pan, Q., O. Shai, et al. (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nat Genet **40**(12): 1413-5.

Pan, Q., O. Shai, et al. (2004). "Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform." Mol Cell **16**(6): 929-41.

Park, J., M.-S. Chen, et al. (1995). "An effective hash-based algorithm for mining association rules." 175-186.

Pilpel, Y., P. Sudarsanam, et al. (2001). "Identifying regulatory networks by combinatorial analysis of promoter elements." Nat Genet **29**(2): 153-9.

Pospisil, H., A. Herrmann, et al. (2004). "EASED: Extended Alternatively Spliced EST Database." Nucleic Acids Res **32**(Database issue): D70-4.

Ramakrishnan, S. and A. Rakesh (1996). "Mining quantitative association rules in large relational tables." SIGMOD Rec. **25**(2): 1-12.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Ruckert, U., L. Richter, et al. (2004). Quantitative association rules based on half-spaces: an optimization approach. Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on.

Salleb-Aouissi, A., C. Vrain, et al. (2007). QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules - CiteSeerX. International Joint Conference on Artificial Intelligence (IJCAI).

Savasere, A., E. Omiecinski, et al. (1995). "An efficient algorithm for mining association rules in large databases." 432-444.

Schaal, T. D. and T. Maniatis (1999). "Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences." Mol Cell Biol **19**(3): 1705-19.

Schuler, G. D. (1997). "Pieces of the puzzle: expressed sequence tags and the catalog of human genes." J Mol Med **75**(10): 694-8.

Shai, O., Q. D. Morris, et al. (2006). "Inferring global levels of alternative splicing isoforms using a generative model of microarray data." Bioinformatics **22**(5): 606-13.

Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-32.

Siepel, A., G. Bejerano, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-50.

Sinha, S., A. S. Adler, et al. (2008). "Systematic functional characterization of cis-regulatory motifs in human core promoters." Genome Res **18**(3): 477-88.

Stadler, M. B., N. Shomron, et al. (2006). "Inference of splicing regulatory activities by sequence neighborhood analysis." PLoS Genet **2**(11): e191.

Stamm, S., J. J. Riethoven, et al. (2006). "ASD: a bioinformatics resource on alternative splicing." Nucleic Acids Res **34**(Database issue): D46-55.

Stamm, S., J. Zhu, et al. (2000). "An alternative-exon database and its statistical analysis." DNA Cell Biol **19**(12): 739-56.

Stanke, M., O. Keller, et al. (2006). "AUGUSTUS: ab initio prediction of alternative transcripts." Nucleic Acids Res **34**(Web Server issue): W435-9.

Stanke, M., A. Tzvetkova, et al. (2006). "AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome." Genome Biol **7 Suppl 1**: S11 1-8.

Stein, L. D. (2004). "Human genome: end of the beginning." Nature **431**(7011): 915-6.

Thanaraj, T. A., S. Stamm, et al. (2004). "ASD: the Alternative Splicing Database." Nucleic Acids Res **32**(Database issue): D64-9.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Tuerk, C. and L. Gold (1990). "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." Science **249**(4968): 505-10.

Turcatti, G., A. Romieu, et al. (2008). "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis." Nucleic Acids Res **36**(4): e25.

Ule, J., A. Ule, et al. (2005). "Nova regulates brain-specific splicing to shape the synapse." Nat Genet **37**(8): 844-52.

Vardhanabhuti, S., J. Wang, et al. (2007). "Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation." Nucleic Acids Res **35**(10): 3203-13.

Voelker, R. B. and J. A. Berglund (2007). "A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing." Genome Res **17**(7): 1023-33.

Wang, K., S. Hock, et al. Interestingness-based Interval Merger for Numeric Association Rules. Proc. 4th Int.Conf. Knowledge Discovery and Data Mining, KDD.

Wang, W., J. Yang, et al. 1 STING: A Statistical Information Grid Approach to Spatial Data Mining.

Wang, Z., M. E. Rolish, et al. (2004). "Systematic identification and analysis of exonic splicing silencers." Cell **119**(6): 831-45.

Webb, G. (2001). Discovering associations with numeric variables. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.

Yager, R. R. (1995). Fuzzy summaries in database mining. CAIA '95: Proceedings of the 11th Conference on Artificial Intelligence for Applications, IEEE Computer Society.

Yeo, G., D. Holste, et al. (2004). "Variation in alternative splicing across human tissues." Genome Biol **5**(10): R74.

Yeo, G. W., E. Van Nostrand, et al. (2005). "Identification and analysis of alternative splicing events conserved in human and mouse." Proc Natl Acad Sci U S A **102**(8): 2850-5.

Yeo, G. W., E. L. Van Nostrand, et al. (2007). "Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements." PLoS Genet **3**(5): e85.

Zaki, M., S. Parthasarathy, et al. (1997). "Parallel Algorithms for Discovery of Association Rules." Data Mining and Knowledge Discovery: 343-373.

Zaki, M. J. and C. J. Hsiao (2005). "Efficient algorithms for mining closed itemsets and their lattice structure." Knowledge and Data Engineering, IEEE Transactions on **17**(4): 462-478.

Zaki, M. J., M. Ogihara, et al. (1996). "Parallel data mining for association rules on shared-memory multi-processors." Supercomputing '96: Proceedings of the 1996 ACM/IEEE conference on Supercomputing (CDROM).

Zavolan, M., S. Kondo, et al. (2003). "Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome." Genome Res **13**(6B): 1290-300.

Zhang, X. H. and L. A. Chasin (2004). "Computational definition of sequence motifs governing constitutive exon splicing." Genes Dev **18**(11): 1241-50.

Zhang, X. H., C. S. Leslie, et al. (2005). "Dichotomous splicing signals in exon flanks." Genome Res **15**(6): 768-79.

Zheng, Z. M., M. Huynen, et al. (1998). "A pyrimidine-rich exonic splicing suppressor binds multiple RNA splicing factors and inhibits spliceosome assembly." Proc Natl Acad Sci U S A **95**(24): 14088-93.

Appendices

Appendix A. Significance of Motif Association Rules

Sergio Alvarez explained the relationship between the chi-square statistic, and the values for support, confidence, and lift of a rule. With a rule "A=>B", we have the following values of support, confidence and lift from the definitions, Supp (A=>B) = P(A∩B), Conf (A=>B) = P(A∩B) / P(A), Lift (A=>B) = P(A∩B) / P(A)P(B).

$$\begin{aligned} &P(A \cap B) = supp, \\ &P(A) = supp / conf, \\ &P(B) = conf / lift \end{aligned} \qquad (1)$$

The contingency tables for the pair of variables (A, B) corresponding to the antecedent and consequent of an association rule A=>B are described as follows.

Table 11 Observed contingency table of a rule A=>B.

	$\overline{B}$	$\overline{B}$
A	$n \cdot supp$	$n \cdot \frac{supp}{conf}(1 - conf)$
$\overline{A}$	$n\left(\frac{conf}{lift} - supp\right)$	$n\left(1 - \frac{supp}{conf}(1 - conf) - \frac{conf}{lift}\right)$

Table 12 Expected contingency table of a rule A=>B.

	$\overline{B}$	$\overline{B}$
A	$n \cdot \frac{supp}{lift}$	$n \cdot \frac{supp}{conf}\left(1-\frac{conf}{lift}\right)$
$\overline{A}$	$n\left(1-\frac{supp}{conf}\right)\frac{conf}{lift}$	$n\left(1-\frac{supp}{conf}\right)\left(1-\frac{conff}{lift}\right)$

The chi-squared statistic is defined in terms of observed values and expected values:

$$\chi^2 = \sum_{0 \le i,j \le 1} \frac{(objserved_{i,j} - expected_{i,j})^2}{expected_{i,j}} \qquad (2)$$

From two contingency tables of a rule A=>B, the chi-squared statistic can be transformed to (see (Alvarez 2003) for detail):

$$\chi^2 = n(lift-1)^2 \frac{supp \cdot conf}{(conf-supp)(lift-conf)} \qquad (3)$$

We use equation (3) to compute the relationship between the support and confidence for a fixed χ^2/n and lift. We set up α=0.05 ($\chi^2 = 3.84$, with 1 degree of freedom) with n=2565 mouse genes and *lift*=1.2. Following the approach described in (Alvarez 2003) we computed the corresponding values for the *maximized support* (=0.032) and the corresponding minimum confidence (=0.195). To accommodate for multiple comparisons, we computed the p-value of each rule, and reported significant rules after Bonferroni adjustment (Bland and Altman 1995).

Appendix B. Discretization – Based Motif Association Rules

With discretization-based method we compute all motif association rules for minimum support = 0.032 (82 genes), minimum confidence = 0.195, and minimum lift = 1.2. These thresholds are computed by maximizing support with α=0.05 and *lift* = 1.2 in chi-square analysis statistic (equation (3) in Appendix A). The following reported rules are significant after Bonferroni adjustment (Bland and Altman 1995)

Table 13 Simple motif association rules by a discretization-based method.

Consequent (exon skipping profile)	Antecedent (hexamer)
BrainHighQuan	7_AGAAAA, 5_CATCAT, 2_TCTAGT, 2_TAAGCT, 5_GCATCA, 3_GAAGAT, 1_GCATCA, 4_GTGGAT, 5_TTTTGT, 2_CTCAAG, 5_TCCACT, 3_CCTCTG, 3_TGTACT, 4_CTTGTG, 5_GTTAGT, 3_ATTCAA, 6_GCAAAG, 2_GGGGAG, 2_TGGGGA, 2_AGATTG, 4_ACTTCT, 4_ACCAAG, 5_CTTTGT, 5_TTTGGC, 4_TGGGGA, 6_AGGAAT, 4_GCTGGG, 2_AACTGT, 7_CAGTCC, 1_GAACAA, 6_CAAAGA, 2_TGGATG, 3_GTGTTG, 2_AAACTG, 2_CAGTCT, 2_GGCCTT, 5_TGGGAG, 3_TATTAT, 2_TAGATT, 4_AGCAGA, 6_CTCCAG, 5_GCCAGG, 4_CAGCTG, 5_GTTTTG, 3_CTTGTA, 4_AAAGAA, 6_AGTAAC, 2_TAACTT, 5_CCTTAC, 4_GGAAAA, 4_TGGCTG, 5_AACTGC, 2_AATCCT, 3_TTGTAG, 2_TGGGAT, 2_ATGGGA,

Table 13 Continued.

	4_AAGATG, 7_GAAGAT, 3_CAGGTG, 2_TGGGTT, 4_ATGTCA, 4_TCCCAT, 3_TGTATT, 5_GCTTTG, 3_TACTCA, 2_GGAAAG, 7_ATTTGG, 6_CCTGGT, 5_CCAGTG, 2_TGAGAA, 4_GTCAGA
BrainLowQuan	2_AATTGA, 1_GTGGAG, 5_ATGTTT, 2_CTTCTT, 2_GGCCAG, 4_ATGGGC, 2_TTCTTC, 2_GGGCGG, 1_CCCCTG, 4_ACAGCC, 3_AACTTA, 5_CATGTT, 3_TCTTAA, 2_TAATGG, 6_GCACAT, 5_GCCAAG, 5_TTGGGC, 2_GCCCAC, 4_CTTTGG, 5_GCTTTC, 4_GCTCAG, 2_ACCATC, 3_GCATGC, 3_CTGGAC, 6_GTTCCA, 3_CAGCAA, 2_TCTTGA, 1_CAGACT, 4_GGACAT, 6_TGCACA, 4_GTGACC, 5_ATGTCA, 4_ATGGAT, 4_ACAAGA, 5_AGCATC, 5_AATGTT, 6_AAATCA, 3_GATTTA, 7_GCTTCC, 4_GGACAC, 4_GACCTC, 6_ATTCTC, 5_AAGGCT, 5_AAAATT, 2_ACTCAG, 5_TTTTAC, 4_TCAGCC, 5_TTTGCA, 2_TCTTCA, 5_AAAACT

HeartHighQuan	5_CAGGGA, 3_TTTGGT, 3_GTTGCT, 5_TGTGCC, 6_AACAAA, 7_AGATGG, 5_AGCCCT, 3_TTGATC, 6_TAAATG, 6_AAAAAA, 7_GAAAAG, 2_AAACTG, 2_CGGTGG, 1_ACCTGA, 4_CTACTG, 6_TTTCTG, 4_CAGAGT, 3_GGGTCC, 3_ACTCAC, 3_CAACAA, 5_GCATTA, 6_TCTTGT, 3_CCTCTG, 2_GGGCTG, 7_CAGTCC, 4_TCCCAT, 7_GGAGTC, 2_GGGGAG, 6_ATGCAG, 5_GGGACA, 6_ATTTAA, 3_GGATTC, 2_AGAGAT, 3_TGGTTG, 5_CTAAAT, 7_CCCAGT,

Table 13 Continued.

	5_TTTGGC, 2_TGGGAG, 6_ATCTTC, 3_TCATGA, 1_GTCCTC, 6_GTCTTG, 3_AAGCCA, 7_CAGTGT, 3_GGGCTA, 4_CAGCAA, 5_GGTCAG, 4_AGCAAG, 2_CGGGAC, 7_TTCTTC, 1_TTGCTG, 4_ACAAGT, 1_GGCAGA, 3_CAACCC, 2_GGCCCG, 2_GAGATC
HeartLowQuan	5_CTGTGC, 7_CAAGAA, 6_TCCAGA, 7_CTCTGG, 2_ACCCAG, 4_CCAGGA, 4_TCCTTT, 3_CTTGGT, 5_GAATGT, 2_AGAGGT, 5_GGAATA, 3_AATGCT, 7_TCTCTC, 5_AAGGCT, 2_AACAAG, 5_TATTTC, 6_AAGCAC, 1_AGGAAA, 5_TGGGGA, 2_GACCCA, 3_TGCTGC, 4_ATGGGC, 4_CCCAGG, 6_CATGTT, 2_GACAGT, 2_CAAGGT, 5_AGTAAG, 4_TGAGGC, 7_CTTGGA, 4_CACCCA, 6_CCACTC, 5_GTGCTA, 5_TCCCAC, 2_AGGCCG, 5_ATTTCA, 1_CAGACT, 2_TGGTCC, 1_AGGCCG, 2_CTGTGG, 5_ACCATT, 5_GCCAAG, 2_GCTAGG, 2_GAGAAT, 1_GGGCCC, 5_CAAAAC, 3_AGGAGT, 7_AACATC, 6_AAATAT,

	1_CATGAG, 4_GTGACC, 2_TTCCAA, 7_TTCTTG, 2_AGGGAG, 1_ATGAGC, 2_ACCATC, 3_ATCTTC, 2_GACCAG, 2_GTAATG, 3_TGCACT, 5_GTATTG, 7_AAGGCT, 6_GTTCCA, 3_GTGGTG, 2_GTGAGG, 2_ACCATG, 1_TGCGGG, 2_CAAGCA
IntestineHighQuan	4_AGCAAG, 2_AGTGGA, 3_TAAAGG, 6_ATTTTC, 4_CACCAT, 5_CCAGTG, 5_TCAGCC, 1_GGCAGA, 3_GATCAG, 4_TGGAGG, 2_TCTGAG, 5_AACTGC, 7_AGTGCA, 1_GTGGTG, 5_TAGCAG, 4_CTGATG,

Table 13 Continued.

	5_TACAGA, 7_GGACAA, 2_TGCTCA, 1_AAAGCA, 3_CCCTCT, 5_ATTGAG, 7_GAACAA, 6_TCCCAA, 2_GAAAGG, 5_CCTTCC, 7_AATGAG, 5_TTGTTT, 6_ATCTTC, 7_CATTCT, 1_CAGCTG, 2_GGCCTT, 5_GAGTGC, 2_CTGAAA, 7_CGGCAG, 2_TGGGAG, 5_TAGTTG, 2_CGCTCC, 6_AGAATA, 4_GACAAG, 4_AAAGGC, 2_CTCAAG, 6_AACAAA, 7_TGCTTC, 4_AGAATG, 5_TTCCTT, 3_TGTGTT, 3_TAAGTA, 7_TCAGAA, 2_ATTGGG, 7_CTTGAA, 1_CAAAGC, 5_GTATCT, 7_TTCTCA, 6_CTATCT, 7_AGATGG, 3_GGGTCC, 4_GACTCT, 5_CCATTG, 6_TGGACA, 4_GGCCCA, 4_AGGCCC, 4_TCCCAT
IntestineLowQuan	2_ACCATC, 3_TGGTAA, 3_TATGGT, 2_CAGGTA, 3_AATATG, 7_TGATGT, 3_GGGCTG, 2_GACTGA, 3_GAAGTT, 5_TGGACC, 7_GCAAGG, 7_TCAAAA, 3_CAGTCA, 5_AAAACT, 5_ATGTTT, 3_TGGCTG,

	3_AGTGCT, 3_CCTTGC, 7_GCTATG, 4_TTACAG, 1_GTGACC, 7_CAAGGA, 4_GCTGTG, 3_CCTTGG, 7_GTCACT, 6_GTTCCA, 2_ACTGAC, 3_AATGCT, 2_GCCCAC, 5_GCCAAG, 4_GTCCTC, 7_ATGAAC, 2_AACAAG, 4_CCCAGG, 7_AGTCAC, 6_CTGTAC, 2_AGCACA, 3_GATTAT, 7_CTCAAG, 5_TGATGG, 7_AAGGCT, 5_GGGGTA, 3_TTGGCT, 5_ACCATC, 3_AGGAGT
KidneyHighQuan	4_CAAAGC, 2_ACACAA, 6_GATAGA, 4_TCCCAT, 1_TGTTCA, 6_AACTAT, 5_CTAAAT, 5_GCATTA,

Table 13 Continued.

	5_GAAATC, 6_GGCATT, 3_ACTGGC, 3_ACCATT, 5_AATCAA, 5_GCTTAT, 7_ATAAAT, 3_TATTAG, 4_GACAAG, 5_TCCAGT, 3_AGCCAG, 2_CGGTGG, 7_GGATGG, 6_AAACTA, 4_CAAAAA, 7_CAGTCC, 3_TGGATG, 2_GGGGAG, 7_TGTTGG, 7_TGCTTC, 1_TGATGG, 6_CCATGA, 5_TAGCAG, 5_TGAAAT, 3_GCCAGA, 6_GAGCCC, 5_CCAGTG, 5_TTTTGT, 6_ATAAAT, 2_TGAGAG, 2_TCAGCC, 7_TCAGAA, 5_TTATTA, 5_TCAGCC, 3_CCAGTG, 6_GTGGGG, 2_GTAGAG, 1_GCTGCA, 4_AGCAAG, 1_GATGTG, 4_AAGATG, 5_AGACAG, 2_CTGAGA, 6_TAAATG, 6_ACAGAA, 4_AAGCCA, 5_AGGAGC
KidneyLowQuan	3_CTGCAG, 1_TGGGGC, 1_GTGGAG, 3_TTTTAT, 5_AAAACT, 4_CACCCA, 3_TTTATG, 2_TGAAAT, 5_GGTATT, 1_ACCAGT, 7_TGCCTT, 3_AGGTCT,

	2_ATTAAA, 5_TTTAGG, 5_TTTTGC, 5_TTGGTG, 7_GGAGTT, 6_TTCTCA, 3_CTATTC, 7_CCAAGT, 4_GTCCTC, 5_TTTGCA, 2_CTTATT, 1_CAGACT, 3_TCTTAA, 5_AACTTT, 2_GGGGTC, 7_AGGGCA, 7_GTGATG, 2_ACCATG, 7_CCCTGG, 6_GGACTT, 3_CCATGT, 6_GCAGGG, 6_GTTCCA, 3_GTGCTC, 2_ATGTTG
LiverHighQuan	5_TAGCAG, 4_TCCCAT, 7_TCTGAG, 5_CTGTTA, 5_AATTTG, 2_CTGCCT, 3_AGCCAG, 2_TCTGAG, 3_TGCAAG, 5_AACCTT, 7_GAAGAT, 2_AGGGGA, 5_TACAGA, 5_TGGGAC, 1_GTGGTG, 3_AAGGTA,

Table 13 Continued.

	3_CAACCC, 2_AGCAAT, 4_AGTTTC, 4_GGAAAA, 4_AGCAAG, 4_GATGAT, 5_CTAAAT, 5_TTTTGT, 5_AAAAAC, 3_AGATTA, 1_TGATGG, 2_TTGGCA, 6_AACTAA, 4_GTTTGA, 7_AGATGG, 3_CTTTTG, 3_TATTAG, 7_GCAAGC, 3_AGAAGG, 1_TTGCTG, 2_TAGAGC, 7_ACAACA, 4_GACAAG, 3_GAACCT, 4_AGCTGA, 2_TCTTGC, 5_TGACCT, 3_ACTGGC, 5_TTCTGT, 2_ACACAA, 3_TAAAGG, 2_AGGAAT, 5_GCATTA, 2_ACCCTT, 2_GTCTGA, 6_CTATCT, 5_TAGGCA, 1_AGGCCC, 2_GTAGAG
LiverLowQuan	6_GGTGGG, 3_AAGTGC, 6_AGAGCT, 3_AGCATC, 7_AGACAC, 1_CATCCA, 5_TGGTGC, 6_CCTTTT, 3_TCTGTG, 5_AAAACT, 5_GCTGGT, 7_CCCCTG, 7_CCCTGG, 7_CCTGAA, 7_TCCTTC, 3_TAAAAA,

	6_GCATCA, 5_TTGGTG, 3_ACCACA, 5_AAAGTC, 5_TGGAAA, 6_TGGGCC, 2_GGCCAG, 5_TGTAAG, 5_CTATAT, 7_GTTGAA, 3_CTATTC, 4_AGATTG, 7_CCCCCT, 3_ATTTAC, 5_TCTTAA, 3_ACAGCA, 1_AGGCCG, 5_AGTAAG, 5_GGTCCT, 7_CAGGGA
LungHighQuan	6_GGAGCC, 5_CCAGTG, 4_AAGTTT, 2_ATTACA, 2_CATTCT, 5_CTGTTA, 7_CCTGTT, 4_TGGTGT, 5_TTTTGT, 3_TAAAGG, 1_CAGCTG, 7_GAGATG, 7_AGATGG, 3_AGATGA, 4_GGCCCA, 1_TGCCTG, 2_AGATTG, 4_CACCAT, 3_GAGGCC, 3_GATTGA, 7_CAGTCC, 3_CAGCAC, 2_TGCTCA, 3_GGGTCC, 6_TCAGTC, 7_GATGGA, 2_TCTTGC, 5_TTTTTG,

Table 13 Continued.

	4_TCCCAT, 5_CCTTCT, 1_ACTGAA, 3_TGTCCC, 6_AACAAA, 2_CTCAAG, 3_ATAGAT, 5_GTTAGT, 7_ACAACA, 3_TTGAAT, 3_ACTTAA, 5_CTAAAT, 2_CTGTGA, 5_TGGGAG, 3_ACTCAC, 6_GTTTGG, 1_TTTCTC, 5_GGTTTT, 4_GGAAAA, 6_GGCACA, 4_AGCAAG, 5_TCATCA, 4_GGAACA, 4_AAACCA, 3_CAACCC, 2_GGCCTT, 5_TTCCTT, 2_ATTCTT
LungLowQuan	6_TTGGGG, 2_ACCATC, 5_ATGATG, 6_TCAATG, 5_ATGGCC, 5_CTTAAT, 5_CTTTCT, 1_CTGGCC, 7_CTTCCT, 5_GGGGTA, 1_CCTGGC, 2_GAAGTC, 2_CAGGCA, 5_TAGTGG, 5_GCTTTC, 3_GGGCTG, 3_GATTAT, 6_AATATG, 4_TTCTTC, 7_AAGGCT, 7_CCCCTG, 6_TGATGG, 6_TTCTCA, 5_GAATGT,

	4_GTCCTC, 1_AAGGAA, 6_CTTCCC, 5_CTATAT, 5_TGATGG, 7_CTCAAG, 3_AGCCTA, 3_CTCAGC, 2_CCACAC, 6_CACATT, 6_AGTGTC, 6_CCCTAG
MuscleHighQuan	4_TGCATC, 6_AAGCTC, 2_ATCCTC, 6_AATGCC, 3_GAAGTA, 2_TCTGTA, 3_ATCTCA, 5_CCACTC, 5_GACCCT, 3_CAACCC, 4_GACAAG, 7_CATTCT, 7_TTCCCT, 3_TTTGTC, 1_ACCAGA, 1_TCAAGA, 7_CAGTCC, 4_ATGCCT, 5_GGAACC, 4_CTGGTG, 3_CAGGCA, 4_GACTGC, 3_ACCAAA, 2_CTCATT, 7_AGTGGT, 3_GGGTCC, 4_TGGTGT, 5_TGACCT, 6_TGTCTT, 5_TTATCT, 7_CTGTGG, 7_ACCAGT, 4_CAGAGT, 5_GGTAGG, 6_ACTATA, 4_CACCAT, 3_CCAAAA, 7_TTCTCA, 4_TCCCAT, 3_CCTGTG,

Table 13 Continued.

	3_GATCTG, 7_GAGTCT, 5_AAGCCT, 4_AAACCA, 5_GGTTAG, 6_CCAAGA, 6_TCCCAA, 5_CTAGGT, 3_ATCTAA, 4_GATGCC, 3_GCCACA, 2_GGCCTT, 7_GGAGTC, 2_ACACAA, 3_GTGTTG, 4_AGCTCC, 1_AGAAAG, 2_CTACCC, 3_TGTGTT, 5_GAACTG, 4_AGCAAG, 1_TCCCAA, 2_GCCCAT, 2_CAAGGG, 5_CCTAGG, 5_TCCAGA, 5_GTTAGT, 4_CTCCAC, 7_TCCCTG
MuscleLowQuan	5_GGAGTC, 3_CACATG, 5_AGAGCC, 3_ATGTAG, 5_CTTCAG, 6_ATGATT, 7_GGGAAC, 3_CTGAAA, 7_CACACT, 3_TAAGAG, 1_CCTGCC, 1_CAGGGG, 2_GTGGTG, 7_TGCTCT, 2_TCAAAA, 3_TTTTCC,

	3_GTGAAC, 5_CAGGCA, 7_AAGGCT, 5_CTTTTT, 7_GGAGCA, 6_TTCTCA, 6_CAAGTA, 1_GGCCAG, 3_TTATGG, 5_GAAGAC, 6_TGATGG, 6_GTGTAG, 3_ACAGCA, 7_AGGACC, 3_AGCCTA, 2_CAGTGT, 4_CTTCAA, 6_CTGGAC, 5_AGGCTA, 7_TGGAGC, 5_GGTATT, 2_ACTCAC, 6_CAGTGT, 3_AGGAGT, 6_CACACC, 7_ATCACC, 2_AACAAG, 3_CCACAG, 3_TGCACT, 6_AAGCCA, 2_AAAATC, 7_TCCCCA, 3_GTTACA, 1_TGCTTC, 5_TGATGG, 7_GTTTGG, 1_AGGCCG, 6_TCTCAG, 7_CCCCTG, 4_CATCCC, 2_TCTACA, 2_GAAGTC, 4_CCCAGG, 7_GATGTC, 6_GCATCA, 5_GTCATG
SalivaryHighQuan	7_CCTGTT, 5_GTTTTT, 6_TCCCAA, 1_AACAGA, 5_TTATCT, 3_ACTCAC, 6_GGGTCT, 2_GTAGAG,

Table 13 Continued.

	1_TGGTGC, 5_TTTTGT, 6_AAGCTC, 5_GTTAGT, 3_TAAAGG, 3_CACTTA, 4_GGCCCA, 1_GCAAAG, 6_CACCAC, 2_CTCAAG, 4_AAGTTT, 5_TCCAGT, 6_ATGCAG, 4_GGAACA, 4_GATGAT, 6_ATCAGT, 6_GGCACA, 4_TGTGAT, 2_TGGGGA, 3_GGGTCC, 4_CACCAT, 7_CCTCCC, 7_AAACAT, 6_TGCCAA, 4_AGCAAG, 5_GCTTAT, 2_GAAAGG, 5_GCATTA, 2_ACACAA, 6_AGTAAC, 3_CTTGTA, 5_CCTCCC, 5_AGCCCT, 3_TTGATC, 2_CTGTGA, 2_AAGGTA, 6_ATAAAT, 1_CAGCTG, 3_CCATTT, 3_AGTTAG, 6_TTTGAA, 6_CTTTGA, 7_AAGATG, 2_TTGAAA,

	5_AGGAGC, 1_TGCCTG
SalivaryLowQuan	1_GGACTT, 1_CTCAGT, 4_TGTGAC, 6_ATCAGC, 1_AGGCCG, 3_TGGCTT, 1_AAGGAA, 4_GACTCC, 7_TTGGCA, 2_GTTGGT, 3_TGCACT, 4_TTCTGG, 5_TGATGG, 7_ATCACC, 5_TGATTT, 3_AGGAGT, 5_GCTTTT, 1_CAGACT, 6_CACTGG, 5_AGGCTA, 2_GAAGTC, 5_GGGGTA, 4_CCAGTT, 7_CCGTGG, 1_ATCCAG, 1_TGACCA, 2_ATGGAA, 6_TGATGG, 3_GATTAT, 5_GCCAAG, 5_TTATAT, 3_TAGAAT, 2_GCCCTT, 3_ATTTAC, 5_TGGCCT, 5_GAAGAC, 5_GATTTC, 6_AAGTTG, 5_GCCAGA, 1_CACCCC, 6_ACTTAA, 3_GTCCTG, 1_CTGGCC, 1_CCAGAG, 2_CAGGCA, 1_TGGCCA, 6_AGCTGT, 7_CTACAA, 7_CTCAAG, 6_AAAGCC, 3_AGCCTA, 3_CTGAAA, 6_GAAGAC, 6_CTTTTC, 1_CGCAGA, 6_TTTTCC,

Table 13 Continued.

	3_TATGGT, 2_ACCATC, 7_AACTGG, 7_CACACT, 2_CCATGA, 1_AGTGGA, 5_AGTTTC, 7_AAGGCT, 4_GTGACC, 7_CCCCTC, 4_GGCTCC, 6_CACTTA, 6_ACTTCC
SpleenHighQuan	6_GCTACA, 4_GAAAGG, 5_AGAAGT, 3_CAGATC, 4_GGAGCA, 7_CAGTCC, 6_AACCCC, 4_CTGGTG, 5_GGTCAG, 2_CATTCT, 7_ACAAAG, 2_TGCTCA, 3_AAAAGT, 6_TCTTGT, 4_AGCAAG, 7_CGGCAG, 5_TACAGA, 2_AGTTAG, 3_GCCTAG, 3_GGGCTA, 4_CCTGGT, 4_CACCAT, 4_GGAAAA, 2_GGCCTT, 7_ATAAAT, 4_TGCATC, 7_TACTGG, 7_GGAGTC, 2_CCCAAC, 1_TGGTGC, 7_GTGAGG, 4_GTTTGA, 5_CCAGTG, 7_AAGCTG, 3_CTGTAC, 3_GGGGCC, 4_TGGGAC, 3_GTCTTA, 4_TGGTGT, 3_TAGCCT, 6_TGGACA, 3_ATATTG, 4_GAGCAG, 5_AGGCAA, 4_GCCTTG, 3_GTTAGG, 4_GGCCCA, 3_GCAGGT, 4_CAGAGT, 6_CTAAGG, 3_GCCTGT, 7_ACCAGT, 4_CAAGAA, 5_GACTCA, 7_AGATGG, 5_CCACTC, 2_TCTTGC, 3_AGGGTC, 4_GCCAAA, 5_GTTAGT, 6_TGCCAA
SpleenLowQuan	6_GTTGGT, 4_AAGGAG, 6_CTAAGT, 6_GGGCAG, 2_CAGGAA, 4_CCTACA, 7_AGCAGG, 3_GGGCTG, 6_TTTCAA, 3_CCTACC, 1_AAGCCA, 3_TTGGCT, 4_GGCAGG, 7_AAGGCT, 5_TGTGGG, 5_GTGATG, 6_GTGAGC, 7_AAGAAG, 3_GAAGTT, 7_CCCCTG, 4_GAGTTT, 7_TGATGT, 4_GTCCTC, 7_AATGCA,

Table 13 Continued.

	2_GCCCAC, 2_CATCCA, 6_GCTAAG, 7_CACACT, 5_TCTATG, 2_GAAGTC, 4_TCTGGT, 3_TGCACT, 3_TCAGCC, 4_GACCAA, 7_GATGTC, 7_CCTGAA, 7_TCCCCA, 5_ATGGCC, 4_TTCTGG, 4_GGAGTT, 4_CATGCT, 3_GATTAT, 5_AGGATG, 1_CAGACT, 4_CTTGTG, 7_AGACAC, 5_GATGGC, 5_TGATTT, 5_TGATGG, 4_TTGTGA, 5_GCCAAG
TestisHighQuan	6_GGGACA, 5_CCAGTG, 6_TGCCAA, 7_AGATGG, 5_CCTGTA, 4_TGCTCA, 3_CACACT, 4_ACTGCT, 3_AGATTG, 5_TTGGTT, 5_TGGGTT, 5_TATCTG, 1_TGGGCT, 4_TGAGCA, 5_TCAGCC, 5_AGTGCT, 3_GAAAGT, 3_CCTGAT, 7_TGAATT, 3_GAAGGT, 4_AAGTTT, 7_GAGATG, 5_GAAGAG, 1_TGGTGC, 5_TGACCT, 5_CCAAAT, 3_CTTTGG, 4_TCCCAT, 2_TAGAGC, 6_GTTTGG, 2_CGCTCC, 7_AATCAG, 5_GCAGTA, 4_CACCAT, 2_ACTCCT, 3_ACCATT, 7_GAATTT, 5_GGACAA, 2_CAAGGG, 4_AGGACC, 4_AAAGAA, 5_GTGGGA, 7_GGAGAT, 5_GAACCT, 4_TAAAGA, 2_TCTGAG, 6_GATTAA, 2_GGCCTT, 4_ACCAGT, 2_AAGGGC, 5_GTGTAG, 4_GGAAAA, 1_AAAGCA, 5_GGGACA, 5_TGGTTG, 4_GACATT, 3_GCCTAG, 4_GGACCC, 4_AGACAT, 1_CTGTGA, 2_GAGATC, 6_AGTAAC, 3_TCAAAT, 2_ACACAA, 7_GCTTCT, 3_TTCCCT, 6_GCTACT, 3_ATAGAT, 2_CTCTCC, 3_CTCTTT, 6_ATTTTT, 4_GCTCAT

Table 13 Continued.

TestisLowQuan	7_TGGAGC, 3_TTGATT, 1_CGGGCT, 6_TGGACC, 5_GGGGAA, 4_AGATTG, 3_GTGCCT, 5_TTGGTG, 1_GCCTGC, 5_GGAAAA, 3_CATGCT, 4_TCTTCA, 3_TGGTAA, 7_GATTCT, 2_CCCAGT, 1_GCCGGC, 7_TTGGCA, 5_AATTTA, 5_GCTAGA, 3_AGACTG, 7_TCTTCC, 2_TGATCT, 2_GGCCTG, 1_ATGAGC, 2_GGCGCT, 1_CCTGGC, 5_TGGCAA, 1_ACCCCT, 5_GTGATG, 5_ATGTCA, 6_GGAACC, 5_AATTAT, 2_GATGGA, 3_CTGGTA, 5_GATTTA, 4_AGCCCT, 6_GAAAGT, 4_CCTGTC, 7_AAGGCT, 5_TTAATG, 1_CGGCTG, 3_GTGTAG, 5_GGCAAG, 5_TTTAGG, 1_GGCGGG, 3_TGTAAG, 4_ATGGGC, 3_TATGGT, 2_TCATCC, 4_ACCAGA, 1_TCTTCC, 7_TGCTCC, 2_TTATTA, 7_TTCCAC, 3_GATTAT, 2_GTTGGT, 2_CATCCA, 7_GTCTCC, 6_TTTCCT, 1_AGGCCG

Table 14 Complex motif association rules.

Consequent (exon skipping profile)	Antecedent (hexamer set)
BrainHighQuan	{5_TTGTTT, 6_TTTCTT}, {5_TTGTTT, 6_TCTTTT}, {5_TTGTTT, 3_TTTTCT}, {3_ATGTTT, 3_TGTTTT}, {3_TTTCTG, 6_TTTTCT}, {3_TTGTTT, 3_TTTTCT}, {3_TGTCTG, 3_TTTTCT}, {3_TGTTGT, 3_TGTTTT}
BrainLowQuan	{5_TTTAAA, 5_TGTTTT}, {3_TTTTTT, 5_TGTTTT}, {3_CTGTGT, 6_TTTCTT}, {6_TTTCTC, 3_TTCTTT}

Table 14 Continued.

HeartHighQuan	{6_TTTTTA, 6_TTTTTC}, {6_TTTTTA, 3_TTCTCT}, {5_TTCCTG, 6_TGTTTT}, {2_GGTGGG, 3_TTTCTT}, {6_GTTTTT, 6_TTTTTT}, {6_TTCTGT, 6_TTTTCT}, {2_GGAGGG, 6_TTTTCT}, {3_TTTCTG, 6_TCTTTT}, {6_TTTCTG, 3_TTTTCT}, {6_TTTTTA, 6_TCTTTT}, {6_TTTCTG, 6_TTTTCT}, {2_CCTGGG, 6_TGTTTT}, {6_TTTCTG, 3_CTTTCT}, {5_TTGTTT, 6_TCTTTT}, {6_TTTCTG, 3_TTTCCT}, {6_TTTATT, 3_TTTTCT}, {6_TTTCTG, 3_TTCTGT}, {6_TTTCTG, 3_TTTCTT}
HeartLowQuan	{3_CTGCAG, 6_TTTCTT}, {3_TTTGCT, 3_TTCTTT}, {5_TTTTTA, 6_TTTTTC}, {6_ACTTTT, 6_TTTTCT}, {6_TGTCTT, 3_TTTCTT}, {6_TGTCTT, 3_TTCTTT}
IntestineHighQuan	{3_TTCTGT, 6_CTTTTT}, {5_TTTTAT, 5_TTGTTT}, {6_TTTTTA, 3_TTCTCT}, {6_TTAAAA, 6_TTTAAA}, {3_TTCTGT, 6_TCTTTT}, {5_TTTGTT, 6_TTGTTT}, {3_TTTCTG, 6_TCTTTT}, {5_GTTTTT, 5_TTGTTT}, {5_TGTGTG, 6_TTCTTT}, {2_GGTGGG, 3_TTTTCT}, {2_GTAAGT, 6_CTTTCT}, {6_TTTTCA, 6_TTTTGT}, {5_TTCTGT, 6_TTTCTT}, {5_TTTTAT, 5_TATTTT}, {5_TTTTGT, 6_TTTTCT}, {6_TTTTTA, 6_TTTTGT}, {5_TTTTTG, 5_TTGTTT}, {5_TTTTTT, 3_TTTTCT}
IntestineLowQuan	{5_CTTTTT, 5_TTTTCT}, {5_CTTTTT, 5_TTTCTT}, {3_TGCTTT, 3_TTTCTT}, {6_CTTTTT, 3_TTTCTT}, {6_CTTTTT, 2_GTGAGT}, {6_CTTTTT, 3_TTCTTT}, {6_TCTTTT, 2_GTGAGT}, {3_CTTTCT, 2_GTGAGT}, {3_CTTTTT, 3_TTTCTT}, {3_CTTTTT, 2_GTGAGT}

Table 14 Continued.

KidneyHighQuan	{6_TGCTTT, 3_TTTCCT}, {5_CTGTGT, 5_TTGTTT}, {5_GTTTTT, 5_TTGTTT}, {5_TTCCTG, 6_TGTTTT}, {5_CTGTCT, 3_TTTCTT}, {5_TTTTTG, 5_TGTTTT}, {2_GTAAGT, 6_CTTTCT}
KidneyLowQuan	{5_ATTTTT, 5_TGTTTT}, {5_TTTAAA, 5_TTTTCT}, {3_TTTTAT, 6_TTTCTT}, {6_CTTTTT, 3_CTTTCT}, {3_TTGCTT, 3_TGTTTT}, {3_TTTTAT, 6_TTCTTT}, {5_TTTAAA, 5_TGTTTT}, {6_TTCTCT, 5_TGTTTT}, {5_TTATTT, 5_TTTAAA}, {5_TTAAAA, 5_TTATTT}, {2_CTTCCC, 6_TTTTCT}
LiverHighQuan	{5_TTTTTG, 5_TGTTTT}, {6_CTTCCT, 6_TTTCTT}, {5_TTCTGT, 6_TTTCTT}, {3_TTTCTG, 6_TCTTTT}, {5_TTTCTT, 6_TTGTTT}, {5_TTTTCT, 6_TTGTTT}, {2_CCTGGG, 6_TGTTTT}, {6_TTTTTA, 3_TTCTCT}, {2_CCTGGG, 3_TGTTTT}, {2_GTAAGT, 6_CTTTCT}
LiverLowQuan	{6_CTTTTT, 3_CTTTCT}, {5_TTTAAA, 5_TTTTAA}, {6_TTTCCT, 3_TGTTTT}, {6_CTTTTT, 3_TTTTTA}, {6_CTTTTT, 3_TTCTTT}, {3_CTGTTT, 3_TTTGTT}, {5_TTAAAA, 5_TGTTTT}, {5_TTAAAA, 5_TTATTT}, {6_CTTTTT, 3_TTTCTT}, {5_TTTAAA, 5_TGTTTT}
LungHighQuan	{5_TTTTTG, 5_TTTGTT}, {5_TTCTGT, 6_TTTCTT}, {5_CTTCCT, 3_TTTTCT}, {5_TTTTTG, 5_TTGTTT}, {5_ATTTTT, 3_TTTTCT}, {2_GTAAGT, 6_CTTTCT}, {3_CTGTCT, 3_TCTTTT}, {5_TTTTAT, 5_TATTTT}, {2_GTAAGT, 6_TTTTGT}, {5_TTTTTG, 5_TGTTTT}
LungLowQuan	{5_CTTTCT, 6_TTTCTT}, {2_TTCTCT, 6_TTTTCT}, {6_AAAATA, 6_TGTTTT}, {6_TGCTTT, 6_TTTTTT},

Table 14 Continued.

	{3_TATTTT, 6_TTCTCT}, {6_TCTCTT, 3_TTTCTT}, {2_TGTTTT, 6_TTTTCT}, {6_TTTCTG, 6_CTTTTT}
MuscleHighQuan	{3_TGTGTT, 6_TTTTCT}, {3_TGTGTT, 6_TTTCTT}, {3_GTGTTT, 3_TTTTGT}, {5_TTTTCT, 6_TTGTTT}, {3_CCTCTG, 6_TTTTCT}, {3_TGTGTT, 6_TTCTTT}, {3_TGTGTT, 6_TCTTTT}, {5_TTTGTT, 6_TTGTTT}, {3_TGTGTT, 3_TTCTGT}, {5_TTTTGT, 5_TTTTCT}, {5_GTGAGT, 6_TCTTTT}
MuscleLowQuan	{5_CTTTCT, 5_TTTTCT}, {6_CTTTTT, 3_TTTCTT}, {5_CTTTTT, 5_TTTTCT}, {5_CTTTTT, 5_TTTCTT}, {3_GTTTTT, 3_CTGTTT}, {2_GTGAGT, 3_TGTTTT}, {3_TGTGTG, 3_TTTCTT}, {6_TTTCAG, 3_TGTTTT}, {2_TGTTTT, 6_TTTTCT}, {3_TTTGTT, 6_TTCTTT}, {3_TGTGTG, 3_TTCTTT}, {5_TTTTTC, 5_CTTTTT}, {6_TATTTT, 3_TGTTTT}
SalivaryHighQuan	{5_GTTTTT, 5_TTGTTT}, {5_CTGTGT, 5_TTGTTT}, {5_TTTGTT, 3_TTTTGT}, {3_AATTTT, 3_TTTTAA}, {3_CTTTCT, 6_TTGTTT}, {5_TTTTTT, 3_TTTTCT}, {5_CTGTCT, 3_TTTTCT}, {5_TTTTTG, 5_GTTTTT}, {5_TTTTTT, 5_TTTGTT}, {5_TTTATT, 6_TTTTCT}, {5_GTTTGT, 5_TTGTTT}, {5_TTTTTG, 5_TGTTTT}, {5_GTTTTT, 5_TTTGTT}, {5_TTTGTT, 5_TTTTTA}, {5_TTTTAT, 5_TTTTTT}, {5_TTTTTG, 5_TTGTTT}, {5_TTGTTT, 3_TTTTCT}, {6_TGTCTT, 6_TGTTTT}, {5_TTTTTG, 5_TTTTTA}, {3_TTTTAA, 3_TTTTGT}
SalivaryLowQuan	{6_TGCTTT, 6_TTTTTT}, {6_TTCTGT, 6_TTTGTT}, {6_TTTTCC, 2_GTGAGT}, {6_CTGTTT, 3_TTTTTT},

Table 14 Continued.

	{2_GTGAGT, 6_TTTTCT}, {2_GTGAGT, 6_TGTTTT}, {6_CTTTTT, 3_TTTCTT}, {6_CTTTTT, 3_TTCTTT}, {6_CTTTTT, 2_GTGAGT}, {6_TGTGTT, 6_CTTTTT}, {6_AAAATA, 6_TGTTTT}
SpleenHighQuan	{5_TTTTTG, 5_TTGTTT}, {3_TGTGTT, 6_TTCTTT}, {5_TTCTGT, 6_TTTCTT}, {3_CTGTCT, 3_TCTTTT}, {2_GTAAGT, 6_TTCTTT}, {5_GTGAGT, 6_TCTTTT}, {2_GTAAGT, 6_CTTTCT}, {5_TCTGTT, 6_TTTTCT}, {3_TGTGTT, 6_TTTCTT}
SpleenLowQuan	{5_CTTTTT, 5_TTTCTT}, {5_TCTTTT, 5_TTTTTA}, {5_TTGTTT, 6_TTCTCT}, {6_TATTTT, 6_TTCTCT}, {3_TGTGTG, 3_TCTTCT}, {3_TGTGTG, 3_TTTCTT}
TestisHighQuan	{3_CTCTTT, 6_TTCTTT}, {3_TTTCTG, 6_TCTTTT}, {6_TTTCTC, 6_TTTTTA}, {3_CCTCTG, 6_TTTTCT}, {3_TTTCTG, 6_TTTTCT}, {3_CTTTTC, 6_TTTTCT}, {6_ATTTTT, 6_TTTTTA}, {3_GTTTTT, 3_TTTAAA}, {6_TTTCTG, 6_TTTTGT}
TestisLowQuan	{5_TTTAAA, 5_TGTTTT}, {2_GTGAGT, 3_TGTTTT}, {2_CTTCCC, 6_TTTTCT}, {3_TGCTTT, 2_GTGAGT}, {3_TTTTTT, 5_TGTTTT}

Appendix C. Distribution – Based Motif Association Rules

Table 15 Brain motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	3_TGACTAG , 0.026, -23.094 3_TTGGTTC 3_TGGTTCT , 0.009, -23.613 4_GCTGGAG , 0.001, -13.545 4_TGCTGGA , 0.004, -16.373 4_TGCTGGA 4_GCTGGAG , 0.018, -19.440 4_TGGGCTG , 0.015, -19.357 6_TTTAAAA 3_TTATTTT , 0.004, -20.216 7_ACCTCAC , 0.018, -18.713
25	3_TGACTAG , 0.017, -23.094 4_GCTGGAG , 0.001, -13.545 4_TGCTGGA , 0.003, -16.373 4_TGCTGGA 4_GCTGGAG , 0.012, -19.440 4_TGGGCTG , 0.010, -19.357 6_TTTAAAA 3_TTATTTT , 0.003, -20.216 7_ACCTCAC , 0.012, -18.713
30	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.002, -16.373 4_TGCTGGA 4_GCTGGAG , 0.009, -19.440 4_TGGGCTG , 0.007, -19.357 7_ACCTCAC , 0.009, -18.713
35	4_CAACAGC , 0.047, -18.049 4_GCTGGAG , 0.000, -13.545

Table 15 Continued.

	4_TGCTGGA , 0.001, 16.373 4_TGGGCTG , 0.005, 19.357 7_ACCTCAC , 0.006, -18.713
40	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.001, -16.373
45	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.001, -16.373
50	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.001, -16.373
55	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.000, -16.373
60	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.000, -16.373
65	4_GCTGGAG , 0.000, -13.545 4_TGCTGGA , 0.000, -16.373
70	4_GCTGGAG , 0.000, -13.545

Table 16 Heart motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	4_TGCTGGA , 0.010, -15.039 4_TGTGGAG , 0.003, -14.416
25	4_TGCTGGA , 0.007, -15.039

Table 16 Continued.

	4_TGTGGAG , 0.002, -14.416 4_GTGGAGT , 0.047, -19.546
30	4_TGCTGGA , 0.005, -15.039 4_TGTGGAG , 0.002, -14.416 4_GTGGAGT , 0.033, -19.546
35	4_TGTGGAG , 0.001, -14.416 4_TGCTGGA , 0.003, -15.039
40	4_TGTGGAG , 0.001, -14.416 4_TGCTGGA , 0.003, -15.039
45	4_TGTGGAG , 0.001, -14.416 4_TGCTGGA , 0.002, -15.039
50	4_TGTGGAG , 0.000, -14.416 4_TGCTGGA , 0.001, -15.039
55	4_TGCTGGA , 0.001, 15.039 4_TGTGGAG , 0.000, -14.416
60	4_TGCTGGA , 0.001, -15.039 4_TGTGGAG , 0.000, -14.416
65	4_TGCTGGA , 0.001, -15.039 4_TGTGGAG , 0.000, -14.416
70	4_TGTGGAG , 0.000, -14.416

Table 17 Intestine motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	4_TGCTGGA , 0.000, -19.731 4_GTGCTGG 4_TGCTGGA , 0.001,- 25.057 4_TGCTGGA 4_GCTGGAG , 0.001,-24.696 4_GCTGGAG , 0.024, -13.480 4_CTGCTGG 4_GCTGCTG , 0.003, -21.049 4_GCTGCTG , 0.032, -13.389 2_GAAGTCC , 0.042, -20.071 4_GACATCA , 0.035, -17.672
25	4_TGCTGGA , 0.000, -19.731 4_CTGCTGG 4_GCTGCTG , 0.002, -21.049 4_GCTGCTG , 0.021, -13.389 2_GAAGTCC , 0.028, -20.071 4_TGCTGGA 4_GCTGGAG , 0.001, -24.696 4_GCTGGAG , 0.016, -13.480 4_GACATCA , 0.024, -17.672
30	4_GGCTGTG , 0.050, -14.498 4_TGCTGGA , 0.000, -19.731 4_CTGCTGG 4_GCTGCTG , 0.001, -21.049 4_TGTGGAG , 0.044, 13.729 4_GCTGCTG , 0.015, 13.389 7_ATGAAAA , 0.042, -15.412 4_TGCTGGA 4_GCTGGAG , 0.000, -24.696 4_GCTGGAG , 0.012, -13.480 4_CTGCTGG , 0.040, -14.224 4_GACATCA , 0.017, -17.672

Table 17 Continued.

35	4_TGTGGAG , 0.032, 13.729 4_GGCTGTG , 0.036, 14.498 4_GCTGCTG , 0.011, -13.389 7_ATGAAAA , 0.030, -15.412 4_TGCTGGA , 0.000, -19.731 4_GCTGGAG , 0.008, -13.480 4_CTGCTGG , 0.029, -14.224 4_GACATCA , 0.012, -17.672
40	2_CGCGCGG , 0.048, +18.975 4_TGTGGAG , 0.023, 13.729 4_GGCTGTG , 0.026, -14.498 4_GCTGCTG , 0.008, -13.389 7_ATGAAAA , 0.022, -15.412 4_TGCTGGA , 0.000, -19.731 4_GCTGGAG , 0.006, -13.480 4_CTGCTGG , 0.021, -14.224
45	2_CGCGCGG , 0.035, +18.975 7_ATGAAAA , 0.016, -15.412 4_TGTGGAG , 0.017, -13.729 4_GGCTGTG , 0.019, -14.498 4_TGCTGGA , 0.000, -19.731 4_GCTGCTG , 0.006, -13.389 4_GCTGGAG , 0.005,- 13.480 4_CTGCTGG , 0.015, -14.224
50	2_CGCGCGG , 0.026, +18.975 7_ATGAAAA , 0.012, 15.412

Table 17 Continued.

	4_TGTGGAG , 0.013, 13.729 4_GGCTGTG , 0.014, -14.498 4_TGCTGGA , 0.000, -19.731 4_GCTGCTG , 0.004, -13.389 4_GCTGGAG , 0.003, -13.480 4_CTGCTGG , 0.011, -14.224
55	7_ATGAAAA , 0.009, -15.412 4_TGCTGGA , 0.000, -19.731 4_GCTGCTG , 0.003, -13.389 4_GCTGGAG , 0.002, -13.480 4_CTGCTGG , 0.008, -14.224 4_TGTGGAG , 0.009, -13.729 4_GGCTGTG , 0.011, -14.498
60	4_TGCTGGA , 0.000, -19.731 4_GCTGCTG , 0.002, -13.389 4_GCTGGAG , 0.002, -13.480 4_CTGCTGG , 0.006, -14.224 4_TGTGGAG , 0.007, -13.729 4_GGCTGTG , 0.008, -14.498
65	4_GCTGGAG , 0.001, -13.480 4_CTGCTGG , 0.005, -14.224 4_TGCTGGA , 0.000, -19.731 4_TGTGGAG , 0.005, -13.729 4_GCTGCTG , 0.002, -13.389
70	4_GCTGGAG , 0.001, -13.480 4_CTGCTGG , 0.004, -14.224

Table 17 Continued.

	4_TGTGGAG , 0.004, -13.729 4_GCTGCTG , 0.001, -13.389

Table 18 Kidney motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	7_TTGCTAA , 0.004, -19.492 4_TGCTGGA , 0.000, -19.273 4_TGCAGAA , 0.003, -15.417 6_AACAGGA , 0.005, -16.567 4_GAGAAGA 4_GGAGAAG , 0.003, -19.899 4_GACATTG , 0.022, -20.627 4_GGAGGTG , 0.002, -17.227
25	4_TGCTGGA , 0.000, -19.273 3_TTTTCTG 3_TTGTTTT , 0.043, +24.570 4_GAGAAGA 4_GGAGAAG , 0.002, -19.899 4_GGAGGTG , 0.001, -17.227 4_TGCAGAA , 0.002, -15.417 6_AACAGGA , 0.003, -16.567
30	4_TGCTGGA , 0.000, -19.273 3_TTTTCTG 3_TTGTTTT , 0.030, +24.570 4_GAGAAGA 4_GGAGAAG , 0.001, -19.899 4_GGAGGTG , 0.001, --17.227 4_TGCAGAA , 0.001, 15.417 6_AACAGGA , 0.002, -16.567

Table 18 Continued.

	4_GCTGGAG , 0.039, -11.838
35	4_GGAGGTG , 0.001, -17.227 4_TGCAGAA , 0.001, -15.417 6_AACAGGA , 0.002, -16.567 4_TGCTGGA , 0.000, -19.273 4_GCTGGAG , 0.028, -11.838
40	4_GGCTGTG , 0.041, -13.555 4_GGAGGTG , 0.000, -17.227 4_TGCAGAA , 0.001, -15.417 6_AACAGGA , 0.001, -16.567 4_TGTGAAG , 0.049, -13.747 4_TGCTGGA , 0.000, 19.273 4_GCTGGAG , 0.020, -11.838
45	6_AACAGGA , 0.001, -16.567 4_GGCTGTG , 0.030, -13.555 4_CTGGTGG , 0.049, 13.066 4_TGTGAAG , 0.036, -13.747 4_GGAGGTG , 0.000, -17.227 4_TGCTGGA , 0.000, -19.273 4_GCTGGAG , 0.015, -11.838 4_TGCAGAA , 0.001, -15.417
50	4_GGCTGTG , 0.022, -13.555 4_GGAGAAG , 0.043, -11.490 4_CTGGTGG , 0.036, -13.066 4_TGTGAAG , 0.027, -13.747 4_TGCTGGA , 0.000, -19.273

Table 18 Continued.

	4_GCTGGAG , 0.011, -11.838 4_TGCAGAA , 0.000, -15.417
55	4_CTGGTGG , 0.027, -13.066 4_TGTGAAG , 0.020, -13.747 4_TGCTGGA , 0.000, -19.273 4_GCTGGAG , 0.008, -11.838 4_GGCTGTG , 0.016, -13.555 4_GGAGAAG , 0.032, -11.490
60	4_CTGGTGG , 0.020, -13.066 4_TGCTGGA , 0.000, -19.273 4_GCTGGAG , 0.006, -11.838 4_GGCTGTG , 0.012, -13.555 4_GGAGAAG , 0.024, 11.490
65	4_GCTGGAG , 0.005, -11.838 4_CTGGTGG , 0.015, -13.066 4_TGCTGGA , 0.000, -19.273 4_GGAGAAG , 0.018, -11.490
70	4_GCTGGAG , 0.003, -11.838 4_GGAGAAG , 0.014, -11.490

Table 19 Liver motif association rules by a distribution-base method.

minsupp	Heptamer set , p-value, mean difference
20	4_TGCTGGA , 0.000, -19.649 4_TGGGCTG 4_CTGGGCT , 0.027, -20.160

Table 19 Continued.

	6_GGTCCAG , 0.004, -21.932 3_GACCTCT 3_TGACCTC , 0.003, -22.357 4_GTGCTGG 4_TGCTGGA , 0.026,- 21.740 2_TCACTCC , 0.029, -19.859 4_TGCTGGA 4_GCTGGAG , 0.007, 22.490 4_GCTGGAG , 0.000, 16.229 6_CTCCTTC 6_CCTCCTT , 0.003, -21.575 4_GAGAAGA 4_GGAGAAG , 0.025, -18.454 4_GACATTG , 0.008, -20.388 4_GGAGGTG , 0.002, -18.561 2_AGGCCTG 2_GGCCTGG , 0.000, -19.685
25	4_TGGGCTG , 0.040, -16.402 4_TGCTGGA , 0.000, -19.649 6_CTCCTTC 6_CCTCCTT , 0.002, -21.575 4_GAGAAGA 4_GGAGAAG , 0.017, -18.454 6_GGTCCAG , 0.002, -21.932 4_GGAGGTG , 0.001, -18.561 2_AGGCCTG 2_GGCCTGG , 0.000, -19.685 3_GACCTCT 3_TGACCTC , 0.002, -22.357 4_TGCTGGA 4_GCTGGAG , 0.005,- 22.490 4_GCTGGAG , 0.000, -16.229
30	4_TGGGCTG , 0.028, -16.402 4_TGCTGGA , 0.000, -19.649 3_TTTGGTC , 0.039, 17.223 4_GAGAAGA 4_GGAGAAG , 0.012, -18.454 4_GGAGGTG , 0.001, -18.561

Table 19 Continued.

	2_AGGCCTG 2_GGCCTGG , 0.000, -19.685 4_TGCTGGA 4_GCTGGAG , 0.004, -22.490 4_GCTGGAG , 0.000, -16.229 4_GACATCA , 0.045, -18.491
35	4_TGGGCTG , 0.021, -16.402 4_GGAGGTG , 0.001, -18.561 2_AGGCCTG 2_GGCCTGG , 0.000, -19.685 3_TTGGTCT , 0.047, -19.969 4_TGCTGGA , 0.000, -19.649 4_GCTGGAG , 0.000, -16.229 4_GACATCA , 0.033, -18.491
40	4_GGCTGTG , 0.050, -13.862 4_GGAGGTG , 0.001, -18.561 2_AGGCCTG 2_GGCCTGG , 0.000, -19.685 4_TGCTGGA , 0.000, -19.649 4_GCTGGAG , 0.000, -16.229
45	4_GGCTGTG , 0.037, -13.862 4_GGAGGTG , 0.000, -18.561 4_TGCTGGA , 0.000, -19.649 4_GCTGGAG , 0.000, -16.229
50	4_TGGCTGT , 0.040, -15.037 4_GGCTGTG , 0.027, -13.862 4_TGCTGGA , 0.000, -19.649 4_GCTGGAG , 0.000, 16.229
55	4_TGCTGGA , 0.000, 19.649

Table 19 Continued.

	4_GCTGGAG , 0.000, -16.229 4_GGCTGTG , 0.020, -13.862
60	4_TGCTGGA , 0.000, -19.649 4_GCTGGAG , 0.000, -16.229 4_GGCTGTG , 0.015, -13.862
65	4_GCTGGAG , 0.000, -16.229 4_TGCTGGA , 0.000, -19.649
70	4_GCTGGAG , 0.000, -16.229

Table 20 Lung motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	4_TGCTGGA , 0.000, -20.773 6_CCTAGTC , 0.002, -22.784 4_TGCTGGA 4_GCTGGAG , 0.000, -25.354 4_GCTGGAG , 0.000, -15.564 3_GAAGAGC 3_AAGAGCA , 0.001, -23.591 2_AGGCCTG 2_GGCCTGG , 0.010, -17.604 6_GTTTTTG 6_TTGTTTT 6_TTTGTTT , 0.024, -21.830
25	7_AGGGAGC , 0.044, +28.086 4_TGCTGGA , 0.000, 20.773 2_AGGCCTG 2_GGCCTGG , 0.007, -17.604 4_TGCTGGA 4_GCTGGAG , 0.000, -25.354 4_GCTGGAG , 0.000, -15.564

Table 20 Continued.

30	4_TGCTGGA , 0.000, -20.773 2_AGGCCTG 2_GGCCTGG , 0.005, -17.604 4_TGCTGGA 4_GCTGGAG , 0.000, -25.354 4_GCTGGAG , 0.000, -15.564
35	2_AGGCCTG 2_GGCCTGG , 0.004, -17.604 4_TGCTGGA , 0.000, -20.773 4_GCTGGAG , 0.000, -15.564
40	2_AGGCCTG 2_GGCCTGG , 0.003, -17.604 4_TGCTGGA , 0.000, 20.773 4_GCTGGAG , 0.000, -15.564
45	4_TGCTGGA , 0.000, -20.773 4_GCTGGAG , 0.000, -15.564
50	4_TGCTGGA , 0.000, -20.773 4_GCTGGAG , 0.000, -15.564
55	4_TGCTGGA , 0.000, -20.773 4_GCTGGAG , 0.000, -15.564
60	4_TGCTGGA , 0.000, -20.773 4_GCTGGAG , 0.000, -15.564
65	4_GCTGGAG , 0.000, -15.564 4_TGCTGGA , 0.000, -20.773
70	4_GCTGGAG , 0.000, -15.564

Table 21 Muscle motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	2_GCCGGGC , 0.034, +15.141 4_GCTGGAG , 0.000, -13.921 2_GGAGCGG , 0.037, +18.662 4_TGTGGAG , 0.007, -14.381
25	4_TTGTGGA 4_TGTGGAG , 0.041, -20.786 1_CCGGAGC , 0.035, +17.970 2_GGAGCGG , 0.025, +18.662 4_TGTGGAG , 0.004, -14.381 2_GCCGGGC , 0.023, +15.141 4_GCTGGAG , 0.000, -13.921
30	1_CCGGAGC , 0.025, +17.970 2_GGAGCGG , 0.017, +18.662 4_TGTGGAG , 0.003, -14.381 2_AGGCCTG 2_GGCCTGG , 0.044, -16.294 2_GCCGGGC , 0.016, +15.141 4_GCTGGAG , 0.000, -13.921
35	4_TGTGGAG , 0.002, -14.381 2_AGGCCTG 2_GGCCTGG , 0.032, -16.294 2_GCCGGGC , 0.012, +15.141 1_CCGGAGC , 0.018, +17.970 2_GGAGCGG , 0.013, +18.662 4_GCTGGAG , 0.000, -13.921
40	4_TGTGGAG , 0.002, -14.381 2_AGGCCTG 2_GGCCTGG , 0.023, -16.294

Table 21 Continued.

	2_GCCGGGC , 0.008, +15.141 1_CCGGAGC , 0.013, +17.970 2_GGAGCGG , 0.009, +18.662 4_GCTGGAG , 0.000, -13.921
45	2_GCCGGGC , 0.006, +15.141 1_CCGGAGC , 0.010, +17.970 4_TGTGGAG , 0.001, -14.381 2_GGAGCGG , 0.007, +18.662 4_TGCTGGA , 0.045, -13.640 4_GCTGGAG , 0.000, -13.921
50	2_GCCGGGC , 0.005, +15.141 1_CCGGAGC , 0.007, +17.970 4_TGTGGAG , 0.001, -14.381 2_GGAGCGG , 0.005, +18.662 4_TGCTGGA , 0.033, -13.640 4_GCTGGAG , 0.000, -13.921
55	2_GCCGGGC , 0.003, +15.141 1_CCGGAGC , 0.005, +17.970 4_TGCTGGA , 0.025, -13.640 4_GCTGGAG , 0.000, -13.921 4_TGTGGAG , 0.001, -14.381 2_GGAGCGG , 0.004, +18.662
60	2_GCCGGGC , 0.003, +15.141 1_CCGGAGC , 0.004, +17.970 4_TGCTGGA , 0.018, -13.640 4_GCTGGAG , 0.000, -13.921

Table 21 Continued.

	4_TGTGGAG , 0.000, -14.381 2_GGAGCGG , 0.003, +18.662
65	2_GCCGGGC , 0.002, +15.141 4_GCTGGAG , 0.000, -13.921 4_TGCTGGA , 0.014, -13.640 4_TGTGGAG , 0.000, -14.381
70	2_GCCGGGC , 0.001, +15.141 4_GCTGGAG , 0.000, -13.921 4_TGTGGAG , 0.000, -14.381

Table 22 Salivary motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	4_TGGGCTG , 0.004, -21.590 7_AGGGAGC , 0.021, +29.095 4_TGCTGGA , 0.001, -17.907 4_TGCTGGA 4_GCTGGAG , 0.003, -24.457 4_GCTGGAG , 0.015, -13.987 4_GGAGGTG , 0.002, -19.290
25	4_TGGGCTG , 0.003, -21.590 7_AGGGAGC , 0.014, +29.095 4_TGCTGGA , 0.001, -17.907 4_GGAGGTG , 0.001, -19.290 4_TGCTGGA 4_GCTGGAG , 0.002, -24.457 4_GCTGGAG , 0.010, -13.987

Table 22 Continued.

30	4_TGGGCTG , 0.002, -21.590 4_TGCTGGA , 0.000, -17.907 4_GGAGGTG , 0.001, -19.290 4_TGCTGGA 4_GCTGGAG , 0.001, -24.457 4_GCTGGAG , 0.007, -13.987
35	4_TGGGCTG , 0.001, -21.590 4_GGAGGTG , 0.001, -19.290 4_TGCTGGA , 0.000, -17.907 4_GCTGGAG , 0.005, -13.987 2_CGCGCGG 2_GCGCGGG , 0.041, +23.934
40	4_GGAGGTG , 0.000, -19.290 4_TGCTGGA , 0.000, -17.907 4_GCTGGAG , 0.004, -13.987
45	4_GGAGGTG , 0.000, -19.290 4_TGCTGGA , 0.000, -17.907 4_GCTGGAG , 0.003, -13.987
50	6_CTTTCCT 6_TTTCCTT , 0.042, +17.415 4_TGCTGGA , 0.000, -17.907 4_GCTGGAG , 0.002, -13.987
55	4_TGCTGGA , 0.000, -17.907 4_GCTGGAG , 0.001, -13.987 6_CTTTCCT 6_TTTCCTT , 0.031, +17.415
60	4_TGCTGGA , 0.000, -17.907 4_GCTGGAG , 0.001, -13.987 6_CTTTCCT 6_TTTCCTT , 0.023, +17.415

Table 22 Continued.

65	4_GCTGGAG , 0.001, -13.987 4_TGCTGGA , 0.000, -17.907
70	4_GCTGGAG , 0.001, -13.987

Table 23 Spleen motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	4_TGCTGGA , 0.000, -18.461 3_AAAATAT 3_TTTGTTT , 0.002, -24.253 2_TTTCTCT 3_TTTCTCT , 0.023, +32.536 3_AAAATAT 3_TTTGTTT 3_TTGTTTT , 0.002, -24.879
25	4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.043, -12.906 1_GCCAAAG , 0.040, -18.186
30	4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.031, -12.906
35	4_GCTGCTG , 0.022, -12.906 4_TGCTGGA , 0.000, -18.461
40	4_GCTGCTG , 0.016, -12.906 4_TGCTGGA , 0.000, -18.461 4_GCTGGAG , 0.046, -11.911 6_GCAGCTG , 0.038, -15.559
45	2_CGCGCGG , 0.042, +18.532

Table 23 Continued.

	4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.012, -12.906 4_GCTGGAG , 0.034, -11.911 6_GCAGCTG , 0.028, -15.559
50	2_CGCGCGG , 0.031, +18.532 4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.009, -12.906 4_GCTGGAG , 0.025, -11.911 6_GCAGCTG , 0.020, -15.559
55	4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.006, -12.906 4_GCTGGAG , 0.018, -11.911
60	4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.005, -12.906 4_GCTGGAG , 0.014, -11.911
65	4_GCTGGAG , 0.010, -11.911 4_TGCTGGA , 0.000, -18.461 4_GCTGCTG , 0.004, -12.906
70	4_GCTGGAG , 0.008, -11.911 4_GCTGCTG , 0.003, -12.906

Table 24 Testis motif association rules by a distribution-base method.

Minsupp	Heptamer set , p-value, mean difference
20	6_ATAAAAT 6_TAAAATG , 0.021, -21.401 3_TTTTTCA 3_TTCATTT , 0.034, -22.472 4_TGCTGGA , 0.016, -15.240 3_ACCCACC 3_CACCCAC , 0.002, -25.001 3_TTGGTCT , 0.045,- 20.042
25	3_TTGGTCT , 0.031, -20.042 4_TGCTGGA , 0.011, -15.240 4_TGAGCTT , 0.045, -20.678
30	4_AGGTGGT , 0.045, -18.570 3_TTGGTCT , 0.022, -20.042 4_TGCTGGA , 0.008, -15.240
35	4_AGGTGGT , 0.033, -18.570 3_TTGGTCT , 0.016,-20.042 4_TGCTGGA , 0.006, -15.240
40	4_TGCTGGA , 0.004, -15.240
45	4_TGCTGGA , 0.003, -15.240
50	4_TGCTGGA , 0.002, -15.240
55	4_TGCTGGA , 0.002, -15.240
60	4_TGCTGGA , 0.001, -15.240
65	4_TGCTGGA , 0.001, -15.240
70	N/A