ABSTRACT

ZHAO,SIHUI. Analysis of Cis-acting Regulatory Motifs Involved in Alternative Splicing. (Under the direction of Dr. Steffen Heber).

Alternative splicing is an important posttranscriptional process in eukaryotes. It dramatically expands the proteome and contributes essentially to the regulation of gene expression. Cis-acting regulatory motifs play a pivotal role in the regulation of alternative splicing. Many human diseases involved with aberrant (alternative) splicing are caused by mutations of splicing regulatory motifs. However, due to the short, degenerate and context-dependent nature, the prediction of cis-acting splicing motifs is a very challenging task.

In this dissertation, we focus on discovery of splicing signals from sequences. This may help to reveal the integrated splicing code and to understand the regulation of gene expression in the resolution of exon level.

In chapter one, we review the up-to-date research development in alternative splicing and its regulation, as well as the experimental and computational approaches in genome-wide alternative splicing analysis.

We describe a large-scale data analysis experiment to discover AS motifs in chapter two. We applied a computational framework to re-analyze a dataset containing about 2,500 cassette exons and skipping rates for regulatory motifs. The alternative spliced events were clustered by their expression profiles to find co-regulated genes. Rather than using a fixed cutoff as cluster boundary, we used systematic sampling to sample sequence clusters and eliminated redundant motifs predicted from overlapping clusters. We conclude that these predicted motifs may be promising candidates responsible for AS regulation by comparison to known motifs and by positional bias.

In chapter three, we describe a new approach to discover short and degenerate AS motifs. We implemented a two-step approach incorporating skipping rates in motif discovery. In the simulation study, we show that this approach is especially suitable to discover short and highly degenerate motifs. Analysis of cassette exons in Central Nervous System tissues produced 15 motifs which are associated with the variation of skipping rates. We discover that Nova and hnRNP A1 binding sites are involved with AS regulation, as well as about ten novel motifs. Moreover, co-operation between predicted motifs are also revealed.

In chapter four, we give the present status of SPRED, a database of cis-acting regulatory splicing elements. The motifs in SPRED are compiled from literature. They are all experimentally validated. The web interface is publically accessible and accompanied with query and similarity search tools. The goal of SPRED is to provide a comprehensive motif dictionary to facilitate the research in AS and its regulation.

Finally, we give the conclusions in chapter five. We also give the perspective for future study and briefly review the potential challenge.

Analysis of Cis-acting Regulatory Motifs Involved in Alternative Splicing

by
Sihui Zhao

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:

Steffen Heber Co-Chair of Advisory Committee	Zhao-Bang Zeng Co-Chair of Advisory Committee
David Bird	Hao Zhang

DEDICATION

To my parents, my wife and my son

BIOGRAPHY

Sihui Zhao was born in Shanghai, China. He got his Bachelor and Master degrees in Genetics from Fudan University. It was during his stay there that he was motivated to pursue an advance degree in biological science with more quantitative discipline. He got his second master degree in bioinformatics from Indiana University, Bloomington where he was not satisfied only using existing tools. Therefore, he came to NC State University for the Ph.D. degree co-major in bioinformatics and statistics. His research focuses on identification of short and degenerate motifs involved in alternative splicing by incorporating splicing profiles. The ultimate goal is to build predictive model to predict splicing outcome from sequences.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Steffen Heber, for his help over the past several years. I have learned a lot from him which can not be learned from textbooks. I want to express my appreciation to Dr. Zhao-Bang Zeng for his generous support and trust for many years. I'm also very grateful to Dr. David Bird, Dr. Hao Zhang and Dr. William Hoffmann for their valuable advises and suggestions. In addition, I want to say thank you to all the faculties, staffs and students in BRC. I really enjoy the time with everybody in BRC. Finally, I want to express my deepest appreciation to my wife and my parents for never giving up faith on me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Background – Alternative Splicing and Posttranscriptional Gene Regulation

Introduction

Alternative splicing (AS) is a ubiquitous biological process through which different mature mRNA isoforms are generated from the same gene by selectively including various combinations of exons. Alternative splicing has been observed in almost every metazoan organism and is especially prevalent in vertebrates. It is estimated that about 40% - 60% human genes are alternatively spliced based on the comparison between expressed sequence tags (ESTs) and genomic sequences, and this ratio can reach to 73% when combined with microarray data [1-3]. Moreover, >80% genes on human chromosomes 21 and 22 are subject to alternative splicing [4]. Most alternative splicing events affect coding regions [2] and half of them causes reading frame shift [5].

Splicing patterns in alternative splicing

There are several different alternative splicing patterns (see Figure 1-1) [1, 6].

The most common pattern is exon skipping that allows either inclusion or exclusion of cassette exons (also called skipped exons) in the mature mRNAs. One famous example of exon skipping is Drosophila sex-lethal (Sxl) gene, which is a switch in sex determination. Skipping of exon 3 of Sxl gene can maintain female differentiation. The exon 3 of Sxl contains a pre-mature stop codon, and inclusion of this exon produces a truncated and probably non-functional protein [7, 8].

Another splicing pattern is mutually exclusive exons, which allows only one of two adjacent exons to be included in the final product. Human fibroblast growth factor receptor 2

(FGFR-2) gene contains exon IIIb and IIIc which are mutually exclusive. The gene product from exon IIIb has much lower binding affinity to fibroblast growth factor [9].

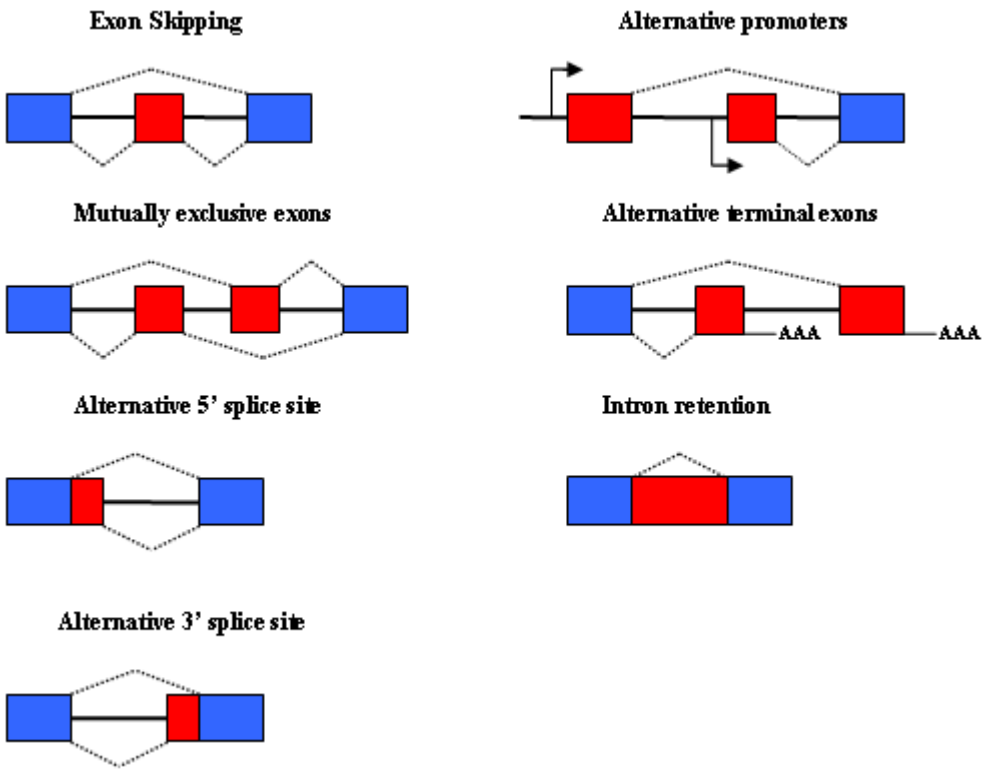


Figure 1-1. Major alternative splicing patterns [2]**.** The blue boxes are constitutive spliced exons and the red boxes are alternatively spliced exons.

Rather than the whole exons, alternative splicing can also splice out part of the exons. Choice of alternative 5' or 3' splice sites generates AS variants with or without an extension that flanks exon. Drosophila fruitless (fru) and double-sex (dsx) genes contain a female-specific alternative splice site, the former at 5' and latter at 3' end. Alternative selection of these splice sites can produce variants with small extensions [10, 11].

Alternative splicing can occur at either end of the transcripts. Alternative terminal exons not only change the inclusion of the last exon but also affect polyadenylation site selection. In many cases, it can lead to a premature stop codon in the last exon and either produce functional truncated polypeptides or cause nonsense-mediated decay (NMD, degradation of mRNAs with termination codons located more than 50-55 bp upstream of the last exon-exon junction) [2, 12, 13]. Calcium-regulating hormone (calcitonin) gene comprises six exons. Mature calcitonin transcript contains the first four exons and uses the polyadenylation site in exon 4, which represents >98% gene products in thyroid C cell. Meanwhile, in brain and other peripheral nervous system, the splice variant with the first three, the fifth and sixth exons encodes the precursor of calcitonin-related peptide and utilizes a downstream adenylation site (CGRP) [14, 15]. Similarly, alternative promoter usage allows choices of different promoters for transcription and generally affects the first exon. Although it is widely regarded as transcriptional regulation, alternative promoter usage is widely correlated with alternative splicing. It has been observed that genes with alternative promoters are more likely to undergo alternative splicing and the numbers of alternative promoters are positively correlative with numbers of alternative spliced variants [16]. Mouse monocarboxylate transporter 2 (MCT2) gene has several alternative promoters, resulting in five unique first exons (1a - 1e). Exon 1c is used in various tissues while others are tissue-specific [17].

Lastly, introns can also participate in alternative splicing. In intron retention, the complete intron can be included or excluded. It was regarded as the rarest pattern in human [1]. However, recent study shows the frequency is much higher (about 15%) in known human genes [18]. Intron retention is more common in plant than in other eukaryotes [19]. For

example, in *Arabidopsis* over 50% AS events are intron retention [20]. The last exon of human FosB (FBJ murine osteosarcoma viral oncogene homolog B) gene contains a 140-bp sequence, which can be spliced out to produce a truncated product ΔFosB. The expression of ΔFosB is observed in animals with chronic drug addiction [21].

Basal splicing machinery

Both alternative and constitutive splicing use the same basal machinery, called spliceosome. The spliceosome recognizes and selects the splice sites (exon-intron junctions) and catalyzes breaking and rejoining of RNA chains. The spliceosome mainly consists of five small nuclear ribonucleoproteins (snRNPs), U1, U2, U4, U5 and U6, which comprises uridine-rich small nuclear RNAs and multiple proteins. They can recognize the splicing signals on pre-mRNA and interact with each other or with other auxiliary splicing factors [22-24].

Three conserved sequence elements are required in splicing. These include canonical or non-canonical splice sites, polypyrimidine tract and branch point (see Figure 1-2) [22, 23]. Splice sites contain conserved short sequences spanning exon-intron junctions. GU and AG dinucleotides are commonly invariant at the 5' and 3' ends of introns, respectively. This type of GU-AG pair in splicing junction is called canonical splice site and exists in over 98% introns in mammalian genomes [25]. Non-canonical splice sites contain dinucleotide pairs such as GC-AG, AT-AC and so on, and are usually rare. Polypyrimidine tract is UC-rich segments located at the 3' end of intron and next to 3' splice site. It is the binding site for several splicing factors, such as U2 snRNP auxiliary factor (U2AF) and polypyrimidine tract

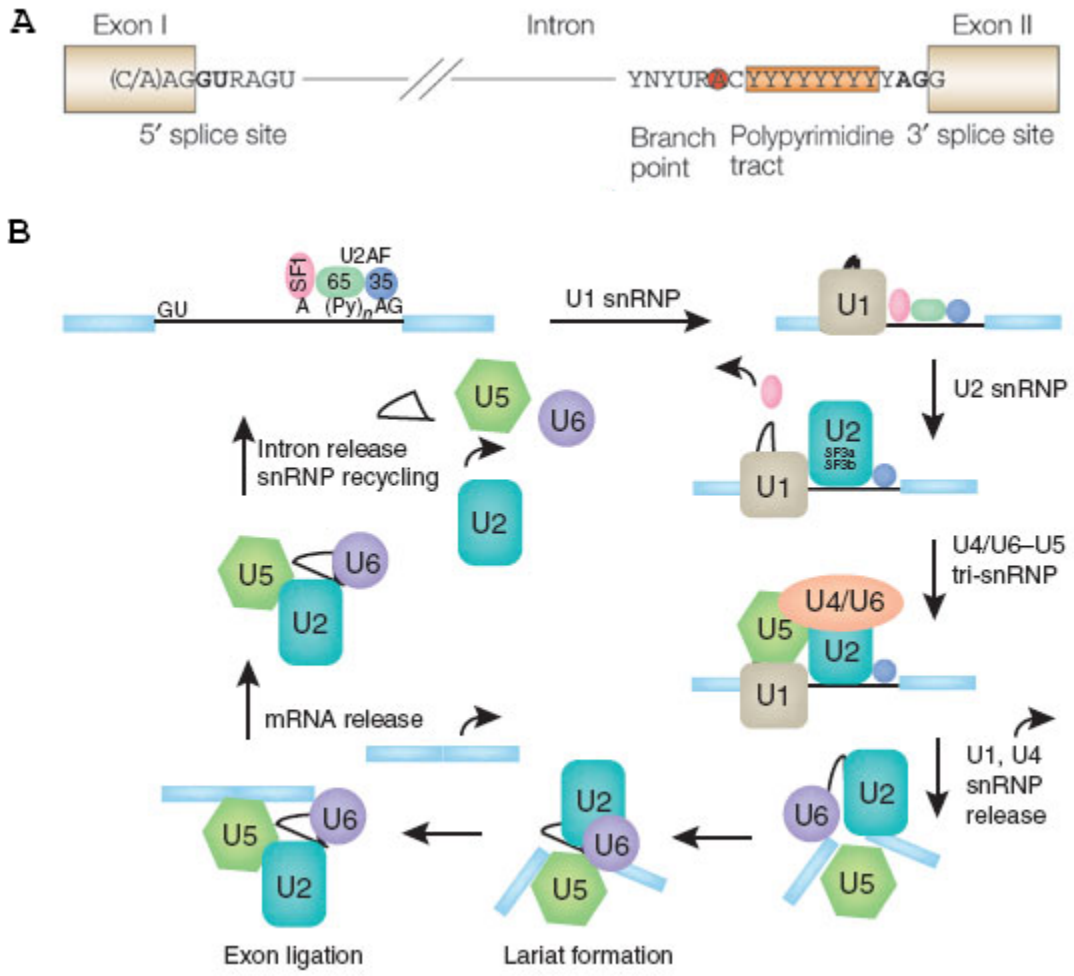


Figure 1-2. Core consensus sequences [2] **and spliceosome cycle** [22, 27]**.** A. Consensus sequences in exons and introns, including canonical splice site, polypyrimidine tract and branch site (or branch point); B. Spliceosome cycle.

binding protein (PTB) [26]. Branch point (also called branch site) is located upstream of polypyrimidine tract. The mean distance between branch point and 3' splice junction is about 30-40 bp in human and mouse [27]. Although the sequence of branch point is variable in

mammals, the mutation of branch point can promote the shift from constitutive to alternative splicing [27].

The spliceosome is assembled in a sequential manner called spliceosome cycle, which involves recognition of sequence elements and addition, rearrangement and release of snRNPs (see Figure 1-2). U1 is the first snRNP to join the spliceosome and binds to 5' splice site. Next, U2 snRNP is recruited by U2AF and bind to branch site. Next, U4/U5/U6 tri-snRNP participates in the pre-spliceosome. After rearrangement of the proteins and RNAs the catalytic spliceosome is generated and catalyzes two transesterification reactions of cutting and rejoining at splice sites. The ligated exon and lariat intron are then released sequentially. The spliceosome becomes unstable and dissociates thereafter. All snRNP particles can be re-used in new spliceosome cycles later [22-24].

Biological roles of alternative splicing

Alternative splicing plays pivotal roles in many different organisms. Knockout of many splicing factor genes regulating alternative splicing in mouse result in either embryonic lethality or severe disorder [28]. Overexpression of splicing factors also has deleterious effects. The correlation between cancer and overexpression of many splicing factors are widely observed in different types of tumors [29].

AS is an important way in post-transcriptional gene regulation, which is involved in cell differentiation and cell death in different developmental stages. In brain development, an unknown genetic switch affects a splicing factor, polypyrimidine tract binding protein (PTB), and replaces it with its closely related paralog, neural PTB (nPTB). This changes the splicing

patterns of numerous splicing target genes, as well as PTB/nPTB and other splicing factors, and therefore affects cell differentiation in determining the fate of cells to be neuron or non-neuron [30, 31]. A large number of apoptotic factors are also regulated through alternative splicing, such as membrane-associated death signal receptors of tumor necrosis factor (TNF), Fas gene, Bcl-2 gene family, Ced-4 gene family and caspase family [32-34]. Bcl-X is a member of Bcl-2 family. The Bcl-X gene contains three exons. Four Bcl-2 homology (BH) domains, BH1 to BH4, are located in exon 2. The long version Bcl-X_L contains the whole exon 2 (therefore includes all four BH domains) and inhibits apoptosis. One of the short isoforms Bcl-X_S uses an alternative 5' splice site in exon 2 and lacks BH1 and 2. Bcl-X_S antagonizes the apoptotic inhibition by Bcl-X_L and results in cell death. Bcl-X_L is highly expressed in tissues containing long-lived postmitotic cells, such as adult brain, while Bcl-X_S is mostly found in cells undergoing turnover, such as developing lymphocytes [33, 35, 36].

Posttranscriptional gene regulation by alternative splicing is commonly through two ways. First, the quantity of gene expression can be lowered through nonsense-mediated decay (NMD). In alternative splicing, intron retention and frame shift in protein coding regions are easy to produce transcripts with translational termination codons. The normal termination codons usually reside within the last exons and not followed by exon-exon junctions. However, transcripts with a premature termination codon >50-55 bp upstream to the following exon-exon junction tend to induce NMD [13]. Underrepresentation of NMD-inducing alternative spliced isoforms suggests that these abnormal transcripts with premature termination codon are commonly removed from the transcript pool [37, 38]. The degradation usually happens in the nucleus [39, 40]. It is estimated about one third of alternative spliced

human genes are subject to degradation through NMD [41]. In one fifth of skipped exons conserved in human and mouse, at least one isoform is subject to NMD [37]. Second, alternative splicing can selectively include exons encoding active protein domains to change gene functions. In a large-scale study on 1,300 alternative spliced events in human, 30% genes show different conserved domains in different mRNA isoforms [42]. Third, majority of the fifty most frequently alternative spliced protein domains are related to protein-protein interactions [43].

Alternative splicing is also a major source of protein diversity. Rather than one gene one protein, the number of proteins is much larger than the number of genes. It has been observed in many high eukaryotes that one gene might have dozens, even thousands of functional products, such as calcium-activated potassium channels and Cadherins in human [44]. An extreme example is Drosophila DSCAM gene, a homolog of human Down syndrome cell adhesion molecule, can potentially generate more than 38,000 isoforms [45]. By alignment and mapping to genomic sequences, many ESTs have been found to cover different portions of same gene, suggesting existence of alternatively spliced mRNA transcripts [44]. Moreover, proteomics study also show evidence that most of the alternative spliced transcripts are actually expressed into protein form in sufficient amount that can be detected [46]. Therefore, a relatively smaller genome can efficiently produce much larger number of proteins through alternative splicing.

Mechanisms of alternative splicing regulation

Regulation of splice site recognition in alternative splicing is a complex dynamic balance among many relevant factors. Two major categories, trans-acting regulatory proteins and cis-acting elements, are generally pivotal in AS regulation. Other sequence signals, such as exon/intron length and core splicing signals, are also involved [1].

Trans-acting splicing factors

Trans-acting splicing factors are proteins involved in (alternative) splicing and splicing regulation. Generally, splicing factors include the core splicing factors in spliceosome assembly, such as snRNPs and U2AF. However, these factors are generally not directly related to alternative splicing regulation. Therefore, we will only discuss the non-snRNP proteins acting as splicing regulators.

Splicing factors can be classified into three major categories: SR-protein family, hnRNP family and others [47]. Serine/arginine-rich (SR) protein family is the most well-studied trans-acting factor family and consists of some remarkably conserved splicing factors (see Table 4-2). The members of SR family generally contain two conserved modules: one to two RNA-recognition motifs (RRMs) on the N-terminal and arginine/serine-rich RS domains with alternating serines and arginines in C-terminal. The RRMs determine the binding specificity to RNA. The RS domain recruits other components in the basal splicing machinery and also interacts with branch site and 5' splice site [48-50]. Binding of SR proteins on pre-mRNA generally promote the inclusion of exons in both constitutive and alternative splicing. Several models have been proposed to explain the mechanism of SR

proteins as splicing enhancer [51, 52]. First, binding of SR proteins can stabilize binding of U1 snRNP and U2AF to 5' and 3' splice sites in exon definition. Second, 5' and 3' splice sites can be juxtaposed in early spliceosome formation through protein-protein interaction between SR proteins and U1 snRNP and U2AF. Third, SR proteins are suggested to facilitate the recruitment of U4/U5/U6 tri-snRNP. Fourth, the inhibition of splicing by hnRNP proteins can be antagonized by SR proteins.

Members of heterogeneous ribonucleoprotein (hnRNP) family belong to another category of trans-acting splicing factor. At least twenty hnRNPs exist in human [53]. Nearly all of the hnRNPs contain one or two RNA-binding domains (RBDs) at the N-terminal, and the C-terminal domains are usually different in different hnRNPs [53, 54]. The auxiliary C-terminal domains may mediate protein-protein interaction such as hnRNP A1-A1 interaction by Glycine-rich domain [55]. It is also related to protein localization, e.g. localization signal in hnRNP C [56]. hnRNPs are believed to antagonize SR proteins and suppress splicing. The antagonistic effect between SF2/ASF and hnRNP A1 to regulate the inclusion of a reporter exon is observed in vivo [57]. The differential ratios of these two antagonists in various tissues are also discovered, which is the key factor of tissue-specific alternative splicing regulation [58]. However, stimulation of alternative splicing by hnRNP A by binding to an intronic binding site on RNA is also observed [59]. This bifold effect in either suppressing or promoting splicing also exist in polypyrimidine tract binding protein (PTB), another member of hnRNP family. PTB suppresses the smooth muscle (SM) exon inclusion in α-actinin through competing with U2AF binding to poly-pyrimidine tract [60, 61]. Nonetheless, PTB

can also bind to an intronic enhancer upstream of exon 4 in calcitonin/calcitonin gene-related peptide to stimulate alternative splicing [62].

Besides SR proteins and hnRNPs, many other splicing factors also play important roles in various cellular mechanisms related to alternative splicing. Neuro-oncological ventral antigen (Nova) protein is a splicing regulator and found to extensively bind to pre-mRNAs in neurons [63, 64]. Nova is believed to be responsible for intron removal. However, suppression or promotion of splicing by Nova depends on the locations of the binding sites. When Nova binds to the binding sites in the flanking introns, it enhances the inclusion of the alternative spliced exon. When binding to exonic binding site, Nova becomes a splicing suppressor by blocking U1 snRNP from correct recognition of 5' splice site [65]. Nova can also block the inhibitory splicing effect from a protein brPTB, which is closely related to PTB and highly expressed in brain [66]. The CUG-BP and ETR3-like factors (CELF) family is another well-studies splicing factor family and consists of seven members in human [67]. The structure of known CELF proteins is highly conserved with three RNA recognition motifs, two at the N-terminal and one at the C-terminal [67, 68]. CELF proteins are involved in alternative splicing of many genes, such as α-actinin, cardiac troponin T, insulin receptor, etc. However, no uniform patterns of suppression or promotion in alternative splicing in different members have been observed. The splicing effect is generally gene-specific. For example, CELF1 (CUG-BP1, CUG binding protein 1) can stimulate splicing in α-actinin [69, 70] and cardiac troponin T [71], but suppress splicing in insulin receptor [72]. The CELF proteins can interact with other splicing factors, such as CELF1 and CELF2 antagonize the inhibitory effect by PTB in the regulation of SM eoxn in α-actinin [69]. Besides splicing

regulation, CELF proteins also play important roles in transcriptional regulation and mRNA stability in cytoplasma [67].

Cis-acting splicing elements

The intronic sequences in the proximity of exons are usually conserved across species. In the conserved alternative spliced exons in both human and mouse, the average length of conserved flanking introns is about 100bp, which is longer than in constitutive spliced exons [73]. This indicates the existence of splicing elements, particularly in alternative spliced genes.

Cis-acting elements reside in pre-mRNAs and are required by both constitutive and alternative splicing [1, 74]. According to their locations and effects on suppressing or promoting splicing, cis-acting elements can be categorized into four classes: exonic splicing enhancer (ESE), exonic splicing silencer (ESS), intronic splicing enhancer (ISE) and intronic splicing silencer (ISS). A cis-acting element can belong to one or multiple categories. For example, Nova binding site can be either ISE or ESS though the motif sequences are identical [65].

Most cis-acting elements are short and highly degenerate [1], e.g. Nova binding site YCAY (Y is C or T) is only four bp long [65]. Therefore, the function of each single element is commonly weak. To compensate for the weak function, many elements present in cluster along the sequence, e.g. Nova binding site [65] and Tra2 [75]. Some are highly repetitive, such as TC repeats of PTB [76] and AC repeat of hnRNP L [77]. Generally, cis-elements must locate within the exons or in the proximity surrounding exons. However, the distal

effect of cis-elements is also observed. For example, an intronic splicing enhancer is identified in the β-casein intron 1 and can stimulate splicing of the downstream exons up to 8 kb [78].

Most of the cis-acting splicing elements are binding sites of trans-acting splicing factors (see Table 4-2 for a list of trans-factors and corresponding cis-elements). By binding to cis-elements, trans-acting splicing factors can regulate alternative splicing by affecting assembly of spliceosome through protein-protein interaction. Moreover, competing for the binding elements may explain the antagonism between two splicing factors. Cis-acting elements can also directly change the secondary structure of pre-mRNAs. For example, in human genes with only one skipped exon, two elements are identified in the flanking introns. One element is C-rich and the other G-rich. The relative position order of these two elements is conserved. Based on the basepairing potential and relative position order, a model of loop structure in the pre-mRNA is proposed to explain the function in alternative splicing regulation [79].

Factors that change splice site choice

The trans-acting factors are required in both constitutive and alternative splicing. Overexpression of many splicing factors has shown the significant effects of splicing factors on exon inclusion. Moreover, antagonism between splicing factors can also affect the fate of exons in splicing. Most of the SR proteins can promote alternative splicing by binding to the cis-element on pre-mRNAs. Several models have been proposed to explain the splicing stimulation mechanisms of SR proteins [51, 52]. Meanwhile, hnRNPs act as antagonists of SR proteins [57]. Therefore, the decision of exon inclusion is mainly determined by the ratios

between SR proteins and hnRNPs in different tissues [58]. Besides SR proteins and hnRNPs, antagonism between other splicing factors has been widely observed. Examples include antagonism between Nova and PTB [66] and between CELF and PTB [69].

Many other pre-mRNA sequence features also affect alternative splicing. Short exons less than 50 bp are subject to be skipped automatically [80]. When surrounding by modest length introns (~500 bp), short exons can be efficiently included if expanded (up to 800 bp). However, when surrounding by large introns (1.5 kb), short exons are generally skipped [81]. Core sequence signals can also affect alternative splicing. Changing from purines to pyrimidines in poly-pyrimidine tract or simply extending it can promote inclusion of short exons [80]. Modification of branch site and 5' splice site to conform to their consensus sequences has the same effect in splicing stimulation [82]. It is widely observed that the alternatively spliced exons are often statistically shorter than other exons and their core splicing signals are weak [1, 52]. Therefore, cis-acting elements, and corresponding splicing factors, are frequently required in correct recognition of splice sites. The compensatory effect of an ESE to 3' weak splice site has been reported in alpha-tropomyosin exon 2 [83]. Also, the requirement of ESE in splicing with weak poly-pyrimidine tract is also observed and related to binding affinity of U2AF to poly-pyrimidine tract [84].

Genome-wide alternative splicing analysis methods

Identification and profiling of alternative splicing events

The most common way to identify alternative splicing events in large-scale is by aligning full-length cDNA or expressed sequence tag (EST) to genomic sequences [1, 85, 86]. Many bioinformatic tools, either common-purpose alignment tools or tools specifically designed for mapping ESTs can be used in the alignment. These include Sim4 [87], BLAT [88], GMAP [89], etc. EST-base method has its limitation and needs to be used with caution. The quality of EST and genomic sequences is the key factor and sequence error may affect the alignment. The sequence similarity to pseudogenes and paralogs may distort the comparison. ESTs may also include some unspliced mRNAs therefore be regarded as intron retention. Bias toward 3' end in ESTs may miss most AS events near 5' end of the transcripts [85]. Also, AS events in genes with low expression usually fail to be detected [90].

Microarray has become one of the most popular techniques in gene expression studies and is widely applied in AS events identification and profiling. Probes of short oligonucleotides are immobilized on chips and the expression of genes can be identified by fluorescent dyes. Three major types of microarray platforms have been reported based on their probe designs: splice junction arrays, exon arrays and tiled genomic arrays [91, 92]. The splice junction arrays use probes within exons (e1, e2 and e3) and spanning exon-exon junction (j1-2, j2-3 and j1-3) (see Figure 1-3). The probes in introns (i1, i2) are used as internal control [2]. The exon-intron structures must be predetermined to design probes, therefore, splicing junction arrays may be subject to errors in the sequence data. Junction arrays are applied in identification of tissue-specific expression patterns in alternative spliced exons [93, 94]. It is

also used to discover the alternative spliced genes and splicing factor correlated with Hodgkin Lymphoma [95]. In exon arrays, probes are designed within annotated and predicted exons (see Figure 1-3) and predetermination of transcript structure is not required. Using Affymetrix exon array in 16 human tissues, numerous new skipped exons are discovered and 17 of them are validated by RT-PCR. Also, 73% detected genes are estimated to undergo alternative splicing in different tissues, which is a little higher than former estimation [96]. Tiled genomic arrays use overlap or non-overlap probes spanning the whole genomic sequences (see Figure 1-3). Therefore, it is possible to detect AS events which are not able to be detected in other arrays. However, the expense and computation difficulty from the added efforts prevent it from wide applications.

Recently, new approaches in expression and splicing analysis have been driven by next-generation sequencing technology, including Roche's 454 sequencing system (http://www.454.com), Solexa/Illumina system (http://www.solexa.com/) and Applied Biosystems' SOLiD system (http://www.appliedbiosystems.com). These systems provide high-throughput sequence data in short reads (up to 50 bp in year 2007 [97] and 400-500 bp currently) with very low cost. Compared to classic microarray platforms in expression analysis, these new technology not only have higher sensitivity in low abundant genes [98] but also provide direct transcript counts rather than signal density of florescent-dye [99]. RNA-Seq approach, coupling high-throughput sequencing and sequence mapping onto genome, has been recently used in detecting splicing variants. Sultan and colleagues used Illumina system to sequence cDNAs from human and mapped the 27-bp reads to human genome. They find 4096 splice junctions which have not been revealed before [100].

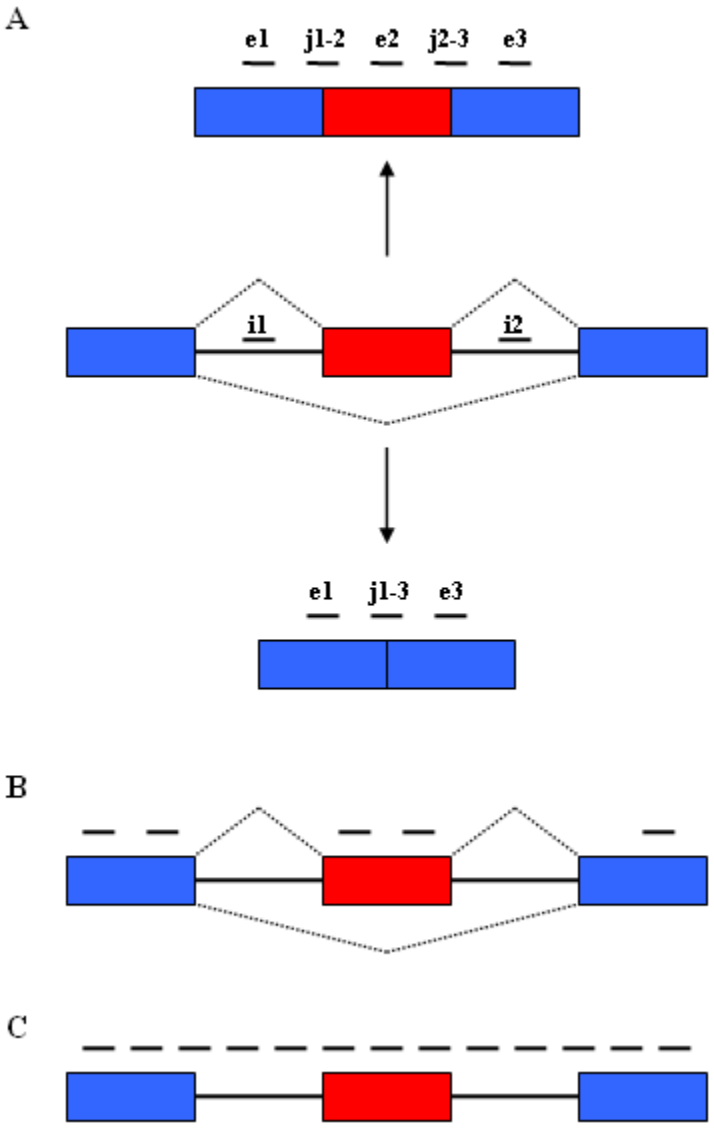


Figure 1-3. Probe design for different microarray platforms [2, 92]**.** A. Splice junction array, probes within exons (e1, e2 and e3), spanning exon-exon junction (j1-2, j2-3 and j1-3) and within introns (i1, i2); B. Exon array, probes all within exons; C. Tile genomic array, overlap or non-overlap probes within genomic sequences.

Motazavi and colleagues applied RNA-Seq to mouse tissues and identified about 3,500 genes undergoing exon skipping [99]. In the near future, as the growth of read size in next-generation sequencing, we may obtain more accurate and cost-efficient expression and splicing map in a very fast pace.

Discovery of cis-acting elements

Another intriguing research topic in alternative splicing is discovery of cis-acting elements. UV-crosslinking is one of the earliest implementation. Radiolabeled RNA fragments with potential elements are incubated with splicing proteins. After UV light treatment, the binding sites can crosslink with proteins and are immune to RNase digestion. The elements protected by splicing factor can be purified and identified after electrophoresis [101, 102]. Cross-linking and immunoprecipitation (CLIP) approach is a recent modification to discover splicing elements in cell extract or live cells and was applied in research of Nova binding sites [65, 103]. The advantage over other crosslinking methods is that CLIP allows the binding in undisturbed live cell environment. After UV crosslinking and RNase digestion, RNA is dephosphorylated, ligated to RNA linker and radioactively labeled. The RNA-protein complex is then purified in immunoprecipitation and electrophoresis. The RNA fragments containing cis-acting elements are amplified in RT-PCR after protein digestion (see Figure 1-4) [103, 104].

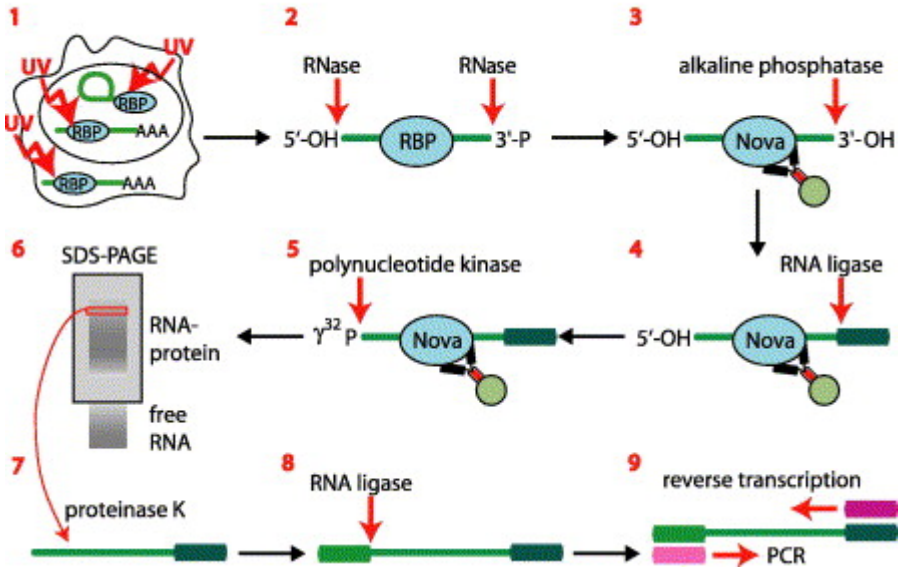


Figure 1-4. Procedure of cross-linking and immunoprecipitation (CLIP) method [104]**.** After UV crosslinking (1) and RNase digestion (2), RNA is dephosphorylated (3), ligated to RNA linker (4) and radioactively labeled (5). The RNA-protein complex is then purified in immunoprecipitation and electrophoresis (6). After protein digestion (7) the RNA fragments are amplified in RT-PCR (9).

Site-directed mutagenesis is another promising approach in cis-acting element discovery. An oligonucleotide library of potential elements is inserted into the exon of a reporter system and then transfected into cells. The cells with spliced transcripts of the reporter gene are then extracted for sequencing. Green fluorescent protein (GFP) is one of the most popular reporter genes to determine splicing silencer [105]. The reporter system of GFP contains three exons. When exon 2 is skipped, exon 1 and 3 encode an active green fluorescent protein. The GFP-

active cells indicate the silencing activity of the oligonucleotides and are extracted for sequencing.

Rather than identifying splicing silencer, systematic evolution of ligands by exponential enrichment (SELEX) *in vivo* or *in vitro* is used to discover exonic splicing enhancer [106, 107]. A mini gene harbors an alternatively spliced exon which requires splicing enhancer to be included in the mature mRNA. A library of random oligonucleotides replaces the natural splicing enhancer. Pre-mRNAs are produced and undergo splicing *in vivo* or *in vitro*. The pool of spliced mRNA is purified and amplified in RT-PCR. The winning fragments which can induce the exon inclusion in each cycle are then used in a new cycle. After several iterations, SELEX yields a limited number of sequence fragments with splicing enhancer activity (see Figure 1-5).

Compare to the experimental approaches described above, computational approaches may be more appropriate in a genome-wide study. Most of computational approaches examine the oligonucleotide usage to find overrepresented patterns associated with (alternative) splicing.

Brudno and colleagues examined all hexanucleotides in introns flanking alternatively spliced exons and compared their frequencies to the control set. The p-value was calculated from resampling from the control sequences. Hexanucleotide UGCAUG, as well as two related pentanucleotide GCAUG and UGCAU, is overrepresented in the downstream introns [108]. Yeo and colleagues used χ^2 values with Yate's correction to compare the frequencies of k-mers (k is 4-6 bp) between alternative spliced exons and control sequences. Calculation of p-value is similar as described by Brudno and colleagues. Many statistically overrepresented k-mers are same as known cis-acting elements [109].

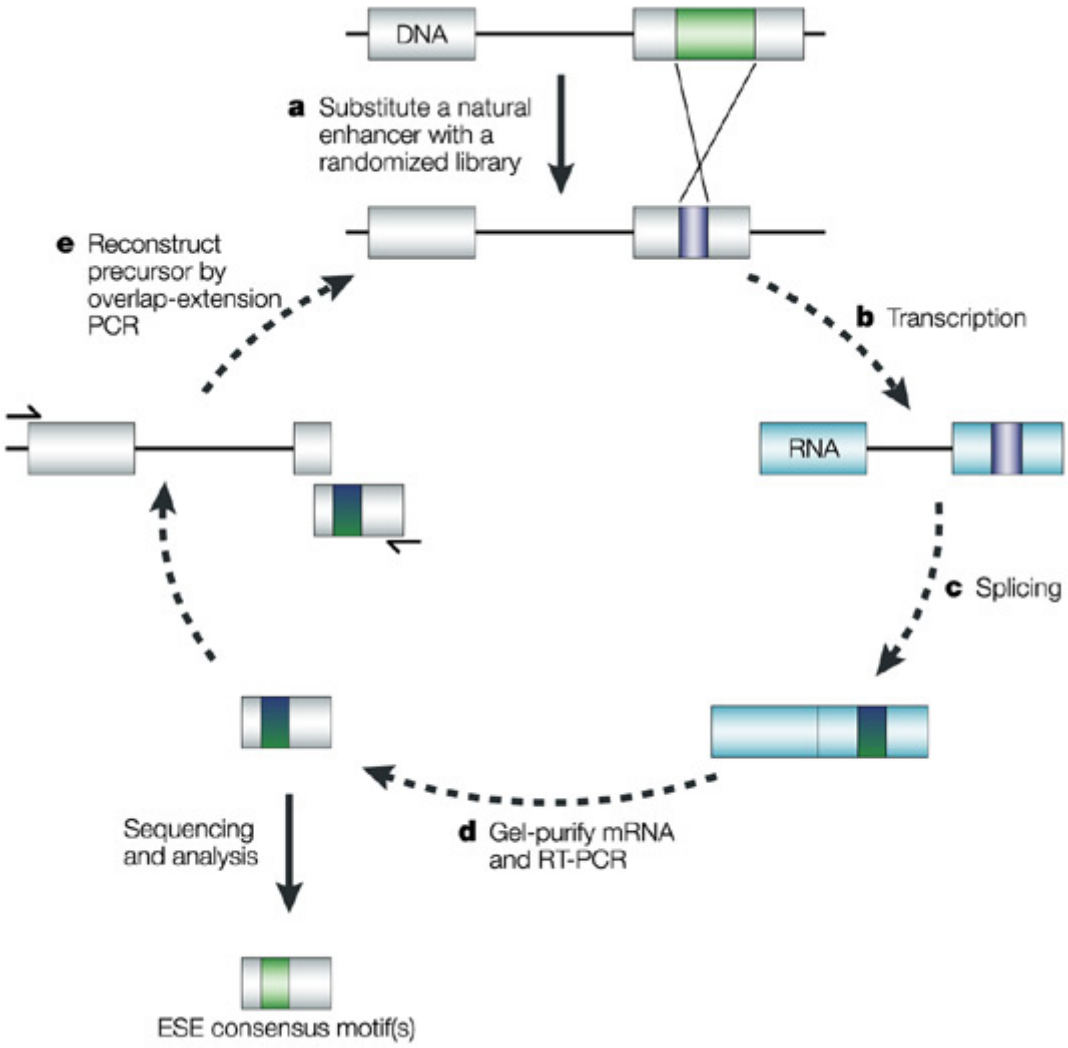


Figure 1-5. Procedure of systematic evolution of ligands by exponential enrichment (SELEX) [107]**.** A library of random oligonucleotides replaces the natural splicing enhancer (a). Pre-mRNAs are transcribed *in vivo* or *in vitro* and undergo splicing (b, c). The pool of spliced mRNA is purified and amplified in RT-PCR (d). The winning fragments of each cycle are then used in a new cycle (e). After several iterations, SELEX yields a limited number of sequence fragments with splicing enhancer activity.

Fairbrother and colleagues used not only overrepresentation but also conservation of splice sites to find hexanucleotides of exonic splicing enhancers. Their RESCUE (Relative Enhancer and Silencer Classification by Unanimous Enrichment) algorithm is based on:

1) Overrepresentation in exons versus introns;
2) Overrepresentation in exons with strong splicing sites versus weak sites because ESE with weak splice sites are under strong selection pressure.

238 hexanucleotides are identified as candidate ESEs and clustered into ten groups. The representative hexanucleotides in each group is proved to have splicing enhancer activity in experiments and point mutations of these sequences cause sharp reduced activities [110]. Zhang and Chasin use similar idea to identify both PESE (Putative ESE) and PESS (Putative ESS). However, to avoid bias in codon usage, they contrast 8-mer frequencies of non-protein coding exons to pseudo-exons and also to 5' untranslated regions of intronless genes. Over 80% of the hexanucleotides identified by Fairbrother can also be found in PESE by Zhang and Chasin [111].

Zhang and colleagues used Support Vector Machine (SVM) to examine the k-mers (k between 4-7 bp) in the 50 bp windows up- and downstream exons. The cis-acting elements are oligonucleotides which can distinguish constitutive spliced internal exons from pseudo-exons. They recursively retrain the SVM using the top half k-mers used in the previous run until the ROC score fall below 90% of that produced from the full k-mer set. Besides core splicing signals, they also find G-rich, C-rich and TG-rich k-mers as candidate splicing elements [112]. Zhang and colleagues then used the same classifier to analyze exons with high or low GC content. Different sets of pentanucleotides are discovered as potential

regulatory elements. This may indicate that exons with different GC content may have different splicing mechanisms and need different splicing factors [113].

Comparative genomics is also used to discover cis-acting regulatory elements. Yeo and colleagues investigated the k-mers (5-7 bp) in the introns 400 bp downstream and 400 bp upstream of human internal exons, respectively. Using the χ^2 value, the k-mers which are highly conserved across four species are regarded as intronic splicing regulatory elements (ISREs) [109]. Goren and colleagues examined the wobble (third) positions of codons encoding proteins which are generally not conserved. Therefore, overrepresented di-codons (hexanucleotide) with conserved wobble positions in human and mouse are identified as exonic splicing regulatory sequences (ESRs) [114].

Concluding remarks

Alternative splicing provides an additional layer of gene regulation in many eukaryotic organisms. Due to the prevalence and important biological processes involved, AS and its regulation has become an intriguing topic to understand the mechanism of genes and gene regulation. Moreover, AS is widely associated with human diseases. It is estimated that over 50% mutations in exons that cause human diseases affect splicing [47]. Numerous human diseases have been observed to be related to alternative splicing, such as several types of malignant tumor, muscular dystrophy, asthma, diabetes, etc [12, 29, 115, 116]. Therefore, study on AS and its regulation may help people to understand the cause and find target and therapy for these diseases. Several interventions, targeting alternative spliced gene products, have been recently applied [12].

Although in the last couple of years new approaches have been implemented to elucidate the mechanisms of alternative splicing, numerous challenges remain. Accurate discovery of cis-acting splicing signals is very difficult. Experimental approaches may give reliable predictions. However, it requires large amount of effort and resource which restricts it from genome-wide studies. On the other hand, the short length, high degeneracy and context-dependency of cis-acting element also make the computational prediction not very trustworthy. Also, prediction of AS and splicing levels from sequence is a challenging task because the sequence signals are commonly context-dependent. Trans-acting splicing factors not only control the splicing of other genes but splicing factors themselves could be splicing targets in regulation [117]. With the accumulation of data for each single component, global analysis may provide integrated splicing code to understand the crucial issues of alternative splicing and its regulation.

Chapter 2

Large-scale Discovery of Regulatory Motifs Involved in Alternative Splicing

Abstract

Alternative splicing is a highly important process in many eukaryotic organisms, but surprisingly little is known about its regulation. Often, this process involves cis-regulatory DNA motifs located within the alternative spliced exons or the flanking introns. To discover such regulatory motifs, we re-analyzed a dataset containing ~3000 skipped exons and their relative expression levels in ten mouse tissues provided by Pan and colleagues [93]. We clustered the skipped exons by their expression profiles and used systematic sampling strategy to sample sequence clusters by setting different number of overall clusters. We applied MEME (Multiple EM for Motif Elicitation) [118] to build a motif dictionary from each cluster. We also investigated the motif instances from different clusters to eliminate the redundant motifs. The discovered motifs were validated by comparing with known regulatory elements and by examining positional bias. By incorporating the levels of alternative splicing, our study provides not only solid targets for further experiments, but also a computational framework applicable in other organisms. This chapter was presented in IEEE 7th International Conference of BioInformatics and BioEngineering (2007) and published in the conference proceeding [119].

Keywords

Alternative splicing, motif discovery; systematic sampling

Introduction

Alternative splicing (AS) generates multiple mature mRNA isoforms from one single gene and plays an important role in cell proliferation, differentiation and apoptosis [1, 2]. Many human diseases are associated with alternative splicing [12, 115] and ~15% of the point mutations resulting in human genetic disease affect splicing [120]. AS is prevalent in many eukaryotic organisms and might generate protein diversity from a relatively small genome. It is estimated that about one to two thirds of human genes undergo alternative splicing [1, 2]. The discrepancy between the limited number of estimated genes and the complexity of proteome indicates a complicated and subtle mechanism between DNA sequences and their products in which AS might be pivotal.

The regulation of alternative splicing is a complicated mechanism, which is affected by many factors, such as splicing sites, exon size, cis-regulatory motifs and trans-factors [1]. Cis-regulatory motifs are short and highly-degenerate sequences located on pre-mRNAs. The function of a single motif is generally weak so they might be present in multiple copies. Mostly the cis-regulatory motifs act as the binding sites for trans-regulatory factors. However, some cis-motifs affect AS by forming loop structure on pre-mRNAs [79]. According to their location and function, they are categorized into four classes: exonic splicing enhancer (ESE), intronic splicing enhancer (ISE), exonic splicing silencer (ESS) and intronic splicing silencer (ISS).

Some computational approaches to identify regulatory motifs involved in AS have been reported [108, 110, 111]. Most commonly, short poly-oligonucleotides (k-mers, k ranges from 5 to 10) are analyzed in order to discover over- or under-representation in comparison

to the background model. Due to the lack of data, levels of alternative splicing were generally not considered in the previous studies. Thus all AS exons were treated as one group to obtain the most common motifs that might regulate hundreds of AS events.

It is widely believed that by using expression profiles to cluster co-regulated genes, the signal-noise ratio in each cluster will increase, improving the quality of discovered motifs [121, 122]. However, determining the cluster boundary remains an open challenge [123]. A too rigid cutoff might exclude sequences containing a motif while a too promiscuous cutoff will introduce noise. Generally, the cluster boundary is determined by setting a fixed overall number of clusters or by optimizing the within- and between-cluster error/distance [121, 122].

We re-analyzed the dataset from a study [93] which measures the relative expression of skipped exons in ten mouse tissues. Motifs were discovered from the potential co-regulated AS events by clustering the expression profiles. Instead of predetermining a fixed number of clusters, we systematically sampled the clusters with more than 10 events under varied setting of overall number of clusters and built our motif dictionary. This sampling strategy might produce very similar motif candidates from nested clusters, which are removed in subsequent post-processing. The significance of motif scores was tested by comparing the occurrences inside the clusters with the background. Overall – after eliminating the redundant and insignificant motifs – our systematic sampling doubled the yield of discovered motifs. We validated our motif dictionary by comparing with experimentally validated regulatory elements, and by assessing their positional bias. Our study provides not only a list of potential regulatory motifs from co-regulated skipped exons but also a computational framework applicable to large-scale motifs discovery in other organisms.

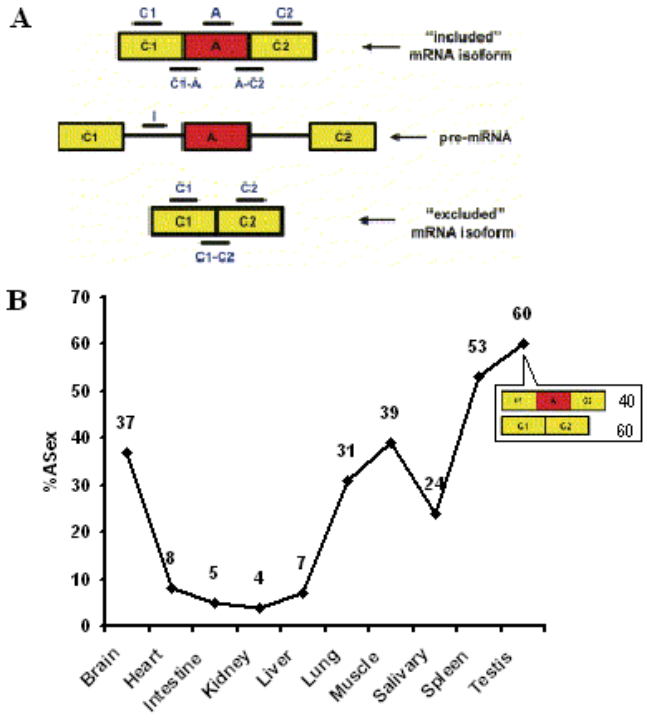


Figure 2-1. Design of probes in microarray and *%ASex* profile. A). Probe design in Pan's quantitative microarray platform. The red box represents skipped exon, yellow boxes are up and downstream exons, six probes are located either within the exon or intron or across the boundaries. B). Exclusion levels in 10 different tissues.

Results

Preparation of the sequence data

The original dataset is from a quantitative microarray platform, which contains 3,126 skipped exons (in 2,647 unique transcripts). The relative abundance (*%ASex*, Alternative Splicing exclusion) of the two different isoforms in 10 tissues, brain, heart, intestine, kidney,

liver, lung, muscle, salivary gland, spleen and testis, were calculated, which is denoted as the percentage of the expression of the short isoform out of the expression of both (see Figure 2-1) [93].

After mapping transcript onto the mouse genome, we got 2566 AS exons, which met the standards of BLAT mapping (see Materials and methods). By excluding those having all *%ASex* values of 0 or 100, we had 2511 skipped exons left. We analyzed three regions of sequences, which are the 3' end of the upstream intron, the AS exon and the 5' end of the downstream intron. The exonic sequences are in full length, while up to 200bp flanking intronic sequences were used (see Figure 2-5).

Discovery of the regulatory motif dictionary

To find co-regulated AS events which might contain similar regulatory motifs, the genes were clustered based on their *%ASex* profiles across ten tissues. The agglomerative hierarchical clustering algorithm was applied with the Ward's method to calculate the dissimilarity between clusters. The Ward's method minimizes the within-cluster variance when choosing which cluster to merge and tends to create compact clusters [124]. Generally, the clustering error is optimized or an arbitrary cutoff is used to determine the number of clusters. However, the optimization does not perform well in a huge dataset with thousands of data points, where the boundaries between each cluster are usually not well-defined due to the high level of noise. Discovering motifs after collecting all possible clusters exhaustively might be one way to solve this problem, but the tremendous computation time restricts it from being widely used in practice. For this reason, we systematically sampled the clusters

containing at least 10 AS events under different settings of overall cluster numbers, ranging from 100, 200, 300, up to 2500. Only one cluster of those containing the same AS events was retained. The number of sequence clusters with at least 10 AS events ranges from 74, when the cutoff was set to 100 overall clusters, to 0, when the cutoff is 2500. Many of these clusters are nested in other clusters due to the hierarchical clustering scheme.

In total we investigated 634 clusters. We discovered 196 motifs by MEME, using a cutoff e-value 0.01. Two-sample t-test was used to test the significance of the motif scores inside the cluster compared with the background for two reasons: 1) to eliminate the motifs which are prevalent both inside the cluster and in the background, such as DNA repeats; 2) to eliminate very weak motifs or motifs covering only a small portion of the sequences within a cluster. The scores of 91 motifs are significantly different as compared to the background with a false discovery rate less than or equal to 0.01.

The discovered motifs might be redundant if they are built from two nested clusters. To eliminate redundancy, we investigated the motif instances on each sequence. On one sequence, if two motif instances, discovered from different but nested clusters, overlap with more than 50% of the motif width, they are considered to be redundant motif instances. And if this redundant motif instances occur in more than 50% of the sequences in the smaller cluster, these two motifs are regarded as redundancy. Only the one with higher/highest information content per column are retained for further analysis. 17 redundant motifs were eliminated, resulting in 74 unique motifs remaining in the final motif dictionary, 29 in upstream introns, 14 in skipped exons, 31 in downstream introns.

We name the discovered motifs as ASM_R*a*_*b*, where ASM stands for alternative splicing motif, R*a* for the region *a* where a motif is discovered (1 for upstream intron, 2 for AS exon and 3 for downstream intron) and *b* for the serial number.

Comparison of PWMs in the motif dictionary

A Pearson's correlation coefficient (PCC) based measure was used to calculate the distance between position weight matrices [125]. The distance was normalized by the length of the shorter PWM, similar to that in Vilo and colleagues [122]. Using Smith-Waterman algorithm, the minimal score of a local alignment is 0, while the maximum will be 1 times the length of the shorter PWM since PCC will not exceed 1 and the length of the best local alignment will be equal to the shorter sequence. This happens when two PWMs are exactly the same or one is nested into the other. By normalization, the distance ranges from 0 to 1, where 0 indicates the exact match of two PWMs or one nested into the other. Motifs were compared only with others in the same region.

Here we distinguish similar motifs discovered from non-nested clusters from those caused by nested clusters. The similar motifs from nested clusters are caused by the hierarchical clustering and systematic sampling scheme. They generally have very similar *%ASex* profiles. The motifs, though share sequence similarity, from non-nested clusters have *%ASex* profiles different enough to be grouped into separate clusters.

8 of the 14 motifs in the skipped exonic region can be grouped into two clusters, whose distances among them are significantly low. Motif ASM_R2_01, ASM_R2_02 and ASM_R2_03 show similarity in 7bp fragments, especially the T/A-G in the second and third

position. The *%ASex* profiles of sequences containing the motif are all below the overall mean (see Figure 2-2). The profiles for the last two motifs are lower than that for the first motif while they are much weaker with lower information contents in that 7bp fragment. In up- and downstream introns, only motifs with low-complexity were discovered to be similar.

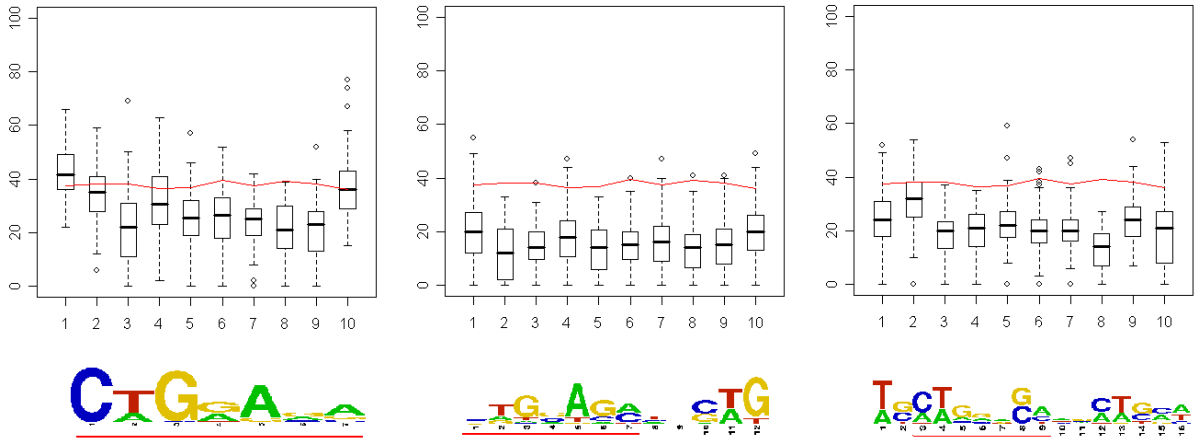


Figure 2-2. Predicted motifs with high similarity to each other. The motifs shown are ASM_R2_01, ASM_R2_02 and ASM_R2_03. Box-plots show the *%ASex* values of genes with the predicted motif in ten tissues and the red line is the mean *%ASex* of all 2511 AS events. The positions with red underlines are the similar fragments in the motif. The tissues from left to right are brain, heart, intestine, kidney, liver, lung, muscle, salivary gland, spleen and testis.

Comparison between discovered motifs and experimentally validated elements

To validate the discovered motifs, we compared our motif dictionary by MAST [126] to experimentally validated regulatory elements which were collected from two resources:

Alternative Exon Database (AEDB) from EBI [127] and the list of intronic elements from Ladd and Copper [1]. Together, 381 regulatory sequences were obtained from AEDB and Ladd's paper.

We compared our motif dictionary to this collection and found 13 candidate motifs for an e-value threshold of 0.01. We will describe three best matches below.

Motif ASM_R1_12 (see Figure 2-3) contains repeats of TC with width 18bp and total information content 17.5. It has strong similarity (e-value1.4E-4) with an intronic enhancer in human Chloride channel (CIC-1), which causes the intron retention in myotonic dystrophy [128]. The *%ASex* values of all the AS events within this cluster dramatically drop in muscle (average 60% compared to ~100% in intestine, kidney, liver, lung, salivary gland, spleen), which indicates the higher inclusion of the skipped exon. This consistent drop of *%ASex* also occurs in brain, heart and testis.

Repeats of TG occur in motif ASM_R3_28 frequently (see Figure 2-3), up to 17 times. This motif has width 20 and total information content 21.7. This motif is similar to an intronic silencer (e-value 3.2E-7) in human cystic fibrosis transmembrane conductance regulator (CFTR). The skipping of exon 9 in CFTR is regulated by poly(TG) and is positively correlated with the amount of poly(TG) [129]. Surprisingly, the transcripts containing motif ASM_R3_28 exhibit low *%ASex* values in brain, liver and testis (*%ASex* ~10%), indicating the proper splicing of the skipped exon. This is in contradiction to the silencing function of the CFTR regulatory element in human. One possible explanation might be owing to the different locations of these two motifs, ASM_R3_28 is located in the downstream while poly(TG) of CFTR in the upstream. Same regulatory elements might act

reversely if residing in different locations in alternative splicing because owing to competing for the trans-factors, thus affecting their correct binding [130].

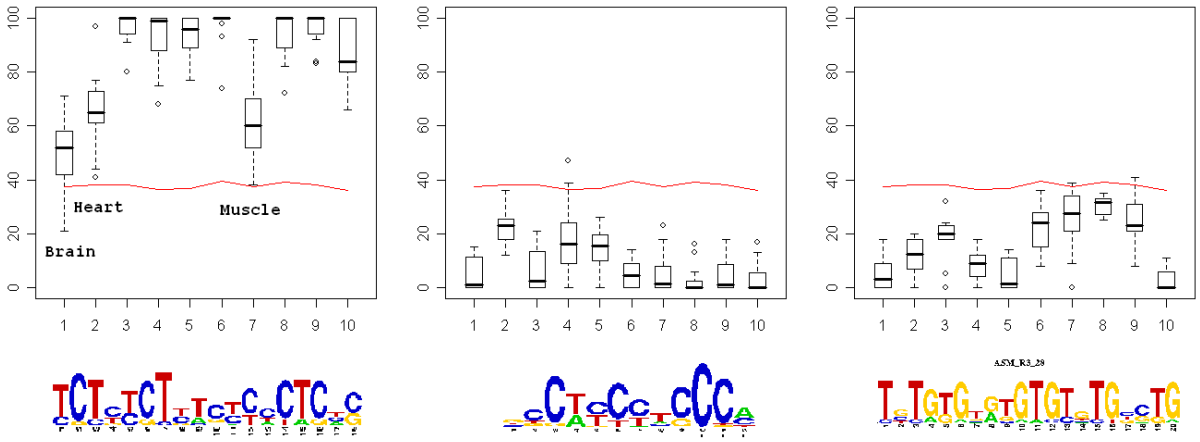


Figure 2-3. Predicted motifs with similarity to known splicing elements. The motifs shown are ASM_R1_12, ASM_R1_17 and ASM_R3_28. Box-plots show the *%ASex* values of genes with the predicted motif in ten tissues and the red line is the mean *%ASex* of all 2511 AS events. The tissues from left to right are brain, heart, intestine, kidney, liver, lung, muscle, salivary gland, spleen and testis.

Motif ASM_R1_17 (see Figure 2-3) has similarity to a rat silencer ISS1 with e-value 2.6E-3. This motif contains a weak repeat of TCCC. The whole width is 12 and the total information content is 12.9. The rat ISS1 silencer regulates two mutually exclusive exons IIIb and IIIc in fibroblast growth factor receptor 2 (FGF-R2) [131]. The repression of the exon inclusion functions via the binding of poly-pyrimidine tract binding protein (PTB). The *%ASex* profiles for the genes containing motif ASM_R1_17 shows very low *%ASex* values (~5%) in brain with a boost (~25%) in heart.

One potential problem for this comparison was caused by sequences with low-complexity. For example, motif ASM_R3_24, which is T-rich but does not contain $(TTTG)_n$ repeats, has strong similarity (e-value 4.1E-4) to a regulatory $(TTTG)_n$ repeat. So for such motifs, more sophisticated sequence analysis is needed.

Positional bias of the discovered motifs

It is widely believed that some binding motifs must be located within certain distance to function properly, such as polypyrimidine-tract binding (PTB) protein binding site which is located in the upstream intron and close to the 5' splice site. To validate our motif dictionary, we investigated the positional bias of the discovered motifs using Kolmogorov-Smirnov (K-S) statistic. Let $l_1, l_2, \ldots, l_n$, be the lengths of sequences with a cluster, the distribution of the motif positions will be uniform if the sequence lengths are constant, assuming a uniform distribution in each sequence. However, sequence lengths are usually not equal, which makes a motif occur less likely when the distance increases. We thus generated 1000 sets of motif positions under uniform assumption of each sequence and pooled them as the reference distribution. Since the p-value for K-S statistic is distribution-free only for continuous random variables, we used a Monte-Carlo strategy by calculating all the K-S statistics between each set of 1000 simulated motif positions and the reference, then obtained the p-value from the ratio of more extreme K-S statistic than that between the motif positions under investigation and the reference. By setting FDR to be 0.01, we have 6 motifs which show positional bias. Interestingly, all of them are located in the upstream intronic regions and are close to the 5' splice sites of the skipped exons. Motif ASM_R1_05, ASM_R1_22

and ASM_R1_29 are all highly TC-rich and are very close to the 5' splice site, even including the 3' AG of the intron in motif ASM122 (see Figure 2-4).

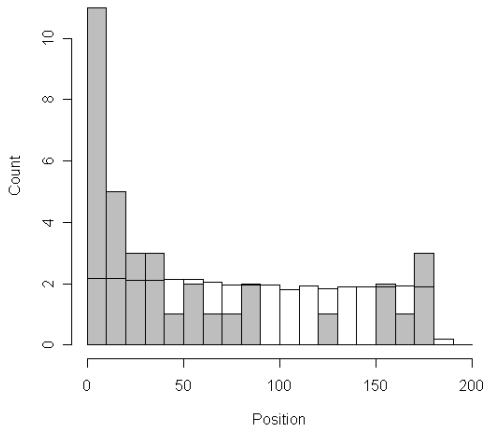


Figure 2-4. Positional bias of motif ASM_R1_29. The grey histogram shows the distribution of the positions of the discovered motif, while the white histogram gives the simulated reference distribution under the uniform assumption. The reference distribution is normalized to have the same number of occurrence as the distribution of the predicted motif.

Discussion

Due to the complexity of the mechanism of alternative splicing, the genome-wide discovery of regulatory motifs still remains as a challenging task. Due to the lack of data, previous studies did not take the level of splicing into account. AS events were often treated as one group, and compared with the background to identify over- or underrepresentation of motifs. Starting from a dataset measuring the relative expression of skipped exons in mouse,

we clustered the skipped exons using their expression profiles to find co-regulated exons. Then we systematically sampled the clusters for motif discovery. Potential regulatory motifs were identified and tested for significance from grouped exons and flanking sequences without pre-determining an optimal number of clusters.

In the context of discovering transcriptional factor binding sites (TFBS), clustering of co-regulated genes has proved to be an effective way to increase the signal-noise ratio, and to improve the quality of motif prediction [121, 122]. However, finding the "right" set of clusters remains an unsolved problem [123]. In our study, we systematically sampled sequence clusters under different settings. The advantage of this strategy is that we do not have to pre-set a cutoff to determine the cluster boundaries, and hopefully identify more biologically meaningful motifs. However, we might re-discover motifs multiple times from nested clusters. We address this problem by comparing the degree of sequence overlap among motif instances.

Our systematic sampling strategy might not cover all possible clusters and miss some of the regulatory motifs. We are still working on better sampling strategies, e.g. exhaustively collecting all possible clusters and keeping those containing $\geq k$ different sequences from others. The challenge is to find a tradeoff between computation time caused by exhaustive search and coverage of clusters containing motifs.

Also, there is no guarantee that every sequence grouped in the same cluster is co-regulated and contains the same motif(s). Although the signal-noise ratio has been dramatically increased by clustering, the robustness of motif discovery tools becomes a pivotal issue to the quality of motif discovery. Reddy and colleagues investigated the impact of noisy decoy

sequences (NDS), which don't contain motifs, in discovery of transcriptional factor binding sites. The average information content decreases exponentially while more NDSs are included. And the percent of positive matches of predicted STE12 TFBS drops from 0.7 to 0.3 by using BioProspector and 0.95 to 0.3 with BioProspector plus motif position information [132]. Although MEME allows zero or one motif in each sequence, its results are still affected by NDS.

In summary, we clustered 2511 skipped exons by their relative expression in ten mouse tissues. We searched 634 individual clusters using MEME. We tested for statistical significance and eliminated redundant occurrences, resulting in a dictionary of 74 regulatory motif candidates. We compared our dictionary to database of experimentally validated regulatory elements and found thirteen significant matches. We also tested for positional bias and found six motifs, most of which are TC-rich, exhibiting positional bias in the upstream introns. In total, we discovered more than 60 new candidates for regulatory motifs. Compared to previous studies, our motif candidates are derived from small groups of AS events, characterized by well distinct AS profiles. We hypothesize that in contrast to regular AS enhancers and silencers, these motifs are involved in a more complex and subtle AS regulation mechanism.

Our study provides not only a dictionary of potential regulatory motifs for further analysis of alternative splicing regulation, but also a computational framework applicable to motif discovery in other organisms.

Materials and methods

Dataset

The 3,126 skipped exons (in 2,647 unique transcripts) are from Pan and colleagues. They used a quantitative microarray platform to measure the expression levels of different isoforms in 10 tissues – brain, heart, intestine, kidney, liver, lung, muscle, salivary gland, spleen and testis. The *%ASex* (percent alternatively spliced exon exclusion) is defined as the percentage of the expression of isoform excluding the skipped exon out of the total expression level of both [93].

To obtain the nucleotide sequences of the skipped exons and the flanking exons/introns, we mapped the transcripts containing skipped exons onto the mouse genome by alignment. Using the Batch Retrieve tool from NCBI, the transcripts containing the AS exons were retrieved by GenBank IDs. The mouse whole genomic sequence was NCBI Build 36 v.1 released in May 2006. We used BLAT to align each transcript to the mouse genomic sequences [88] after trimming polyA tail by trimest from EMBOSS suite [133]. We defined a match as ≥95% identity along the whole transcript. Each transcript might contain multiple partial matches to the genomic sequence, indicating potential exons. Each partial match separated by ≤5bp was merged to form longer exons. By comparing the relative position on the transcripts between the skipped exons and the whole set of potential exons from BLAT alignment, we mapped the AS exons onto the mouse genome if the difference was ≤5bp on both ends.

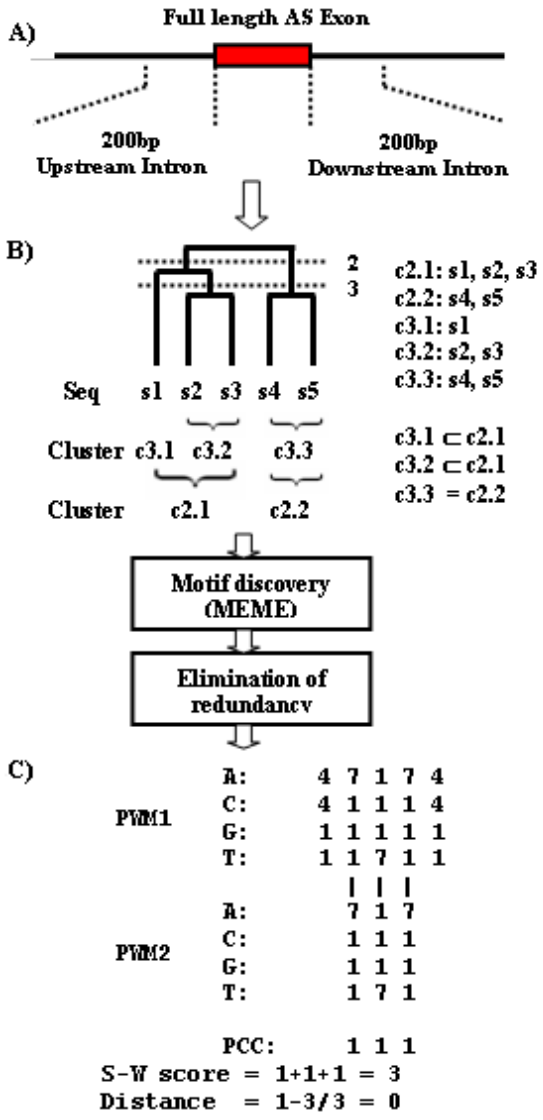


Figure 2-5. Workflow of the computational framework. A) Three regions examined in our study: upstream intron, AS exon and downstream intron. B) Hierarchical clustering by setting different numbers of total clusters, the clusters c3.1, c3.2 are nested in c2.1. C) Comparison of two PWMs which have a perfect match of distance 0.

Clustering of the *%ASex* profiles

By clustering AS events based on their expression profiles, the signal-noise ratio is often increased assuming the AS events are co-regulated and contain similar regulatory motifs. We used Euclidean distance and hierarchical clustering to group skipped exons with similar *%ASex* profiles across ten tissues (see Figure 2-5). We systematically sampled clusters with more than 10 AS events by setting different overall numbers of clusters, ranging from 100, 200, 300, and up to 2500. Only one set was kept if multiple sets contained exactly the same skipped exons.

Discovery of motif dictionary

The DNA sequences of skipped exons in each cluster were analyzed by MEME [118] to discover potential regulatory motifs. We analyzed three regions of the DNA sequences flanking the AS exons: 3' end of the upstream intron, AS exon and 5' end of the downstream intron (numbered region 1, 2, 3). The width of motifs was set to 5-20bp. Each sequence within the cluster was not required to contain a motif (by –mod zoops option in MEME). The candidate motifs are those with e-value less than 0.01.

We used two-sample t-test to test if the log-odds scores of the candidate motifs were significantly different from the background (motif scores of sequences outside the cluster). The best score in every sequence was calculated by summing the log-odds score of each position of the motif. The false discovery rate [134] of 0.01 was used to adjust the p-value in the multiple testing.

To eliminate redundancy caused by nested clusters, we investigated the motif instances on each sequence. On one sequence, two motif instances, discovered from different but nested clusters, are defined to be redundant if their positions overlap with more than 50% of the motif width. And if this redundant motif instances occur in more than 50% of the sequences in the smaller cluster, these two motifs are regarded as redundancy. Only the one with higher/highest information content per column were retained for further analysis.

Comparison of Position Weight Matrices

To find similar motifs, we compared PWMs of the discovered motifs. We used Pearson's correlation coefficient (PCC) to calculate the scores between columns of Position Weight Matrices (PWMs) [125]. Let $\mathbf{C_1} = (p_{1A}, p_{1C}, p_{1G}, p_{1T})^T$ and $\mathbf{C}_2 = (p_{2A}, p_{2C}, p_{2G}, p_{2T})^T$ be two columns in PWMs, the match score for two columns is

$$s(\mathbf{C_1}, \mathbf{C}_2) = \frac{\mathrm{cov}(\mathbf{C_1}, \mathbf{C}_2)}{\sqrt{\mathrm{var}(\mathbf{C_1})\,\mathrm{var}(\mathbf{C}_2)}} \qquad \textbf{(2.1)}$$

We applied Smith-Waterman algorithm to find the optimal local alignment between two PWMs with a large gap penalty (100) to prohibit gaps.

We computed the distance *d* between PWMs by

$$d(\mathrm{PWM}_1, \mathrm{PWM}_2) = 1 - \frac{sw(\mathrm{PWM}_1, \mathrm{PWM}_2)}{\min(l_{\mathrm{PWM}_1}, l_{\mathrm{PWM}_2})} \qquad \textbf{(2.2)}$$

where *sw* is the match score of the optimal local alignment from Smith-Waterman algorithm, *l* is the length of a PWM. The cutoff was determined by the 1% quantile of distances between all pairs of PWMs.

Test of motif positional bias

The position of motif occurrence, relative to the splice site, ranges from 1 to *l-w+1*, where *l* is the length of sequence and *w* is the width of motif. For a set of *n* sequences with length l_1, l_2, …, l_n, assuming the occurrence of a motif is uniformly distributed along the sequence, we randomly generated one occurrence of motif for each sequence in the cluster and repeated for 1000 times. We pooled all the 1000*n* motif positions to obtain the reference distribution. To test whether the distribution of positions of the discovered motif is different from the reference distribution, we used Kolmogorov-Smirnov (K-S) statistic, which is the biggest difference between two empirical cumulative distributions:

$$D = \sup_x \left| F(x) - F_0(x) \right| \qquad \textbf{(2.3)}$$

where *x* is the positions of the motif, *F(x)* and $F_0(x)$ are the position distributions of the discovered motif and the reference. Instead of obtaining the p-value from K-S test, which is distribution-free only if continuous, we first found the distribution by calculating K-S statistic between each set of randomly generated positions and the reference. The p-value is calculated as the number of occurrence of larger of equal K-S statistic over 1000. The false discovery rate of 0.01 is used as the cutoff (see Figure 2-6).

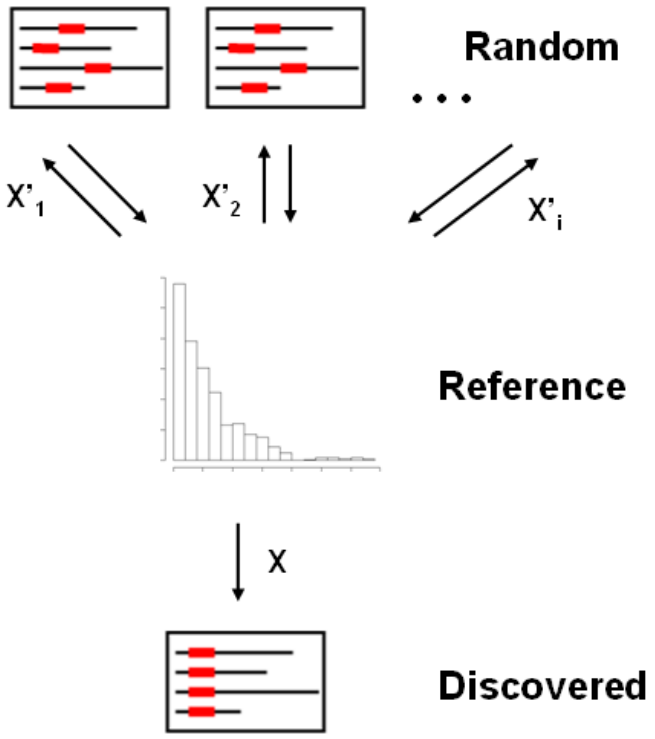


Figure 2-6. Test for positional bias of discovered motif. 1000 random set of motifs were generated under uniform distribution and pooled as reference. The Kolmogorov-Smirnov statistic was calculated between each set of random motif and the reference to get an empirical distribution. The p-value between the positions of the discovered motif and the reference was obtained from the empirical distribution.

Chapter 3

Analysis of Cis-regulatory Motifs of Cassette Exons in Central Nervous System by Incorporating Exon Skipping Rates

Abstract

Identification of cis-regulatory motifs has long been a hotspot in the study of alternative splicing. We propose a two-step approach: we first identify k-mer seed motifs by testing for enrichment and significant differences in exon skipping rate, then a local stochastic search is applied to refine the seed motifs. Our approach is especially suitable to discover short and degenerate motifs. We applied our method to a dataset of CNS-specific cassette exons in mouse and discovered 15 motifs. Two of these motifs are highly similar to validated motifs, Nova and hnRNP A1 binding sites. Four motifs show positional bias relative to the splice sites. Our study provides a dictionary of sequence motifs involved in the regulation of alternative splicing in CNS tissues, and a novel tool to detect such motifs. This chapter is accepted by the 5th International Symposium on Bioinformatics Research and Applications (2009) and will be published in Lecture Notes in Bioinformatics.

Keywords

Alternative splicing, motif discovery, exon skipping rate

Introduction

Alternative splicing (AS) may produce multiple mature mRNA isoforms from one single gene, contributing essentially to protein diversity [1, 2]. AS has a pivotal role in many biological processes, such as cell proliferation, differentiation and death. It is believed that AS is also involved in the regulation of gene expression via nonsense-mediated decay and

selective inclusion of protein domains [42, 43]. About 50% of all mutations in exons causing human disease affect splicing [47].

Cis-regulatory motifs in combination with other factors such as splicing sites, exon size, etc, play an important role in the regulation of alternative splicing [1]. Often, cis-regulatory motifs are degenerate short sequences on pre-mRNAs, acting as the binding sites for trans-regulatory factors or basepairing to form loop structure [79]. According to their location and function, they are categorized into exonic splicing enhancer/silencer (ESE/ESS) or intronic splicing enhancer/silencer (ISE/ISS).

Several computational approaches to identify cis-regulatory motifs in (alternative) splicing have been reported [108, 110, 111, 135]. Most of them focus on enumeration of k-mers (5-10 bp) to find the statistically overrepresented patterns. However, k-mers have limited flexibility to represent AS motifs due to the higher degeneracy of these motifs [1].

Although adding biological features may improve the quality of the predictions (see ref. [136] for a comprehensive review), only a few approaches for transcriptional factor binding sites (TFBSs) have been implemented. Bussemaker and colleagues identified k-mers with significant effect on mRNA abundance in a linear regression experiment [137]. Conlon and colleagues identified candidate TFBSs using MDScan in combination with model selection in linear regression [138]. Smith and colleagues analyzed TFBS via weighted log-odds scores using CHIP-chip data [139].

This paper provides a novel approach to discover cis-regulatory motifs involved in the regulation of alternative splicing which makes use of exon skipping rate measurements. Similar measurements become more and more abundant with the growing use of microarray

platforms targeting AS. It is believed that DNA motifs involved in alternative splicing are generally short and highly degenerate [1]. A simulation study demonstrates that our approach is especially suitable for detecting such motifs. We applied our method to CNS-specific (Central Nervous System) cassette exon data in mouse [94], and compared the predicted motifs with the Conserved Domain Database [140] to ensure that no protein domains are wrongly reported as motifs in coding regions.

Results

Motif re-discovery in simulated sequence data

We compared our approach to the approach combining sequence grouping by expression and MEME analysis in simulated sequence data.

We combined all artificial sequences with skipping rates greater than 0.5 and used MEME for motif re-discovery. This procedure is similar to the traditional approach – selecting sequences by similar expression profiles and doing motif discovery thereafter. Skipping rate of 0.5 is the expected mean in the simulated data. In the non-overlap group, MEME only analyzed the sequences containing an implanted motif. In the overlap group, the chance of also including sequences without motif into the MEME analysis is about 16%. This corresponds to the (more realistic) case that the start sequence set is not perfect. To compare the performance between our approach and MEME, we calculated the distance between the implanted and re-discovered motifs using formula 3.3.

Table 3-1. Mean distances between the re-discovered and implanted motifs using different methods. The distance is calculated as described in Materials and methods and is normalized by width of implanted motifs. The standard deviation is given in the parenthesis. The numbers in bold (upper-right) show the better performance of our approach in the prediction of short and degenerate motifs.

Width (bp)	Skipping Rate	Information Content 1.15 New	1.15 MEME	0.95 New	0.95 MEME	0.75 New	0.75 MEME	0.55 New	0.55 MEME
4	overlap	**0.386 (0.190)**	**1.687 (0.448)**	**0.524 (0.147)**	**1.588 (0.716)**	**0.540 (0.108)**	**1.953 (0.563)**	**0.578 (0.069)**	**1.855 (0.618)**
	non-overlap	**0.410 (0.172)**	**1.937 (0.586)**	**0.527 (0.140)**	**1.559 (0.789)**	**0.574 (0.091)**	**1.837 (0.548)**	**0.577 (0.080)**	**1.805 (0.492)**
6	overlap	**0.236 (0.051)**	**0.633 (0.591)**	**0.300 (0.100)**	**1.104 (0.622)**	**0.357 (0.093)**	**1.252 (0.430)**	**0.431 (0.061)**	**1.294 (0.366)**
	non-overlap	**0.243 (0.037)**	**0.561 (0.565)**	**0.313 (0.087)**	**1.155 (0.433)**	**0.398 (0.101)**	**1.302 (0.409)**	**0.410 (0.061)**	**1.317 (0.420)**
8	overlap	0.240 (0.048)	0.102 (0.026)	**0.282 (0.073)**	**0.581 (0.472)**	**0.341 (0.069)**	**0.892 (0.355)**	**0.354 (0.051)**	**0.991 (0.300)**
	non-overlap	0.240 (0.051)	0.112 (0.043)	**0.285 (0.062)**	**0.666 (0.514)**	**0.345 (0.072)**	**0.892 (0.318)**	**0.371 (0.047)**	**0.941 (0.292)**
10	overlap	0.224 (0.071)	0.074 (0.014)	0.284 (0.066)	0.147 (0.144)	**0.331 (0.056)**	**0.892 (0.334)**	**0.336 (0.039)**	**0.779 (0.218)**
	non-overlap	0.245 (0.063)	0.075 (0.016)	0.282 (0.051)	0.122 (0.071)	**0.331 (0.049)**	**0.892 (0.294)**	**0.333 (0.036)**	**0.730 (0.223)**

Table 3-1 gives the mean and standard deviation of the distance (normalized by the width of implanted motifs) in 100 replicates under different combinations of parameters. Generally, the mean distance increases with decreasing motif conservation (information content). Similarly, the mean distance decreases with increasing motif width.

MEME, as a well-known and mature motif discovery tool, performs very well in the re-discovery of longer and more conserved motifs (the left bottom corner of Table 3-1). Under these conditions, MEME performs similar to or slightly better than our approach. However, when analyzing shorter and more degenerate motifs, our approach outperforms MEME. Overall, our approach performs steadily in all combinations of settings, even when the implanted motifs are only 4 bp long. Although the short width and degeneracy also deteriorate the performance of our approach, the magnitude is much smaller than to MEME.

Overview of the CNS-specific cassette exon data

We downloaded a dataset from Fagnani and colleagues [94] which targets exon skipping rate of more than 3,000 cassette exons (also called skipped exons) using a splicing junction microarray platform. The exon skipping rate (called *%ASex*) was measured in 23 tissues (bladder, brain, cerebellum, cortex, eye, heart, hindbrain, intestine, kidney, liver, lung, mammary, midbrain, muscle, ovary, salivary, spinalcord, spleen, stomach, striatum, teeth, testis, tongue) and is denoted as the abundance of the short isoform out of the abundance of both isoforms. Furthermore, they also provided about 100 CNS (Central Nervous System)-specific cassette exons which showed significant difference of skipping rates in seven CNS tissues (brain, cerebellum, cortex, hindbrain, midbrain, spinalcord and striatum) compared to

all other tissues. We mapped these CNS-specific skipped exons to the mouse genomic sequences using BLAT and got the sequence of cassette exon, up- and down-stream exons/introns in 75 of them. We did motif discovery using our approach in seven regions, including skipped exon and flanking exons and introns.

Motif discovery in CNS-specific exon skipping events

We applied our motif discovery approach to the DNA sequences of seven regions of 75 CNS-specific exon skipping events from Fagnani and colleagues [94]. By incorporating logit-transformed skipping rates, we identified 15 candidate motifs (see Table 3-2) with width between five and ten basepairs.

To investigate the effect of a motifs on the exon skipping rate, we performed a regression of the motif log-odds scores against the exon skipping rates using the model $y = a+b\times log\text{-}odds+e$. We first determined a more accurate and stringent motif cutoff by shuffling the sequences and calculated the log-odds scores for all possible sliding windows. The 99.5% quantile of all the log-odds scores was used as the cutoff to make sure the random occurrence of motif match was less than one per sequence. We then calculated the sum of log-odds score for the matches to the predicted motifs in each sequence. Eleven predicted motifs show a significant effect (α=5%) on the skipping rates in the simple linear regression for each motif separately. Two motifs have a significance level between 0.05 and 0.1. The corresponding R^2 values are between 6% and 23% for each significant motif. By using all eleven significant motifs in multiple linear regression, the overall R^2 can reach 62% and the adjusted R^2 is 55%.

This is slightly better or comparable to a similar approach searching for transcriptional factor binding sites [137, 138].

The occurrence and log-odds score sums of several predicted motifs are strongly correlated even though they were predicted independently in different regions. The scores of motif 9 (CNYGK) and motif 14 (VYCAK) show positive correlation (Spearman's correlation coefficient 0.50 with p-value 0, see Figure 3-1). The p-value for the Spearman's correlation coefficient is calculated by shuffling between the motif scores and sequences for 1000 times and computing the chance of getting more extreme values. Both motifs act as intronic splicing enhancer. More interestingly, motif 9 and motif 14 reside in the down-stream intron, but motif 9 in the 5' end and motif 14 in the 3' end. This might suggest the cooperative role of these two motifs. Motif 11 (CHDCNBHB) and motif 14 are also positively correlated. They both reside in the 3' end of the down-stream intron and act as intronic splicing silencer. The Spearman's correlation coefficient between them is 0.41 with p-value equal to 0. Negative correlation between predicted motifs is also observed (see Figure 3-1). Motif 11 and 14 are both negatively correlated with motif 4 (HWKATTWWTD) with Spearman's correlation coefficients -0.37 and -0.33 (p-value 0 and 0.002). The former two motifs are intron enhancer while the latter one acting as silencer in up-stream intron.

In summary, we identified 15 candidate motifs. Three of them occur in exons, one in the up-stream, one in the cassette, and one in the down-stream exon. The remaining motifs occur in introns. Candidate motifs show different correlation with the skipping rates. Seven have a positive correlation with the skipping rate, indicating that they are splicing silencer. The remaining eight are negatively correlated and might be splicing enhancer.

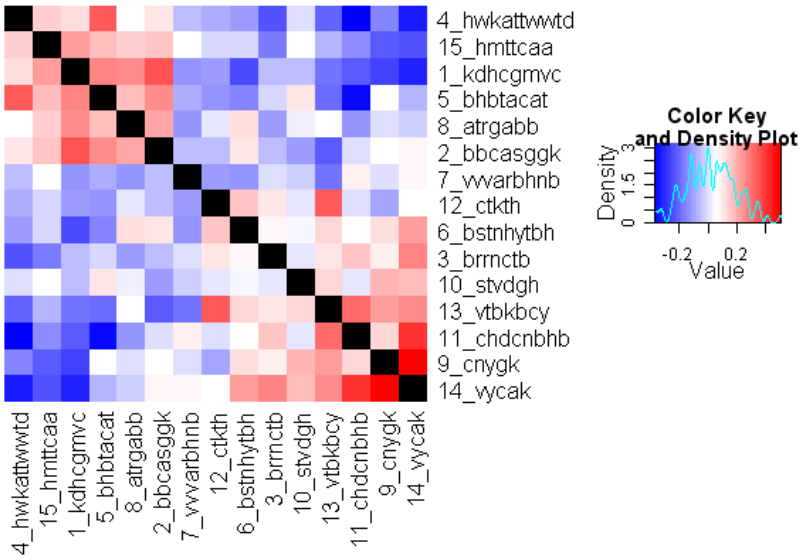


Figure 3-1. Correlation between the predicted motifs. The colors encode the correlation between motifs. Red indicates a positive correlation and blue negative correlation. The diagonal, the correlation of each motif to itself, is in black. The ID gives the motif number and consensus sequence in IUPAC symbols.

Positional bias of the discovered motifs

DNA motifs involved in (alternative) splicing are usually located in the vicinity of 5' or 3' splice sites. Therefore, positional bias is often used as a validation of predicted AS motifs [109, 141]. We used Kolmogorov-Smirnov (K-S) statistic and a Monte Carlo method to test the positional bias within the sequences used for motif discovery. Because the sequence lengths are unequal, the background will not be a flat line (see Figure 3-2).

Using a cutoff false discovery rate of 0.1, we found four motifs showing positional bias relative to the splice sites. Motif 6, 12 and 13 show a preference occurring close to the 3' splice site while all of them are TC-rich and located in the 3' end of the flanking introns. This might indicate the similarity between these motifs and poly-pyrimidine tract binding sites. Motif 10 is located in the 5' end of the downstream intron and has positional bias toward the 5' splice site.

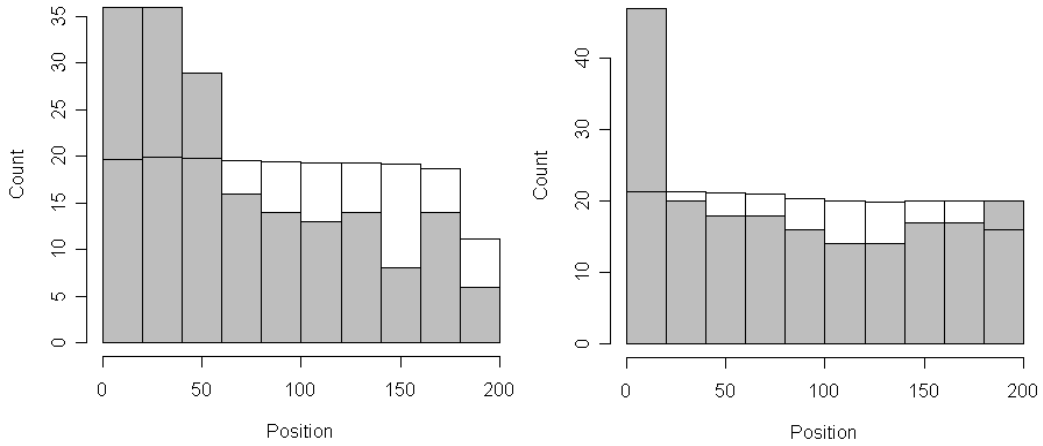


Figure 3-2. Positional bias of the predicted motifs 6 (left) and 12 (right). The grey histogram shows the observed positions of the predicted motif, while the white histogram gives the background distribution under the uniform assumption. The background distribution is normalized to have the same total number of occurrences as the predicted motifs.

Comparison to protein domains

Genes containing same active protein domains may be co-regulated and have similar splicing pattern in certain tissues or developmental stages, thus erroneously regarded as splicing motifs correlated with skipping rates. Therefore, we checked the Conserved Domain Database (CDD) to verify that the predicted motifs do not coincide with known protein domains. We retrieved protein sequences for each gene and searched CDD in NCBI [140]. Using the default cutoff of e-value 0.01, we did not find any significant enrichment of protein domains in the sequences containing predicted AS motifs (no more than two occurrences). Therefore, we conclude that there is no significant overlap between our motifs and protein domains in CDD.

Predicted motifs with match to known AS motifs

Compared to transcriptional factor binding sites, relatively little information about AS motifs is known and validated. However, some AS motifs have been reported [65, 142, 143], and by comparing our predictions to these motifs we found several close matches.

Motif 14 (see Table 3-2) is discovered in the 3' end of the down-stream intron. It is negatively correlated with the exon skipping rate, suggesting a role as intronic splicing enhancer. The motif has a consensus of VYCAK or [ACG][CT]CA[GT] which is similar to the binding sites of Nova. Nova is a neuron-specific alternative splicing factor which affects spliceosome assembly and removal of introns, thus regulates the inclusion of exons [65, 103]. The binding sites of Nova is YCAY, which is almost identical to VYCAK except the last position (Y partially matches with K because Y is C or T and K is G or T). The location of a

Nova binding site affects its biological function. If YCAY occurs within or in immediate up-stream of AS exons, Nova acts as a silencer by blocking the binding of U1 snRNP. However, when YCAY occurs in the down-stream intron, Nova becomes an intronic enhancer which boosts the inclusion of AS exons [65]. The location and biological function in the latter case, coincides with our motif 14. Most (61/75) of the sequences in our dataset contain motif 14. About half of the sequences have multiple copies (2 - 4). Not only the strength of motif but also the number of copies affects exon skipping. In the simple linear regression, the contribution of motif 14 to the variation of skipping rates is 11% (p-value 3.8E-3). The enhancing effect of Nova is believed to be caused by controlling removal of introns harboring YCAY. It is interesting that the log-odds scores of motif 9 are positively correlated with motif 14, indicating the co-occurrence of these two motifs in the same intron sequences (motif 9 at the 5' end and motif 14 at the 3' end). This, along with the same enhancing effects of both motifs, may suggest their cooperative role during intron removal.

Motif 2 (see Table 3-2) has a consensus sequence of BBCASGGK which is similar to the binding site of heterogeneous nuclear ribonucleoproteins A1 (hnRNP A1). The binding sites of hnRNP A1 are either TAGGGT or CAGG[GA]T [142, 143], while starting from position three motif 2 is CA[GC]GG[GT]. The family of hnRNP counteracts the effect of SR protein in splicing regulation. The hnRNP A1 can act as splicing silencer by affecting the selection of splice sites [142, 143]. Motif 2 is positively correlated with exon skipping rate, suggesting the role of a splicing silencer. The motif occurs in the 5' end of the up-stream intron. Unlike motif 14 which occurs in most of the 75 CNS genes with multiple copies, only thirteen genes

contain this motif and most with single copy. The corresponding R^2 value of motif 2 in a linear regression is 0.17 with a p-value 2.7E-4.

Conclusion and Discussion

In this paper, we describe a two-step approach to predict regulatory motifs involved in alternative splicing. Our approach makes essential use of exon skipping rate measurements to overcome the inaccuracy caused by short motifs with high degree of degeneracy. We start from identifying motif seeds that both are enriched and have significant effect on skipping rate. Subsequently, the motif seeds are extended and refined by a local stochastic search. Our algorithm optimizes an objective function which combines both skipping rates and motif log-odds scores. In finding short and less conserved motifs, a simulation study indicates that our approach performs better than an approach that groups sequences by expression data and uses MEME for motif discovery. This makes our method especially suitable for AS motif discovery where motifs are typically short and highly degenerate [1]. MEME shows a similar or better performance for longer and more conserved motifs. One possible explanation for this behavior is that our algorithm restricts the search space to discrete IUPAC symbols, and cannot refine the corresponding PWMs beyond this resolution. We suggest testing this hypothesis in future work by implementing a more flexible stochastic search procedure. Compared to other k-mer based motif finders, e.g. YMF [144] and Weeder [145], our approach refines the k-mer seeds by association with skipping rates and thus shows more flexibility, particularly in motif length and degenerate symbols allowed.

Table 3-2. Predicted motifs using our new approach. Motifs are predicted in seven regions: up-stream exon, 5' and 3' end of the up-stream intron (up to 200 bp), cassette exon, 5' and 3' end of the down-stream intron (up to 200 bp) and down-stream exon. R^2 and corresponding p-values for single linear regression for each predicted motif are given. The correlation between the motif scores and skipping rates indicates the splicing effect. Negative correlation suggests splicing enhancer and positive correlation suggests splicing silencer. Logos are also given for each predicted motif. The height of each position is proportional to information content.

Acronyms: ESE/ESS, exonic splicing enhancer/silencer; ISE/ISS, intronic splicing enhancer/silencer. IUPAC symbols: R: A/G, Y: C/T, M: A/C, K: G/T, S: C/G, W: A/T, B: C/G/T, D: A/G/T, H: A/C/T, V: A/C/G and N: A/C/G/T.

Table 3-2 Continued

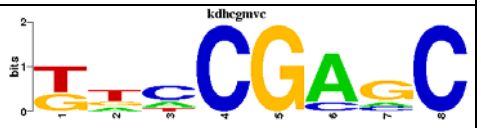
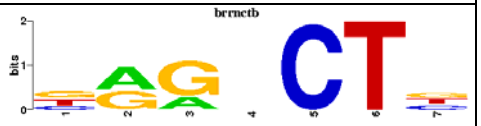
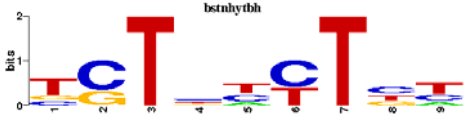
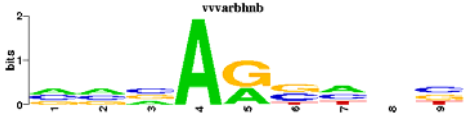
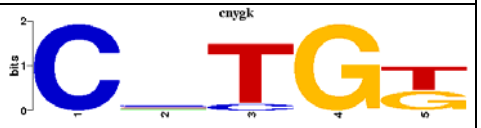
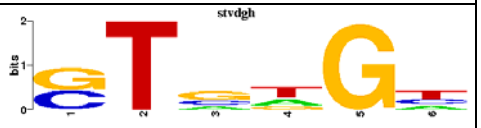
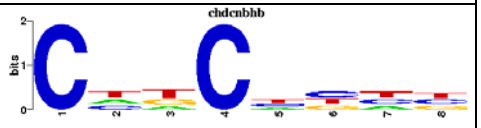
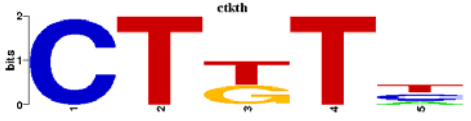
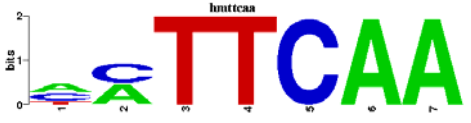
Motif	Width (bp)	Region	Consensus	P-value in regression	R^2 in regression	Correlation (Effect)	Logo
1	8	Up-stream Exon	KDHCGMVC	1.1E-5	0.23	+ ESS	
2	8	Up-stream Intron 5’ end	BBCASGGK	2.7E-4	0.17	+ ISS	
3	7	Up-stream Intron 5’ end	BRRNCTB	0.027	0.065	- ISE	
4	10	Up-stream Intron 5’ end	HWKATTWWTD	0.041	0.056	+ ISS	
5	8	Up-stream Intron 3’ end	BHBTACAT	8.5E-4	0.14	+ ISS	
6	9	Up-stream Intron 3’ end	BSTNHYTBH	0.098	0.037	- ISE	
7	9	Cassette Exon	VVVARBHNB	0.16	0.027	+ ESE	

Table 3-2 Continued

8	7	Down-stream Intron 5' end	ATRGABB	1.6E-4	0.18	- ISS	
9	5	Down-stream Intron 5' end	CNYGK	9.6E-4	0.14	- ISE	
10	6	Down-stream Intron 5' end	STVDGH	0.45	0.0078	- ISE	
11	8	Down-stream Intron 3' end	CHDCNBHB	4.3E-3	0.11	- ISE	
12	5	Down-stream Intron 3' end	CTKTH	0.075	0.043	- ISE	
13	7	Down-stream Intron 3' end	VTBKBCY	1.2E-4	0.18	- ISE	
14	5	Down-stream Intron 3' end	VYCAK	3.8E-3	0.11	- ISE	
15	7	Down-stream Exon	HMTTCAA	5.7E-4	0.15	+ ESS	

We applied our approach to a dataset containing expression information for 75 exon skipping events in multiple CNS tissues. We identified in total fifteen motifs located in cassette exon as well as flanking exons and introns. The comparison of the predicted motifs to the Conserved Domain Database does not provide any evidence that any of our predicted motifs could be explained by conserved protein domains. Each single motif accounts for the variation of skipping rate between 6 and 20%, while the overall R^2 value is about 60%. Four motifs show positional bias relative to the splice sites. Two of our predicted motifs show high sequence similarity to known and well investigated AS motifs – NOVA and hnRNP A1 binding sites. Also the observed effects on exon skipping rates and the motif positions of our motifs coincide with those of NOVA and hnRNP A1. We find co-occurrences of several predicted motifs, suggesting possible cooperative roles of these motifs in exon skipping. Interestingly, we find a so far unreported motif (motif 9) co-occurring with Nova binding sites. This suggests that Nova might have a more complex role in the CNS-specific splicing code than expected before.

Originally, Fagnani and colleagues performed an *ab initio* motif discovery study by comparing k-mer usage, as well as searching experimentally validated motifs in the sequence dataset [94]. We compared our motif set with the consensus sequences of the motifs reported in Fagnani's study. Similar to their results, we also identified C/U-rich motifs involved in CNS-specific AS. In addition, we discovered binding sites which were not found in Fagnani's *ab initio* motif discovery (but in their later scanning for known motifs), e.g. NOVA binding site. Ten of our predictions may be novel findings. We hypothesize that this might be attributed to our motif discovery approach which makes essential use of exon

skipping rate information. Based on the evidences we presented above, we believe our motifs are part of a complex CNS-specific splicing code, and that they are promising candidates for future validation experiments.

Materials and Methods

Motif discovery algorithm

A two-step motif discovery approach which incorporates exon skipping rate information is implemented (see Box 3-1).

Box 3-1. Workflow of the motif discovery approach.

Step 1: Find seed motifs
- Generate all k-mers
- Remove k-mers occurring in less than 10% of the sequences
- Identify k-mers which significantly affect skipping rates
- Remove similar seeds
  - Repeat
    - Record the k-mer with the highest significance of skipping rate difference
    - Eliminate k-mers with more than 3bp match
  - until no more k-mers

Step 2: Motif refinement via local stochastic search
- Repeat
  - Extend, eliminate or change one position using IUPAC symbols
  - Compute the position weight matrix from corresponding k-mers
  - Evaluate the objective function
  - Accept and update the motif if objective function improves
- until no further improvement

In step one, we identify seed motifs. We check exhaustively all possible k-mers with k ranging between 4 and 7. K-mers which occur in less than 10% of the gene sequences are removed. For the remaining k-mers we compare the skipping rates between sequences with and without the k-mer via a t-test, and eliminate k-mers which do not show a significant (p-value<0.005) skipping rate difference. To remove highly similar k-mer seeds, we sort the k-mers based on their significance level. We select the k-mer with the highest significance among the remaining k-mers, store this k-mer, and eliminate all k-mers with three or more basepairs match. We repeat this procedure until no more k-mers are left.

In step two, we use stochastic search to explore the neighborhood of each seed in the sequences to either get a more flexible representation or extend/shrink the motif using IUPAC symbols. Motifs made from IUPAC symbols correspond to discrete PWMs. They provide a flexible motif representation while simultaneously keeping the search space limited. Starting from a motif seed, we extend, shrink or modify one motif position to get a new candidate motif. Log-odds scores for each sliding window in the sequences are computed based on the position weight matrix which corresponds to the candidate motif. We then use the skipping rates as weights in the following objective function:

$$f(M,Y)=\sum_{i=1}^{n} y_i \sum_{j=1}^{|S_i|-|w_{ij}|+1} z\log\left(\frac{\Pr(w_{ij}\mid M)}{\Pr(w_{ij}\mid B)}\right) \qquad \textbf{(3.1)}$$

Here, S denotes a set of sequences, S_1, S_2, …, S_n, *and* Y is the associated skipping rates, y_1, y_2, …, y_n. M and B indicates whether a string is a motif or from background, respectively. w_{ij} is the string which corresponds to the j-th sliding window in sequence i, $\log(\Pr(w_{ij}|M)/\Pr(w_{ij}|B))$ is the log-odds score of string w_{ij}, and z is an indicator variable for the case that the

log-odds score is greater than 0. This objective function is a modified version from Smith and colleagues [139]. It calculates the weighted sum of log-odds scores. For each string in *j*-th sliding window on sequence *i*, we compute the log-odds score. The indicator variable *z* discards the string with score less than 0. We then sum the scores over all sequences and all strings weighted by the skipping rates. We accept a motif change if the objective function improves. We repeat this procedure until no improvement can be achieved for any change.

Simulation Study

We tested our approach in a simulation study. We simulated DNA sequences, implanted artificial motifs, generated corresponding exon skipping rates, and evaluated the performance of our algorithm. Each analyzed sequence set consists of 100 simulated sequences, ranging from 100 to 300 basepairs, using a uniform nucleotides distribution. We implanted motifs in 50 sequences with different combinations of parameters: motif width (4, 6, 8 and 10 basepairs) and motif conservation, measured by information content (1.05, 0.95, 0.75, 0.55 ±0.025 per column). The definition of information content of a column in motif is as follow [139, 146]:

$$\begin{aligned} I_i &= \sum_{j\in\{A,C,G,T\}} p_{ij}\log_2\left(p_{ij}/q_j\right) \\ &= 2+\sum_{j\in\{A,C,G,T\}} p_{ij}\log_2\left(p_{ij}\right) \end{aligned} \qquad \textbf{(3.2)}$$

where *i* is the position of each column and p_{ij} is the probability of nucleotides in motif and q_j is the background probabilities. The second equation only applies when the background probabilities of nucleotides are all equal.

We generated two scenarios for exon skipping rates. A non-overlap group with skipping rates 0.80±0.15 ([0.65, 0.95]) for sequences with motifs implanted and 0.20±0.15 ([0.05, 0.35]) without motifs; an overlap group with skipping rates 0.60±0.15 and 0.40±0.15, respectively. We performed 100 simulations for each combination of parameters and calculated the mean and standard deviation of distance between the predicted and implanted motifs. To evaluate the performance of our approach, we also analyzed the artificial sequences with skipping rates greater than 0.5 by MEME [118] for comparison. This mimics the more traditional way of selecting sequences by skipping rates and then doing motif discovery.

Comparison of position weight matrices

The column metric is based on Euclidean distance and is a modified version from Tsai and colleagues [147]. Let $\mathbf{C_1} = (p_{1A}, p_{1C}, p_{1G}, p_{1T})^T$ and $\mathbf{C_2} = (p_{2A}, p_{2C}, p_{2G}, p_{2T})^T$ be two columns in matrices 1 and 2, where p is the probability of a nucleotide. The distance between two columns is calculated as

$$d(\mathbf{C_1}, \mathbf{C_2}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j \in \{A,C,G,T\}} \left(p_{1j} - p_{2j} \right)^2} \qquad \textbf{(3.3)}$$

Two PWMs are globally aligned without internal gaps while end gaps are allowed. The distance between the column and the background nucleotide distribution is used as penalty of overhang. We define the distance between two PWMs as the score of the optimal global alignment.

Collection of sequence data

We analyzed the mouse skipped exons (the most common type of alternative splicing) from Fagnani and colleagues [94]. A quantitative microarray platform created by the same lab was used to measure the skipping rate (*%ASex*, expression of isoform without the skipped exon divided by the total expression of both) [93]. The whole-length transcripts were downloaded from NCBI. We used BLAT to map each transcript to the mouse genomic sequences [88]. We did motif discovery in seven regions, which are up-stream exon, 5' and 3' end of the up-stream intron (up to 200 bp), cassette exon, 5' and 3' end of the down-stream intron (up to 200 bp) and down-stream exon. Since the skipping rate is a ratio, we transform the associated skipping rate by logit transformation.

$$y = \log\left(\frac{\%ASex}{1-\%ASex}\right) \tag{3.4}$$

Positional bias of motif occurrences

We use the Kolmogorov-Smirnov (K-S) statistic to test for non-uniform occurrences of the predicted motifs within the sequences used for motif discovery rather than the whole introns. The K-S statistic is defined as

$$D = \sup_x \left|F(x) - F_0(x)\right| \tag{3.5}$$

where $F(x)$ and $F_0(x)$ are the motif and background distributions.

We use a Monte Carlo approach to estimate the empirical distribution of K-S statistic. Let $L_1, L_2, \ldots, L_n$ be the lengths of sequences from which a motif is predicted, we randomly generate one motif for each sequence assuming that motif can occur at any position with

equal probability and repeat for 1000 times. Similarly, we generate a large number of random motif positions to obtain a reference background distribution. We then calculate the K-S statistic between each set of randomly generated positions and the reference. The p-value is the number of larger or equal K-S statistics divided by 1000.

Chapter 4

SPRED: Splicing Regulatory Element Database

Abstract

Cis-acting regulatory elements play an important role in the regulation of (alternative) splicing. Our newly developed Splicing Regulatory Element Database (SPRED) contains comprehensive information about cis-acting splicing regulatory elements compiled from literature. The web interface is available at http://weir.statgen.ncsu.edu/asmotif/home.html. A web query tool allows keyword queries and similarity searches for the user-provided inputs. The goal of SPRED is to provide a comprehensive dictionary of cis-regulatory splicing elements. SPRED does not only facilitate the study of RNA motifs involved in (alternative) splicing, but also is a first step towards understanding the tissue- and condition-specific splicing code.

Keywords

Alternative splicing, splicing regulatory element

Introduction

RNA splicing is an essential step in post-transcriptional process. Together with other splicing signals, cis-acting regulatory elements play an important role in the regulation of (alternative) splicing [1]. There are several databases about alternative splicing [127, 148-151]. However, most of them focus on splicing events and sequences in multiple organisms and very few are about the cis-regulatory splicing elements. Alternative Exon Database (AEDB) [151] is one of the pioneer databases in which splicing elements are included as a sub-database. The regulatory elements in AEDB are experimentally validated, but some of

them may also include non-functional neighboring sequence. Hollywood database [148] contains a collection of hexanucleotides (6 bp oligonucleotides) which are candidate splicing elements predicted either by RESCUE-ESE algorithm [110] or by screening random oligonucleotide library in a fluorescent reporting system [105].

We here present the current status of SPRED, a database of cis-acting regulatory elements involved in splicing (particularly in alternative splicing). We also describe the potential applications and future directions of SPRED.

Content of SPRED

SPRED contains a compilation of cis-acting regulatory nucleotide elements involved in (alternative) splicing. These elements are commonly the binding sites of trans-acting splicing factors. All the elements are from published research articles. Many of them are validated by Systematic Evolution of Ligands through EXponential enrichment (SELEX) experiments. Others are discovered by RNA-protein crosslinking or by sequence analysis. For each splicing factor, multiple elements may be included as unique records from different article or from same article with multiple elements provided.

Each record of the elements consists of the following fields with tags (see Figure 4-1): "AC" gives the accession number that acts as the unique key of each record. The accession number starts with C or M and followed by five digits, representing consensus or profile respectively. "BF" gives the name of splicing factor that binds to the element. All the factor names are acronyms. Alternative names of splicing factors are also included in the "BF" field

```
AC  M00100

BF  SC35, SRp30b

CA  SR protein

OR  Human

FU  Exonic Splicing Enhancer

MA  0.067 0.033 0.867 0.033
MA  0.233 0.000 0.600 0.167
MA  0.200 0.467 0.033 0.300
MA  0.000 0.567 0.000 0.433
MA  0.200 0.333 0.300 0.167
MA  0.100 0.467 0.367 0.067
MA  0.200 0.300 0.033 0.467
MA  0.267 0.000 0.733 0.000

CO  GRYYCSYR

RF  Liu H-X, et al. Mol Cell Biol. 2000, 20(3): 1063-1071

RM  Matrix is from the multiple alignment of element sequences.

DT  12222008
```

Figure 4-1. An example of the record in SPRED. The element could be in either profile/matrix (MA tag) or in consensus (CO tag). Tags: AC, accession number; BF, splicing factor; CA, category of splicing factor; OR, organism; FU, splicing effect and location; MA, element in profile/matrix form; CO, element in consensus form; RF, reference; RM, remarks; DT, date of modification.

and separated by commas. “CA” represents the categories which splicing factors belong to. Currently there are three categories: SR protein, hnRNP protein and other. “OR” gives the organism in which the element was discovered and validated in the experiments. “FU” indicates the splicing effect of the elements (suppression or promotion of splicing), as well as the locations where elements reside in. This can be classified mainly into four categories:

Exonic Splicing Enhancer, Exonic Splicing Silencer, Intronic Splicing Enhancer and Intronic Splicing Silencer. The locations of several splicing elements are unknown, therefore, only splicing enhancer or silencer will be given. "CO" gives the consensus sequence of the elements. We use IUPAC symbols to represent multiple nucleotides occurring in the same position. For those elements that are highly repetitive yet with undefined length, we use (XXX)n where XXX is the core site. For example, we use (GAA)n to represent repetitive element of Tra2a. "MA" represents the profile of the elements. The consensus sequences are extracted directly from literature and the profiles are compiled from the alignment of sequences provided with literature. "RF", "RM" and "DT" give reference, remark and date of modification, respectively. Since the number of known and validated elements is still limited, we keep all the records in a flat text file with different tags in each line indicating different fields.

Currently, there are totally 63 records in SPRED, 47 in consensus form and 16 in profile (see Table 4-2 for a list of the consensus sequences and corresponding splicing factors). The number of unique corresponding splicing factors is 29. Some well-studied splicing factors may correspond to multiple elements from different articles or from same articles, e.g. SC35, ASF/SF2. Most elements are derived from human or mouse, and very few are identified from other organisms (5 elements). Majority (49) of the elements act as enhancers and 19 as silencers. Due to the context-dependent nature five elements have double roles in splicing and can be either enhancer or silencer. The numbers of exonic elements are about doubled compared to intronic elements (31 exonic elements vs. 18 intronic elements). The average

width of all elements is 7.0 bp and standard deviation 2.4 bp excluding the highly-repetitive elements.

Table 4-1. Statistics of SPRED

Overall statistics	
Total no. of records	63
No. of consensus sequences	47
No. of profiles	16
No. of unique splicing factors	29
Mean (SD) of element width (excluding repetitive elements)	7.0 (2.4)
Category of splicing factor	
SR protein	34
hnRNP protein	14
Other	15
Organism coverage	
Human	55
Mouse/Rat	11
Other	5
Splicing effect	
Enhancer	49
Silencer	19
Location	
Exonic	31
Intronic	18

Keyword Query

Your keyword:
Category - SR protein
Binding factor - SC35

ID	Category	Factor	Organism	Effect	Consensus or Profile	Reference
C00100	SR protein	SC35, SRp30b	Human	Exonic Splicing Enhancer	GRYYCSYR	Liu H-X, et al. Mol Cell Biol 2000, 20 (3): 1063-1071
C00110	SR protein	SC35, SRp30b	Mouse	Exonic Splicing Enhancer	AGSAGAGTA	Tacke R, et al. EMBO J. 1995, 14(14): 3540-3551
P00100	SR protein	SC35, SRp30b	Human	Exonic Splicing Enhancer	A 0.067 0.233 0.200 0.000 0.200 0.100 0.200 0.267 C 0.033 0.000 0.467 0.567 0.333 0.467 0.300 0.000 G 0.867 0.600 0.033 0.000 0.300 0.367 0.033 0.733 T 0.033 0.167 0.300 0.433 0.167 0.067 0.467 0.000	Liu H-X, et al. Mol Cell Biol. 2000, 20 (3): 1063-1071
P00110	SR protein	SC35, SRp30b	Mouse	Exonic Splicing Enhancer	A 0.875 0.042 0.042 0.750 0.000 0.750 0.125 0.250 0.500 C 0.042 0.042 0.417 0.167 0.000 0.083 0.000 0.042 0.083 G 0.042 0.917 0.542 0.083 1.000 0.000 0.833 0.208 0.125 T 0.042 0.000 0.000 0.000 0.000 0.167 0.042 0.500 0.292	Tacke R, et al. EMBO J. 1995, 14(14): 3540-3551

Figure 4-2.An example of keyword query. The upper panel depicts the interface of keyword query tool. Keywords “SR protein” in category field and “SC35” in binding factor field are used. The query result is given in the bottom panel. It includes four records, two in consensus and two in profile format. Alternative factor name, organism in which the element was identified, sequence or matrix and reference are also given.

Keyword query

We provide several online services to search our database. Keyword query is a primary method to retrieve the information in the database (see Figure 4-2). Currently, we allow query using keywords in these fields: category, binding factor, organism, splicing effect (enhancer or silencer) and location (exonic or intronic). The query is simply done by text-matching of user-defined keywords in the fields. Also, keywords in multiple fields are allowed which gives the intersection of all keyword matches. Therefore, if mutually exclusive keywords are given, the result will give no record. For example, no records will show if using keywords "SR protein" and "hnRNP A1". If no keywords are given, the list of all elements in SPRED will be shown.

Similarity search

We provide similarity search of user-defined elements as well (see Figure 4-3). The input can be in either a consensus sequence or a profile. For the consensus sequence, users can provide either regular expression with multiple nucleotides in the same position, e.g. G[AG][CT][CT]C[CG][CT][AG], or use IUPAC symbols, e.g. GRYYCSYR. Users may give multiple elements for a batch search. Also, the input can be in profile form and each line gives the probabilities of nucleotides in order of A, C, G and T for one position. If multiple elements are given, identifiers for each element are required.

The cutoff criteria: P-value 0.01

>1
GAAGAA

ID	Category	Factor	Organism	Effect	Alignment	Metric	Reference
C00900	SR protein	Tra2a, Tra2alpha	Human	Splicing Enhancer	**Query** GAAGAA A 0.000 1.000 1.000 0.000 1.000 1.000 C 0.000 0.000 0.000 0.000 0.000 0.000 G 1.000 0.000 0.000 1.000 0.000 0.000 T 0.000 0.000 0.000 0.000 0.000 0.000 **Sbjct** (GAA)n A 0.000 1.000 1.000 0.000 1.000 1.000 C 0.000 0.000 0.000 0.000 0.000 0.000 G 1.000 0.000 0.000 1.000 0.000 0.000 T 0.000 0.000 0.000 0.000 0.000 0.000	0.000 (p<0.001)	Tacke R, et al. Cell 1998, 93 (1):139-148
C01100	SR protein	SRm160	Human	Exonic Splicing Enhancer	**Query** GAAGAA A 0.000 1.000 1.000 0.000 1.000 1.000 C 0.000 0.000 0.000 0.000 0.000 0.000 G 1.000 0.000 0.000 1.000 0.000 0.000 T 0.000 0.000 0.000 0.000 0.000 0.000 **Sbjct** (GAA)n A 0.000 1.000 1.000 0.000 1.000 1.000 C 0.000 0.000 0.000 0.000 0.000 0.000 G 1.000 0.000 0.000 1.000 0.000 0.000 T 0.000 0.000 0.000 0.000 0.000 0.000	0.000 (p<0.001)	Eldridge AG, et al. Proc Natl Acad Sci 1999, 96 (11):6125-6130; Cheng C, et al. Mol Cell Biol 2006, 26(1): 362-370

Figure 4-3.An example of similarity search. The upper panel shows the interface of similarity search tool. A consensus "GAAGAA" is used to search SPRED. The distance metric used is "Euclidean distance" and the cutoff for significant match is p-value 0.01. The bottom panel depicts the search result. The alignment of matches is shown in matrix. The distance and corresponding p-value are also given.

We compare elements through profiles. If a consensus sequence is given by users, we first convert it into matrix form by assigning probabilities to each nucleotide and do then profile comparison. For example, if A or G occurs in one position of an element, we assign 0.5 to A and G and 0 to C and T. To calculate the distance between input element and the highly repetitive elements, we expand the elements to the minimal length that greater than the user-defined elements before doing comparison. For example, if the input is 5 bp long, we will calculate the distance metric between input element and GAAGAA for Tra2.

To calculate distance metric, we use two different column metrics – Euclidean distance and Pearson's correlation coefficient. Let $\mathbf{C}_1 = (p_{1A}, p_{1C}, p_{1G}, p_{1T})^T$ and $\mathbf{C}_2 = (p_{2A}, p_{2C}, p_{2G}, p_{2T})^T$ be two columns in profile 1 and 2, where p is the probability of a nucleotide. The metric of Euclidean distance between two columns is calculated as

$$d(\mathbf{C_1}, \mathbf{C}_2) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j \in \{A,C,G,T\}} \left(p_{1j} - p_{2j} \right)^2} \qquad (\mathbf{4.1})$$

which is a modified version from Tsai and colleagues [147]. The column metric based on Pearson's correlation coefficient is modified from Pietrokovski [125] and defined as

$$d(\mathbf{C_1}, \mathbf{C}_2) = \left(1 - \frac{\mathrm{cov}(\mathbf{C_1}, \mathbf{C}_2)}{\sqrt{\mathrm{var}(\mathbf{C_1})\,\mathrm{var}(\mathbf{C}_2)}} \right) \Big/ 2 \qquad (\mathbf{4.2})$$

Both metrics are normalized to range from 0 to 1. The smaller the metric is the more similar two columns are.

Two profiles are globally aligned without internal gaps. The column metrics between the overhang and the background nucleotide distribution on either end is used as penalty. We define the distance between two profiles as the score of the optimal global alignment (see Figure 4-4).

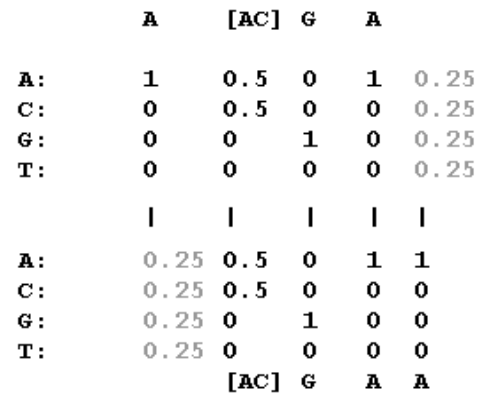


Figure 4-4. An example of alignment between two elements. Two elements A[AC]GA and [AC]GAA are converted into profiles first and then aligned. Overhangs occur on both ends in the optimal alignment. The distance between the column and background nucleotide distribution assuming equal distribution of nucleotides (in grey) is used as penalty. Therefore the distance based on Euclidean distance is

$$d_{euc} = 0.61 + 0 + 0 + 0 + 0.61 = 1.22$$

and the distance based on Pearson's correlation coefficient is

$$d_{pcc} = 0.5 + 0 + 0 + 0 + 0.5 = 1.$$

In similarity search, users can get either the best one, three or five matches to their elements or all the matches with a user-defined cutoff of p-value. We use a resampling

strategy to estimate the p-value of the distance metric. We generate 1000 random matrices by shuffling all the profile columns in our database and calculate all the distance metrics to estimate its empirical distribution of random matches. The p-value is denoted as the chance of having smaller or equal distance metrics.

Conclusion and future direction

In this study, we compile a database, named SPRED, of experimentally validated cis-acting regulatory splicing elements. All the elements are extracted or compiled from literature. Annotated features for each element are also available. SPRED has significant advantages compared to similar resources:

1) The data of regulatory elements are from wet-lab experiments in literatures, which provide significant reliability;
2) Regulatory elements in consensus or in profile give significant flexibility compared to k-mers, thus facilitate the comparison of sequence similarity;
3) SPRED is coupled with search tools and is open accessible, which can help genome-wide studies through internet.

SPRED will continue to expand while more regulatory elements are available by experiments. As the number of elements and related features grow, we will implement SPRED in a relational database management system to meet the increasing demand of information retrieval. Additionally, sequence scan tool will be incorporated into SPRED web interface for scanning of existing elements in user-provided sequences.

Table 4-2. The consensus sequences of cis-acting elements and corresponding trans-acting splicing factors. The table includes the information about name of splicing factor, organisms where the cis-acting elements are discovered, splicing effect and consensus sequences. Oligonucleotides in parenthesis denote the unit in repetitive elements. IUPAC symbols: R: A/G, Y: C/T, M: A/C, K: G/T, S: C/G, W: A/T, B: C/G/T, D: A/G/T, H: A/C/T, V: A/C/G and N: A/C/G/T.

SR protein			
SC35, SRp30b	Human, Mouse	Exonic Splicing Enhancer	GRYYCSYR [152] AGSAGAGTA [153]
ASF/SF2, SRp30a	Human, Mouse	Exonic Splicing Enhancer	RGAAGAAC [153] SRSASGA [101] CSSCSSR [154]
SRp40	Human	Exonic Splicing Enhancer	ACDGS [101] TGGGAGCRGTYRGCTCGY [155]
SRp55	Human	Exonic Splicing Enhancer	TSCGKM [101]
SRp20, RBP1	Human	Exonic Splicing Enhancer	WCWWC [156]
9G8	Human	Exonic Splicing Enhancer	AGACKACGAY [156]
SRp30c	Human	Exonic Splicing Enhancer	AGSAS [157]
SRp38	Human	Splicing Silencer	ACAAAGACAAA [158]
Tra2a	Human	Splicing Enhancer	$(GAA)_n$ [75]
Tra2b	Human	Splicing Enhancer	$(GAA)_n$ [75]
SRm160	Human	Exonic Splicing Enhancer	$(GAA)_n$ [159, 160]

Table 4-2 Continued

hnRNP protein			
hnRNP A1	Human Papillomavirus, Human	Splicing Silencer	TAGGGW [142] CAGGRT [143]
hnRNP A2, hnRNP B1	Rat	Intronic Splicing Enhancer	$(TTAGGG)_n$ [161]
hnRNP C	Unknown	Splicing Enhancer	TTTTT [162]
hnRNP F	Human	Exonic Splicing Silencer	G-rich [163]
hnRNP G, RBMX	Human	Splicing Enhancer	AAGT [164]
hnRNP H	Human	Exonic Splicing Silencer	G-rich [163]
hnRNP L	Human	Intronic Splicing Enhancer	$(AC)_n$ [77, 165] $(ACAT)_n$ [77]
PTB, hnRNP I	Human	Intronic Splicing Silencer	TTCTCT [166] CTCTCT [76] TCTT [167]
nPTB	Human	Intronic Splicing Silencer	CTCTCT [168]

Table 4-2 Continued

Other			
PSF	Human, Rat, Mouse, Chicken, Frog, Fly	Splicing Enhancer	TGGAGAGGAAC [169]
P54nrb	Human, Rat, Mouse, Chicken, Frog, Fly	Splicing Enhancer	GAGAGGAAC [169]
Fox-1, A2BP1	Human	Intronic Splicing Enhancer, Intronic Splicing Silencer	GCATG [170] TGCATG [171]
Fox-2, Rbm9	Human	Intronic Splicing Enhancer	TGCATG [171]
Nova	Human, Mouse	Intronic Enhancer, Exonic Splicing Silencer, Intronic Splicing Silencer	YCAY [65]
CUGBP1, CELF1	Human	Intronic Splicing Enhancer	$(TGT)_n$ [172]
CUGBP2, CELF2, ETR-3	Human	Intronic Splicing Enhance	$(TG)_n$ [173]
Mbn11	Human, Chicken	Intronic Splicing Enhancer, exonic Splicing Silencer	YGCTTY [174]
Quaking, QKI	Mouse	Splicing Silencer	NACTAAY [175]

Chapter 5

Conclusions — From a Parts List to an Integrated Splicing Code

Conclusions

In most eukaryotes genes are interrupted by intervening introns which are removed in the final mRNA transcripts. Although exons are often short and embedded in the much longer introns, they can be recognized with remarkable precision. Moreover, in alternative splicing different sets of exons within the same gene can be spliced in to produce multiple variants. The regulation of alternative splicing is very complex and requires the involvement of numerous splicing proteins and recognition of the mRNA sequence signals. Therefore, discovery of the cis-regulatory elements can contribute to predict the isoform expression patterns of genes subject to alternative splicing. It can also help to understand the gene regulation network on exon level in different tissues and different developmental stages. However, it is a big challenge to predict the cis-acting splicing motifs due to their short, degenerate and context-dependent nature.

Our first attempt to identify AS motif is described in chapter two. Due to the lack of data, expression information from microarray was not commonly applied to help motif prediction in alternative splicing. Based on the experience gained from similar experiments in transcriptional factor binding site prediction, it is likely that prediction quality may be dramatically improved after clustering potential co-regulated genes by expression profile [121, 122, 136]. However, choosing the boundary of clusters is challenging and it must be a tradeoff between exclusion of signals and inclusion of noise. Rather than giving a fixed value to define cluster boundary, we use systematic sampling to sample gene clusters and remove redundant motifs thereafter. We applied our computational framework to analyze about 3,000 skipped exons with corresponding skipping rates and provide a promising motif dictionary.

Although clustering co-regulated genes significantly increases the signal/noise ratio, most motif finders rely on overrepresented patterns and still may miss short and degenerate motifs. This could be a problem of motif discovery in alternative splicing based on the nature of AS motifs. Therefore, our algorithm described in chapter three combines seed identification and motif refinement. Both steps fully utilize skipping rates and sequence pattern information. A simulation study shows that our approach performs better in predicting short and degenerate motifs in comparison to a conventional motif discovery approach which relies on searching overrepresented patterns in clusters of co-regulated genes. We applied the new approach to CNS-specific exon skipping data. We show that two known motifs, Nova and hnRNP A1 binding sites, and ten novel motifs are involved with alternative splicing regulation in CNS tissues. Potential cooperation between different motifs is revealed, particularly, the co-operative role of a new motif with Nova binding site has not been reported before.

Besides our effort in cis-acting motif prediction, other studies also provide experimentally validated motifs. However, it is difficult to find these results. Researchers have to search the literature to compare with known motif information case by case, and very often this might be the most time-consuming step. To overcome this bottleneck, we compiled an open-accessible database, SPRED, containing existing validated splicing motifs in order to simplify data access and retrieval. Potential application of SPRED include:

1) Motif similarity search to validate newly discovered AS motifs.
2) Sequence motif scan to find potential genes whose splicing are regulated by the corresponding splicing factors.
3) Integration of known splicing signal to predict alternative splicing outcome.

In summary, we developed computational approaches by incorporating expression profiles. Our approaches are specifically designed to improve the motif prediction quality in datasets with low signal/noise ratio. The application of our methods to a mouse cassette exon dataset results in a dictionary of cis-acting splicing regulatory motifs. The dictionary could be used for experimental validation. Along with other validated motifs in SPRED database, we provide a parts list which may help to understand the integrated splicing code.

Future directions

It has been about thirty years since alternative splicing was proposed to produce multiple mRNA isoforms from a single gene [176]. Numerous efforts have been made, listing genes undergoing alternative splicing and cis- or trans-acting regulators involved in AS regulation.

Several attempts to predict the splicing outcome in specific gene groups have been published. Ule and colleagues scanned the genes for YCAY clusters and identified 51 candidate Nova-regulated skipped exons. 20 out of 41 new predicted exons were validated in mouse brain. The enhancing or silencing effect in all 30 (10 old plus 20 new) Nova-regulated skipped exons was correctly predicted [65]. Sorek and colleagues used seven features to classify cassette exons and constitutively spliced exons. The features include:

1) Exon length,

2) Divisibility by three in exon,

3) Percent identity of exon in human and mouse,

4-7) Lengths of best local alignment and percent identity in up- and down-stream flanking introns.

The classifier learned the best rule which can identify the maximal number of cassette exons while making no false positives in the training dataset. In the five-fold cross-validation, the average sensitivity is about 32% and the specificity is 99% [177]. Baek and Green tried to predict exclusion rate in human cassette exons by multiple linear regression. The significant explanatory variables include exon length, splice site score and intron conservation between human and mouse. Although each explanatory variables are very significant, the overall predictive power is not high (R^2 0.26). They argued that the low predictive power was due to the over-simplified linear model which did not take the splicing motifs into account [37].

Similar to the study from Baek and Green, the datasets we analyzed contain cassette exons and associated skipping rates from splicing junction array. It may be applicable to predict the splicing outcome from the sequence signals in our study as well. In our preliminary attempt using linear model in CNS-specific AS events, the R^2 can reach 62% (adjusted R^2 55%) using eleven significant motifs.

Our next challenges will be integrating all possible splicing signals to develop ability of predicting splicing outcome from sequence data only. To achieve a higher predictive power, we may need to use all possible cis-acting splicing signals. Exon and flanking intron lengths, splice site score, branch site score and measurements of poly-pyrimidine tract strength are proved to affect alternative splicing (see Chapter 1). Cis-acting regulatory motifs are another pivotal category of features and must be included in AS prediction. It is interesting that intron conservation is significantly associated with skipping rates [37], indicating the existence of regulatory motifs. However, the positive correlation between conservation and skipping rates seems to suggest that only exon exclusion need splicing motifs. Also, the divisibility by three

in exons may account for the bias of skipping rates caused by degradation in nonsense-mediated decay [38]. The biggest challenge in incorporating all these splicing signals into the predictive model is that each single signal has very weak effect and may contribute only a small portion to explain the variance in dependent variables [1, 52]. The context-dependency of cis-acting motifs and possible cooperative manner between signals may also make the model very complicated. Finally, although the correlation between strength of splicing signals and splicing levels has been observed, the linear relationship between motif binding affinity and binding outcome is still controversial [178].

Combined with the experience gained from predictive models of gene expression from transcriptional factor binding sites (TFBS), the choice of dependent variables is also important. Beer and colleagues clustered yeast genes based on the expression profiles and identified 49 categories containing totally 2,500 co-regulated genes. They predicted the category labels for each gene by TFBSs using Bayesian network. They achieved 73% accuracy in ten-fold cross-validation [179]. Smith and colleagues identified genes that are significantly elevated or inhibited in different tissues and predicted elevation/inhibition of expression from TFBSs. The best prediction has an accuracy of 67% [180]. In the above studies, the dependent variables are discrete, such as class labels. Meanwhile, the dependent variables could also be continuous. Conlon and colleagues predicted continuous gene expression level in yeast genes whose expression changed after 30 minute of amino acid starvation. After excluding insignificant TFBSs and using forward model selection, 25 motifs remained in the model and the R^2 in multiple linear regression model is about 20% [138]. It

seems the predictive power by using continuous variables, even without cross-validation, may be lower compared to dichotomous or discrete variables.

Despite the challenges described above, decoding the regulation of alternative splicing is of high importance. It may help understanding the regulation of gene expression on the exon level, and it may also help finding therapeutic approaches for human diseases. The research on this level may result in the start of a new “exonomics” era [47].

REFERENCE

1. Ladd AN and Cooper TA. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* 3: reviews0008.

2. Matlin AJ, Clark F and Smith CW. (2005) Understanding alternative splicing: Towards a cellular code. *Nat Rev Mol Cell Biol* 6: 386-398.

3. Sorek R, Shamir R and Ast G. (2004) How prevalent is functional alternative splicing in the human genome?. *Trends Genet* 20: 68-71.

4. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H and Gingeras TR. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 14: 331-342.

5. Clark F and Thanaraj TA. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* 11: 451-464.

6. Black DL. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291-336.

7. Bell LR, Maine EM, Schedl P and Cline TW. (1988) Sex-lethal, a drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. *Cell* 55: 1037-1046.

8. Granadino B, Penalva LO, Green MR, Valcarcel J and Sanchez L. (1997) Distinct mechanisms of splicing regulation in vivo by the drosophila protein sex-lethal. *Proc Natl Acad Sci U S A* 94: 7343-7348.

9. Dell KR and Williams LT. (1992) A novel form of fibroblast growth factor receptor 2. alternative splicing of the third immunoglobulin-like domain confers ligand binding specificity. *J Biol Chem* 267: 21225-21229.

10. Boggs RT, Gregor P, Idriss S, Belote JM and McKeown M. (1987) Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the transformer gene. *Cell* 50: 739-747.

11. Lam BJ, Bakshi A, Ekinci FY, Webb J, Graveley BR and Hertel KJ. (2003) Enhancer-dependent 5'-splice site control of fruitless pre-mRNA splicing. *J Biol Chem* 278: 22740-22747.

12. Garcia-Blanco MA, Baraniak AP and Lasda EL. (2004) Alternative splicing in disease and therapy. *Nat Biotechnol* 22: 535-546.

13. Nagy E and Maquat LE. (1998) A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. *Trends Biochem Sci* 23: 198-199.

14. Amara SG, Jonas V, Rosenfeld MG, Ong ES and Evans RM. (1982) Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature* 298: 240-244.

15. Leff SE, Evans RM and Rosenfeld MG. (1987) Splice commitment dictates neuron-specific alternative RNA processing in calcitonin/CGRP gene expression. *Cell* 48: 517-524.

16. Xin D, Hu L and Kong X. (2008) Alternative promoters influence alternative splicing at the genomic level. *PLoS ONE* 3: e2377.

17. Zhang SX, Searcy TR, Wu Y, Gozal D and Wang Y. (2007) Alternative promoter usage and alternative splicing contribute to mRNA heterogeneity of mouse monocarboxylate transporter 2. *Physiol Genomics* 32: 95-104.

18. Galante PA, Sakabe NJ, Kirschbaum-Slager N and de Souza SJ. (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA* 10: 757-765.

19. Barbazuk WB, Fu Y and McGinnis KM. (2008) Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. *Genome Res* 18: 1381-1392.

20. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R and Fluhr R. (2004) Intron retention is a major phenomenon in alternative splicing in arabidopsis. *Plant J* 39: 877-885.

21. Marinescu V, Loomis PA, Ehmann S, Beales M and Potashkin JA. (2007) Regulation of retention of FosB intron 4 by PTB. *PLoS ONE* 2: e828.

22. Lodish H, Berk A, Zipursky LS, Matsudaira P, Baltimore D and Darnell J. (2000) *Molecular Cell Biology,* W. H. Freeman and Company, New York

23. Weaver RF. (2002) *Molecular Biology,* McGraw-Hill, New York

24. Brow DA. (2002) Allosteric cascade of spliceosome activation. *Annu Rev Genet* 36: 333-360.

25. Burset M, Seledtsov IA and Solovyev VV. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 28: 4364-4375.

26. Rymond B. (2007) Targeting the spliceosome. *Nat Chem Biol* 3: 533-535.

27. Kol G, Lev-Maor G and Ast G. (2005) Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet* 14: 1559-1568.

28. Moroy T and Heyd F. (2007) The impact of alternative splicing in vivo: Mouse models show the way. *RNA* 13: 1155-1171.

29. Grosso AR, Martins S and Carmo-Fonseca M. (2008) The emerging role of splicing factors in cancer. *EMBO Rep* 9: 1087-1093.

30. Boutz PL, Stoilov P, Li Q, Lin CH, Chawla G, Ostrow K, Shiue L, Ares M,Jr and Black DL. (2007) A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* 21: 1636-1652.

31. Goymer P. (2007) Alternative splicing switches on the brain. *Nat Rev Neurosci* 8: 576-576.

32. Jiang Z and Wu JY. (1999) Alternative splicing and programmed cell death. *Proceedings of the Society for Experimental Biology and Medicine* 220: 64-72.

33. Schwerk C and Schulze-Osthoff K. (2005) Regulation of apoptosis by alternative pre-mRNA splicing. *Molecular Cell* 19: 1-13.

34. Shin C and Manley JL. (2004) Cell signalling and the control of pre-mRNA splicing. *Nat Rev Mol Cell Biol* 5: 727-738.

35. Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, Turka LA, Mao X, Nunez G and Thompson CB. (1993) Bcl-X, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* 74: 597-608.

36. Akgul C, Moulding DA and Edwards SW. (2004) Alternative splicing of bcl-2-related genes: Functional consequences and potential therapeutic applications. *Cell Mol Life Sci* 61: 2189-2199.

37. Baek D and Green P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci U S A* 102: 12813-12818.

38. Magen A and Ast G. (2005) The importance of being divisible by three in alternative splicing. *Nucleic Acids Res* 33: 5574-5582.

39. Maquat LE. (2004) Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* 5: 89-99.

40. McGlincy NJ and Smith CW. (2008) Alternative splicing resulting in nonsense-mediated mRNA decay: What is the meaning of nonsense?. *Trends Biochem Sci* 33: 385-393.

41. Lewis BP, Green RE and Brenner SE. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* 100: 189-192.

42. Loraine AE, Helt GA, Cline MS and Siani-Rose MA. (2003) Exploring alternative transcript structure in the human genome using blocks and InterPro. *J Bioinform Comput Biol* 1: 289-306.

43. Resch A, Xing Y, Modrek B, Gorlick M, Riley R and Lee C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res* 3: 76-83.

44. Black DL. (2000) Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* 103: 367-370.

45. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE and Zipursky SL. (2000) Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101: 671-684.

46. Tress ML, Bodenmiller B, Aebersold R and Valencia A. (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol* 9: R162.

47. Blencowe BJ. (2006) Alternative splicing: New insights from global analyses. *Cell* 126: 37-47.

48. Kramer A. (1993) Mammalian protein factors involved in nuclear pre-mRNA splicing. *Mol Biol Rep* 18: 93-98.

49. Bruzik JP. (1996) Splicing glue: A role for SR proteins in trans splicing?. *Microb Pathog* 21: 149-155.

50. Kramer A. (1996) The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu Rev Biochem* 65: 367-409.

51. Long JC and Caceres JF. (2009) The SR protein family of splicing factors: Master regulators of gene expression. *Biochem J* 417: 15-27.

52. Sanford JR, Ellis J and Caceres JF. (2005) Multiple roles of arginine/serine-rich splicing factors in RNA processing. *Biochem Soc Trans* 33: 443-446.

53. Dreyfuss G, Matunis MJ, Pinol-Roma S and Burd CG. (1993) hnRNP proteins and the biogenesis of mRNA. *Annu Rev Biochem* 62: 289-321.

54. Varani G and Nagai K. (1998) RNA recognition by RNP proteins during RNA processing. *Annu Rev Biophys Biomol Struct* 27: 407-445.

55. Cobianchi F, Karpel RL, Williams KR, Notario V and Wilson SH. (1988) Mammalian heterogeneous nuclear ribonucleoprotein complex protein A1. large-scale overproduction in escherichia coli and cooperative binding to single-stranded nucleic acids. *J Biol Chem* 263: 1063-1071.

56. Swanson MS, Nakagawa TY, LeVan K and Dreyfuss G. (1987) Primary structure of human nuclear ribonucleoprotein particle C proteins: Conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. *Mol Cell Biol* 7: 1731-1739.

57. Caceres JF, Stamm S, Helfman DM and Krainer AR. (1994) Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science* 265: 1706-1709.

58. Hanamura A, Caceres JF, Mayeda A, Franza BR,Jr and Krainer AR. (1998) Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA* 4: 430-444.

59. Martinez-Contreras R, Fisette JF, Nasim FU, Madden R, Cordeau M and Chabot B. (2006) Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol* 4: e21.

60. Lin CH and Patton JG. (1995) Regulation of alternative 3' splice site selection by constitutive splicing factors. *RNA* 1: 234-245.

61. Matlin AJ, Southby J, Gooding C and Smith CW. (2007) Repression of alpha-actinin SM exon splicing by assisted binding of PTB to the polypyrimidine tract. *RNA* 13: 1214-1223.

62. Lou H, Gagel RF and Berget SM. (1996) An intron enhancer recognized by splicing factors activates polyadenylation. *Genes Dev* 10: 208-219.

63. Li Q, Lee JA and Black DL. (2007) Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci* 8: 819-831.

64. Buckanovich RJ and Darnell RB. (1997) The neuronal RNA binding protein nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol Cell Biol* 17: 3194-3201.

65. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ and Darnell RB. (2006) An RNA map predicting nova-dependent splicing regulation. *Nature* 444: 580-586.

66. Polydorides AD, Okano HJ, Yang YY, Stefani G and Darnell RB. (2000) A brain-enriched polypyrimidine tract-binding protein antagonizes the ability of nova to regulate neuron-specific alternative splicing. *Proc Natl Acad Sci U S A* 97: 6350-6355.

67. Barreau C, Paillard L, Mereau A and Osborne HB. (2006) Mammalian CELF/Bruno-like RNA-binding proteins: Molecular characteristics and biological functions. *Biochimie* 88: 515-525.

68. Han J and Cooper TA. (2005) Identification of CELF splicing activation and repression domains in vivo. *Nucleic Acids Res* 33: 2769-2780.

69. Gromak N, Matlin AJ, Cooper TA and Smith CW. (2003) Antagonistic regulation of alpha-actinin alternative splicing by CELF proteins and polypyrimidine tract binding protein. *RNA* 9: 443-456.

70. Suzuki H, Jin Y, Otani H, Yasuda K and Inoue K. (2002) Regulation of alternative splicing of alpha-actinin transcript by bruno-like proteins. *Genes Cells* 7: 133-141.

71. Philips AV, Timchenko LT and Cooper TA. (1998) Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science* 280: 737-741.

72. Savkur RS, Philips AV and Cooper TA. (2001) Aberrant regulation of insulin receptor alternative splicing is associated with insulin resistance in myotonic dystrophy. *Nat Genet* 29: 40-47.

73. Sorek R and Ast G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res* 13: 1631-1637.

74. Pozzoli U and Sironi M. (2005) Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* 62: 1579-1604.

75. Tacke R, Tohyama M, Ogawa S and Manley JL. (1998) Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell* 93: 139-148.

76. Chan RC and Black DL. (1997) The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. *Mol Cell Biol* 17: 4667-4676.

77. Hwang B, Lim JH, Hahm B, Jang SK and Lee SW. (2009) hnRNP L is required for the translation mediated by HCV IRES. *Biochem Biophys Res Commun* 378: 584-588.

78. Lenasi T, Peterlin BM and Dovc P. (2006) Distal regulation of alternative splicing by splicing enhancer in equine beta-casein intron 1. *RNA* 12: 498-507.

79. Miriami E, Margalit H and Sperling R. (2003) Conserved sequence elements associated with exon skipping. *Nucleic Acids Res* 31: 1974-1983.

80. Dominski Z and Kole R. (1991) Selection of splice sites in pre-mRNAs with short internal exons. *Mol Cell Biol* 11: 6075-6083.

81. Sterner DA, Carlo T and Berget SM. (1996) Architectural limits on split genes. *Proc Natl Acad Sci U S A* 93: 15081-15085.

82. Dominski Z and Kole R. (1992) Cooperation of pre-mRNA sequence elements in splice site selection. *Mol Cell Biol* 12: 2108-2114.

83. Dye BT, Buvoli M, Mayer SA, Lin CH and Patton JG. (1998) Enhancer elements activate the weak 3' splice site of alpha-tropomyosin exon 2. *RNA* 4: 1523-1536.

84. Graveley BR. (2000) Sorting out the complexity of SR protein functions. *RNA* 6: 1197-1211.

85. Modrek B and Lee C. (2002) A genomic view of alternative splicing. *Nat Genet* 30: 13-19.

86. Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, Jin L, Motono C, Hata H, Isogai T, Nagai K, Otsuki T, Kuryshev V, Shionyu M, Yura K, Go M, Thierry-Mieg J, Thierry-Mieg D, Wiemann S, Nomura N, Sugano S, Gojobori T and Imanishi T. (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res* 34: 3917-3928.

87. Florea L, Hartzell G, Zhang Z, Rubin GM and Miller W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 8: 967-974.

88. Kent WJ. (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12: 656-664.

89. Wu TD and Watanabe CK. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21: 1859-1875.

90. Gupta S, Zink D, Korn B, Vingron M and Haas SA. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics* 5: 72.

91. Cuperlovic-Culf M, Belacel N, Culf AS and Ouellette RJ. (2006) Microarray analysis of alternative splicing. *OMICS* 10: 344-357.

92. Moore MJ and Silver PA. (2008) Global analysis of mRNA splicing. *RNA* 14: 197-203.

93. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, Frey BJ and Blencowe BJ. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* 16: 929-941.

94. Fagnani M, Barash Y, Ip J, Misquitta C, Pan Q, Saltzman A, Shai O, Lee L, Rozenhek A, Mohammad N, Willaime-Morawek S, Babak T, Zhang W, Hughes T, van der Kooy D, Frey B and Blencowe B. (2007) Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol* 8: R108.

95. Relogio A, Ben-Dov C, Baum M, Ruggiu M, Gemund C, Benes V, Darnell RB and Valcarcel J. (2005) Alternative splicing microarrays reveal functional expression of neuron-specific regulators in hodgkin lymphoma cells. *J Biol Chem* 280: 4779-4784.

96. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A and Blume JE. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol* 8: R64.

97. Hall N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* 210: 1518-1525.

98. 't Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ and den Dunnen JT. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 36: e141.

99. Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5: 621-628.

100. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H and Yaspo ML. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956-960.

101. Liu HX, Zhang M and Krainer AR. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12: 1998-2012.

102. Wu S and Green MR. (1997) Identification of a human protein that recognizes the 3' splice site during the second step of pre-mRNA splicing. *EMBO J* 16: 4421-4432.

103. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A and Darnell RB. (2003) CLIP identifies nova-regulated RNA networks in the brain. *Science* 302: 1212-1215.

104. Ule J, Jensen K, Mele A and Darnell RB. (2005) CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods* 37: 376-386.

105. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M and Burge CB. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell* 119: 831-845.

106. Holste D and Ohler U. (2008) Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. *PLoS Comput Biol* 4: e21.

107. Cartegni L, Chew SL and Krainer AR. (2002) Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat Rev Genet* 3: 285-298.

108. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I and Conboy JG. (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res* 29: 2338-2348.

109. Yeo G, Holste D, Kreiman G and Burge C. (2004) Variation in alternative splicing across human tissues. *Genome Biol* 5: R74.

110. Fairbrother WG, Yeh RF, Sharp PA and Burge CB. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007-1013.

111. Zhang XH and Chasin LA. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 18: 1241-1250.

112. Zhang XH, Heller KA, Hefter I, Leslie CS and Chasin LA. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res* 13: 2637-2650.

113. Zhang XH, Leslie CS and Chasin LA. (2005) Dichotomous splicing signals in exon flanks. *Genome Res* 15: 768-779.

114. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T and Ast G. (2006) Comparative analysis identifies exonic splicing regulatory sequences--the complex definition of enhancers and silencers. *Mol Cell* 22: 769-781.

115. Faustino NA and Cooper TA. (2003) Pre-mRNA splicing and human disease. *Genes Dev* 17: 419-437.

116. Wang GS and Cooper TA. (2007) Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 8: 749-761.

117. Barberan-Soler S and Zahler AM. (2008) Alternative splicing regulation during C. elegans development: Splicing factors as regulated targets. *PLoS Genet* 4: e1000001.

118. Bailey TL and Elkan C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2: 28-36.

119. Zhao S, Kim J and Heber S. (2007) Large-scale discovery of regulatory motifs involved in alternative splicing. *Proc IEEE Int Conf BioInformatics BioEngineering* 1399-1403.

120. Krawczak M, Reiss J and Cooper DN. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences. *Hum Genet* 90: 41-54.

121. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM. (1999) Systematic determination of genetic network architecture. *Nat Genet* 22: 281-285.

122. Vilo J, Brazma A, Jonassen I, Robinson A and Ukkonen E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc Int Conf Intell Syst Mol Biol* 8: 384-394.

123. Eden E, Lipson D, Yogev S and Yakhini Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* 3: e39.

124. Jakel J and Martin N. (2004) Validation in the cluster analysis of gene expression data. Workshop on the Fuzzy Systen and Computational Intelligence, pp. 13–32, 2004.

125. Pietrokovski S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 24: 3836-3845.

126. Bailey TL and Gribskov M. (1998) Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* 14: 48-54.

127. Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL and Thanaraj TA. (2006) ASD: A bioinformatics resource on alternative splicing. *Nucleic Acids Res* 34: D46-55.

128. Charlet-B N, Savkur RS, Singh G, Philips AV, Grice EA and Cooper TA. (2002) Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing. *Mol Cell* 10: 45-53.

129. Niksic M, Romano M, Buratti E, Pagani F and Baralle FE. (1999) Functional analysis of cis-acting elements regulating the alternative splicing of human CFTR exon 9. *Hum Mol Genet* 8: 2339-2349.

130. Kanopka A, Muhlemann O and Akusjarvi G. (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* 381: 535-538.

131. Carstens RP, Wagner EJ and Garcia-Blanco MA. (2000) An intronic splicing silencer causes skipping of the IIIb exon of fibroblast growth factor receptor 2 through involvement of polypyrimidine tract binding protein. *Mol Cell Biol* 20: 7388-7400.

132. Reddy TE, Shakhnovich BE, Roberts DS, Russek SJ and DeLisi C. (2007) Positional clustering improves computational binding site detection and identifies novel cis-regulatory sites in mammalian GABAA receptor subunit genes. *Nucleic Acids Res* 35: e20.

133. Rice P, Longden I and Bleasby A. (2000) EMBOSS: The european molecular biology open software suite. *Trends Genet* 16: 276-277.

134. Benjamini Y and Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.Series B (Methodological)* 57: 289-300.

135. Yeo GW, Van Nostrand EL and Liang TY. (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet* 3: e85.

136. Vingron M, Brazma A, Coulson R, van Helden J, Manke T, Palin K, Sand O and Ukkonen E. (2009) Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol* 10: 202.

137. Bussemaker HJ, Li H and Siggia ED. (2001) Regulatory element detection using correlation with expression. *Nat Genet* 27: 167-171.

138. Conlon EM, Liu XS, Lieb JD and Liu JS. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A* 100: 3339-3344.

139. Smith AD, Sumazin P, Das D and Zhang MQ. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 21 Suppl 1: i403-12.

140. Marchler-Bauer A and Bryant SH. (2004) CD-search: Protein domain annotations on the fly. *Nucl Acids Res* 32: W327-331.

141. McCullough AJ and Berget SM. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol* 17: 4562-4571.

142. Burd CG and Dreyfuss G. (1994) RNA binding specificity of hnRNP A1: Significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *EMBO J* 13: 1197-1204.

143. Zhao X, Rush M and Schwartz S. (2004) Identification of an hnRNP A1-dependent splicing silencer in the human papillomavirus type 16 L1 coding region that prevents premature expression of the late L1 gene. *J Virol* 78: 10888-10905.

144. Sinha S and Tompa M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 31: 3586-3588.

145. Pavesi G, Mereghetti P, Mauri G and Pesole G. (2004) Weeder web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* 32: W199-203.

146. Stormo GD. (2000) DNA binding sites: Representation and discovery. *Bioinformatics* 16: 16-23.

147. Tsai HK, Huang GT, Chou MY, Lu HH and Li WH. (2006) Method for identifying transcription factor binding sites in yeast. *Bioinformatics* 22: 1675-1681.

148. Holste D, Huo G, Tung V and Burge CB. (2006) HOLLYWOOD: A comparative relational database of alternative splicing. *Nucleic Acids Res* 34: D56-62.

149. Huang HD, Horng JT, Lin FM, Chang YC and Huang CC. (2005) SpliceInfo: An information repository for mRNA alternative splicing in human genome. *Nucleic Acids Res* 33: D80-5.

150. Ji H, Zhou Q, Wen F, Xia H, Lu X and Li Y. (2001) AsMamDB: An alternative splice database of mammals. *Nucleic Acids Res* 29: 260-263.

151. Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O and Zhang MQ. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol* 19: 739-756.

152. Liu HX, Chew SL, Cartegni L, Zhang MQ and Krainer AR. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* 20: 1063-1071.

153. Tacke R and Manley JL. (1995) The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J* 14: 3540-3551.

154. Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ and Krainer AR. (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* 15: 2490-2508.

155. Tacke R, Chen Y and Manley JL. (1997) Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: Creation of an SRp40-specific splicing enhancer. *Proc Natl Acad Sci U S A* 94: 1148-1153.

156. Cavaloc Y, Bourgeois CF, Kister L and Stevenin J. (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* 5: 468-483.

157. Paradis C, Cloutier P, Shkreta L, Toutant J, Klarskov K and Chabot B. (2007) hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. *RNA* 13: 1287-1300.

158. Shin C and Manley JL. (2002) The SR protein SRp38 represses splicing in M phase cells. *Cell* 111: 407-417.

159. Eldridge AG, Li Y, Sharp PA and Blencowe BJ. (1999) The SRm160/300 splicing coactivator is required for exon-enhancer function. *Proc Natl Acad Sci U S A* 96: 6125-6130.

160. Cheng C and Sharp PA. (2006) Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. *Mol Cell Biol* 26: 362-370.

161. Moran-Jones K, Wayman L, Kennedy DD, Reddel RR, Sara S, Snee MJ and Smith R. (2005) hnRNP A2, a potential ssDNA/RNA molecular adapter at the telomere. *Nucleic Acids Res* 33: 486-496.

162. Gorlach M, Burd CG and Dreyfuss G. (1994) The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. *J Biol Chem* 269: 23074-23078.

163. Coles JL, Hallegger M and Smith CW. (2009) A nonsense exon in the Tpm1 gene is silenced by hnRNP H and F. *RNA* 15: 33-43.

164. Nasim MT, Chernova TK, Chowdhury HM, Yue BG and Eperon IC. (2003) HnRNP G and Tra2beta: Opposite effects on splicing matched by antagonism in RNA binding. *Hum Mol Genet* 12: 1337-1348.

165. Hui J, Hung LH, Heiner M, Schreiner S, Neumuller N, Reither G, Haas SA and Bindereif A. (2005) Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. *EMBO J* 24: 1988-1998.

166. Ashiya M and Grabowski PJ. (1997) A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: Evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. *RNA* 3: 996-1015.

167. Perez I, Lin CH, McAfee JG and Patton JG. (1997) Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. *RNA* 3: 764-778.

168. Markovtsov V, Nikolic JM, Goldman JA, Turck CW, Chou MY and Black DL. (2000) Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol* 20: 7463-7479.

169. Peng R, Dye BT, Perez I, Barnard DC, Thompson AB and Patton JG. (2002) PSF and p54nrb bind a conserved stem in U5 snRNA. *RNA* 8: 1334-1347.

170. Jin Y, Suzuki H, Maegawa S, Endo H, Sugano S, Hashimoto K, Yasuda K and Inoue K. (2003) A vertebrate RNA-binding protein fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J* 22: 905-912.

171. Ponthier JL, Schluepen C, Chen W, Lersch RA, Gee SL, Hou VC, Lo AJ, Short SA, Chasis JA, Winkelmann JC and Conboy JG. (2006) Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. *J Biol Chem* 281: 12468-12474.

172. Marquis J, Paillard L, Audic Y, Cosson B, Danos O, Le Bec C and Osborne HB. (2006) CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochem J* 400: 291-301.

173. Faustino NA and Cooper TA. (2005) Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment. *Mol Cell Biol* 25: 879-887.

174. Ho TH, Charlet-B N, Poulos MG, Singh G, Swanson MS and Cooper TA. (2004) Muscleblind proteins regulate alternative splicing. *EMBO J* 23: 3103-3112.

175. Galarneau A and Richard S. (2005) Target RNA motif and target mRNAs of the quaking STAR protein. *Nat Struct Mol Biol* 12: 691-698.

176. Gilbert W. (1978) Why genes in pieces? *Nature* 271: 501.

177. Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G and Shamir R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res* 14: 1617-1623.

178. Das D, Banerjee N and Zhang MQ. (2004) Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A* 101: 16234-16239.

179. Beer MA and Tavazoie S. (2004) Predicting gene expression from sequence. *Cell* 117: 185-198.

180. Smith AD, Sumazin P, Xuan Z and Zhang MQ. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A* 103: 6275-6280.

-