CHAPTER 6

Searching for Splicing Motifs

Lawrence A. Chasin*

Abstract

Intron removal during pre-mRNA splicing in higher eukaryotes requires the accurate identification of the two splice sites at the ends of the exons, or exon definition. The sequences constituting the splice sites provide insufficient information to distinguish true splice sites from the greater number of false splice sites that populate transcripts. Additional information used for exon recognition resides in a large number of positively or negatively acting elements that lie both within exons and in the adjacent introns. The identification of such sequence motifs has progressed rapidly in recent years, such that extensive lists are now available for exonic splicing enhancers and exonic splicing silencers. These motifs have been identified both by empirical experiments and by computational predictions, the validity of the latter being confirmed by experimental verification. Molecular searches have been carried out either by the selection of sequences that bind to splicing factors, or enhance or silence splicing in vitro or in vivo. Computational methods have focused on sequences of 6 or 8 nucleotides that are over- or under-represented in exons, compared to introns or transcripts that do not undergo splicing. These various methods have sought to provide global definitions of motifs, yet the motifs are distinctive to the method used for identification and display little overlap. Astonishingly, at least three-quarters of a typical mRNA would be comprised of these motifs. A present challenge lies in understanding how the cell integrates this surfeit of information to generate what is usually a binary splicing decision.

Splice Site Sequences Are Necessary but Not Sufficient

In the process of converting a pre-mRNA molecule to a mature mRNA, introns are removed by the spliceosome, a very large protein-RNA complex that contains five small nuclear RNA molecules and scores of proteins (refer to chapter by Matlin and Moore). During this reaction, the two bordering exons must be brought close together, much as two substrates in a synthetic reaction of intermediary metabolism. But in the latter case, each of the two substrates usually consists of a population of identical molecules, whereas the two ends of the intron have a varied composition. The enzyme that is the spliceosome must bring together the GU and AG (which are almost always identical) in the midst of some variety among the adjacent nucleotides.

The Splice Sites

The adjacent nucleotides at each splice site are far from random; they comprise two easily distinguished consensus sequences of nine bases for the 5' splice site and about 15 bases for the 3' splice site. Position specific scoring matrices (PSSM) compiled from thousands of introns reflect the relative contributions of each base at each position and allow any given sequence to be quantitatively evaluated for its degree of agreement to a consensus. One widely used such index is the consensus value (CV), which ranges from 100 (perfect consensus) to 0 (the worst consensus).[1,2] The median CVs of human 5' and 3' splice sites are 82 and 80, respectively, and the distribution of

*Lawrence A. Chasin—Department of Biological Sciences, Columbia University, New York, New York 10027, USA. Email: lac2@columbia.edu

Alternative Splicing in the Postgenomic Era, edited by Benjamin J. Blencowe and Brenton R. Graveley.

scores is wide: cutoffs of 78 for 5' splice sites and 75 for 3' splice sites capture only three-quarters of the sites. Interestingly, the consensus sequences themselves do not represent a majority among splice sites. For instance, in a set of 5000 randomly chosen constitutive exons, less than 5% contain 5' splice sites that perfectly match the consensus ((C or A)AG/GT(A or G)AGT) and the consensus sequences do not represent the four most common 5' splice site sequences (Fig. 1A). 5' Splice sites that contain only three of the seven variable bases are not uncommon; the data in Figure 1B suggest that about 20,000 such mismatched 5' splice site sequences are present in the human transcriptome.

The protein factors that recognize the 5' and 3' splice sites need to bind to many distinct sequences. As an example, U2AF,[65] which binds to the polypyrimidine tract of 3' splice sites, can accommodate a wide variety of pyrimidine rich sequences in its binding site.[3] Thus, this degree of diversity might be tolerable if introns and exons had evolved to lack sequences that resemble the splice site consensuses, so that the splice sites would be easily recognizable despite their degeneracy. But just the opposite is the case: pseudo splice site sequences (false splice site sequences that are not used) are about an order of magnitude more abundant than the real splice sites in large transcripts and are present at a frequency similar to or greater than that expected by chance (see Fig. 1C,D for the human *HPRT* gene). Moreover, many pseudo 5' splice sites exist that perfectly match the sequence of real splice sites; in these cases factors other than intrinsic strength[4] must play a role in distinguishing between the real and pseudo sites.

Different splice site sequences have different strengths and this strength generally correlates with the CV score. Thus the words "strong" and "weak" usually refer to the CV and not to a splicing measurement. Indeed, splicing regulation takes advantage of this strength—on average, alternative splice sites are slightly weaker than constitutive splice sites.[5-7] However, the correlation between splice site strength and splicing is far from perfect. For instance, Eperon and colleagues placed different 5' splice sites in competition with a constant globin 5' splice site and measured the proportion of splicing at the test site.[8] Correlation coefficients between splicing efficiency and agreement to the consensus were respectable (0.68 to 0.76) but far from perfect. Strength experiments have usually been set up as competitions between two nearby splice sites,[8] a situation that is not always the case for endogenous splice sites. That is, a weak splice site may be recognized efficiently if no nearby competitor is present. Inefficient splicing of a splice site in a heterologous context implies that in the natural context a splice site communicates with other nearby sequence elements. Splice site sequences may even have to be tailored to their context. For examples, mutation of a *DHFR* 5' splice site from AGA/GTAAGT (CV 79.6) to AGG/GTCAGT (CV 80.9) preserved the CV and the predicted ability to form a duplex with U1 snRNA, yet reduced splicing efficiency from 100% to 3%.[9] More sophisticated methods, such as treating the PSSM as probabilities,[10] using maximal dependence decomposition (MDD),[11] or a support vector machine (SVM),[12] may marginally improve splice site predictions. However, such enhancements do not change the conclusion that many real weak splice sites must be efficiently recognized, while many strong pseudo splice sites must be ignored in the course of splicing a typical pre-mRNA.

The Branch Point

A third element that plays a central role in pre-mRNA splicing is the branch point. The human branchpoint consensus is YNYCRAY, although this sequence was derived from the biochemical characterization of a small number of branchpoints.[13] The conserved adenosine attacks the 5' splice site and is usually located 18 to 40 nt upstream of the 3' splice site, although it can be more distant.[14] The variable distance and poor conservation of the branchpoint makes it a poor predictor of real 3' splice sites. For instance, including the branchpoint in a computational search for real 3' splice sites in the *HPRT* gene did not increase the accuracy of 3' splice site predictions.[15]

Exon Definition

The excision of an intron requires the pairing of splice sites at the ends of the intron, which can be considered "intron definition". However, the initial recognition of most splice sites probably involves

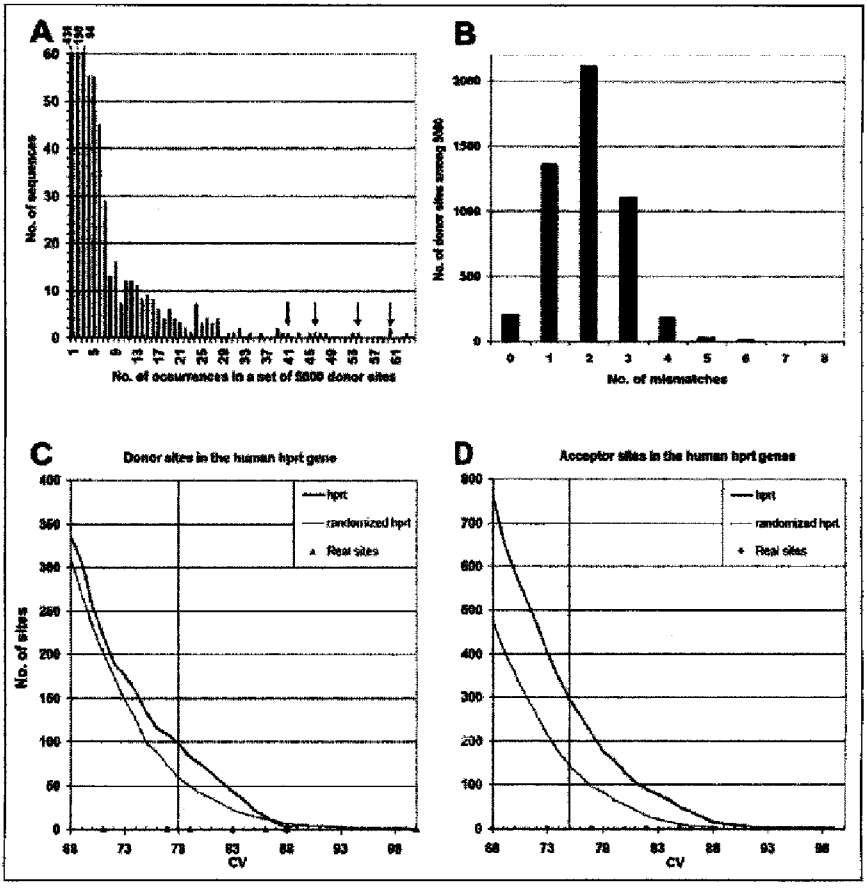

Figure 1. A) Histogram of the number of occurrences of each unique splice donor site sequence found among a set of 5000. The arrows show the points representing the four consensus sequences. For instance, there are 2 sequences that each appear 61 times in this set of 5000, one of which is a consensus sequence. B) Distribution of the number of mismatches to the 4 donor site consensus sequences among 5000 human donor sites. C) Frequency of pseudo donor splice site sequences in the 40,000 nt human *HPRT* transcript having the indicated minimum CV score. Also shown is the same analysis of randomized versions of the HPRT transcript (average of 10 randomizations). The symbols along the abscissa indicate the values for the eight real splice sites. The vertical line indicates the third quartile score for donor sites of real exons (i.e., one-quarter of real donor sites have CV scores below that value). D) As C, but for acceptor sites.

"exon definition", the identification of the two splice sites across the exon. There is plentiful genetic evidence supporting this idea, in that the usual consequence of mutating one splice site is skipping of the exon—the remaining wild type splice site on the other end of the exon is not used.[9,16] Similarly, a downstream 5' splice site greatly enhances splicing to an upstream 3' splice site in vitro.[17,18] Terminal exons are defined by interactions between factors that recognize the 5' cap and U1 snRNP for the first exon[19,20] and polyadenylation factors and splicing factors that recognize the 3' splice site for the last exon.[21] Despite the widespread acceptance of exon definition, the molecular basis for the

implied communication has been seldom studied,[18,22] with experimental designs favoring 2-exon RNA molecules and interpretations emphasizing spliceosomal interactions across introns.

Internal human exons have an average size of about 120 nt and less than 5% of exons are greater than 250 nt in length.[23] If one adds the constraint that a potential 3' splice site must be followed by a 5' splice site within 250 nt, then the number of false 3' splice sites is substantially reduced, but the number of false 5' splice sites is not, as they become the limiting factor. The pseudo exons that are defined by these false 3' and 5' splice sites which are assumed to be never used, outnumber the real exons by more than an order of magnitude.[15] If we accept exon definition as the usual case, then the problem becomes that of distinguishing real exons from these pseudo exons.

Additional Sequence Information Lies within Exons and Introns

Early experiments implicated exons as a source of information necessary for alternative splicing.[24,25] In 1993, Shimura and colleagues defined an exonic enhancer sequence as a short purine rich sequence.[26] Since that discovery there has been a steady stream of descriptions of analogous regulatory sequences. These splicing regulatory sequences fall into four categories based on their location and their mode of action: exonic splicing enhancers and silencers (exonic splicing enhancers [ESEs] and exonic splicing silencers [ESSs]) and their intronic counterparts (intronic splicing enhancers [ISEs] and intronic splicing silencers [ISSs]). There are myriad examples of ESEs and the great majority of these have been identified from studies of alternatively spliced exons. More recently, it has become evident that constitutively spliced exons require ESEs as well.[27,28] In general, ESEs bind members of the SR protein family (refer to chapter by Lin and Fu).[29] All SR proteins have an arginine-serine (RS)-rich domain that can interact with other proteins[30] and with the RNA itself.[31] They also contain one or more RNA recognition motifs (RRMs). Most of the SR protein RRMs bind to a highly degenerate set of RNA sequences, yet display enough specificity so as to be distinguishable from one another.[32]

Although less broadly studied, a number of ESSs have been identified in alternative exons.[33] ESSs are typically bound by heterogeneous nuclear ribonuclear proteins (hnRNPs) (refer to chapter by Martinez-Contreras et al), such as hnRNPA1 and hnRNPI (polypyrimidine binding protein, PTB). Like SR proteins, hnRNPs show preferences for particular sequence motifs while binding many other sequences with less, yet notable, affinity.[34] Fewer ISEs and ISSs have been described, but some of these have been extensively characterized.[35] Comprehensive lists of mammalian alternative exons subject to both enhancement and silencing (both by ESSs and ISSs) together with many of their mediators have recently been compiled.[33,36]

Global Approaches for Defining Sequence Motifs for Splicing

A powerful approach to understanding how splice sites are recognized and regulated is to use bioinformatics or experimental approaches to define all the cis-elements that are implicated in splice site recognition. The hope here is that general rules will become evident as one uncovers the "splicing code".[37] The global approaches have been principally two: (1) statistical analysis of genomic sequences to find motifs associated with enhancement or silencing; and (2) molecular selection to define all the sequence motifs that will enhance or silence splicing in a particular context and/or in response to a particular splicing factor, or to find the sequences that bind best to a purified splicing factor. The remainder of this chapter will focus on such global approaches. Understanding the splicing code will allow for a more exhaustive identification of exons and therefore of genes and proteins, with wide implications for genomics and medicine. In addition, it will help us predict the patterns of alternative splicing and understand the mechanism and regulation of splicing.

Exonic Splicing Enhancers (ESEs) Predicted by Computation

ESEs lie by definition within exons and most exons differ from the rest of the genome in containing sequences that must code for proteins. Thus a search for motifs that are abundant in exons vs. other regions would be confounded by the inevitable emergence of common codon sequences. This problem has been dealt with in several different ways.

Fedorov et al[38] compared the frequencies of tetramers and pentamers in exons to those in intronless genes, reasoning that while both code for proteins, the former require splicing signals but the latter do not. Twenty-three sequences were identified that were significantly more abundant in exons, ranging from 17% to 42% overrepresentation. The number of intronless genes used in this study was relatively small,[63] perhaps limiting significance scores; and the biological activity of the sequences found was not tested.

Fairbrother et al[39] got around the protein coding problem by comparing exons to exons, thus neutralizing the effect of protein coding. They reasoned that ESEs should be more abundant in exons with weak splice sites than in exons with strong splice sites. Using all 4096 possible hexamers, they identified motifs for which this frequency difference was high, treating donor sites and acceptor sites separately. To sharpen the selection, they added another criterion: the motifs must also be more abundant in exons compared to flanking intronic regions. Using a cutoff of >2.5 standard deviations for both criteria yielded a combined set of 238 hexamers, or 5.8% of all possibilities. About a quarter of these were common to 5' and 3' splice sites. Many of these motifs were shown to be active in functional assays demonstrating the validity of this approach. Thus most of these hexamers are capable of acting as ESEs and are known as "RESCUE-ESEs". Since the success rate of the validity tests was high, one must conclude that this selection was stringent and that additional hexamers falling below the cutoffs may also act as ESEs. Even at a selection rate of 5.8%, about 23% of randomized exon sequences would correspond to RESCUE-ESEs (Table 1). Thus this study suggested that ESEs are abundant motifs. RESCUE-ESE sequences can be found at http://genes.mit.edu/burgelab/rescue-ese/ESE.txt.

In all motif selection experiments (computational or molecular) there are caveats due to biases inherent in each selection strategy. For example, in the RESCUE-ESE approach, by focusing on exons with weak splice sites, there may have been a biased selection for ESEs associated with alternatively spliced exons, since as a whole they exhibit weaker splice site sequences than constitutive exons.[5-7] A more subtle bias arises from the fact that the transcriptome has an intrinsically high A + T content of 57%.[2] When that content is reasonably used as a background to calculate splice site PSSM scores, G + C-rich splice site sequences will tend to stand out in information content as "strong" (more distinct) whereas A + T-rich sequences will appear "weaker" (less distinct from background). A search for weak splice sites using PSSM values will thus favor A + T-rich sequences for this reason alone and these sites will be associated with A + T rich genes (in A + T rich isochores) and consequently A + T-rich ESE candidates. This argument could explain why RECUE-ESEs have a relatively high (61%) A + T content (Table 2).

The validity of RESCUE-ESEs was subsequently tested by examining evolutionary conservation. SNP density at synonymous sites within these motifs is lower than expected, especially when located nearer to splice sites, supporting the idea that these motifs have been subject to purifying selection and thus are functional.[40]

A second study used the same general approach, but different criteria to search for ESE candidates. Our laboratory[41] circumvented the protein coding problem by limiting the analysis to nonprotein-coding exons. Forty percent of human genes contain noncoding first exons[42] and there are also a substantial number of genes with translation initiation sites located in the 3rd exon, or an exon that is further downstream. The latter represent a pool of about 2000 internal noncoding exons, of which about 500 were chosen that were less likely to have originated from sequencing errors. We searched for all possible octamers in this exon set, allowing a single mismatch per octamer in order to obtain a sufficient number of hits. Octamers were identified that were present at a much higher frequency in the noncoding exons compared to the sequences of two different negative control sets: (1) pseudo exons from the same genes; and (2) the 5' untranslated regions (UTRs) of intronless genes. Neither of these sequences code for proteins and the intronless UTR sequences may contain information for stability, transport and translation that should also be present in the noncoding exons and so would be filtered out. Motifs that fell above 2.8 standard deviations from the mean were considered putative ESEs (PESEs) and numbered 2069 of 65536 possibilities, or 3.2%. This comparison also allowed the identification of motifs that are

Table 1. Exon coverage by predicted splicing motifs

	Mean Hits per Exon[1]	Hits per nt per Motif	Mean Density[2]	nt per 127 Base Exon[3]	Mean Density in Randomized Exons[4]	Real / Randomized Density[1]	% of Exons with No Hits	Pseudo Exon Mean Density[5]
PESE	9.3	3.8×10^{-5}	0.305	39	0.169	1.80	10.6	0.167
RESCUE-ESE	13.6	4.7×10^{-4}	0.319	41	0.231	1.38	3.2	0.219
PESE + RESCUE-ESE	22.9		0.472	60	0.322	1.47	1.2	0.304
Goren ESR	14.9	4.3×10^{-4}	0.526	67	0.335	1.57	0.4	0.444
PESS	1.5	1.2×10^{-5}	0.067	9	0.084	0.80	36.3	0.174
FAS-hex3	2.0	1.6×10^{-4}	0.068	9	0.094	0.72	19.0	0.117
PESS + FAS-hex3	3.5		0.127	16	0.161	0.79	9.1	0.263
All 5	41.3		0.740	94	0.595	1.24	0.02	

[1]Based on 5000 randomly chosen constitutive exons, 50 to 250 nt in length. [2]Mean nt per exonic nt for each exon. [3]The average size of the 5000 exons tested. [4]All nts in each exon were shuffled randomly. [5]In a set of 1803 randomly chosen pseudo exons 50 to 250 in length with acceptor/donor site CVs > 75/78.

rare in real exons compared to pseudo exons and the 5′ UTRs of intronless genes and these were considered putative ESSs (PESSs). Eight of eight PESEs enhanced splicing in a functional assay and single base mutations that reduced the PESE scores to near neutrality reduced this activity. Of 58 examples of mutations reported in the literature to reduce splicing, 33% could be explained by the disruption of a PESE (and 28% could be explained be the creation of a PESS). Again, because the success rate of the validation tests was high, it is likely that additional PESEs would be found among octamers with somewhat lower scores than those chosen. At the conservative threshold of 3.2%, about 17% of a randomized exon sequence would be represented by PESEs (Table 1); so like RESCUE-ESEs, these motifs are abundant. The average internal (coding) constitutive exon of 120 nt contains nine PESEs, often in overlapping clusters. PESEs are 2-fold more abundant in exons compared to introns. A full set of PESE sequences can be found at http://www.columbia.edu/cu/biology/faculty/chasin/xz3/pese262.rxt. A list of the scores for each of the two criteria for all 65,536 octamers can be found at http://www.columbia.edu/cu/biology/faculty/chasin/xz3/octamers.txt.

Again, biases could have influenced the types of motifs that were selected. The 5′ UTRs of intronless genes that were used as an ESE-under-represented data set are often situated within regions that are rich in CpG sequences, since the CpG islands that lie upstream of numerous genes often extend as much as 2 kb into the gene.[43] Moreover, noncoding exons used as the positive examples are low in CpG content relative to coding exons.[44] For both of these reasons, CpG-containing ESE motifs may have been under represented in this selection, since they may not be enriched over the relatively high background of CpG-containing octamers in the 5′ UTR of intronless genes. In the earlier comparison of exons to intronless genes mentioned above,[38] most of the candidate ESS pentamers identified as being relatively scarce in exons contained CpG dinucleotides. Although they have a CpG content similar to that of exons as a whole (Table 2, compare columns 2 and 12), PESEs do not include many ESEs predicted from molecular selections and these tend to have very high CpG contents (Table 2, columns 5 to 8). Another possible weakness in the selection method described above stems from the assumption that noncoding genes do not contain protein coding information. In fact, such exons may have coded for proteins in the evolutionary past and maintained a vestige of this nonrandomness. In this case, PESE candidates that merely overlap with highly used dicodons could have been isolated as false positives. However, such sequences may tend to be ESEs nonetheless and the high validation rates of PESEs argues against this possibility.

PESEs were subsequently tested in a more rigorous fashion.[28] Six real mammalian exons (five constitutive and one alternative) were computationally scanned for PESEs. About four PESE clusters per 100 nt were found. By knocking out each individual PESE cluster with single base substitutions and assaying splicing in vivo, 18 of the 22 predicted PESEs were shown to be functional. In addition to functionality, this result showed that each exon required nearly all of its ESEs to work in concert to promote efficient splicing; i.e., most were not redundant. A similar test has been carried out using a minigene containing an alternatively spliced alpha-tropomyosin exon.[45] Eleven PESEs or PESSs were mutated to reduce their absolute scores and in 10 of the 11 trials the splicing levels responded accordingly (J. Coles and C.W. Smith, personal communication). As well as providing additional validation of PESEs, this study provides the first such experimental test of PESSs.

Although entirely different criteria were used to select RESCUE-ESEs and PESEs, they show considerable overlap, a fact that further supports the validity of both sets (Table 3). At the same time, the two sets contain distinct ESEs. As can be seen in Table 1 (column 4), each motif set covers about 30% of the nucleotides in a collection of 5000 human exons, but together they cover 47%, only slightly less than would be expected if they were randomly associated (~52%). Thus these two conservatively derived sets of ESEs already cover half of an average exon and there are several additional motif sets yet to be discussed.

Table 2. Base compositions of motif sets

1 Sequence Source:	2 RESCUE-ESEs	3 PESEs[4]	4 ESRs	5 Func. SELEX In Vivo ESEs	6 Func. SELEX (nuc. ext.)	8 Func. SELEX (nuc. ext.)	7 Func. SELEX ESEfinder (4 SRs)	9 Func. SELEX (ASF/SF2)[2]	10 PESSs[4]	11 Sironi ESSs[3]	12 Func. SELEX In Vivo ESSs	13 Real Exons[1]	14 Pseudo Exons[1]
Ref.:	Fairbrother, 2002[42]	Zhang, 2004[44]	Goren, 2006[51]	Coulter, 1997[83]	Tian, 1995[76], 2001[77]	Schaal, 1998[107]	Liu[78,79]; Cartegni, 2003[80]	Smith, 2006[49]	Zhang, 2004[44]	Sironi, 2004[53]	Wang, 2004[85]	1	1
No. of motifs	238	2060	285	30	101	46	104	60	1018	NA	103		
Sequence space	4096	65536	3721	4^{14}	4^{20}	4^{18}	4^{20}	4^{7}	65536	NA	4^{10}		
% of sequence space	5.8	3.1	7.7	10^{-5}	10^{-8}	10^{-7}	10^{-8}	0.4	1.6	NA	10^{-4}		
A%	48	31.2	28	25	34	12	22	20	29	15	17	26	26
C%	14	24.4	23	36	21	37	23	42	9	9	9	25	22
G%	25	28.5	25	30	31	35	36	30	15	53	36	27	23
T%	13	15.9	25	10	14	16	19	7	47	22	38	22	29
GC%	39	52.9	47	66	52	72	60	72	24	62	45	52	44
AT%	61	47.1	53	34	48	28	40	28	76	38	55	48	56
CpG%[5]	2.6	3.4	0.7	9.3	10.8	12.3	8.8	18.3	0.6	NA	2.6	3.2	1.3

NA: not available

[1] data from 5000 randomly chosen exons and 1803 randomly chosen pseudo exons with CV scores in the top 3 quartiles.

[2] 60 7mers from 7mers and 14mers

[3] Extracted from positional scoring matrix

[4] Using corrected data (http://www.columbia.edu/cu/biology/faculty/chasin/xz3/octamers.txt)

[5] Expectation by chance is about 6% (1/16)

Table 3. Overlaps between motif sets[1]

	Motif Set Size	PESEs[2]	Expected by Chance[3]	PESSs[2]	Expected by Chance[3]	Motif Reference
RESCUE-ESE[42]	238	74%	36%	11%	23%	45
FAS-hex3[85]	103	10%	37%	54%	23%	87
Goren ESR[51]	285	53%	44%	19%	21%	54

[1] Percentage of the indicated hexamer set members that can be found within PESE or PESS octamers.
[2] PESE, PESS: ref. 44, as amended at http://www.columbia.edu/cu/biology/faculty/chasin/xz3/octamers.txt
[3] calculated by simulation. Goren ESRs based on 3721 allowed hexamers.

Exonic Splicing Silencers (ESSs) Predicted by Computation

Global computational searches for ESS motifs have also been carried out. Sironi et al[46] collected a subset of pseudo exons that was rich in predicted ESEs and then searched for overrepresented hexamers as candidates for ESSs. A second criterion, overabundance in pseudo exons compared to sequences flanking the pseudo exons, was applied to normalize for possible base compositional differences between pseudo exon and exon regions. This second criterion also sharpened the search to make it test the hypothesis that ESSs function to prevent the splicing of pseudo exons, as opposed to simply being avoided in real exons. The winning motifs were clustered into families to generate three consensus sequences. One of the three ((T/G)G(T/A)GGGG) reduced exon inclusion about five-fold in a functional assay. This G-rich motif was overrepresented in a test set of pseudo exons compared to real exons.

A large set of putative ESSs emerged from our statistical analysis described above for PESEs.[41] By searching for octamers that were underrepresented in real exons compared to both pseudo exons and the 5' UTRs of intronless genes, the influence of codons was avoided and the influence of other nonsplicing signals residing in mRNA was minimized. A set of 974 PESSs was identified, grouped into families and a sampling tested in functional assays. Eleven of 12 PESSs increased exon skipping and single base mutations reversed this skipping. Sixteen of 58 exonic splicing mutations in the literature could be explained by PESS formation, a number comparable to those that could be explained by PESE disruption. The PESS set represents about 1.5% of all octamers. These sequences are very T-rich (47%) and C-poor (Table 2, column 9). PESSs are 3.5-fold more abundant in introns compared to exons overall and show an additional increase just downstream of real 5' splice sites, suggesting that they may function to facilitate accurate recognition of the real sites. They are also found at a higher frequency in the vicinity of pseudo exons, suggesting a use in repressing false splice sites. The combination of PESEs and PESSs increases the discrimination between real and pseudo exons: the ratio of PESEs to PESSs in real exons is 5.5 as opposed to 0.6 for pseudo exons. This difference has been used as a guide to suggest whether a given sequence is a exon or a pseudo exon (e.g., see Fig. 6 in reference 45). However, the frequency distribution of ESEs and ESSs in exons and pseudo exons overlap considerably, making them less than a reliable predictor of real exons. A list of PESSs is located at http://www.columbia.edu/cu/biology/faculty/chasin/xz3/pess262.txt .

Exonic Splicing Regulators (ESRs) Predicted by Computation

Yet another computational strategy to search for splicing regulatory motifs was devised by Goren et al.[47] Reasoning that splicing signals would be both conserved and abundant in exons, they ranked hexamers that stood out in these two respects. To get around the protein coding problem conservation was scored only at synonymous sites. Overabundant hexamers were chosen as dicodons that appeared more frequently than expected if codons were paired randomly; here again only codons differing at synonymous sites were compared so as to avoid the influence of protein

coding. Hexamers with high scores for both criteria were collected, resulting in a set of 285 (7.7% of all hexamers considered) that represented the best combination of scores. Ten of these sequences were tested in functional assays and most of these were shown to either increase or decrease exon inclusion while none of nine control hexamers with low abundance and conservation scores significantly affected splicing. One might have thought that a selection based on abundance and conservation would favor ESEs, but both enhancer and silencer effects were observed, depending on the hexamer and on the host exon. The authors followed up on this dichotomy by placing each of two winning hexamers at 26 different positions within an 81 nt test exon. Here again, both enhancing and silencing results were obtained, this time dependent on position. Finally, when four winning hexamers resident in real exons were mutated so as to lose their high score, a mixture of positive and negative effects was observed while mutation of nonwinners had no effect. The authors thus called these motifs ESRs, for exonic splicing regulators, since their effects could be either positive or negative depending on the context. They further suggested that putative ESEs and ESSs identified by others but untested for a position effect should be similarly regarded. A list of all hexamers surveyed and their scores in terms of p-values for the significance of their deviation from mean frequencies can be found at http://www.tau.ac.il/~gilast/sup_mat7.htm.

The hexamers selected here will be biased toward sequences that harbor synonymous codons. Extreme examples are hexamers that contain a stop codon as either the first or last three positions; these 375 hexamers are removed from consideration. A second limitation is that the criteria used do not specifically target splicing motifs but apply to any function implicit in mRNA (transport, stability, etc.). However, the fact remains that greater splicing effects were seen with many of these sequences compared to controls. Some of the substitutions made to test for splicing phenotypes also caused sequence changes in overlapping endogenous PESE or RESCUE-ESE motifs (not shown), an outcome that is not surprising given the coverage figures shown in Table 1 and since any hexamer substitution changes 10 overlapping hexamers (or 13 octamers). Nevertheless, there were so many ESRs tested here at so many different positions, that it is likely that the ESR set does contain many novel motifs that affect splicing.

An interesting idea that this work gives rise to is the possibility that the same motif can act as an enhancer in one context and a silencer in another. The efficacy of some enhancers is known to be dependent on the distance from a target splice site. For example, the enhancement of splicing at the *dsx* 3' splice site has been shown to drop off when ESEs are placed more than 150 to 200 nt downstream.[48] On a chemical level, the ability of an RS domain to crosslink to a splice site also falls off at distances greater than 100 nt.[49] However, while decreasing efficiency, these far positionings do not reverse the effect of an enhancer. Individual natural ESEs have also been shown to be able to act negatively when placed within an intron[50] and there are examples in which a splicing factor acts positively at one splice site and negatively at another (e.g., hnRNP H/F and SRrp86[35]), or a single sequence element is a target for both positive and negative factors.[51-53] One could argue that there is a need for ESSs in exons in order to silence internal pseudo splice sites. But there are few of these in constitutive exons: summed pseudo 5' and 3' splice sites numbered less than 0.2 per 120 nt of exon in a set of 5000 examined (with a size limit of 250 nt) and over 80% of exons had neither such site. (Exons that contain alternative 5' or 3' splice sites obviously have more than one splice site per exon, but these should be considered real sites, not pseudo sites.)

The cautionary note sounded by Goren et al[54] presents a serious challenge, as there have been few systematic studies of the effect of position on motifs isolated by global searches. For the most part, however, experiments have not shown a context-dependent effect on activity. For example, eight PESSs that were originally found to be effective when inserted just downstream of a 3' splice site in a first test exon were equally effective when inserted just upstream of a 5' splice in a second test exon.[41] ESSs isolated by molecular selection similarly acted consistently as silencers in several different contexts.[55] Still, in these studies position was not systematically varied within a single context. Analysis of natural occurrences of splicing motifs may be more relevant and here too, the results have so far been consistent with prediction. For example, when we knocked out predicted ESEs in a beta-globin exon 2, five of the six disruptions decreased splicing and the exception did

not significantly increase splicing. If some of the predicted ESEs were really ESSs, as predicted by the ESR idea, some of the knockouts should have increased splicing. Finally, almost all of a panel of human mutations affecting splicing can be explained by the disruption of predicted ESEs or the creation of predicted ESSs,[56] in accord with the expected behavior of the motif. It is possible that these motifs function as predicted when in their natural contexts but that their normal activity can be subverted when experimentally placed in ectopic positions. An analogous result, known as transcriptional interference, has been seen with experimentally manipulated promoters.[57]

Molecular Selections

Most molecular selections have targeted ESEs. Almost all of these types of experiments are based on the Systematic Evolution of Ligands by Exponential Enrichment (SELEX), originally designed to select for nucleic acid sequences that bind to a given protein or small molecule.[58] SELEX has been applied in two ways: (1) determining the sequences that can be recognized by a given RNA binding protein (binding SELEX); and (2) isolating sequences that functionally enhance splicing (functional SELEX).

Protein Binding SELEX

In protein-binding SELEX, a complex pool of RNA molecules containing a randomized region 8 to 20 nt long is incubated with a purified RNA-binding protein or domain. The RNAs that are bound by the protein are isolated, converted to cDNA, amplified by PCR and then transcribed into RNA for a subsequent round of selection. This process is repeated several times to enrich for RNA molecules with high affinity for the RNA binding protein. In this way, the binding specificities of several SR proteins, hnRNPs and other splicing factors have been determined.[32,34,59-67] The sequences are then analyzed to determine a consensus binding site(s).

The consensus motifs that have emerged from these binding SELEX experiments[68] (refer to chapters by Lin and Fu, Martinez-Contreras et al, and Ule and Darnell) illustrate that each protein binds to a distinct set of sequences but at the same time can recognize a diverse repertoire of sequences.[68] For instance, Cavaloc et al[63] sequenced over 90 oligonucleotides bound by the SR protein SC35 and found five distinct consensuses. However, such degeneracy is not always the case. Only a single long consensus was found by Tacke and Manley in binding SELEX experiments performed with SRp40 (ref. 66). Why more than one consensus sequence appears in many of these experiments is not clear. In cases of proteins with more than one RRM, it is possible that each binds a distinct sequence. Yet, when SELEX was applied to a single RRM derivative of ASF/SF2, the one resulting consensus sequence differed from the two consensus sequences yielded by the intact, wild type protein.[59] Alternatively, a single binding site may be endowed with some flexibility to accommodate a specific set of different sequences.[69] A certain dichotomy in binding behavior might also be related to the multiple roles SR proteins play in the splicing process:[70] initial recognition, exon definition,[71] spliceosome assembly[71,72] and perhaps the catalytic steps themselves.[49,73-75] (refer to chapter by Lin and Fu for a more detailed description of the functions of SR proteins in splicing.) It is also likely that the diversity of sequences recognized is related to the fact that many RNA binding proteins must bind to protein coding exons. Thus, the range of sequence motifs that can exist in a given exon is confined by the protein sequence encoded in that exon.

The advantage of binding SELEX experiments is that they probe and define the binding specificity of a purified protein, in the absence of possible interference by other factors. As such they provide a valuable starting point in the interpretation of the roles of SR proteins and the motifs dictating their action. On the other hand, SELEX may identify RNA sequences that bind to any surface of a purified protein, not necessarily the natural RNA binding site and could therefore include sequences that are not functionally relevant in a biological context.

Functional SELEX

In this strategy motifs that are able to influence splicing activity are selected. The iterative isolation and amplification steps of SELEX are used here to select for short sequences that, when inserted into an exon, enhance pre-mRNA splicing. In these studies, a pre-mRNA pool is

first synthesized that contains a weak test exon containing a localized randomized region. This pre-mRNA pool is used in in vitro or in vivo splicing assays and the successfully spliced mRNAs are isolated and amplified by RT-PCR. As with the binding SELEX experiments, the winning RNA sequences from the first round are recycled through several additional rounds, enriching for RNA sequences that best enhance splicing.

The first experiments of this kind were carried out in vitro by Tian and Kole.[76] They found that the winning sequences were quite heterogeneous and could be divided into two classes: a majority that were purine-rich, typically consisting of short runs of 5-6 purine nucleotides and a significant minority (15% to 30%) that were not rich in purines. Almost all of the retested sequences produced efficient splicing, whereas only 1 in 10 of the unselected sequences promoted splicing. In a subsequent refinement of this procedure, shorter versions of the winning sequences were produced and these yielded a less heterogeneous group with a consensus GACGAC...CAGCAG (the core being of variable length) that was shown to bind SRp30.[77]

Two larger studies used random sequences inserted into the second exon of a 2-exon transcript spliced in vitro. Liu et al[78,79] selected sequences that responded to one of four different SR proteins by using S100 extracts for splicing. These extracts lack all SR proteins, but splicing activity can be restored upon supplementation with individual SR proteins.

Using this approach, the authors selected 20-mer sequences that promoted splicing in response to SRp40, ASF/SF2, SC35 and SRp55. Each of these selections yielded a consensus sequence that was distinct from the others. However, there was considerable heterogeneity within each group, the consensuses were often short (5 to 8 nucleotides) and contained many ambiguous positions. Some of the degeneracy might be explained by assigning a role to the sequences flanking each test motif, as not all the motifs could promote splicing when tested on their own. Nevertheless, it was possible to assemble a PSSM for each class of motifs and this information has been incorporated into an ESE searching program "ESEfinder" (http://rulai.cshl.edu/tools/ESE/) that scores sequences for their ability to match each of the SR protein-specific consensuses.[80] High-scoring motifs are found at a significantly higher frequency within exons as opposed to introns,[79,81] although the difference is modest (10%-20%). These motifs are also more strongly associated with weak compared to strong 3' splice sites.[82]

This experimental approach was refined in a later study that focused on motifs that mediate the activity of ASF/SF2.[56] Here, the random oligomer pools were restricted to either 7 or 14 nucleotides and were used to replace a 7-nucleotide natural ESE in *BRCA1* exon 18. In this construct, exon 18 was present as the central exon of a 3-exon transcript, an internal exon situation that is more commonly found in nature. Once again a degenerate though obvious sequence preference was evident from these experiments. Most of these sequences not only enhanced inclusion of BRCA1 exon 18, but also functioned in the heterologous context of exon 7 of the *SMN1* gene. A PSSM was derived using not only the relative prevalence of bases at each position, but also, in a novel approach, taking into account the degree to which each sequence enhanced splicing. The PSSM derived from the 7 nt oligomers (consensus of CCCCGCA) proved to be the better predictor of enhancement than the PSSM derived from the 14 nt oligomers. This consensus differed from that of the earlier derived ASF/SF2 consensus sequences (CACACGA) from the functional SELEX experiment employing a 2-exon pre-mRNA and both of these consensus sequences differ from the consensus sequences (AGGACAGAGC and RGAAGAAC) derived from binding SELEX experiments.[59] The authors combined the matrices from the two functional selections and showed that the resulting PSSM (with the consensus CGCACGA) was able to predict, with a statistically significant frequency, the outcome of a set of exonic mutations known to affect splicing.

Schaal and Maniatis[83] took a slightly different approach to defining consensus sequences that mediate SR protein function. As in the study by Liu et al described above, they inserted random oligomers into the second exon of a 2-exon transcript but assayed for splicing in vitro using nuclear extracts instead of S100 extracts. Sequences that enhanced splicing were identified after multiple rounds of selection using the same procedure as in Liu et al. The winning sequences were subsequently screened for their ability to respond to specific SR proteins in S100 extracts and the

sequences were grouped according to their response. Here again heterogeneity and distinctness characterized the sequences identified. Most SR proteins displayed some sequence preferences but in general these sequences do not match the Liu et al consensuses mentioned above, nor do they match well with the results of binding SELEX experiments. That different motifs emerge from the functional SELEX experiments could be due to the effect of context, including the position of the insert or substitution, or related to the another aspect of the test exon used, e.g., small size, mutational weakening of splice sites, natural weakness of an alternative splice site, etc.

Functional SELEX for splicing was performed in vivo by Coulter et al,[84] who inserted random 14-mers into a poorly spliced second exon in a 2-exon construct. After several rounds of selection based on transient transfection, the winning clones were sequenced. These fell into three categories, purine-rich, adenine- plus cytosine-rich (ACE) and neither. Within the first two categories, the identifiable motifs were quite degenerate. An ACE motif in the human *CD44* gene was subsequently shown to act as an ESE and to be bound by the nonSR protein YB-1.[85] This last result represents a cautionary tale: by testing only for responsiveness or binding to SR proteins, other, possibly more significant, mediators targeting the isolated motifs may be overlooked.

In an interesting variation on the functional SELEX scheme, Woerfel et al tested the activity of ~50 bp fragments of the *CD44* mRNA in an enhancer-dependent exon in vitro.[86] About half of the active sequences tested enhanced splicing and many of these mapped to a specific region within the *CD44* mRNA. Most of the sequences contained a short AC-rich motif whereas others contained purine-rich runs. This starting material was quite limited compared to random oligomers and the fact that a restricted subset of ESEs was overrepresented suggests that the test exon likely has a preference for particular ESEs.

ESSs

Wang et al[87] used an elegant genetic selection to isolate sequences that could cause exon skipping in vivo. When the central exon is included, it interrupts the reading frame of GFP; when skipped, functional GFP is expressed, allowing the positive cells to be isolated by FACS. After insertion of a set of random decamers, 133 unique sequences promoting GFP expression were isolated from this screen. Many of the most common hexamers caused skipping in a heterologous exon. This collection of 103 common ESS hexamers is known as FAS-hex3. By avoiding the iterative enrichment process of SELEX, any sequences that inhibit splicing were isolated, rather than only those that have the strongest ESS activity. FAS-hex3 sequences are 2-fold overrepresented in introns vs. exons and are especially overrepresented in exonic regions located between alternative 5' or 3' splice sites.[55] Like the computationally selected PESSs, they show a peak just upstream and downstream of 3' and 5' splice sites, perhaps to prevent neighboring pseudo sites from being used. Mutation of these sequences in natural exons increased the use of the proximal sites and resulted in increased exon inclusion, attesting to the silencing function of these sequences in natural alternative splicing. Alternative intron retention events were also inhibited by FAS-hex3 sequences, often in favor of skipping of the retained intron along with its flanking exons. Interestingly, FAS-hex3 sequences that could inhibit intron retention by increased skipping tended to be different from those that acted to decrease the use of an alternative splice site, suggesting that different ESSs act via different mechanisms.

ISEs and ISSs

Whereas there has been extensive investigation of the effect of intronic sequences on the alternative splicing of individual genes (e.g., 35, 88-90), there has been relatively little global searching for intronic splicing regulatory motifs. Early studies by Nussinov[91] identified G-runs as overrepresented in introns near both the 3'and 5' ends of exons. This work was extended by McCullough et al[92,93] to show that these sequences can enhance splicing at 5' splice sites by enhancing the binding of U1 snRNP. Statistical analyses and comparative genomics showed that intronic flanks of exons harbor short runs of G, C or T.[43,94] Evidence for a role of intronic flanks in splicing regulation comes from the finding that the flanks of alternatively spliced exons are more conserved than those of constitutive exons.[6,95]

Subsets of the genome have been searched for ISE motifs. Statistical analysis of sequences at the ends of short introns in several different organisms produced pentamers that could be used to enhance the accuracy of splice site prediction in such introns.[96] For humans, 8 of the 10 top pentamers were rich in G-triplets. Individual motifs have also been strongly associated with alternative splicing in the brain. Brain-specific ISE and ISS candidate motifs were identified statistically by analyzing 25 brain-specific alternatively spliced exons[97]; in contrast to introns overall, G-triplets were underrepresented downstream of these exons. One of the ISE sequences, UGCAUG, was subsequently shown to bind the Fox-1 and Fox-2 splicing factors and to be associated with regulated alternative splicing in different tissues, including brain and muscle in a number of species.[98-102] A downstream intronic G-tetramer motif was found to be associated with exon-skipping in alternatively spliced exons in the brain by Han et al and, interestingly, this element was shown to function in conjunction with an exonic UAGG to effect silencing.[103]

Our own laboratory has used machine learning to assess whether information for splice site recognition is present in sequences flanking constitutive exons.[12] A support vector machine (SVM) found that sequences residing within ~50 bases of the splice sites can help distinguish real exons from pseudo exons and identified overrepresented (ISE candidates) or underrepresented (ISS candidates) pentameric motifs that best aided the distinction. These included some novel motifs as wells as G-triplets mentioned above and, despite its degeneracy, branchpoint-like sequences, with a clear peak 24 nt upstream of the 3' splice sites. This work was followed up by a statistical test for pentamers overrepresented in human exon flanks compared to pseudo exons flanks.[104] A conservation filter was also applied here: only those pentamers that were also present within a 50 nucleotide region flanking the orthologous mouse exons were retained. The resulting ISE candidates fell into two distinct groups based on G + C content. A survey of 100,000 constitutive exons showed that their 50-nt flanks in general fell into distinct GC-rich or GC-poor categories; remarkably, the extent of this dichotomy was much greater that that exhibited by the host genes overall (i.e., due to residence in a particular isochore). Thus the factors that recognize these putative ISEs are probably different for GC-rich genes and AT-rich genes, leaving open the possibility that distinct mechanisms operate for these two gene classes. The GC-rich exons differed from the AT-rich exons in other ways: the GC-rich ISEs tended to have a complementary ISE in the opposite flank whereas the AT-rich ISEs tended to have the same ISE in the opposite flank and predicted base pairing between the flanks and the exon tended to be avoided for GC-rich exons but not for AT-rich exons.

Although our predicted ISE/S motifs were not specifically tested, we did show that intronic sequences are often important for efficient splicing.[104] Specifically, we found that splicing of an exon is often inefficient when it is not flanked by the 50 nt intronic sequence beyond its splice site sequence (i.e., −63 to −14 and +7 to +56). In addition, two of three exons lacking their flanking introns exhibited decreased or aberrant splicing when moved to ectopic positions within the same intron. These studies show that intronic sequences proximal to exons contribute to splice site recognition in an exon-specific manner. The interplay between exonic elements and presumed ISEs could also be seen in our test of natural PESEs. In transcripts (from the *HBB-2* and *THBS4-13*) tested for ESEs by mutational analysis, the wild type exons exhibited 50% to 80% inclusion when their flanking intronic sequences were removed; in this situation, the removal of any one of several PESEs reduced inclusion several-fold. However, if the flanking intronic sequences were retained, splicing was refractory to such single PESE knockouts.[28] A similar interaction was seen with a PESS knockout, the mutated exon in this case shedding its intron flank requirement (X. Zhang, unpublished result).

Functional SELEX for Splice Site Sequences

Branchpoint sequences are active participants in catalysis rather than ISEs, but they do present an analogous tension between specificity and degeneracy. Functional SELEX has also been used for the selection of effective branchpoint sequences. After seven rounds of selection at a fixed position the consensus that emerged was TACTAAC,[105] which can optimally pair with U2 snRNA. However, when only a single round of selection was carried out, a wide variety of effective

sequences were found, many with only 3 to 5 bases capable of pairing with U2 snRNA.[105,106] If the starting transcript had a weak polypyrimidine tract, or if an ESE was removed, a better match to U2 snRNA was selected, another indication that a balance of compensatory factors determines a splicing outcome.

Functional SELEX has also been used to define splice site sequences, including the polypyrimidine tract.[105-107] Here again, multiple rounds of selection converged on the consensus sequences and so no additional insight was provided into how exons with weak splice sites are recognized. Interestingly, the 5' splice site consensus was also selected in extracts lacking the U1 snRNA 5' end,[107] suggesting that this sequence may be recognized by protein components as well as by RNA-RNA hybridization.

Comparison of Computationally Predicted and Functional SELEX Selected Exonic Motifs

As shown in Table 3, RESCUE-ESEs show considerable overlap with PESEs and avoid overlap with PESSs. Similarly, fas3-hex3 silencers overlap with PESSs and avoid PESEs. Goren ESRs, which exhibit either enhancer or silencer activities, overlap less with PESEs and not significantly with PESSs (Table 3). Thus each set contains common and unique information. In contrast, ESEs defined by functional SELEX[56,76-79,83] appear quite distinct from their computationally-derived ESEs. These differences can be seen at the level of base composition and particularly in CpG content, which is remarkably high in SELEX winners obtained in four different laboratories (Table 2, columns 5 to 9). Binding SELEX motifs also usually differ from or show only weak similarities to those yielded by functional SELEX,[68] although certain of the former also display a high CpG content.[63,66] Given the methylated status of most CpGs in genomic DNA, one is tempted to speculate on connections to transcription: the slowing of transcriptional elongation at methylated CpGs[108] may provide time for the reloading of splicing factors onto the C-terminal domain (CTD) of RNA polymerase,[109] as well as increased time for the association of factors with weak splicing signals independent of any association with the CTD (refer to chapter by Kornblihtt).

One way to compare computationally derived motifs with those obtained from functional SELEX is to use ESEfinder, a Web-based program[80] that scores sequences using a PSSM derived from motifs selected for responsiveness to four different SR proteins.[78,79] About 40% of the PESE octamers contain sequences that fall above the threshold for at least one of the four SR proteins, a proportion that is not unreasonable given that PESEs presumably include binding sites for all SR proteins whereas ESEfinder covers only four. However, 28% of a random set of octamers also achieves this rather undemanding benchmark. ASF/SF2 motifs were the most common among PESEs at 15%. Looking more directly for overlaps, Wang et al also concluded that there was no significant overlap between ESEfinder motifs and RESCUE-ESES or PESEs, with the exception of ASF/SF2 motifs with PESEs.[81]

An Embarrassment of Riches?

The multi-pronged global attacks on defining splicing regulatory motifs summarized above have promoted optimism that the "splicing code" may soon be solved.[110,111] At this moment however, the number of effective motifs that have been experimentally defined or predicted may have reached the point of diminishing returns. The use of four computationally selected sets of ESEs and one genetically selected set of ESSs covers about three-quarters of the nucleotides in a typical exon (Table 1). Moreover, there is no reason to think that all ESEs and ESSs have been identified. While it is probably true that some of the predicted but untested ESE candidates will turn out to be inactive, it is even more certain that many more as yet unidentified motifs will have enhancer activity. These motifs are to be found in the sequence space below the conservative thresholds that have been used in selecting predicted ESEs. Hard evidence for such new sequences can be found in the tests of the predicted sequences. The specificity of both RESCUE-ESEs and PESEs was assessed by mutating the predicted enhancer to a sequence that did not score highly. In almost every case, the predicted decrease in splicing was indeed observed. But in half these tests the decrease

was modest, with more than 50% of the enhancer activity remaining and a several fold splicing enhancement was still in evidence.[39,41] Thus the percentage of sequence space assigned to ESEs in these two studies must undoubtedly be increased by at least 50% and probably more. Stadler et al[112] have now formalized a search for additional ESEs and ESSs by developing an algorithm, called Neighborhood Inference, that searches for new ESEs on the basis of sequence similarity to known ESEs and dissimilarity from ESSs and vice versa. A sample of high-scoring hexamers identified by this algorithm proved to act as ESEs or ESSs as predicted. The authors conclude that the list of splicing regulatory motifs is much greater than previously thought.

An unreported reservoir of ESEs is also evident in most SELEX experiments. For example of the 28 ASF/SF2-responsive heptamer motifs isolated in a functional selection, no two were identical.[56] A repeat of this experiment would therefore be expected to turn up many additional unique heptamers.

The extensiveness of the ESE list results in a ubiquity of these elements. Although present at lower densities in introns than in exons, computationally defined ESEs nevertheless heavily populate pseudo exons (Table 1, column 9), and ESEs predicted by functional SELEX occur in introns at about 80% to 90% of the level found in exons.[81] While from a strictly bioinformatics point of view one can take solace in the high statistical significance of these differences, the fact remains that the overall differences are modest. Thus we are once again challenged with figuring out how, in the face of this richness of signals, the cell distinguishes real splice sites from pseudo sites and real exons from pseudo exons. This present situation has engendered the less optimistic view that we have indeed reached a point "right on the edge of chaos".[113]

One consequence of the prevalence of splicing motifs is technical: one must be careful in drawing conclusions from mutational perturbations of pre-mRNA sequences. A single base change can impinge on many possible resident motifs; and insertions, deletions and new junctions increase the number of collaterally emergent sequences considerably. It may be necessary to turn to very simple exons or to aim at easily characterized or isolated regions to minimize ambiguity in the results. It will also be prudent to take this accumulated list of ESEs into account when interpreting the result of any selection experiment (computational or SELEX based). An examination of the results of three types of ESEs (sequences underlying ESEfinder, binding SR proteins or discovered in individual genes) showed three-quarters to be populated by at least one RESCUE-ESE or PESE (data not shown). The second consequence is conceptual. How can we explain how so many sequences can act together to produce the binary decision that is made for the great majority of exons? How can the perceived modest yet real differences in ESE and ESS densities between exons and pseudo exons be leveraged to produce that binary decision? To say that it involves a "balance of combinatorial factors" is not much better than saying that it depends on "context", in that it describes the situation we see without really explaining it.

One easy way out is to invoke secondary structure acting to "present" some of these motifs but not others (e.g., as loops; see ref. 114) or to mask some motifs and not others (e.g., see ref. 115). One technical problem in evaluating the role of secondary structure on the splicing of natural exons is that we do not really know how transcripts fold in vivo. In particular, many conformations that would be considered too unstable to contribute to predicted equilibrium structures could be kinetically trapped during transcription and last long enough to influence splicing outcomes. It is clear that secondary structures do play role in many splicing decisions (reviewed in ref. 116) (refer to chapter by Park and Graveley) but whether this influence is pervasive is not yet clear. If secondary structure is not invoked then we need models that seek to explain why so many ESEs are present.

It is reasonable to think that the majority of these splicing motifs are playing a role in exon definition.[17,18] The main feature of this model is that both ends of an exon must be recognized before splicing at either end can occur. The implied communication between the two ends of the exon could be realized via a bridge of proteins, each a necessary link in a flow of information mediated by a series of allosteric transitions. This is a rather complicated model but does explain the necessity of multiple ESEs to construct this bridge. Moreover, a fitting combination of proteins may be needed to ensure proper signal propagation, although this need not be a unique assemblage

(Fig. 2A). Consistent with this model is the finding that different pre-mRNA molecules associate with different sets of nuclear proteins.[117] Arguments against such a model are that random sequences can be inserted into exons often without dire consequences. Thus insertion of bacterial sequences from 100[118] to 1000[119] nucleotides long into the central exon of a 3-exon construct did not impair splicing despite the probability that they do not have high densities of ESEs. Although we found that one-third of a randomly chosen set of human genomic sequences of about 100 nt decreased splicing when inserted into an exon, the other two-thirds had little effect.[118] An argument against a requirement for specific proteins lies in the fact that insertion of any of a wide variety of predicted ESE sequences can enhance the splicing of a crippled exon.[39,41,112] Indeed it should not be necessary to invoke a continuous bridge for the two ends of an exon to communicate in exon definition: an ESE at each end could suffice to recruit splicing factors that could then interact by simply forming a loop (Fig. 2B).

An alternative model that necessitates large numbers of ESEs puts the emphasis on ESSs. If ESSs are fairly common and if the binding of an inhibitory factor to a single ESS is sufficient to inhibit splicing[120,121] then it may be necessary to have enough ESEs per exon to prevent even a single silencing protein from binding (Fig. 2C-E). If there are no ESSs this is not a problem as may be the case of the bacterial inserts mentioned above. But the protein coding requirements of the exon may not allow the complete exclusion of ESSs, and these sequence elements may have additional downstream roles even in constitutively spliced exons, such as in translation.[122,123] Furthermore, splicing inhibitors such as hnRNP A1 can undergo multimerization leading to cooperative binding[124] with the consequent displacement of many more distant ESE binding factors (Fig. 2F-H). Extensive coverage of the exon with ESEs and their binding factors may prevent proteins like hnRNP A1 from gaining a foothold. The ESE in the immunoglobulin M2 exon acts in this way by disrupting the association of PTB with an ESS.[120] Of course neither of these anti-ESS models excludes a positive role for other ESEs in promoting splicing at the same time.

It is possible that ESSs provide more of the information for discriminating real exons from pseudo exons than ESEs (for a review see ref. 46). As can be seen in Table 1, RESCUE-ESEs plus PESEs are about 60% more abundant in real exons than pseudo exons, but PESSs and FAS-hex3 silencers are 2-fold more abundant in pseudo exons than real exons. There are now several examples

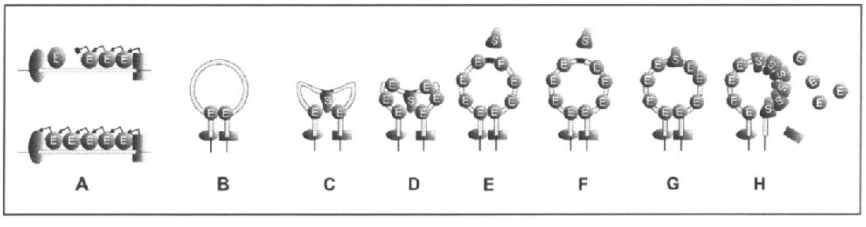

Figure 2. Models incorporating a role for extensive exon coverage in exon definition. Thick lines, exons; thin lines, introns; E, enhancer binding proteins; S, silencer binding proteins, rectangle and oval, spliceosomal or prespliceosomal complexes. A) A bridge of proteins is required in order to sense that both ends of the exon have been recognized as splice sites. Complete exon coverage is required for efficient transmission of this information via allosteric transitions. B) The opposite case, in which enhancers help recruit splicing factors to the splice sites and then interact with each other directly to convey the information that both sites have been recognized. C) A single silencer disrupting the interaction in B, so that splicing does not occur and the exon is skipped. D) Even if many splicing activators are bound to many enhancers, the binding of a single inhibitory protein to an available silencer can still inhibit splicing. E) Coverage of the exon is extensive enough for steric hindrance to prevent the binding of even a single silencer protein. F-H) Despite many enhancers, leaving a single silencer unobstructed allows binding by an inhibitory protein (e.g., hnRNP A1). Once bound, the inhibitory protein multimerizes, leading to the displacement of splicing factors.

of intronic mutations that create new exons apparently by the inactivation of silencers elements (e.g., see ref. 125). Such events have led to the exonization of Alu sequences[126] and could underlie the evolution of new genes in general.[127] There is little doubt that ESSs can play a role in the silencing of pseudo exons, but whether this mechanism represents a global role has not been established.

The Future

Structure determines function in biology and the structure of RNA within an RNP is ultimately dictated by its sequence. What proteins bind to what RNAs,[128] how the position and order of motifs influence factor binding and splicing and how positive and negative intron sequences factor into the equation are all questions that are approachable experimentally. Answers to these questions will help move us from lists to mechanisms. From bioinformatics we can expect yet more global information defining ISEs and ISSs as well as refining ESEs and ESSs. Relationships between motifs (e.g., see ref. 103) and between motifs and expressed splicing factors[129-131] are beginning to be revealed using combined computational and molecular approaches. The problem of how the cell integrates information from a large set of overlapping signals is not unique to splicing. In transcriptional regulation the choice of true promoters from among pseudo signals has parallels in the identification of true splice sites, and even the cellular decisions that are made during embryonic development are somewhat analogous in that slight differences in a morphogenetic gradient are amplified to produce a binary response. For splicing we now know many of the players, a necessary step in order to decipher the rules of the game.

Acknowledgements

I thank Mauricio Arias, Shengdong Ke and Xiang Zhang for useful discussions and Xiang Zhang for the databases used in calculations performed here. Adrian Krainer generously provided the 20-mer sequences underlying ESEfinder. Splicing research in our laboratory is supported by a grant from the NIH.

References

1. Senapathy P, Shapiro MB, Harris NL. Splice junctions, branch point sites and exons: sequence statistics, identification and applications to genome project. Methods Enzymol 1990; 183:252-278.
2. Zhang XH, Leslie CS, Chasin LA. Computational searches for splicing signals. Methods 2005; 37(4):292-305.
3. Sickmier EA, Frato KE, Shen H et al. Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. Mol Cell. Vol 23; 2006:49-59.
4. Roca X, Sachidanandam R, Krainer AR. Intrinsic differences between authentic and cryptic 5' splice sites. Nucleic Acids Res 2003; 31(21):6321-6333.
5. Thanaraj TA, Stamm S. Prediction and statistical analysis of alternatively spliced exons. Prog Mol Subcell Biol 2003; 31:1-31.
6. Zheng CL, Fu XD, Gribskov M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. RNA 2005; 11(12):1777-1787.
7. Itoh H, Washio T, Tomita M. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. RNA 2004; 10(7):1005-1018.
8. Lear AL, Eperon LP, Wheatley IM et al. Hierarchy for 5' splice site preference determined in vivo. J Mol Biol. Vol 211; 1990:103-115.
9. Carothers AM, Urlaub G, Grunberger D et al. Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. Mol Cell Biol 1993; 13(8):5085-5098.
10. Schneider TD. Information content of individual genetic sequences. J Theor Biol 1997; 189(4):427-441.
11. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997; 268(1):78-94.
12. Zhang XH, Heller KA, Hefter I et al. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res 2003; 13(12):2637-2650.
13. Green MR. Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Annu Rev Cell Biol 1991; 7:559-599.
14. Smith CW, Nadal-Ginard B. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. Cell 1989; 56(5):749-758.

15. Sun H, Chasin LA. Multiple splicing defects in an intronic false exon. Mol Cell Biol 2000; 20(17):6414-6425.
16. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 1992; 90(1-2):41-54.
17. Berget SM. Exon recognition in vertebrate splicing. J Biol Chem 1995; 270(6):2411-2414.
18. Robberson BL, Cote GJ, Berget SM. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 1990; 10(1):84-94.
19. Lewis JD, Izaurralde E, Jarmolowski A et al. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. Genes Dev 1996; 10(13):1683-1698.
20. Zeng C, Berget SM. Participation of the C-terminal domain of RNA polymerase II in exon definition during pre-mRNA splicing. Mol Cell Biol 2000; 20(21):8290-8301.
21. Cooke C, Hans H, Alwine JC. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. Mol Cell Biol 1999; 19(7):4971-4979.
22. Query CC, McCaw PS, Sharp PA. A minimal spliceosomal complex A recognizes the branch site and polypyrimidine tract. Mol Cell Biol 1997; 17(5):2944-2953.
23. Zhang MQ. Statistical features of human exons and their flanking regions. Hum Mol Genet 1998; 7(5):919-932.
24. Cooper TA, Ordahl CP. Nucleotide substitutions within the cardiac troponin T alternative exon disrupt pre-mRNA alternative splicing. Nucleic Acids Res 1989; 17(19):7905-7921.
25. Streuli M, Saito H. Regulation of tissue-specific alternative splicing: exon-specific cis-elements govern the splicing of leukocyte common antigen pre-mRNA. EMBO J 1989; 8(3):787-796.
26. Watakabe A, Tanaka K, Shimura Y. The role of exon sequences in splice site selection. Genes Dev 1993; 7(3):407-418.
27. Schaal TD, Maniatis T. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol Cell Biol 1999; 19(1):261-273.
28. Zhang XH, Kangsamaksin T, Chao MS et al. Exon inclusion is dependent on predictable exonic splicing enhancers. Mol Cell Biol 2005; 25(16):7323-7332.
29. Graveley BR. Sorting out the complexity of SR protein functions. RNA 2000; 6(9):1197-1211.
30. Kohtz JD, Jamison SF, Will CL, et al. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. Nature 1994; 368(6467):119-124.
31. Shen H, Green MR. RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. Genes Dev. Vol 20; 2006:1755-1765.
32. Tacke R, Manley JL. Determinants of SR protein specificity. Curr Opin Cell Biol 1999; 11(3):358-362.
33. Pozzoli U, Sironi M. Silencers regulate both constitutive and alternative splicing events in mammals. Cell Mol Life Sci 2005; 62(14):1579-1604.
34. Abdul-Manan N, Williams KR. hnRNP A1 binds promiscuously to oligoribonucleotides: utilization of random and homo-oligonucleotides to discriminate sequence from base-specific binding. Nucleic Acids Res 1996; 24(20):4063-4070.
35. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003; 72:291-336.
36. Zheng ZM. Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. J Biomed Sci 2004; 11(3):278-294.
37. Fu XD. Towards a splicing code. Cell 2004; 119:736-738.
38. Fedorov A, Saxonov S, Fedorova L et al. Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. Nucleic Acids Res 2001; 29(7):1464-1469.
39. Fairbrother WG, Yeh RF, Sharp PA et al. Predictive identification of exonic splicing enhancers in human genes. Science 2002; 297(5583):1007-1013.
40. Fairbrother WG, Holste D, Burge CB et al. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2004; 2(9):E268.
41. Zhang XH, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 2004; 18(11):1241-1250.
42. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. Nat Genet 2001; 29(4):412-417.
43. Majewski J, Ott J. Distribution and characterization of regulatory elements in the human genome. Genome Res 2002; 12(12):1827-1836.
44. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 2006; 103(5):1412-1417.
45. Grellscheid SN, Smith CW. An apparent pseudo-exon acts both as an alternative exon that leads to nonsense-mediated decay and as a zero-length exon. Mol Cell Biol 2006; 26(6):2237-2246.

46. Sironi M, Menozzi G, Riva L et al. Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res 2004; 32(5):1783-1791.
47. Goren A, Ram O, Amit M et al. Comparative analysis identifies exonic splicing regulatory sequences—The complex definition of enhancers and silencers. Mol Cell. Vol 22; 2006:769-781.
48. Graveley BR, Hertel KJ, Maniatis T. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J 1998; 17(22):6747-6756.
49. Shen H, Green MR. RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. Genes Dev 2006; 20(13):1755-1765.
50. Kanopka A, Muhlemann O, Akusjarvi G. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. Nature 1996; 381(6582):535-538.
51. Wu JY, Kar A, Kuo D et al. SRp54 (SFRS11), a Regulator for tau Exon 10 Alternative Splicing Identified by an Expression Cloning Strategy. Mol Cell Biol 2006; 26(18):6739-6747.
52. Cartegni L, Hastings ML, Calarco JA et al. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. Am J Hum Genet 2006; 78(1):63-77.
53. Kashima T, Manley JL. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. Nat Genet 2003; 34(4):460-463.
54. Goren A, Ram O, Amit M et al. Comparative analysis identifies exonic splicing regulatory sequences—The complex definition of enhancers and silencers. Mol Cell 2006; 22(6):769-781.
55. Wang Z, Xiao X, Van Nostrand E et al. General and specific functions of exonic splicing silencers in splicing control. Mol Cell 2006; 23(1):61-70.
56. Smith PJ, Zhang C, Wang J et al. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 2006; 15(16):2490-2508.
57. Eszterhas SK, Bouhassira EE, Martin DI et al. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. Mol Cell Biol 2002; 22(2):469-479.
58. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 1990; 249(4968):505-510.
59. Tacke R, Manley JL. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. EMBO J 1995; 14(14):3540-3551.
60. Kim S, Shi H, Lee DK et al. Specific SR protein-dependent splicing substrates identified through genomic SELEX. Nucleic Acids Res 2003; 31(7):1955-1961.
61. Hui J, Hung LH, Heiner M et al. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J 2005; 24(11):1988-1998.
62. Amarasinghe AK, MacDiarmid R, Adams MD et al. An in vitro-selected RNA-binding site for the KH domain protein PSI acts as a splicing inhibitor element. RNA 2001; 7(9):1239-1253.
63. Cavaloc Y, Bourgeois CF, Kister L et al. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA 1999; 5(3):468-483.
64. Wang J, Dong Z, Bell LR. Sex-lethal interactions with protein and RNA. Roles of glycine-rich and RNA binding domains. J Biol Chem 1997; 272(35):22227-22235.
65. Buckanovich RJ, Darnell RB. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Mol Cell Biol 1997; 17(6):3194-3201.
66. Tacke R, Chen Y, Manley JL. Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. Proc Natl Acad Sci USA 1997; 94(4):1148-1153.
67. Faustino NA, Cooper TA. Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment. Mol Cell Biol 2005; 25(3):879-887.
68. Bourgeois CF, Lejeune F, Stevenin J. Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. Prog Nucleic Acid Res Mol Biol 2004; 78:37-88.
69. Sickmier EA, Frato KE, Shen H et al. Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. Mol Cell 2006; 23(1):49-59.
70. Sanford JR, Ellis J, Caceres JF. Multiple roles of arginine/serine-rich splicing factors in RNA processing. Biochem Soc Trans 2005; 33(Pt 3):443-446.
71. Boukis LA, Liu N, Furuyama S et al. Set/Arg-rich protein-mediated communication between U1 and U2 small nuclear ribonucleoprotein particles. J Biol Chem 2004; 279(28):29647-29653.
72. MacMillan AM, McCaw PS, Crispino JD et al. SC35-mediated reconstitution of splicing in U2AF-depleted nuclear extract. Proc Natl Acad Sci USA 1997; 94(1):133-136.
73. Shen H, Green MR. A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. Mol Cell 2004; 16(3):363-373.
74. Shen H, Kan JL, Green MR. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Mol Cell 2004; 13(3):367-376.

75. Chew SL, Liu HX, Mayeda A et al. Evidence for the function of an exonic splicing enhancer after the first catalytic step of pre-mRNA splicing. Proc Natl Acad Sci USA 1999; 96(19):10655-10660.
76. Tian H, Kole R. Selection of novel exon recognition elements from a pool of random sequences. Mol Cell Biol 1995; 15(11):6291-6298.
77. Tian H, Kole R. Strong RNA splicing enhancers identified by a modified method of cycled selection interact with SR protein. J Biol Chem 2001; 276(36):33833-33839.
78. Liu HX, Chew SL, Cartegni L et al. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol Cell Biol 2000; 20(3):1063-1071.
79. Liu HX, Zhang M, Krainer AR. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 1998; 12(13):1998-2012.
80. Cartegni L, Wang J, Zhu Z et al. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 2003; 31(13):3568-3571.
81. Wang J, Smith PJ, Krainer AR et al. Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. Nucleic Acids Res 2005; 33(16):5053-5062.
82. Wu Y, Zhang Y, Zhang J. Distribution of exonic splicing enhancer elements in human genes. Genomics 2005; 86(3):329-336.
83. Schaal TD, Maniatis T. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol Cell Biol 1999; 19(3):1705-1719.
84. Coulter LR, Landree MA, Cooper TA. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol Cell Biol 1997; 17(4):2143-2150.
85. Stickeler E, Fraser SD, Honig A et al. The RNA binding protein YB-1 binds A/C-rich exon enhancers and stimulates splicing of the CD44 alternative exon v4 EMBO J 2001; 20(14):3821-3830.
86. Woerfel G, Bindereif A. In vitro selection of exonic splicing enhancer sequences: identification of novel CD44 enhancers. Nucleic Acids Res 2001; 29(15):3204-3211.
87. Wang Z, Rolish ME, Yeo G et al. Systematic identification and analysis of exonic splicing silencers. Cell 2004; 119(6):831-845.
88. Wagner EJ, Garcia-Blanco MA. Polypyrimidine tract binding protein antagonizes exon definition. Mol Cell Biol 2001; 21(10):3281-3288.
89. Dredge BK, Darnell RB. Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. Mol Cell Biol 2003; 23(13):4687-4700.
90. Singh NN, Androphy EJ, Singh RN. In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. RNA 2004; 10(8):1291-1305.
91. Nussinov R. Conserved signals around the 5' splice sites in eukaryotic nuclear precursor mRNAs: G-runs are frequent in the introns and C in the exons near both 5' and 3' splice sites. J Biomol Struct Dyn 1989; 6(5):985-1000.
92. McCullough AJ, Berget SM. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol 1997; 17(8):4562-4571.
93. McCullough AJ, Berget SM. An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. Mol Cell Biol 2000; 20(24):9225-9235.
94. Louie E, Ott J, Majewski J. Nucleotide frequency variation across human genes. Genome Res 2003; 13(12):2594-2601.
95. Sorek R, Ast G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 2003; 13(7):1631-1637.
96. Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci USA 2003; 98(20):11193-11198.
97. Brudno M, Gelfand MS, Spengler S et al. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res 2001; 29(11):2338-2348.
98. Minovitsky S, Gee SL, Schokrpur S et al. The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucleic Acids Res 2005; 33(2):714-724.
99. Auweter SD, Fasan R, Reymond L et al. Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. EMBO J 2006; 25(1):163-173.
100. Zhou HL, Baraniak AP, Lou H. A role for Fox-1/Fox-2 in mediating the neuronal pathway of calcitonin/CGRP alternative RNA processing. Mol Cell Biol 2006.
101. Nakahata S, Kawamoto S. Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. Nucleic Acids Res 2005; 33(7):2078-2089.
102. Jin Y, Suzuki H, Maegawa S et al. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. EMBO J 2003; 22(4):905-912.
103. Han K, Yeo G, An P et al. A combinatorial code for splicing silencing: UAGG and GGGG motifs. PLoS Biol 2005; 3(5):e158.

104. Zhang XH, Leslie CS, Chasin LA. Dichotomous splicing signals in exon flanks. Genome Res 2005; 15(6):768-779.
105. Lund M, Tange TO, Dyhr-Mikkelsen H et al. Characterization of human RNA splice signals by iterative functional selection of splice sites. RNA 2000; 6(4):528-544.
106. Buvoli M, Mayer SA, Patton JG. Functional crosstalk between exon enhancers, polypyrimidine tracts and branchpoint sequences. EMBO J 1997; 16(23):7174-7183.
107. Lund M, Kjems J. Defining a 5' splice site by functional selection in the presence and absence of U1 snRNA 5' end. RNA 2002; 8(2):166-179.
108. Lorincz MC, Dickerson DR, Schmitt M et al. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol 2004; 11(11):1068-1075.
109. Listerman I, Sapra AK, Neugebauer KM. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat Struct Mol Biol 2006; 13(9):815-822.
110. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 2005; 6(5):386-398.
111. Fu XD. Towards a splicing code. Cell 2004; 119(6):736-738.
112. Stadler MB, Shomron N, Yeo GW et al. Inference of splicing regulatory activities by sequence neighborhood analysis. PLoS Genet 2006; 2(11):e191.
113. Buratti E, Baralle M, Baralle FE. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. Nucleic Acids Res 2006; 34(12):3494-3510.
114. Varani G, Nagai K. RNA recognition by RNP proteins during RNA processing. Annu Rev Biophys Biomol Struct 1998; 27:407-445.
115. Buratti E, Muro AF, Giombi M et al. RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. Mol Cell Biol 2004; 24(3):1387-1400.
116. Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 2004; 24(24):10505-10514.
117. Bennett M, Pinol-Roma S, Staknis D et al. Differential binding of heterogeneous nuclear ribonucleoproteins to mRNA precursors prior to spliceosome assembly in vitro. Mol Cell Biol 1992; 12(7):3165-3175.
118. Fairbrother WG, Chasin LA. Human genomic sequences that inhibit splicing. Mol Cell Biol 2000; 20(18):6816-6825.
119. Chen IT, Chasin LA. Large exon size does not limit splicing in vivo. Mol Cell Biol 1994; 14(3):2140-2146.
120. Shen H, Kan JL, Ghigna C et al. A single polypyrimidine tract binding protein (PTB) binding site mediates splicing inhibition at mouse IgM exons M1 and M2. RNA 2004; 10(5):787-794.
121. Zahler AM, Damgaard CK, Kjems J et al. SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. J Biol Chem 2004; 279(11):10077-10084.
122. Bonnal S, Pileur F, Orsini C et al. Heterogeneous nuclear ribonucleoprotein A1 is a novel internal ribosome entry site trans-acting factor that modulates alternative initiation of translation of the fibroblast growth factor 2 mRNA. J Biol Chem 2005; 280(6):4144-4153.
123. Valcarcel J, Gebauer F. Posttranscriptional regulation: the dawn of PTB. Curr Biol 1997; 7(11):R705-708.
124. Zhu J, Mayeda A, Krainer AR. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol Cell 2001; 8(6):1351-1361.
125. Pagani F, Buratti E, Stuani C et al. A new type of mutation causes a splicing defect in ATM. Nat Genet 2002; 30(4):426-429.
126. Sorek R, Lev-Maor G, Reznik M et al. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. Mol Cell 2004; 14(2):221-231.
127. Zhang XH, Chasin LA. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proc Natl Acad Sci USA 2006; 103(36):13427-13432.
128. Jurica MS, Licklider LJ, Gygi SR et al. Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. RNA 2002; 8(4):426-439.
129. Blanchette M, Green RE, Brenner SE et al. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. Genes Dev 2005; 19(11):1306-1314.
130. Ule J, Jensen KB, Ruggiu M et al. CLIP identifies Nova-regulated RNA networks in the brain. Science 2003; 302(5648):1212-1215.
131. Ule J, Stefani G, Mele A et al. An RNA map predicting Nova-dependent splicing regulation. Nature 2006; 444(7119):580-586.