Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases?

Linda Hartmann1,4, Stephan Theiss2, Dieter Niederacher3,4, Heiner Schaal1,4

1Heinrich-Heine-University Duesseldorf, Institute for Virology, D-40225 Duesseldorf, Germany, 2Result GmbH, Friedenstrasse 39, D-40219 Duesseldorf, Germany, 3Heinrich-Heine-University Duesseldorf, Dept. of Obstetrics and Gynecology, D-40225 Duesseldorf, Germany, 4Heinrich-Heine-University, Duesseldorf, Center for Biological and Medical Research (BMFZ), Germany

TABLE of CONTENTS

1. Abstract
2. Introduction
3. Identification of pathogenic splicing mutations and the dilemma of diagnosis
   3.1. Mutation detection techniques
   3.2. Interpretation of sequence variants
4. Splice sites and cis-active regulatory elements
   4.1. 5’ Splice site recognition
   4.2. 3’ Splice site recognition
   4.3. Cis-active regulatory elements
5. Splice site strength and identification of regulatory motifs
   5.1. Splice site strength algorithms
   5.2. 5’ Splice site strength algorithms
   5.3. 3’ Splice site strength algorithms
   5.4. Computational identification of exonic regulatory elements
   5.5. Computational identification of intronic regulatory elements
6. Conclusions
7. Acknowledgements
8. References

1. ABSTRACT

Pathogenic splicing alterations caused by point mutations in both splice sites and auxiliary cis-regulatory elements are increasingly recognized as an important mechanism through which gene mutations cause human disease. Unfortunately, in routine genetic diagnostic settings, splicing mutations may escape identification, due to the lack of RNA samples. Since most patients are genotyped only, any computational prediction of mutation effects on splicing can be beneficial for the human geneticist. Here, we review common techniques to identify human point mutations and delineate the molecular basis for splice site recognition. Moreover, this article provides basic insights into web-tools predicting splice sites and cis-regulatory elements and discusses their benefits for judgment of clinically identified sequence variants of disease-specific genes.

2. INTRODUCTION

Genetic factors play a prominent role in common diseases and cancer syndromes including breast, colorectal, skin, prostate and ovarian cancer. Increased cancer risk due to mutations, e.g. in known tumor suppressor genes, explains a significant portion of hereditary cancers in families with these syndromes. Genetic testing advanced into clinical practice through identification of disease-specific genes and supports a variety of clinical decisions: risk assessment for future disease (predictive genetic testing), confirmation of diagnosis, and more recently, therapeutic selection and prognosis.

Clinically identified sequence variants of disease-specific genes are characterized as either known deleterious (often protein-truncating) mutations, recognized polymorphisms (assumed to be) neutral with respect to
null
Diagnostics of pathogenic splicing mutations

immobilized DNA hybridization arrays still have significant false positive signals, as well as high costs per assay.

Denaturing high performance liquid chromatography (DHPLC) has generated increased interest in clinical genetics in recent years, because of its potential for automation and its ease of application. The technique, which is based on heteroduplex detection, allows for automated identification of single-nucleotide substitutions and small deletions or insertions. Heteroduplex profiles are easily distinguished from homoduplex peaks (14). Furthermore, DHPLC has been shown to clearly resolve mutations in various genes with detection rates ranging from 92.5 to 100% (15, 16). Direct DNA sequencing of PCR fragments using dye terminators has often been reported as the gold standard with a sensitivity of almost 100%, with automation of all steps and high throughput capacity, which can still be increased using capillary sequencers.

3.2. Interpretation of sequence variants

For the interpretation of an observed sequence variant, the causal role of the alteration in the loss of protein function and/or in the pathogenesis of the disease has to be established. Particularly in genes with a large heterogeneity of different rare sequence alterations dispersed throughout the gene sequences, published data may not be available. In the absence of experimental evidence, the pathological significance of an observed sequence variant has to rely on plausibility considerations. Mutations formally interfering with proper protein synthesis like nonsense and frame-shift mutations, or missense mutations with experimental proof of their protein function impairment (mutations in functionally relevant protein motifs), have most likely pathological consequences for the protein function. Mutations which are considered to lead to aberrant splicing of the mRNA generally have to be checked by mRNA/cDNA biochemical assays providing more definitive results regarding a potential mutation’s impact on the length or the stability of the mRNA transcript.

These experiments routinely involve RT-PCR based assays to compare transcripts in patients and controls. An informative result could be the detection of an aberrant cDNA fragment (e.g. indicating exon skipping) with increased prevalence in the patient (17). Due to the possible presence of nonsense mediated RNA decay (NMD) it is important to clearly assess both allelic contributions to cDNA fragments amplified from transcripts (18). Heterozygous polymorphisms in the coding region may be useful markers to show that the normally spliced RNA is produced solely by one allele (most likely the wild type allele), if heterozygosity of this polymorphism is lost in the appropriate RT-PCR fragment indicating an NMD based degradation of the mutant allele transcript. Another informative result may be obtained detecting an aberrant product transcribed from the mutant allele. There are some published reports where this approach successfully corrected the misinterpretation of BRCA splicing mutations (19, 20, 21).

If RNA of a carrier of a putative splice site mutation is not available, other biological assays like the use of splicing models analyzing transcripts of artificial minigene reporter constructs might be performed. These extensive experimental approaches are challenging and require special experiences with the models applied in research settings. In particular, it has been shown that the splicing outcome of a minigene can be influenced by the flanking sequences including splice sites of the exon-trapping vector selected for the splice site mutation analysis [(22), L.H., H.S. unpublished observation].

As discussed above, most functional assays to assess the pathogenic nature of a putative splice site mutation can not be part of a routine diagnostic service. Therefore, clinical laboratories offering genetic testing have to rely on available published data, often deposited in locus-specific databases. For instance, BRCA1 and BRCA2 are probably the most extensively analyzed cancer predisposition genes studied in clinical genetic diagnostics and research settings. More than 70,000 women worldwide have been tested for BRCA1 and BRCA2 mutations to assess their risk for hereditary breast and ovarian cancer. Over 50,000 of these patients, representing family members from unrelated families, have received comprehensive whole-gene mutation analysis, because mutations in these genes are distributed throughout the BRCA gene sequences (23). In order to support the detection and interpretation of BRCA mutations and to make the results available to the diagnostic community, a central repository of mutations, polymorphisms and VUS, the Breast Cancer Information Core (BIC), was created and is being maintained by an international collaborative effort hosted by NHGRI (http://research.nhgri.nih.gov/bic). About 4% of all BRCA1 and BRCA2 alterations submitted to the BIC database are reported as splice site alterations. For most of them, however, there is no further knowledge about their effect at the cDNA level. Although considered to be a useful resource the data submitted to the BIC is not validated and care should be taken when referring to it.

4. SPlice SITES AND CIS-ACTIVE REGULATORY ELEMENTS

The high fidelity of splicing is critically dependent on the recognition of the signals that mark exon–intron boundaries (Figure 1). The two strongest contributing signals are the donor or 5’ splice site (5’ ss) and the acceptor or 3’ splice site (3’ ss), which are frequent targets of mutations in genetic diseases and cancer.

A statistical description of annotated human 5’ ss or 3’ ss can be obtained by aligning a large number of those, yielding a splice site motif specific for any given (exon) data set. In such a motif, sequence conservation in fixed positions is indicated by one or two predominant nucleotide(s), while outside the conserved region the nucleotides are statistically distributed (with “background probability” of approx. 25% for G, T, A and C). Correspondingly, the splice site’s consensus sequence is determined by picking the most frequent nucleotide in each conserved position. Consensus sequences can be identified
Figure 1. Schematic close-up of exon–intron boundaries of a eukaryotic gene. A spliceable gene (top) consists of exons (orange) and at least one intron (green), which is removed from the pre-mRNA by the splicing machinery. The exon-intron boundary is called the splice donor site or 5' splice site (5' ss), while the intron-exon boundary is termed splice acceptor site or 3' splice site (3' ss). The 5' ss of the pre-mRNA is recognized by the free 5' end of the U1 snRNA via base-pairing. Binding of U1 snRNP (U1) to the 5' ss and can be influenced among others by the RS-domain of SR-proteins (SR) or members of the hnRNP family (hnRNP). The splicing regulatory factors can bind to an exonic or intronic position of the transcript (middle). Depending on both the position of their target sequence and their splicing regulatory function, the respective cis-acting sequences are called exonic (E) or intronic (I) splicing (S) enhancer (E) or silencer (S), e.g. ESE for exonic splicing enhancer. The large subunit of the U2 snRNP auxiliary factor (U2AF65) recognizes the polypyrimidine tract and recruits the U2 snRNP to the branch site. The smaller subunit U2AF35 (35) recognizes the most 3' intronic dinucleotide AG. Sequence motifs of typical cis-acting splicing regulatory sequences are depicted at the bottom as sequence logos. For an explanation of such logos see legend to figure 2.

For both 5' ss and 3' ss. On a molecular level, however, the recognition of 5' ss and 3' ss strongly differs.

4.1. 5' Splice site recognition

The 5' ss in mRNA precursors is recognized early during splicing catalysis in the nucleus by the free 5' end of the U1 snRNA by complementary base paring (24). This RNA duplex formation is necessary for the splicing and binding of U1 snRNP, and at least in some instances, also protects pre-mRNA against nuclear degradation (25), as evident from human 5' ss mutations leading to RNA degradation rather than to aberrant splicing (26, 27).

For human 5' ss, the consensus sequence MAG/GURAGU (where R = purine, M = C or A, and / denotes the exon-intron border) includes positions −3 to +6 (i.e., the last 3 nucleotides [nt] of the upstream exon and the first 6 nt of the intron). However, nucleotides capable of participating in U1 snRNA:pre-mRNA interaction have been shown to include positions −3 to +8 of the 5' ss and all 11 nt constituting the single-stranded 5' end of U1 snRNA (28).

Indeed, an alignment of 46,308 annotated canonical human 5' ss does not display a significant bias towards position +7 and +8 (Figure 2). Nevertheless, a contiguous stretch of only six U1 snRNA complementary nucleotides can be functional for splicing, as demonstrated by CAGGTAnnnnn and nnnGTAAGTnn (n=non-complementary; (25)). Thus, an 11 nucleotides wide motif from a 5' ss alignment superimposes splice sites with very different – exon or intron centered – regions of U1 snRNA complementarity, obliterating splice site motif information in downstream positions +7 and +8. This becomes evident when 5' ss subsets are selected according to the U1 snRNA complementarity of the exonic nucleotides. Within the subset of 10,796 5' ss sequences with full U1 snRNA (Watson-Crick-) complementarity in the three exonic positions (upper left), the intronic complementarity is less pronounced. Vice versa, the 3,830 5' ss with no complementarity in the exonic positions (lower left) clearly display a bias (12 and 8 percentage points) towards complementary bases even in position +7 and +8.

Furthermore, the stability of the RNA duplex does not seem to be exclusively determined by its complementarity to U1 snRNA, but also by additional interactions of protein components with the pre-mRNA in the vicinity of the 5' ss, including the U1-specific proteins, U1-C, U1-A and U1 70K, which bind to the loop I of U1
Figure 2. Sequence logos of annotated canonical human 5' ss. At each position the height of a base is proportional to its frequency in that position. Bases are ordered top to bottom in decreasing frequency. A sequence logo of 46,308 5' ss compiled from human chromosomes 6, 7, 9, 10, 13, 14, 20, 22 and X is shown in the upper right corner, with percentages of the “top” (most frequent) bases given below the intronic positions. This data set is split into three subsets: 10,796 sequences with full U1 snRNA (Watson-Crick-) complementarity in the three exonic positions (upper left), 3,830 with no complementarity in the exonic positions (lower left), and the remainder (mixed complementarity). To indicate the effect of the exonic (non-) complementarity on the intronic nucleotide frequencies within the splice site, the “top” intronic bases’ frequency change is given in percentage points above the logos.

snRNA (29, 30, 31). In addition, the recognition of 5' ss seems to be subjected to a proofreading mechanism since conformational rearrangement during spliceosome assembly results in the displacement of U1 by U6 snRNA which base-pairs to positions +2 to +6 of the 5' ss through an invariant ACAGA-box sequence in U6 snRNA, and by U5 snRNA (32, 33, 34, 35, 36).

4.2. 3' Splice site recognition

The 3' ss is a multipart signal comprising a less conserved branchpoint consensus YNYURAY (Y = pyrimidine, R = purine, N = any nucleotide, branch point is underlined), and a stretch of pyrimidines (known as the polypyrimidine tract or PPT) adjacent to the invariant 3' ss AG (37).

The distances between 3' ss signals are highly variable. The branch point sequence (BPS) is usually located 18-40 nucleotides upstream of the 3' ss AG, but may also reside up to several hundred nucleotides further upstream (38, 39, 40). Accordingly, polypyrimidine tracts vary in length and sequence composition. In particular, those polypyrimidine tracts composed of long uridine stretches promote the use of adjacent 3' ss (41, 40). However, natural polypyrimidine tracts are frequently interrupted by cytosines or purines (42). This is reflected by the essential pre-mRNA splicing factor U2AF65, which guides splice site selection by recognition of the polypyrimidine tract near the 3' ss AG. U2AF (U2 auxiliary factor) is a heterodimer comprising a large subunit, U2AF65, and a small subunit, U2AF35 (43, 44). U2AF65
might distinguish purines (adenine and guanine) from pyrimidines (uracil and cytosine) on the basis of their size, but more likely on the basis of their unique patterns of hydrogen bond donors and acceptors (45). Since U2AF65 preferentially binds uridine-rich RNA segments, polypyrimidine tracts with long uridine stretches are stronger than those with interruptions of other nucleotides (46). These weak polypyrimidine tracts require an additional U2AF35-3' ss AG interaction for their recognition (40, 47). The branch point, which often bears little resemblance to the consensus motif, appears to be specified independently of the 3' ss AG by its immediate sequence context and by its proximity to the polypyrimidine tract (48).

Consequently, the 3' ss recognition can be subdivided into several steps. In a first recognition step, mBBP/SF1 (mammalian branchpoint binding protein / splicing factor 1) specifically recognizes both the branch site sequence and the branch site adenosine through its KH (hnRNP K homology) domain (53, 54). SF1 cooperatively interacts with the splicing factor U2AF65 (55), which binds to the adjacent polypyrimidine tract through its RNA recognition motif (RRM) (56, 57, 45). Mutational analysis and in vitro genetic selection indicate that U2AF35 has a sequence-specific RNA-binding activity that recognizes the 3' ss consensus, AG/G (58, 47, 59). The recognition of the 3' ss is proofread by DEK, a chromatin- and RNA-associated protein, which has to be phosphorylated for intron removal and prevents binding of U2AF65 to pyrimidine tracts not followed by AG (60).

Concurrently, U2AF65 recruits the U2 snRNP via binding to the U2 snRNA associated protein SF3b155. SF3b155 represents a subunit of the heteromic splicing factor SF3b (61, 62), which interacts with the 5'-half of the U2 small nuclear RNA (U2 snRNA), whereas SF3a associates with the 3'-portion of U2 snRNA (63). There is evidence that sequence-independent binding of the highly conserved SF3a/SF3b subunits upstream of the branch site is essential for anchoring U2 snRNP to the pre-mRNA (64). The U2 snRNP base pairs with the branch point region while the nucleophilic branch site adenosine does not base pair with the U2 snRNA, but rather bulges out of the recognition helix (65, 66, 67). Binding of mBBP/SF1 is mutually exclusive with the U2 snRNP, thus the U2 snRNP replaces mBBP/SF1 (68). Upon stable integration of the U2 snRNP into the spliceosome, a 14 kDa protein (p14) interacts directly with the branch adenosine (69). Most probably, p14 is positioned within the inner cage of the SF3b structure (61).

As soon as both splice sites are recognized, the tri-snRNP complex of U4/U6/U5 snRNP enters and interaction between the U2 snRNP and U6 snRNP generates the catalytic core of the spliceosome (70, 71, 72, 73, 74). In addition to the snRNP spliceosomal “DExD/H box”, ATPases are required for promoting RNA rearrangements and many non-snRNP protein factors are involved in proofreading the steps of splicing within the spliceosome that is composed of as many as 300 distinct proteins (75, 76, 32).

4.3. Cis-active regulatory elements

Accurate splice site recognition further depends on cis-regulatory elements in the pre-mRNA that modulate splice site selection and allow to discriminate between real and pseudo splice sites (77, 78) (Figure 1). Most exons contain exonic splicing enhancers (ESEs), which define them as recognition units promoting the use of their splice sites (79, 80, 81). In addition, exons also contain functional splicing suppression units known as exonic splicing silencers (ESSs) (82, 83). Moreover, intronic splicing enhancers (ISEs) or intronic splicing silencers (ISSs) enhance or repress the use of nearby 5' or 3' ss (84, 85, 86, 87, 88). These cis-acting splicing regulators are short degenerate RNA sequences, which occur frequently in the genome.

Enhancer motifs are frequently found by the group of serine/arginine rich (SR) proteins, which mostly exerts a positive effect on splice site recognition and stimulates spliceosome assembly (89, 90, 91, 92, 93, 94, 95, 96, 97, 98). These positive effects can be antagonized by heterogeneous nuclear ribonucleoproteins (hnRNPs) that usually bind to silencer elements (99, 100, 101, 102, 103, 104). However, it should be noted that the same sequence motif can act as an enhancer or silencer, depending on its position with respect to the splice sites (105, 106). The activities of cis-acting elements were shown to be context specific and there is compelling evidence that SR proteins can suppress splicing when bound to sequences located within the intron, and there are also examples of members of the hnRNPs exhibiting stimulating effects on splicing (107, 108, 109, 110, 111, 98). HnRNPs recognize the RNA via their KH (K homology) and RRM RNA-binding domains and RGG and glycine-patch domains. The multiple α-helices and antiparallel β-strands bind short motifs of 4-7 nucleotides in single-stranded DNA or RNA. Moreover, the β-sheet surface on the RRM domain of many SR proteins recognizes specific RNA sequences through base stacking, hydrophobic, polar and electrostatic interactions (112, 113, 114, 115, 116). The majority of KH and RRM proteins contain more than one copy of each RNA recognition domain engaging a range of different motifs leading to ‘fuzzy’ identity of cis-active regulatory elements (117).

Specific splice site regulation, despite frequent occurrence of the degenerate target motifs, is achieved by clusters of degenerate RNA motifs bound by several different activator and repressor proteins. In addition, competition between SR proteins and hnRNP or between these proteins and general splicing factors modulate splice site selection (118). Furthermore, the activity of SR proteins as splicing factors depends on the phosphorylation
status of the serine residues in their RS domains, which can lead to a movement into a different subcellular localization (such as from the nucleus to the cytoplasm), where they are unable to affect splicing (119, 118, 120). The RS domains of SR proteins engage in protein-protein interactions promoting interactions between the components of the spliceosome to define exons or interactions across the intron during spliceosome assembly (121). Binding of RS domains to RNA presumably shields negative charges facilitating annealing of complementary RNA strands during numerous base-pairing rearrangements required for spliceosome assembly and catalysis (122, 123, 118).

5. SPLICE SITE STRENGTH AND IDENTIFICATION OF REGULATORY MOTIFS

Mutations, even single nucleotide changes, can modify splicing in various ways: they can strengthen, weaken or even destroy an existing proper splice site or cis-regulatory element, or create a new one. Such splicing signal modifications may or may not lead to observable phenomena like exon skipping, activation of cryptic or de novo splice sites, or intron retention. Most patients, however, are genotyped only, and diagnostic RNA-level information about aberrant splicing is usually not available. Therefore, any computational prediction of DNA mutation effects on splicing (for an overview see Table 1) can be beneficial for the human geneticist. Such predictions can be obtained from algorithms scoring the functionality of a given splice site and/or cis-regulatory element.

Ab initio gene prediction mostly employs probabilistic algorithms like Hidden Markov Models (HMM) (124), dividing the gene structure into interconnected submodels (“states”) for components as promoters, splice sites, start and stop codons (GENSCAN, AUGUSTUS) (http://genes.mit.edu/GENSCAN.html (125, 126), http://augustus.gobics.de/submission (127, 128, 129)). It is a forte of HMM to accommodate both unobserved (“hidden”) variables as well as observable ones – the actually generated nucleotides –, with transition probabilities governing their relations (130, 131). Moreover, ab initio methods often use additional, highly non-local information as reading frame and protein content (132). In contrast, splice site prediction methods try to use only the information available to the biological splicing machinery, and are primarily based on computational models for the neighborhood of the conserved dinucleotides GT and AG (133, 134).

5.1. Splice site strength algorithms

The “splice site strength” is a useful and central concept in judging the possible effect of a splicing signal mutation. Together with a “threshold” for splice site functionality, comparing strengths of wild type and mutant signal could yield reliable predictions of splicing effects (22). However, although widely used in the literature, the term “splice site strength” does not refer to a unique definition. In principle, any measure of “functional splicing signal strength” should quantitatively describe, why a given splice site is preferred over competing nearby potential (“pseudo”, “mock” or “decoy”) splice sites under cell-specific conditions. It should take into account not only the proper 5’ or 3’ ss sequence, but also its context of cis-regulatory elements and pseudo splice sites, and even the cellular environment of SR proteins. In practice, this ambitious comprehensive concept (“the splicing machinery itself”) has not yet been implemented in silico and is approximated by more limited computational procedures. It comes natural that a wide variety of concepts from computational physics, artificial intelligence and machine learning have been applied to this problem.

In principle, two types of computational methods for splice site detection can be distinguished: those that are trained only by positive examples (real splice sites) – e.g. Weight Matrix/Array Models and Maximum Dependency Decomposition –, and those additionally requiring a training data set of negative examples (decoy splice sites). Locally, several different algorithms calculate a splice site’s intrinsic strength from a narrow region of nucleotides around the respective consensus dinucleotides (GT or AG), irrespective of its wider sequence context. A splice site’s relative strength then refers to the difference (or ratio) of its intrinsic strength to the neighboring pseudo sites, thus depending on the splice site context. The meaningful combination of cis-regulatory elements and relative splice site strength into a single functional strength measure still remains an open question, although a first step towards combining splice site scores and those of cis-regulatory elements has been taken by the splicing simulation software ExonScan, which independently adds up log-odds-scores of individual components to obtain one overall score (http://genes.mit.edu/exonscan/ (82, 83)).

However, all local primary sequence methods are bound to misdiagnose splice sites, due to the huge overlap of sites in the real and decoy data sets. This property yields an upper bound on the accuracy of splice site prediction, since predicting a real splice site as real also admits the decoys with the same sequence, and vice versa. E.g., 98% of the real 5’ ss sequences in the Yeo-Burge data set (http://genes.mit.edu/burgelab/maxent/ssdata/) are also contained in the decoy set, so that for this data the maximum accuracy for 5’ ss is found to correspond to a correlation coefficient C = 0.676 (135).

The performance of a splice site scoring algorithm on two given data sets of real and decoy splice sites is usually evaluated by receiver-operating characteristic (ROC) curves, plotting sensitivity versus 1−specificity while varying the prediction threshold X between predicted positive sites with score >X and predicted negative sites with score <X. The prediction sensitivity is the rate of predicted positives among all real splice sites, while the specificity (= 1−false-positive ratio) is the rate of predicted negatives among all decoy splice sites. By construction, the ROC-graph lies above the unit square diagonal, and increases monotonously from the lower left corner (high X, low sensitivity, high specificity) to the upper right corner (low X, high sensitivity, low specificity). A greater true positive rate at a given false positive rate indicates a more accurate scoring method. The area under the ROC-graph (AUC), a number between 0.5 and 1, serves
Diagnoses of pathogenic splicing mutations

Table 1. Currently available web tools for splice site assessment and identification of putative cis-regulatory elements

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Basic principle</th>
<th>Link</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro &amp; Senapathy</td>
<td>Position-specific weight matrix (PSWM); Coincidence with consensus sequence</td>
<td><a href="http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm">http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm</a></td>
<td>5' ss motif (9-mer) position -3 to +6; 3' ss motif (15-mer) position -14 to +1</td>
<td>S&amp;S Score (0-100)</td>
</tr>
<tr>
<td>Weight matrix model</td>
<td>Quantification of the relative likelihood of candidate splice site sequence with respect to the background nucleotide distribution from a training set of splicing signals</td>
<td><a href="http://genes.mit.edu/burgelab/maxent/Xma">http://genes.mit.edu/burgelab/maxent/Xma</a> xentscan_scoreseq.html</td>
<td>5' ss motif (9-mer) position -3 to +6; 3' ss motif (23-mer) position -20 to +3</td>
<td>Numerical Score</td>
</tr>
<tr>
<td>Maximum Dependence</td>
<td>Iterative decision-tree approach; captures dependencies between neighboring and non-neighboring positions by WAM (weight array model) and WMM</td>
<td><a href="http://genes.mit.edu/burgelab/maxent/Xma">http://genes.mit.edu/burgelab/maxent/Xma</a> xentscan_scoreseq.html</td>
<td>5' ss motif (9-mer) position -3 to +6; 3' ss motif (23-mer) position -20 to +3</td>
<td>Numerical Score</td>
</tr>
<tr>
<td>First-order Markov</td>
<td>Statistical approach which considers dependencies between adjacent positions</td>
<td><a href="http://genes.mit.edu/burgelab/maxent/Xma">http://genes.mit.edu/burgelab/maxent/Xma</a> xentscan_scoreseq.html</td>
<td>5' ss motif (9-mer) position -3 to +6; 3' ss motif (23-mer) position -20 to +3</td>
<td>Numerical Score</td>
</tr>
<tr>
<td>Maximum Entropy Model</td>
<td>Statistical approach representing the least biased approximation for the distribution of sequence motifs, consistent with a set of constraints estimated from available data (real and decoy splice sites), incorporates local adjacent and non-adjacent position dependencies</td>
<td><a href="http://genes.mit.edu/burgelab/maxent/Xma">http://genes.mit.edu/burgelab/maxent/Xma</a> xentscan_scoreseq.html</td>
<td>5' ss motif (9-mer) position -3 to +6; 3' ss motif (23-mer) position -20 to +3</td>
<td>Numerical Score</td>
</tr>
<tr>
<td>H-Bond Model</td>
<td>Hydrogen bond pattern between the 5' ss and all 11 nt of the free 5' end of the U1 snRNA; hydrogen bond formation at individual positions, models nucleotide interdependence beyond nearest neighborhood relationships; experimental evidence</td>
<td><a href="http://uni-duesseldorf.de/rna">http://uni-duesseldorf.de/rna</a></td>
<td>5' ss motif (11-mer) position -3 to +8;</td>
<td>Numerical Score</td>
</tr>
<tr>
<td>Neural Network (NN)</td>
<td>Machine learning approach that recognizes sequence patterns once it is trained with sets of DNA sequences encompassing authentic and decoy splice sites</td>
<td><a href="http://www.fruitfly.org/seq_tools/">http://www.fruitfly.org/seq_tools/</a> splice.html</td>
<td>long sequence stretches</td>
<td>Score between 0 and 1</td>
</tr>
<tr>
<td>Branch Site Tool</td>
<td>Algorithm that locates both the BPS based on its consensus sequence together with the PPT by searching known combination of BPS and PPT. The PPT borders are determined by a heuristic method based on experimental evidence</td>
<td><a href="http://ast.bioinfo.tau.ac.il/BranchSite.htm">http://ast.bioinfo.tau.ac.il/BranchSite.htm</a></td>
<td>long sequence stretches</td>
<td>Numerical Score</td>
</tr>
<tr>
<td>ESEfinder</td>
<td>Prediction of SR protein specific putative ESE, based on an in vitro SELEX approach dependent on addition of individual SR proteins</td>
<td><a href="http://rulai.cshl.edu/tools/ESE/">http://rulai.cshl.edu/tools/ESE/</a></td>
<td>long sequence stretches</td>
<td>ESE motif score</td>
</tr>
<tr>
<td>RESCUE-ESE</td>
<td>Statistical approach based upon different distribution of hexamers in exons and introns with different properties, e.g. weak and strong splice sites</td>
<td><a href="http://genes.mit.edu/burgelab/rescue-ese">http://genes.mit.edu/burgelab/rescue-ese</a></td>
<td>long sequence stretches</td>
<td>Z-score (picks out extremely over- or under-represented hexamers)</td>
</tr>
<tr>
<td>PESX server</td>
<td>Statistical approach based on over-representation of octamers in exons versus pseudoexons or versus the 5' UTR (untranslated regions) of intronless genes; Predicts enhancers (PESE) and silencers (PESS)</td>
<td><a href="http://cubweb.biology.columbia.edu/pesx/">http://cubweb.biology.columbia.edu/pesx/</a></td>
<td>long sequence stretches</td>
<td>P-score, l-score (see Z-score)</td>
</tr>
<tr>
<td>ESR search</td>
<td>Evolutionary conservation of wobble positions with statistically significant overabundance of dicodons (hexamers) relative to the expected frequency of their independent individual codons</td>
<td><a href="http://ast.bioinfo.tau.ac.il/ESR.htm">http://ast.bioinfo.tau.ac.il/ESR.htm</a></td>
<td>long sequence stretches</td>
<td>Z-score</td>
</tr>
</tbody>
</table>

as an overall parameter of description accuracy (for an example ROC see Figure 3).

5.2. 5' Splice site strength algorithms

The most widely-used intrinsic strength concept simply measures the 5' splice site’s similarity with a consensus motif. Initially, Shapiro and Senapathy (S&S) developed a position-specific weight matrix (PSWM) for 5' ss, which reflects the degree of sequence conservation of the known 5' ss from position ~3 (the third nucleotide from the 3' end of the upstream exon) to +6 (the sixth nucleotide in the intron) in an alignment of 1,446 5' ss (42, 136). From
Diagnostics of pathogenic splicing mutations

![Figure 3](image)

**Figure 3.** Exemplary receiver-operating characteristic (ROC) curve comparing sensitivity and specificity of three 5' ss scores, maxent (red), HBond (green) and Shapiro-Senapathy (black), on a given data set of 7,839 canonical real and 56,408 exonic decoy 5' ss (not used in alternative transcripts) from human chromosome 6. For each score (prediction) threshold X, 5' ss with score >X are predicted positive, and those with score <X are predicted negative. The prediction sensitivity is the rate of predicted positives among all real 5' ss, while the specificity (= 1 – false-positive ratio) is the rate of predicted negatives among all decoy 5' ss. The ROC plots the sensitivity as function of the false-positive ratio, obtained by variation of the assumed threshold X. Only the sensitivity range 0.7 – 1.0 is plotted here. By construction, the ROC-graph lies above the unit square diagonal, and increases monotonously from the lower left corner (high X, low sensitivity, high specificity) to the upper right corner (low X, high sensitivity, low specificity). A greater true positive rate at a given false positive rate indicates a more accurate 5' ss scoring method. The area under the ROC-graph (AUC), a number between 0.5 and 1, serves as an overall parameter of description accuracy. Note that both maxent and HBond score significantly outperform S&S score. ROC curves for maxent and HBond scores are very close; only in the HBond score range 10.45 to 11.85 does maxent outperform HBond.

This matrix they derived the S&S score in the range 0–100, with score 100 representing full coincidence with the consensus sequence, and score 0 obtained, if every position is occupied by the least likely nucleotide. All positions in the 5' ss are assumed independent by the S&S score, as with every weight matrix model.

Traditionally, splice sites with a high degree of resemblance to the consensus have been considered as strong splice sites, whereas non-consensus splice sites have been assumed to be intrinsically weak. Although this is still widely accepted, significance of such a consensus sequence remains arguable, because resemblance to frequency-based consensus matrices of independent nucleotides turned out to be insufficient for reliable prediction of 5' ss (137). Moreover, many matches to each consensus are present along pre-mRNAs, but the vast majority of these sequences are pseudo or decoy splice sites never selected for splicing (78).

Weight matrix models (WMM) represent an extension to the S&S score, indicating the relative importance of each base at every position: they quantify the relative likelihood of a given candidate splice site sequence with respect to the background nucleotide distribution from a training set of splice signals, but they still fail to incorporate nucleotide interdependencies. To score a given sequence, WMM add up the logarithm of the independent likelihood in each motif position and yield a log-odds score that is not normalized. Although no decoy splice sites enter into the construction of WMM, real and decoy sites can be ranked according to the log-odds score, and WMM can be evaluated by receiver operating characteristics (ROC), plotting sensitivity versus 1 – specificity for different cutoffs between predicted real and predicted decoy sites (Figure 3). Weight array methods (WAM) slightly increase the discriminative power of WMM by using conditional probabilities for pairs of neighboring nucleotides (138, 139). Mathematically, both WMM and WAM are special cases of first order Markov chains. First-order Markov models (MM) only consider dependencies between adjacent positions, representing the sequence as a chain of “states” with transition probabilities from each position to its successor.

A complementary approach to the description by numerical weights can be implemented as a “dictionary procedure”, looking up sequences in different dictionaries (data sets) for real and decoy splice sites. The primary sequence rank (PSR) method ranks all sequences in the training data set according to their frequency of occurrence in the real data set versus the entire (real + decoy) set. To accommodate unknown sequences not contained in the training data set, the ranking is smoothed by adding pseudo-counts from neighboring sequences, permitting to calculate ranks for all training sequences plus their single and double point mutations (http://rna.williams.edu/). For the Yeo-Burge data set, the maximum correlation coefficient obtained for a smoothed 5' ss PSR model is C = 0.668, close to the theoretical upper limit C = 0.676, which is < 1 due to the considerable overlap of real and decoy splice sites (135).

An improvement for 5' ss prediction has been achieved by considering dependencies between bases of the 5' ss. Burge and colleagues developed three different algorithms that take into account dependencies between positions −3 to +6 of the 5' ss motif (140): these algorithms apply probabilistic approaches to large datasets of known RNA splicing signals. The maximum dependence decomposition model (MDD) is an iterative decision-tree approach that captures the strongest dependencies – also between non-neighboring positions – in the early branches of the tree by WAM, and uses WMM for nearly independent positions.

The maximum entropy model (MEM) performs better than previous models and is based on the maximum
Diagnostics of pathogenic splicing mutations

entropy distribution (MED). In statistical theory, this approach represents the least biased approximation for the distribution of sequence motifs, consistent with a set of constraints estimated from available data – known real and decoy signal sequences. It makes no further assumptions about the distribution than consistency with this empirical distribution, and different sets of constraints generate different models. The MEM incorporates local adjacent and non-adjacent position dependencies consistent with low-order marginal constraints for “few” nucleotides estimated from available data (MaxENTScanning algorithm: http://genes.mit.edu/burgelab/maxent/). These algorithms use input sequences of constant length – a 9-mer in case of the 5' ss and 23-mer for the 3' ss – and assign each sequence a numerical score reflecting the likelihood of the sequence being a true splice site. According to ROC curve analysis, the currently most successful 5' ss maxent model is me2x5 at a correlation coefficient C = 0.659, which is close to the theoretical maximum for this data set, and additionally has the lowest number of proximal intrinsic higher-scoring decoys. Comparing me2x5 to MDD and WMM, the sequence ranks are found to strongly differ between different models. Top-scoring sequences are usually well-correlated between models, while lower-scoring sequences vary much more.

While weight matrix/array models require the selection of relevant information features by hand, machine learning techniques automatically deduce a classification function (“rule”) that optimizes a given criterion in distinguishing training data sets of positive and negative sequences (real and decoy splice sites). For example, the neural network method (NN) is a machine learning approach that recognizes sequence patterns once it is trained with sets of DNA sequences encompassing authentic and decoy splice sites (http://www.fruitfly.org/seq-tools/splice.html) (141, 142, 143). It employs a backpropagation feedforward neural network with one hidden layer, and produces an output score between 0 and 1 for each splice site candidate. Interestingly, decoy GT sites close to a real 5' ss have weaker neural network scores than those farther away, which seems consistent with the concept of relative splice site strength, comparing a real 5' ss with decoy sites in its neighborhood.

Support vector machines (SVM) also belong to the category of machine learning systems that infer a classification function from a training data set. By using an appropriate representation for features of real and decoy sequences, specific splice site patterns can be obtained from the discrimination function of such models. Typically, only a small fraction of the large number of features, represented by a high-dimensional feature vector, are relevant for the classification and are mutually independent. Genetic algorithms have been successfully applied in the selection of such a “minimal feature set” with best classification performance. Estimation of distribution (EDA) algorithms have been shown to improve on these, most importantly providing normalized “feature weights” as ranking criterion (144, 145, 146).

SVM algorithms were also applied to detect splicing features in the human genome: 2,200 real and 2,300 pseudo exons including flanking intronic sequences were divided into five non-overlapping sequence compartments. The strongest features searched in words of length 4–7 nucleotides were the presence or absence of 4-mers and 5-mers, consistent with motifs identified by other methods, and at comparable sensitivities and specificities (147).

With a view to the biological function of the 5' ss as a recognition site for the U1 snRNP early in spliceosome assembly, it seems obvious to determine a 5' splice site’s intrinsic strength regarding this interaction. Indeed, stable RNA duplex formation between the U1 snRNA and the 5' ss is a prerequisite for spliceosome formation, and it has been shown that the stability of the U1 snRNA duplex has strong influence on the selection between two nearby 5' ss (148, 149, 28). From a thermodynamic viewpoint, the 5' ss:U1 snRNA duplex stability can be quantified by its free energy ΔG, using the nearest-neighbor RNA base-pairing parameters reported by the Turner laboratory (150). These empirically fitted formulae are based on measurements with synthetic oligoribonucleotides and reflect the contribution of hydrogen bonding, base stacking, mismatches, and Watson-Crick or G-U base pairs (151). The nearest-neighbor approximation works very well for Watson-Crick base pairs, satisfactorily well for G-U base pairs flanked by Watson-Crick base pairs, but is less reliable for mismatches. Moreover, undetermined energy corrections at the ends of a short RNA duplex may impose limits on the accuracy of the free energy calculations (28, 152, 153). Therefore, approximate free energies, calculated e.g. by popular computational web tools like DynAlign (154), HyTher (155) and Bindigo (156), seem insufficient for a reliable description of U1 snRNA duplex contribution to 5' ss strength.

In a complementary approach to experimentally determine intrinsic 5' ss strength in a model system, U1 snRNA duplex formation has been monitored within a retroviral-derived model transcription unit (28). It is well known that stable U1 snRNA duplex formation with 5' ss can protect pre-mRNA against degradation prior to splicing, and also initiates formation of the spliceosome. In combination with functional splicing assays, this protection mechanism has been used to obtain biological evidence of duplex stability. This experimental evidence was supplemented with a computational hydrogen bond weight model, translating the hydrogen bond pattern between the 5' ss and all 11 nt of the free 5' end of the U1 snRNA into a numerical HBond score (available at the web-interface http://www.uni-duesseldorf.de/rna/html/hbond_score.php). Beyond hydrogen bond formation at individual positions, the HBond algorithm also partially models nucleotide interdependence beyond nearest neighbor relationships. Contrary to purely statistical approaches currently ignoring nucleotides beyond position +6 due to lack of information content, the HBond algorithm fully takes positions +7 or +8 into account, with experiments confirming the dependency of the U1 snRNA duplex on these nucleotides. This observation is consistent with in vitro selection experiments.
Diagnostics of pathogenic splicing mutations

to isolate functional 5' ss from pools of random sequences, where those 5' ss with the best complementarity to U1 snRNA were selected most efficiently, even if base-pairing to U1 snRNA extended to positions +7 and +8 (157).

Moreover, mutual relationships between nucleotide positions within the 5' ss motif have been confirmed by human-mouse comparative genomics, and the nucleotide positions within the 5' ss motif have been examined extensively by in vitro 5' ss competition assays of the human β-globin gene (158, 152, 153). Studies with this gene revealed that the authentic 5' ss of the first exon lies in the vicinity of a cryptic 5' ss located 16 nucleotides upstream, which is only activated when the authentic one is sufficiently weakened by mutation (159, 152). In this case, the cryptic splice site can outweigh the mutant authentic one and be selected for splicing. Six 5' ss scores, including free energy ΔG, S&S, MM and MAXENT, were compared regarding their ability to explain these in vitro splicing analyses. However, no discriminating score threshold could be determined for any score that stringently separated activated from unused potential splice sites. Correlation (Pearson’s r) between experimentally determined percentage of splicing activation and scores was maximal for MAXENT, MM and ΔG in different competition schemes, suggesting mechanisms captured by different score algorithms. Indeed, both authentic and weakened 5' ss (reference sequences) have complementary nucleotides in positions +7 and +8, while the test sites do not. All examined 5' ss scores ignore these positions, which may be accountable for the lack of stringent differentiation. Interestingly, there was no correlation between the extent of complementarity of the 5' ss with U6 snRNA, which is in accordance with the observation that hyperstabilization of the 5' ss:U1 snRNA interaction does not inhibit replacement of the U1 snRNP by the U6 snRNP in higher eukaryotes (160, 152).

5.3. 3’ Splice site strength algorithms

The description of the inherent strength of 3’ ss is more complicated due to sequence constraints of the 3’ ss motif including the AG dinucleotide, the presence of the polypyrimidine tract (PPT) and the branch point sequence (BPS) upstream of the 3’ ss. In addition, the distances between 3’ ss signals are highly variable.

Algorithms that describe the intrinsic strength of 3’ ss are based on nucleotide frequency matrices, machine learning approaches, neural networks, and on information contents of individual nucleotides, or apply probabilistic approaches considering dependencies between adjacent and non-adjacent positions (142, 143, 161, 162, 42, 136, 140). The Shapiro and Senapathy matrix counts base frequencies at positions -14 to +1 of the 3’ ss motif, whereas the first-order Markov (MM) and maximum entropy model (MaxEntScore) use a wider sequence range of 3’ ss positions from -20 to +3 (AG consensus at positions –1 and –2). Since the 3’ ss sequence motif is much longer than the 5' ss, in a first step the maxent approach breaks up the 3’ ss sequences into 3 consecutive non-overlapping fragments of length seven each, excluding the invariant AG dinucleotide. This splitting, however, ignores the dependencies across fragment boundaries. To avoid that, six additional partially overlapping subfragments are introduced, and the final maxent likelihood is calculated from the appropriate ratio of individual segment distributions using second-order marginal constraints in each segment. While this second-order Markov model is superior to a first-order model, performance is decreased again for third-order models. Long-range dependencies across several “skipped” nucleotides are neglected in these models, but introducing additional dependencies does not significantly improve the performance beyond two-nucleotide-separation.

The optimal performance among the examined models was achieved by an me2x2-model, skipping up to two nucleotides, with a maximum correlation coefficient of C = 0.6291, which only slightly exceeds the simpler modified me2x0-model (C = 0.6172). Comparison of the splice site strength using current prediction algorithms showed that the maximum entropy model class allowed the best discrimination between authentic and mutation-induced aberrant 3’ ss (134).

Ast and colleagues developed an algorithm which combines pairs of PPT and BPS to identify the location of functional BPS, since consensus scores alone are not sufficient to locate the BPS in introns due to frequent occurrence of high score motifs in exons and introns (http://ast.bioinfo.tau.ac.il/) (163). This algorithm is based on the BPS consensus calculated by Burge (164) and locates both the BPS and the PPT together by searching known combinations of BPS and PPT. The PPT borders are determined by a heuristic method based on experimental evidence (41, 165).

Their approach is contrasted by an algorithm which is primarily based on AG dinucleotide exclusion zones between the 3’ ss AG and the BPS for branch point prediction (38). This algorithm incorporates exons with distant BPS extending the usual search for probable branch points within a fixed distance of the 3’ ss. Nevertheless, prediction of cryptic and de novo 3’ ss is still a difficult task (166).

In general, local primary sequence data is insufficient to determine, if a given sequence will be a splice site, already due to the substantial overlap of real and decoy splice site sets. Additional information away from the exon/intron junction is needed to put a definitive label on a splice site. Besides splicing enhancers and silencers treated in the next section, RNA secondary structure may be an important non-local splicing determinant.

5.4. Computational identification of exonic regulatory elements

Even translationally silent mutations can act on RNA level by altering cis-regulatory elements and thereby disrupt splicing. Hence, the combination of intrinsic or relative splice site strength with context information on the identity and distribution of cis-regulatory elements may improve the prediction of aberrant splicing in disease
Diagnostics of pathogenic splicing mutations

mutations, and may further clarify its pathway (exon skipping or activation of cryptic splice sites).

Exonic enhancers, frequently bound by SR proteins, have been identified through the analysis of disease alleles and by site-directed mutagenesis of minigene constructs, traditionally on a single transcript basis, whereas more recent purely computational approaches, followed by experimental validation, have contributed to general definition of splicing regulatory motifs.

ESE protein-binding sites on RNA are typically modeled by some form of position-specific scoring matrix (PSSM), constructed from aligned sets of experimentally determined binding sequences (167). For every position along the pre-mRNA strand, a log-odds score ("ESE motif score") is calculated from these PSSM, measuring the probability that the sequence at that position is an instance of the considered ESE motif, compared to the global background nucleotide frequencies. This score, however, treats different positions independently and cannot accommodate interactions. In a typical analysis, for each ESE an individual score threshold (detection sensitivity) is given, and only scores above the threshold are considered to belong to putative ESE positions.

Position-specific scoring matrices can only be successfully derived from sets of binding sites that are homogeneous with respect to the binding protein, which excludes non-specific screening assays based on cis-regulatory activity of one or more putative elements (168). Derivation of PSS matrices from experimental data is typically performed by alignment algorithms, such as the Gibbs Motif sampler (169), sometimes followed by clustering of the detected motifs to reduce the number of independent patterns, e.g. by CLUSTALW (170).

Most early experimental work on enhancers focused on purine-rich exonic elements, but it soon became clear that a high purine content by itself is not sufficient to define an ESE, as the precise sequence of the element, which might contain interspersed pyrimidines, is also important (171). Functional SELEX (Systematic evolution of ligands by exponential enrichment) experiments have confirmed the existence of several types of ESE that include both purine-rich and non-purine-rich sequences, and have uncovered a new broad class of adenosine-cytosine-rich elements (172). These SELEX methods isolate ESEs from a complex pool of random sequences by iteratively selecting and amplifying the fraction of molecules that can function as ESEs in a highly specific reporter assay. Splicing that is dependent on the addition of individual SR proteins to cytoplasmic extract allowed the selection of distinct motifs for SF2/ASF, SC35, SRp40 and SRp55 (94). The identified motifs are short (6-8 nt), degenerate, and some overlap partially.

To predict the location of SR protein specific putative ESEs, the web-based program ESEfinder (http://rulai.cshl.edu/tools/ESE/) permits the calculation of ESE motif scores along pre-mRNA sequences (79, 173). ESEfinder can also be useful to predict mutation effects: e.g., a number of disease-associated point mutations resulting in exon-skipping reduced high-score motifs to below threshold values (174), and on the other hand, a mutation that results in activation of a cryptic 5' ss due to increased SC35 binding to an ESE leads to a higher score (175). Unfortunately, to date ESE motifs for only four SR proteins have been identified, and sequences corresponding to the RNA binding specificities of other SR proteins still remain to be found. Furthermore, interactions of different enhancers and/or silencers have to be taken into account: the presence of a high score motif in a sequence does not necessarily identify that sequence as an exonic splicing enhancer in its native context, since nearby or overlapping silencer elements may interfere.

Iterative in vivo selection of exonic sequences through splicing in an intact cell may have an advantage over in vitro SELEX methods by recapitulating a true process of molecular evolution. An advanced in vivo SELEX has recently been used in a proof-of-concept experiment for SMN1 exon 7 analysis, employing partial randomization of an entire exon of 54 nucleotides (176, 177). Based upon the mutability of residues, two negative elements near the 5' and 3' ends of SMN1 exon 7, and one positive element in the middle were identified. Interestingly, these elements were considerably larger than current computationally predicted enhancer or silencer motifs, which is consistent with the fact that most splicing factors are capable of forming multi-component complexes with higher RNA specificity than their individual components, e.g. SR proteins. Since by this in vivo method the relative significance of every exonic position was tested, the variable impact of individual nucleotides within an identified cis-regulatory element could be determined. Beyond uncovering key positions within the exon, mutability values also provided clues about the critical roles of terminal stem-loop RNA structures. As the chosen example demonstrates, a forte of the method is the identification of cis-regulatory elements in the context of the entire exon. With the proof-of-concept in place, it will be a natural extension to examine cis-regulatory elements in more complex, alternatively spliced exons, and to incorporate flanking intronic sequences.

There are several statistical approaches to identify cis-regulatory elements based upon the different distribution of oligomers in exons and introns with different properties, e.g. weak and strong splice sites. Calculating statistically normalized Z-scores, they pick out extremely over- or underrepresented oligomers as candidate sequences. Evidently, such Z-scores are highly dependent upon the strength concept chosen, and suboptimal underlying splice site scores will lead to less useful Z-scores and motifs.

In the RESCUE-ESE (relative enhancer and silencer classification by unanimous enrichment) approach, hexamers were identified in constitutively spliced human exons by enrichment in exons versus introns, and in exons with weak splice sites versus exons with strong splice sites—separately for 5' and 3' ss (178). Here, the splice site
Diagnostics of pathogenic splicing mutations

strength was measured by a log-odds-score, derived from a position-specific weight matrix measuring the splice site similarity with the consensus motif for nucleotides –3 through +6. Hexamers that were significantly enriched in exons with weak splice sites are assumed to act as enhancers, consistent with the view that exons with weak splice sites are not accurately spliced without the aid of additional enhancer elements. The identified $238$ distinct hexamers clustered into ten motifs on the basis of sequence similarity. Two of them were specific to 5' ss, five were specific to 3' ss, and three associated with both splice sites. For each cluster a representative hexamer was validated by its ability to “rescue” exon recognition in a splicing reporter construct and the robustness of the approach with respect to differences in the local context was demonstrated by extended exemplars of the hexamers.

This study revealed an average of 5.2 ESE "clumps" (multiple overlapping ESE hexamers) per exon, with most exons having between 3 and 7 ESE clumps. The RESCUE-ESE method has recently been applied to three additional species (mouse, zebrafish and pufferfish) and is also available as an online ESE analysis tool (http://genes.mit.edu/burgelab/rescue-ese) (80). Similar caveats as with the ESEfinder motifs apply to the RESCUE-ESE candidates, since context effects and interactions with adjacent silencer motifs, as well as secondary structure could not be taken into account.

By following a similar rationale, RESCUE-ISE predicts as ISEs (intronic splicing enhancer) hexamers that share two properties: significant enrichment in introns relative to exons and significant enrichment in introns with weak 5' ss or 3' ss relative to introns with strong splice sites. Applying this method to large datasets of human and mouse introns identified the triplet motif GGG and a C-rich motif, respectively (179). When the RESCUE-ISE approach was applied to datasets of Fugu introns, the analysis revealed dramatic differences in candidate ISEs between mammals and fish, which may be important with regard to the evolution of the splicing process.

There is a sharp transition in sequence composition between exons and introns because of the fact that most exons code for protein, whereas introns do not. Since it is not clear to what proportion this information is used as “ESE” in addition to protein coding, the PESX algorithm compares frequencies of octamers in constitutively spliced noncoding exons with those in pseudo exons and the 5' untranslated regions (UTRs) of intronless genes. Sequences overrepresented in the noncoding exons were designated as putative ESEs, and underrepresented sequences were designated as putative ESS (exonic splicing silencer) (180). Similar to the RESCUE-ESE hexamer approach, for each octamer there are two normalized Z-scores, describing the joint probability of occurrence in the different data sets: the P-score measures overrepresentation in exons versus pseudo exons, while the L-score compares to the 5' UTR of intronless genes. For a given significance threshold (0.2% probability of chance occurrence), 2,069 (=3% of all possible) octamers were found as putative exonic enhancers and 974 as exonic silencers. These PESE/PESS can be grouped into families of related sequences, and consequentially tend to occur in clusters. For representative octamers of each class, regulatory effects were confirmed by tests in minigene constructs. To examine effects of PESE/PESS in their natural context, they were knocked out by site-directed mutagenesis in six exons naturally containing them (181). Eighteen of the 22 mutations actually disrupted splicing or reduced splicing efficiency by a factor of at least 2. The high success rate of mutagenesis implies a remarkable lack of redundancy among multiple ESEs, and in most cases several ESEs must act in concert to ensure splicing, underlining the importance of ESE interaction. There is some overlap between RESCUE-ESE-, ESEFINDER- and PESE-motifs, but each method identifies ESEs missed by the other two. Several experimental results point to ISEs as alternatives to ESEs (181). The identification of PESEs and PESSs is also permitted through a web site interface (http://cubweb.biology.columbia.edu/pesx/) (181).

A comparative genomics method based upon 46,103 human-mouse orthologous exons extracted 285 significant exon splicing regulatory sequences (ESRs) by combining evolutionary conservation of wobble positions with statistically significant overabundance of dicodons (hexamers) relative to the expected frequency of their independent individual codons. For a given reading frame, wobble positions are codon positions which, due to the degeneracy of the genetic code, do not result in amino acid changes. Conservation of nucleotides in such wobble positions in homologous exons between mouse and human may reflect a selective pressure, indicating that mutations in these nucleotides did alter appropriate splicing (105). ESRs were found conserved to a higher degree in alternatively spliced exons, while having higher abundance in constitutively spliced exons, which may suggest that constitutive exons can afford changes in ESR elements due to multiple redundant splicing signals, while alternatively spliced exons are more sensitive. The experimental validation of ten putative ESRs, measuring exon inclusion levels in two suboptimal minigene constructs, revealed that the same ESR can have positive as well as negative effect on splicing, depending on different exon context and even their locations within the same exon. These findings are consistent with previous reports of regulatory factors having enhancing or repressing effects, depending on the location of their binding sites in different RNAs (106).

Exonic silencers have also been systematically searched by ‘fluorescence-activated screen for exonic splicing silencers’ (FAS-ESS), an elaboration of an in vivo functional selection approach using cell-transfection assays (83). A library of random decamers was cloned into a test exon, stably transformed cells were selected, and cells that produced eGFP were selected by FACS. PCR of genomic DNA then allowed 141 FAS-ESS decamer sequences to be produced eGFP were selected by FACS. PCR of genomic DNA then allowed 141 FAS-ESS decamer sequences to be derived, 133 of which were unique. Most of the 133 unique FAS-ESS decamers could be clustered into groups by multiple alignment with CLUSTALW, yielding seven putative ESS motifs, three of which resemble known motifs bound by hnRNPs H and A1. The silencer activity of over a
Diagnostics of pathogenic splicing mutations

dozens of these sequences was also experimentally confirmed in a heterologous exon/intron context, and statistical overrepresentation in the set of recovered decamers was used as a criterion to identify 103 specific “FAS-hex3” hexamers occurring at least three times in the decamer set and likely to possess intrinsic ESS activity. In addition, the identified FAS-ESSs seem to play a role in splice site definition and pseudo-exon suppression (82). Most FAS-hex3 hexamers were significantly enriched in pseudo exons and skipped exons relative to constitutive exons, and many also in exons with strong relative to those with weak splice sites. The 103 FAS-hex3 hexamers partially overlapped with the 974 PESS 8-mers (53%).

FAS-ESS were also included into the first generation splicing simulation software EXONSCAN, which integrates signal scoring information assumed to be used by the splicing machinery, but ignores non-local information like sequence composition, conservation or reading frame. EXONSCAN derives a composite score for each potential exon by independently adding up log-odds-score contributions from discrete sequence elements like 5’ and 3’ ss sequences and cis-regulatory elements (G triplets as ISE, RESCUE-ESE and FAS-ESS hexamers). By including different types of ESEs and ESSs into this composite score and calculating exon prediction accuracy, their individual contribution to exon recognition can be estimated. 10,891 internal human exons were screened by EXONSCAN – in different versions, incorporating either splice site recognition alone, or additional enhancer or silencer motif recognition. In this comparison, using FAS-hex3 resulted in the greatest improvement due to a single silencer motif recognition. In this comparison, using FAS-hex3 resulted in the greatest improvement due to a single

Most recently, a new, general method termed Neighborhood Inference (NI) combined the results of the RESCUE-ESE hexamer approach with PESEs and PESSs octamers as well as with FAS-ESSs screened in a fluorescence-based assay (168). From two given disjoint sets of “trusted” cis-regulatory elements – ESEs and ESSs – of length k, the NI approach counts the numbers of ESEs and ESSs in a well-defined neighborhood of the examined k-mer. The k-mer’s normalized NI-score with values between –1 and +1 is calculated from the carefully weighted ESE- and ESS-distribution in neighborhoods of given radius, and reflects, whether there are relatively more trusted ESEs or ESSs at a given sequence space distance from the k-mer. Thus, the NI-score is based upon the two input (training) sets, and consequently k-mers in the ESE (ESS) set have scores +1 (-1). In contrast to PSSM approaches, NI does not require that the set of known sites be of homogeneous origin (same binding protein) or specifically aligned, and multiple potentially overlapping motifs can be modeled simultaneously, including both positively and negatively acting elements.

As input sets, 666 trusted ESE hexamers were extracted from RESCUE-ESEs and PESE octamers, and 386 trusted ESS hexamers were obtained from FAS-ESS and PESS octamers. Sequences that are inactive in splicing are hard to detect reliably, and no “inactive” datasets have been reported to date, so that NI accuracy was assessed exclusively on the sets of trusted ESS and EHE hexamers, with one acting as negative control for the other. Excellent separation of trusted ESSs from ESSs was achieved in 10-fold cross-validation analysis, when both ESSs and ESEs were used as an input.

Hundreds of novel candidate ESE and ESS hexamers were identified at NI score cutoff ± 0.8, where the numbers of misclassifications of trusted ESE/ESSs in the cross-validation analyses were negligible. 153 of 313 NI-predicted ESE hexamers (49%) were obtained in excess of RESCUE-ESEs from updated data sets and using maxent scores, although these “new” RESCUE-ESE hexamers still contained 85% of the original RESCUE-ESEs. A total of 27 NI-predicted ESEs and five NI-predicted ESS were also found in a set of exonic regulatory sequences based on overrepresented and conserved dicodons in orthologous human and mouse exons (105).

Moreover, NI scoring was able to quantitatively predict the magnitudes of effects on splicing point mutations introduced into exons, experimentally measured by change in exon inclusion levels. 24 previously uncharacterized hexamers, evenly distributed across the NI-range [-1,+1] were examined in a splicing reporter assay engineered to detect both splicing enhancement and repression. The correlation between hexamer NI-score and test exon inclusion confirmed that hexamers with NI-score > 0.8 (< -0.8) had enhancing (silencing) effects, while those between –0.8 and +0.8 were mostly splicing-neutral. Future improvements of the NI method may include weighting of individual sequences of known activity according to their abundance in experimentally determined sequence sets, according to their binding affinity, or according to their phylogenetic conservation, as well as more sophisticated k-mer distance measures incorporating e.g. position-specific weights for mismatches.

5.5. Computational identification of intronic regulatory elements

To date, intronic splicing regulatory elements (ISRE) have been identified to a conceivably lesser extent than exonic elements. Studies relating to ISREs exploit two possible strategies: evolutionary conservation of non-coding intronic sequences flanking conserved exons indicating regulatory function, and possible regulation of differential expression of tissue-specific alternatively spliced cassette exons by intronic sequences. The latter assumes that critical regulatory elements are most prominent in introns adjacent to exons with a highly specific regulation pattern. Although it is only recently that they have received attention, ISREs can be considered cornerstones in understanding tissue-specific alternative splicing (182).
Diagnostics of pathogenic splicing mutations

In a comparative analysis of human and mouse genomes, intronic sequences with a high degree of conservation (88% upstream and 80% downstream) were identified proximal to the enclosed exons: 77% of conserved alternatively spliced exons were flanked on both sides by conserved intronic sequences, as opposed to only 17% of constitutive exons. These conserved sequences are hypothesized to be involved in alternative splicing regulation (183).

Cis-regulatory elements were searched for in introns adjacent to 25 brain-specific alternatively spliced cassette exons by word contrast algorithms, comparing oligomer frequencies to introns next to control exons. Statistically overrepresented oligomers were mostly found within 100–300 nucleotides up- and downstream of the cassette exon, confirming the hypothesis that intronic regulatory elements are localized close to the exon. In the downstream intron, the brain tissue overrepresented hexamer UGCAUG, pentamers GCAUG and UGCAU, and the underrepresented GGG triplet have been identified with the highest significance. The upstream introns showed a remarkably similar oligomer distribution for brain and other tissues, with UUUUUU and known PTB consensus binding sites as most significant motifs (184).

A well-known and particularly successful approach in identification of cis-regulatory elements in transcription, correlation of motif occurrence with expression has been used to identify ISREs in tissue-specific alternative splicing. Based upon human exon microarray data, 56 cassette exons with significantly higher normalized expression level in muscle tissue compared to other adult tissue were identified. Hexamers were identified as ISRE candidates by statistical significance of log-linear regression of hexamer frequency in 200 bases long flanking intronic sequences with gene-normalized expression ratios. The most prominent downstream ISREs were Fox1- and CELF-binding sites, and a branchpoint-like element ACUAAC, while pyrimidine-rich elements like PTB-binding sites occurred upstream. These factors were found to act both independently and collaboratively in muscle-specific splicing (185) (http://vision.lbl.gov/People/ddas/NAR_SPLICE1).

A comparative genomics approach across four mammalian species searched for highly conserved words of length 5–7 in 400 bases long intronic sequences flanking conserved exons. ISRE candidates were scored by a statistical χ²-parameter measuring their conservation rate, and subsequently grouped into families containing an average number of five words. This procedure finally identified 158 ISREs in downstream and 156 ISREs in upstream introns, 84% (94%) of which suppressed proximal 5’ (3’) ss in competing splice reporter constructs in human cells. Up to 50% of these ISREs were enriched near alternative exons, in particular near tissue-specific alternative events, and included nearly all binding sites of known alternative splicing factors. A subset of ~40–45% of these ISREs may play a dual role as exonic splicing silencers (182).

6. CONCLUSIONS

Gene mutations disrupting the splicing mechanism are increasingly recognized as having a strong disease causing potential, since accurate pre-mRNA splice site recognition in the nucleus is a mandatory prerequisite for correct cellular function. Splicing regulatory elements are sensitive targets of nucleotide alterations: even single DNA mutations can strengthen, weaken or destroy a splice site or cis-regulatory element, or create a new one, and may thus lead to observable phenomena on RNA level like aberrant splicing (exon skipping, activation of cryptic or de novo splice sites, or intron retention).

Through identification of disease-specific genes, genetic testing has found its way into clinical routine and supports a variety of clinical decisions in many common diseases and cancer syndromes. Most patients, however, are genotyped only, and diagnostic RNA-level information about aberrant splicing is usually not available. Therefore, computational predictions of in vivo DNA mutation effects on splicing can be beneficial for the human geneticist – but presents a major challenge for bioinformatics, due to the complex interplay of splice site-defining sequence elements.

Today, in silico implementation of the comprehensive splicing machinery is still limited to a variety of independent algorithms scoring splice sites and/or cis-regulatory elements. Indeed, these dedicated scores for 5’ ss or 3’ ss, as well as exonic or intronic splice enhancers or silencers, have been applied to the prediction of individual factors with considerable success. However, it has been recognized that splicing regulatory elements act in concert, and their interactions and dependencies play an important role in splice site functionality, but the meaningful combination of cis-regulatory elements and splice site scores into a single functional measure still remains to be achieved.

Nevertheless, currently available web-based tools for the scoring of splice sites and cis-regulatory elements, as described in this review, may provide the human geneticist with valuable information for estimating DNA mutation effects on splicing. Although currently bioinformatics does not yet cover all bases, appropriate combination of the different available algorithms may present the next step towards a reliable diagnostic tool that one day will become an integral part of clinical diagnosis.

7. ACKNOWLEDGEMENTS

We thank Jennifer Ashurst and Eduardo Eyras for their continuous support, providing datasets of human annotated 5’ ss. This work was supported by DFG grant SCH A 909/2-2 (H.S.) and by grants from the Stiftung für AIDS-Forschung, Düsseldorf (H.S.) and the Cancer Society North-Rhine-Westfalia (H.S.; D.N.). L.H. was supported through a Ph.D. scholarship from the Jürgen Manchot Stiftung, Düsseldorf.
8. REFERENCES


Diagnoses of pathogenic splicing mutations

42. Senapathy, P., Shapiro, M. B. & Harris, N. L.: Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183, 252-278 (1990)
Diagnostics of pathogenic splicing mutations

Diagnostics of pathogenic splicing mutations

129. Stanke, M., Tzvetkova, A. & Morgenstern, B.: AUGUSTUS at EGASP: using EST, protein and genomic
Diagnostics of pathogenic splicing mutations


**Key Words:** Pathogenic Splicing Mutation; Aberrant Splicing; Splice Site Strength; Splicing Regulatory Elements; Bioinformatics; Computational Tools; Algorithms; Consensus Sequence, Review

**Send correspondence to:** Dr. Heiner Schaal, Institute for Virology, Heinrich-Heine-University Duesseldorf, University Street 1, Bld. 22.21, D-40225 Duesseldorf, Germany, Tel: 492118112393, Fax: 492118112227, E-mail: schaal@uni-duesseldorf.de

http://www.bioscience.org/current/vol13.htm