[Frontiers in Bioscience 13, 3252-3272, May 1, 2008]

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

Linda Hartmann[1,4], Stephan Theiss[2], Dieter Niederacher[3,4], Heiner Schaal[1,4]

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

TABLE of CONTENTS

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

disease risk, or *variants of unknown significance* (VUS). It is obvious that in particular VUS pose problems for genetic counseling, since tested individuals and their families are given a seemingly ambiguous result. This purely phenomenological characterization has different counterparts on molecular level.

From a *protein coding* viewpoint, sequence variations in the coding region are classified as either *frame-shift*, *nonsense*, *missense* or *synonymous*. Frame-shift or nonsense mutations produce truncated protein isoforms, whereas missense mutations affect amino acids that may be important for structure and function of a protein. Translationally synonymous mutations – allelic polymorphisms or so-called nucleotide variations – are considered to be neutral. However, from a *transcript* viewpoint, translationally neutral DNA alterations might very well affect RNA processing by altering an RNA stability element, or by aberrant splicing due to (in-) activation of a *cis*-regulatory splicing element. Both mutations in the proper splice site consensus sequences and in auxiliary *cis*-regulatory splicing elements are known to disrupt splicing, mostly by exon-skipping or activation of cryptic splice sites, and can thereby change the overall splicing pattern of the mutant transcript and thus its open reading frame (ORF). Ignoring this pre-mRNA splicing pathway presents a major shortcoming of protein-addressing classification of mutations.

Pathogenic splicing alterations are increasingly recognized as a widespread mechanism through which gene mutations cause disease. In the Human Gene Mutation Database (www.hgmd.org, as of 2007-10-01), single base-pair substitutions within exon/intron boundaries of a total of 2,768 human genes constitute ~10% of a total of 73,411 mutations causing human inherited diseases. For some genes the number of known splicing mutations even exceeds or is as high as the number of all other identified mutations.

Pathogenic splicing mutations may escape identification or correct interpretation by genomic DNA based assays, such as high-throughput sequencing or screening approaches, because they may not be distinguishable from neutral splice-site polymorphisms (1, 2). Frequently, RNA based assays to identify splicing defects cannot be used due to practical difficulties in RNA extraction from cell lines or tissues, and since the availability of biopsies is often limited and insufficient for the laboratory (3). Moreover, they often fail to give unambiguous results due to sources of variability in individual patients.

Thus, in assessing a splice site mutation's pathogenicity, reliable *in silico* prediction of its *in vivo* splicing outcome can increase the efficiency of genomic DNA based mutation detection assays and possibly resolve diagnostic dilemmas in patients. Today, several web-based software tools are available to assess the impact on aberrant splicing of splice donor or acceptor sequence alterations. It is the challenge of this review to provide basic knowledge to use these prediction tools, and judge power and limitations of the underlying bioinformatics algorithms.

3. IDENTIFICATION OF PATHOGENIC SPLICING MUTATIONS AND THE DILEMMA OF DIAGNOSIS

In general, routine genetic diagnostics is based on mutation detection techniques using genomic DNA extracted from blood leucocytes as the source of choice. Sometimes, RNA samples are additionally collected from the same patient to confirm the effect of putative splicing mutations. In routine clinical settings RNA/cDNA is not used as a template for mutation screening, because RNA is much less stable than DNA, and technical problems during specimen transportation, RNA isolation or reverse transcription may artificially alter the quantity and distribution of cDNA fragments. In contrast, applying mutation detection techniques to DNA templates yields equal quantities of both alleles providing hetero- or homozygote sequence information. Analysis of RNA/cDNA may be hampered by the occurrence of alternative splicing products or a different distribution of allele transcripts e.g. due to nonsense mediated decay of mRNA eliminating or reducing transcripts containing premature translation termination codons (PTCs) (4).

3.1. Mutation detection techniques

Numerous techniques are available for mutation detection like direct sequencing (DS), protein truncation test (PTT) (5), single-strand conformation polymorphism (SSCP) (6), dideoxy fingerprinting assay (DDF) (7), denaturing gradient gel electrophoresis (DGGE) (8), two-dimensional gene scanning (TDGS) (9), conformation-sensitive gel electrophoresis (CSGE) (10), enzymatic mutation detection (EMD) (11) allele-specific oligonucleotide hybridization (ASO) (12), and immobilized DNA hybridization assays.

The widely used SSCP analysis is economical and simple, but has low sensitivity, ranging from 60 to 90%. However, the major disadvantage of SSCP is that non-appearance of a band-shift does not prove the absence of a mutation. ("*Absence of evidence is not evidence of absence.*") DDF is a combination of a Sanger sequencing reaction with multiple-fragment SSCP and is more sensitive than SSCP alone, but still labor intensive. With PTT only sequence alterations leading to a truncated protein can be detected. Other alterations (missense, in-frame deletions, and insertions) escape the detection of the PTT assay. Other methods, including DGGE and CSGE, have been used with some success, but despite their improved sensitivity over SSCP, these methods are technically rather challenging and formatted for manual use only. TDGS is a method for analyzing multiple DNA fragments in parallel for all possible sequence variations, using extensive multiplex PCR and two-dimensional electrophoretic separation on the basis of size and melting temperature. One source of error is the interpretation of the complex spot patterns produced by this method. Also, the sensitivity of TDGS is impaired by frequent preferential amplification of the non-mutant allele, resulting in very "light" heteroduplices that obscure accurate reading of the gels (13). EMD methods for mutation scanning still lack the sensitivity and specificity of the chemical cleavage of the mismatch method. Other approaches such as

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

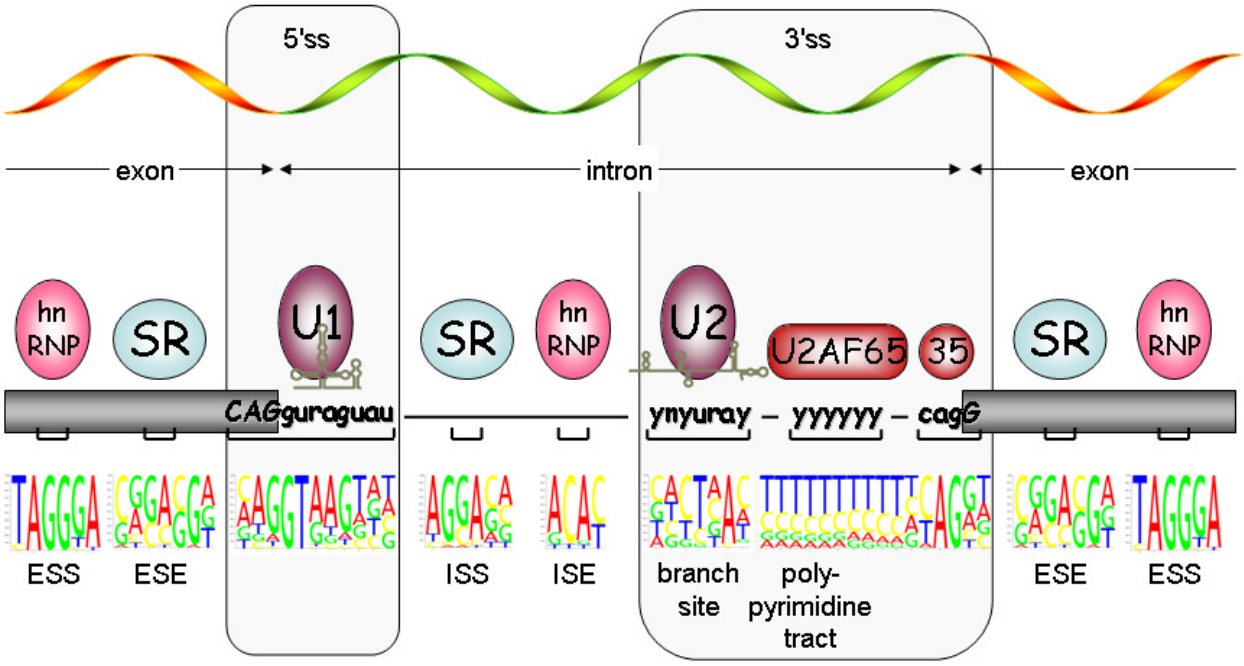

Figure 1. Schematic close-up of exon–intron boundaries of a eukaryotic gene. A spliceable gene (top) consists of exons (orange) and at least one intron (green), which is removed from the pre-mRNA by the splicing machinery. The exon-intron boundary is called the splice donor site or 5' splice site (5' 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

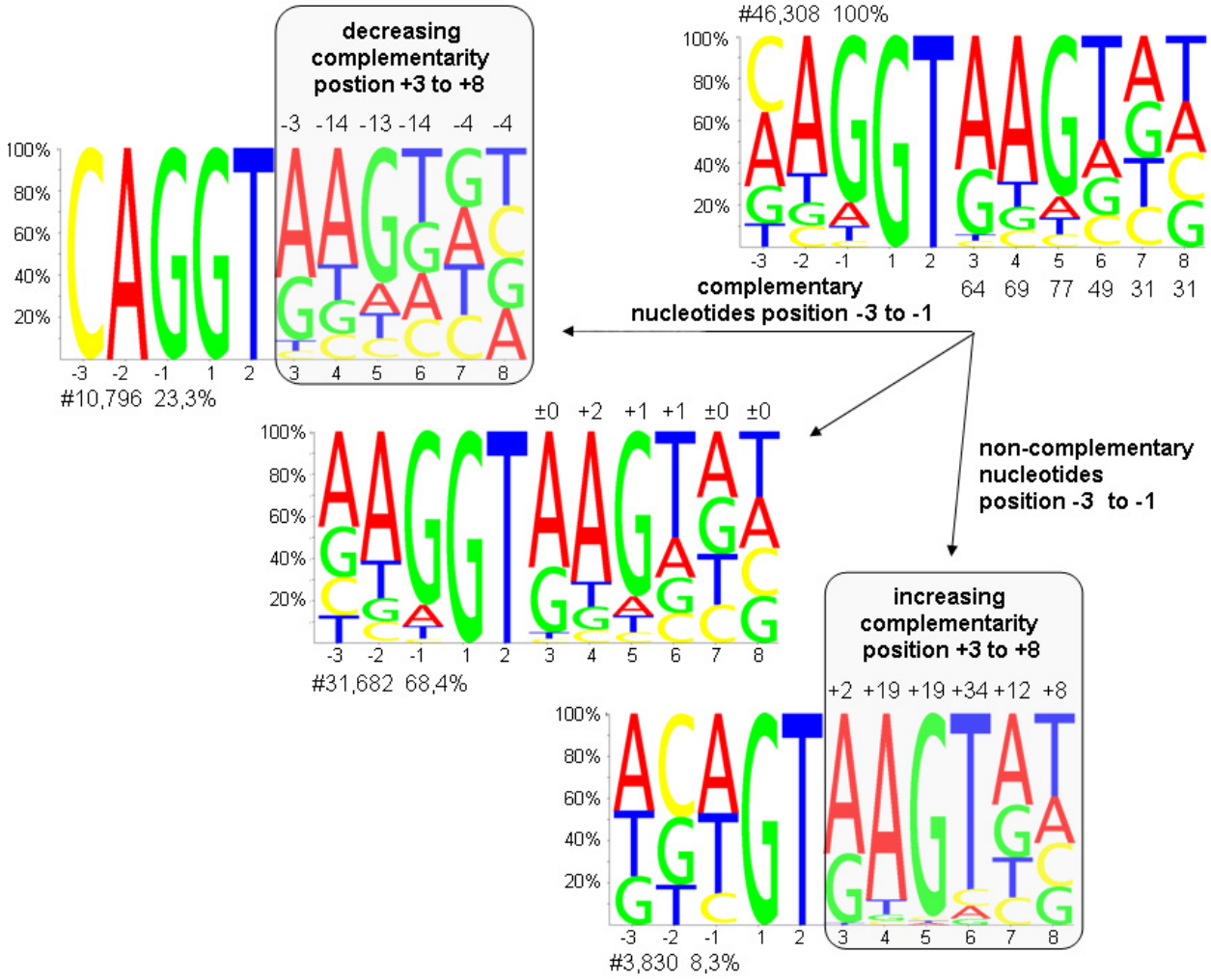

Figure 2. Sequence logos of annotated canonical human 5' ss. At each position the height of a base is proportional to its frequency in that position. Bases are ordered top to bottom in decreasing frequency. A sequence logo of 46,308 5' ss compiled from human chromosomes 6, 7, 9, 10, 13, 14, 20, 22 and 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). The 3' ss itself seems to be recognized in a scanning process for the first AG dinucleotide downstream of the branchpoint/polypyrimidine tract. Interestingly, CAG, UAG and AAG triplets were efficient 3' ss whereas GAG was not used at all (49, 50). This was also shown for 'tandem' (NAGNAG) 3' ss that effectively compete with each other (51). Exceptions of the scanning process occurred, if the AG resides very close to the BPS and then can be bypassed (52, 38).

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 heteromeric 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 snRNPs 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 bound 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 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 hnRNPs 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 promotors, 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

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

Algorithm	Basic principle	Link	Input	Output
Shapiro&Senapathy Score (S&S)	Position-specific weight matrix (PSWM); Coincidence with consensus sequence	http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm	5' ss motif (9-mer) position -3 to +6; 3' ss motif (15-mer) position -14 to +1	S&S Score (0-100)
Weight matrix model (WMM)	Quantification of the relative likelihood of candidate splice site sequence with respect to the background nucleotide distribution from a training set of splicing signals	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	5' ss motif (9-mer) position -3 to +6 ; 3' ss motif (23-mer) position -20 to +3	Numerical Score
Maximum Dependence Decomposition Model (MDD)	Iterative decision-tree approach; captures dependencies between neighboring and non-neighboring positions by WAM (weight array model) and WMM	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	5' ss motif (9-mer) position -3 to +6;	Numerical Score
First-order Markov Model	Statistical approach which considers dependencies between adjacent positions	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	5' ss motif (9-mer) position -3 to +6 ; 3' ss motif (23-mer) position -20 to +3	Numerical Score
Maximum Entropy Model (MEM)	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	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	5' ss motif (9-mer) position -3 to +6; 3' ss motif (23-mer) position -20 to +3	Numerical Score
H-Bond Model	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	http://uni-duesseldorf.de/rna	5' ss motif (11-mer) position -3 to +8;	Numerical Score
Neural Network (NN)	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	long sequence stretches	Score between 0 and 1
Branch Site Tool	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	http://ast.bioinfo.tau.ac.il/BranchSite.htm	long sequence stretches	Numerical Score
ESEfinder	Prediction of SR protein specific putative ESE, based on an *in vitro* SELEX approach dependent on addition of individual SR proteins	http://rulai.cshl.edu/tools/ESE/	long sequence stretches	ESE motif score
RESCUE-ESE	Statistical approach based upon different distribution of hexamers in exons and introns with different properties, e.g. weak and strong splice sites	http://genes.mit.edu/burgelab/rescue-ese	long sequence stretches	Z-score (picks out extremely over- or under-represented hexamers)
PESX server	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)	http://cubweb.biology.columbia.edu/pesx/	long sequence stretches	P-score, I-score (see Z-score)
ESR search	Evolutionary conservation of wobble positions with statistically significant overabundance of dicodons (hexamers) relative to the expected frequency of their independent individual codons	http://ast.bioinfo.tau.ac.il/ESR.htm	long sequence stretches	Z-score

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

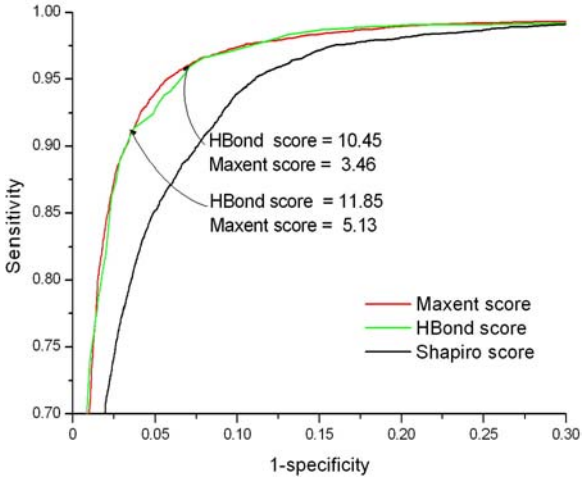

Figure 3. Exemplary receiver-operating characteristic (ROC) curve comparing sensitivity and specificity of three 5' ss scores, maxent (red), HBond (green) and Shapiro-Senapathy (black), on a given data set of *7.839* canonical real and *56.408* exonic decoy 5' ss (not used in alternative transcripts) 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 $C_{max, Y-B} = 0.676$, 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

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 (MaxENTScan algorithm: http://genes.mit.edu/burgelab /maxent/Xmaxentscan_scoreseq.html). 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 intronic 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

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 contribution of individual 5' ss nucleotides to the intrinsic strength of human 5' ss has 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 me2s0-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

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

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 I-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 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

dozen 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 type of regulatory element, suggesting that ESS may be at least as important as enhancer elements. The 103 FAS-hex3 hexamers were found overrepresented by a factor of at least ~1.5 in the 200 intronic compared to the 70 exonic nucleotides next to constitutive 5' and 3' ss, which contributes further evidence for the FAS-ESS' activity in suppression of decoy splice sites upstream and downstream of constitutive exons.

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 ESE hexamers, with one acting as negative control for the other. Excellent separation of trusted ESEs 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 ESEs/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).

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 χ^2-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 SCHA 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

1. Claes, K., Poppe, B., Machackova, E., Coene, I., Foretova, L., De Paepe, A. & Messiaen, L.: Differentiating pathogenic mutations from polymorphic alterations in the splice sites of BRCA1 and BRCA2. *Genes Chromosomes Cancer* 37, 314-320 (2003)
2. Messiaen, L. M. & Wimmer, K.: Pitfalls of automated comparative sequence analysis as a single platform for routine clinical testing for NF1. *J Med Genet* 42, e25 (2005)
3. Baralle, D. & Baralle, M.: Splicing in action: assessing disease causing sequence changes. *J Med Genet* 42, 737-748 (2005)
4. Usuki, F., Yamashita, A., Kashima, I., Higuchi, I., Osame, M. & Ohno, S.: Specific inhibition of nonsense-mediated mRNA decay components, SMG-1 or Upf1, rescues the phenotype of Ullrich disease fibroblasts. *Mol Ther* 14, 351-360 (2006)
5. Hogervorst, F. B., Cornelis, R. S., Bout, M., van Vliet, M., Oosterwijk, J. C., Olmer, R., Bakker, B., Klijn, J. G., Vasen, H. F., Meijers-Heijboer, H. & .: Rapid detection of BRCA1 mutations by the protein truncation test. *Nat Genet* 10, 208-212 (1995)
6. Markoff, A., Savov, A., Vladimirov, V., Bogdanova, N., Kremensky, I. & Ganev, V.: Optimization of single-strand conformation polymorphism analysis in the presence of polyethylene glycol. *Clin Chem* 43, 30-33 (1997)
7. Lancaster, J. M., Berchuck, A., Futreal, P. A. & Wiseman, R. W.: Dideoxy fingerprinting assay for BRCA1 mutation analysis. *Mol Carcinog* 19, 176-179 (1997)
8. Fodde, R. & Losekoot, M.: Mutation detection by denaturing gradient gel electrophoresis (DGGE). *Hum Mutat* 3, 83-94 (1994)
9. van Orsouw, N. J., Dhanda, R. K., Elhaji, Y., Narod, S. A., Li, F. P., Eng, C. & Vijg, J.: A highly accurate, low cost test for BRCA1 mutations. *J Med Genet* 36, 747-753 (1999)
10. Markoff, A., Sormbroen, H., Bogdanova, N., Preisler-Adams, S., Ganev, V., Dworniczak, B. & Horst, J.: Comparison of conformation-sensitive gel electrophoresis and single-strand conformation polymorphism analysis for detection of mutations in the BRCA1 gene using optimized conformation analysis protocols. *Eur J Hum Genet* 6, 145-150 (1998)
11. Taylor, G. R. & Deeble, J.: Enzymatic methods for mutation scanning. *Genet Anal* 14, 181-186 (1999)
12. Richter, S. & Seth, A.: One step direct detection of recurrent mutations in the breast cancer susceptibility gene, BRCA1. *Int J Oncol* 12, 1263-1267 (1998)
13. Eng, C., Brody, L. C., Wagner, T. M., Devilee, P., Vijg, J., Szabo, C., Tavtigian, S. V., Nathanson, K. L., Ostrander, E. & Frank, T. S.: Interpreting epidemiological research: blinded comparison of methods used to estimate the prevalence of inherited mutations in BRCA1. *J Med Genet* 38, 824-833 (2001)
14. Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L. & Oefner, P. J.: Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* 7, 996-1005 (1997)
15. Xiao, W., Stern, D., Jain, M., Huber, C. G. & Oefner, P. J.: Multiplex capillary denaturing high-performance liquid chromatography with laser-induced fluorescence detection. *Biotechniques* 30, 1332-1338 (2001)
16. Xiao, W. & Oefner, P. J.: Denaturing high-performance liquid chromatography: A review. *Hum Mutat* 17, 439-474 (2001)
17. Hoffman, J. D., Hallam, S. E., Venne, V. L., Lyon, E. & Ward, K.: Implications of a novel cryptic splice site in the BRCA1 gene. *Am J Med Genet* 80, 140-144 (1998)
18. Pyne, M. T., Brothman, A. R., Ward, B., Pruss, D., Hendrickson, B. C. & Scholl, T.: The BRCA2 genetic variant IVS7 + 2T-->G is a mutation. *J Hum Genet* 45, 351-357 (2000)
19. Fackenthal, J. D., Cartegni, L., Krainer, A. R. & Olopade, O. I.: BRCA2 T2722R is a deleterious allele that causes exon skipping. *Am J Hum Genet* 71, 625-631 (2002)
20. Robledo, M., Osorio, A., Sentis, C., Albertos, J., Estevez, L. & Benitez, J.: The 12 base pair duplication/insertion alteration could be a regulatory mutation. *J Med Genet* 34, 592-593 (1997)
21. Scholl, T., Pyne, M. T., Ward, B. & Pruss, D.: Biochemical and genetic characterisation shows that the BRCA1 IVS20 insertion is a polymorphism. *J Med Genet* 36, 571-572 (1999)
22. Sahashi, K., Masuda, A., Matsuura, T., Shinmi, J., Zhang, Z., Takeshima, Y., Matsuo, M., Sobue, G. & Ohno, K.: *In vitro* and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. *Nucleic Acids Res* [Epub ahead of print] (2007)
23. Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., de Silva, D., Zharkikh, A. & Thomas, A.: Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 43, 295-305 (2006)
24. Zhuang, Y. & Weiner, A. M.: A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* 46, 827-835 (1986)
25. Kammler, S., Leurs, C., Freund, M., Krummheuer, J., Seidel, K., Tange, T. O., Lund, M. K., Kjems, J., Scheid, A. & Schaal, H.: The sequence complementarity between HIV-1 5' splice site SD4 and U1 snRNA determines the steady-state level of an unstable env pre-mRNA. *RNA* 7, 421-434 (2001)
26. Kirschner, L. S., Carney, J. A., Pack, S. D., Taymans, S. E., Giatzakis, C., Cho, Y. S., Cho-Chung, Y. S. & Stratakis, C. A.: Mutations of the gene encoding the protein kinase A type I-alpha regulatory subunit in patients with the Carney complex. *Nat Genet* 26, 89-92 (2000)
27. Wijk, R., van Wesel, A. C., Thomas, A. A., Rijksen, G. & van Solinge, W. W.: *Ex vivo* analysis of aberrant splicing induced by two donor site mutations in PKLR of a patient with severe pyruvate kinase deficiency. *Br J Haematol* 125, 253-263 (2004)
28. Freund, M., Asang, C., Kammler, S., Konermann, C., Krummheuer, J., Hipp, M., Meyer, I., Gierling, W., Theiss, S., Preuss, T., Schindler, D., Kjems, J. & Schaal, H.: A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res* 31, 6963-6975 (2003)

29. Heinrichs, V., Bach, M., Winkelmann, G. & Luhrmann, R.: U1-specific protein C needed for efficient complex formation of U1 snRNP with a 5' splice site. *Science* 247, 69-72 (1990)
30. Surowy, C. S., van, S., V, Scheib-Wixted, S. M. & Spritz, R. A.: Direct, sequence-specific binding of the human U1-70K ribonucleoprotein antigen protein to loop I of U1 small nuclear RNA. *Mol Cell Biol* 9, 4179-4186 (1989)
31. Nagai, K., Oubridge, C., Jessen, T. H., Li, J. & Evans, P. R.: Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature* 348, 515-520 (1990)
32. Staley, J. P. & Guthrie, C.: Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell* 92, 315-326 (1998)
33. Newman, A. J. & Norman, C.: U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell* 68, 743-754 (1992)
34. Kandels-Lewis, S. & Seraphin, B.: Involvement of U6 snRNA in 5' splice site selection. *Science* 262, 2035-2039 (1993)
35. Lesser, C. F. & Guthrie, C.: Mutations in U6 Snrna That Alter Splice-Site Specificity - Implications for the Active-Site. *Science* 262, 1982-1988 (1993)
36. Wassarman, D. A. & Steitz, J. A.: Interactions of small nuclear RNA's with precursor messenger RNA during *in vitro* splicing. *Science* 257, 1918-1925 (1992)
37. Moore, M. J.: Intron recognition comes of AGe. *Nat Struct Biol* 7, 14-16 (2000)
38. Gooding, C., Clark, F., Wollerton, M. C., Grellscheid, S. N., Groom, H. & Smith, C. W.: A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol* 7, R1 (2006)
39. Helfman, D. M. & Ricci, W. M.: Branch point selection in alternative splicing of tropomyosin pre-mRNAs. *Nucleic Acids Res* 17, 5633-5650 (1989)
40. Reed, R.: The Organization of 3' Splice-Site Sequences in Mammalian Introns. *Genes Dev* 3, 2113-2123 (1989)
41. Coolidge, C. J., Seely, R. J. & Patton, J. G.: Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res* 25, 888-895 (1997)
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)
43. Zamore, P. D. & Green, M. R.: Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl Acad Sci U S A* 86, 9243-9247 (1989)
44. Zamore, P. D., Patton, J. G. & Green, M. R.: Cloning and domain structure of the mammalian splicing factor U2AF. *Nature* 355, 609-614 (1992)
45. Sickmier, E. A., Frato, K. E., Shen, H., Paranawithana, S. R., Green, M. R. & Kielkopf, C. L.: Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Mol Cell* 23, 49-59 (2006)
46. Singh, R., Valcarcel, J. & Green, M. R.: Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science* 268, 1173-1176 (1995)
47. Wu, S., Romfo, C. M., Nilsen, T. W. & Green, M. R.: Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* 402, 832-835 (1999)
48. Smith, C. W., Chu, T. T. & Nadal-Ginard, B.: Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol* 13, 4939-4952 (1993)
49. Smith, C. W. J., Porro, E. B., Patton, J. G. & Nadalginard, B.: Scanning from An Independently Specified Branch Point Defines the 3' Splice Site of Mammalian Introns. *Nature* 342, 243-247 (1989)
50. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G.: The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* 300, 1288-1291 (2003)
51. Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R. & Platzer, M.: Single-Nucleotide Polymorphisms in NAGNAG Acceptors Are Highly Predictive for Variations of Alternative Splicing. *Am J Hum Genet* 78, 291-302 (2006)
52. Chua, K. & Reed, R.: An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Mol Cell Biol* 21, 1509-1514 (2001)
53. Berglund, J. A., Fleming, M. L. & Rosbash, M.: The KH domain of the branchpoint sequence binding protein determines specificity for the pre-mRNA branchpoint sequence. *RNA* 4, 998-1006 (1998)
54. Peled-Zehavi, H., Berglund, J. A., Rosbash, M. & Frankel, A. D.: Recognition of RNA branch point sequences by the KH domain of splicing factor 1 (mammalian branch point binding protein) in a splicing factor complex. *Mol Cell Biol* 21, 5232-5241 (2001)
55. Berglund, J. A., Abovich, N. & Rosbash, M.: A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev* 12, 858-867 (1998)
56. Banerjee, H., Rahn, A., Gawande, B., Guth, S., Valcarcel, J. & Singh, R.: The conserved RNA recognition motif 3 of U2 snRNA auxiliary factor (U2AF(65)) is essential *in vivo* but dispensable for activity *in vitro*. *RNA* 10, 240-253 (2004)
57. Guth, S. & Valcarcel, J.: Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J Biol Chem* 275, 38059-38066 (2000)
58. Merendino, L., Guth, S., Bilbao, D., Martinez, C. & Valcarcel, J.: Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF(35) and the 3 ' splice site AG. *Nature* 402, 838-841 (1999)
59. Zorio, D. A. R. & Blumenthal, T.: Both subunits of U2AF recognize the 3 ' splice site in Caenorhabditis elegans. *Nature* 402, 835-838 (1999)
60. Soares, L. M. M., Zanier, K., Mackereth, C., Sattler, M. & Valcarcel, J.: Intron removal requires proofreading of U2AF/3 ' splice site recognition by DEK. *Science* 312, 1961-1965 (2006)
61. Golas, M. M., Sander, B., Will, C. L., Luhrmann, R. & Stark, H.: Molecular architecture of the multiprotein splicing factor SF3b. *Science* 300, 980-984 (2003)
62. Spadaccini, R., Reidt, U., Dybkov, O., Will, C., Frank, R., Stier, G., Corsini, L., Wahl, M. C., Luhrmann, R. & Sattler, M.: Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in

branch point definition during pre-mRNA splicing. *RNA* 12, 410-425 (2006)
63. Kramer, A., Gruter, P., Groning, K. & Kastner, B.: Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J Cell Biol* 145, 1355-1368 (1999)
64. Gozani, O., Feld, R. & Reed, R.: Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev* 10, 233-243 (1996)
65. Berglund, J. A., Rosbash, M. & Schultz, S. C.: Crystal structure of a model branchpoint-U2 snRNA duplex containing bulged adenosines. *RNA* 7, 682-691 (2001)
66. Query, C. C., Moore, M. J. & Sharp, P. A.: Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. *Genes Dev* 8, 587-597 (1994)
67. Zhuang, Y. & Weiner, A. M.: A compensatory base change in human U2 snRNA can suppress a branch site mutation. *Genes Dev* 3, 1545-1552 (1989)
68. Staley, J. P.: Hanging on to the branch. *Nat Struct Biol* 9, 5-7 (2002)
69. Will, C. L., Schneider, C., MacMillan, A. M., Katopodis, N. F., Neubauer, G., Wilm, M., Luhrmann, R. & Query, C. C.: A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J* 20, 4536-4546 (2001)
70. Sander, B., Golas, M. M., Makarov, E. M., Brahms, H., Kastner, B., Luehrmann, R. & Stark, H.: Organization of core spliceosomal components U5 snRNA loop I and U4/U6 Di-snRNP within U4/U6.U5 tri-snRNP as revealed by electron cryomicroscopy. *Mol Cell* 24, 267-278 (2006)
71. Valadkhan, S. & Manley, J. L.: Splicing-related catalysis by protein-free snRNAs. *Nature* 413, 701-707 (2001)
72. Valadkhan, S. & Manley, J. L.: Intrinsic metal binding by a spliceosomal RNA. *Nat Struct Biol* 9, 498-499 (2002)
73. Valadkhan, S. & Manley, J. L.: Characterization of the catalytic activity of U2 and U6 snRNAs. *RNA* 9, 892-904 (2003)
74. Yean, S. L., Wuenschell, G., Termini, J. & Lin, R. J.: Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature* 408, 881-884 (2000)
75. Jurica, M. S. & Moore, M. J.: Pre-mRNA splicing: Awash in a sea of proteins. *Mol Cell* 12, 5-14 (2003)
76. Nilsen, T. W.: The spliceosome: the most complex macromolecular machine in the cell? *Bioessays* 25, 1147-1149 (2003)
77. Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G. P., Bresolin, N., Giorda, R. & Pozzoli, U.: Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res* 32, 1783-1791 (2004)
78. Sun, H. & Chasin, L. A.: Multiple splicing defects in an intronic false exon. *Mol Cell Biol* 20, 6414-6425 (2000)
79. Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. & Krainer, A. R.: ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* 31, 3568-3571 (2003)
80. Fairbrother, W. G., Yeo, G. W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P. A. & Burge, C. B.: RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* 32, W187-W190 (2004)
81. Selvakumar, M. & Helfman, D. M.: Exonic splicing enhancers contribute to the use of both 3' and 5' splice site usage of rat beta-tropomyosin pre-mRNA. *RNA* 5, 378-394 (1999)
82. Wang, Z., Xiao, X., Van, N. E. & Burge, C. B.: General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* 23, 61-70 (2006)
83. Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M. & Burge, C. B.: Systematic identification and analysis of exonic splicing silencers. *Cell* 119, 831-845 (2004)
84. Carlo, T., Sterner, D. A. & Berget, S. M.: An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. *RNA* 2, 342-353 (1996)
85. Ponthier, J. L., Schluepen, C., Chen, W., Lersch, R. A., Gee, S. L., Hou, V. C., Lo, A. J., Short, S. A., Chasis, J. A., Winkelmann, J. C. & Conboy, J. G.: Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. *J Biol Chem* 281, 12468-12474 (2006)
86. Tange, T. O., Damgaard, C. K., Guth, S., Valcarcel, J. & Kjems, J.: The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. *EMBO J* 20, 5748-5758 (2001)
87. Modafferi, E. F. & Black, D. L.: A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon. *Mol Cell Biol* 17, 6537-6545 (1997)
88. Kashima, T., Rao, N. & Manley, J. L.: An intronic element contributes to splicing repression in spinal muscular atrophy. *Proc Natl Acad Sci U S A* 104, 3426-3431 (2007)
89. Berget, S. M.: Exon Recognition in Vertebrate Splicing. *J Biol Chem* 270, 2411-2414 (1995)
90. Caputi, M., Freund, M., Kammler, S., Asang, C. & Schaal, H.: A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. *J Virol* 78, 6517-6526 (2004)
91. Carlo, T., Sierra, R. & Berget, S. M.: A 5' splice site-proximal enhancer binds SF1 and activates exon bridging of a microexon. *Mol Cell Biol* 20, 3988-3995 (2000)
92. Fu, X. D., Mayeda, A., Maniatis, T. & Krainer, A. R.: General splicing factors SF2 and SC35 have equivalent activities *in vitro*, and both affect alternative 5' and 3' splice site selection. *Proc Natl Acad Sci U S A* 89, 11224-11228 (1992)
93. Liu, H. X., Chew, S. L., Cartegni, L., Zhang, M. Q. & Krainer, A. R.: Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* 20, 1063-1071 (2000)
94. Liu, H. X., Zhang, M. & Krainer, A. R.: Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12, 1998-2012 (1998)
95. Manley, J. L. & Tacke, R.: SR proteins and splicing control. *Genes Dev* 10, 1569-1579 (1996)
96. Zahler, A. M., Neugebauer, K. M., Stolk, J. A. & Roth, M. B.: Human SR proteins and isolation of a cDNA encoding SRp75. *Mol Cell Biol* 13, 4023-4028 (1993)

97. Zahler, A. M., Neugebauer, K. M., Lane, W. S. & Roth, M. B.: Distinct functions of SR proteins in alternative pre-mRNA splicing. *Science* 260, 219-222 (1993)
98. Sun, Q., Mayeda, A., Hampson, R. K., Krainer, A. R. & Rottman, F. M.: General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. *Genes Dev* 7, 2598-2608 (1993)
99. Caputi, M., Mayeda, A., Krainer, A. R. & Zahler, A. M.: hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. *EMBO J* 18, 4060-4067 (1999)
100. Caputi, M. & Zahler, A. M.: SR proteins and hnRNP H regulate the splicing of the HIV-1 tev-specific exon 6D. *EMBO J* 21, 845-855 (2002)
101. Crawford, J. B. & Patton, J. G.: Activation of alpha-tropomyosin exon 2 is regulated by the SR protein 9G8 and heterogeneous nuclear ribonucleoproteins H and F. *Mol Cell Biol* 26, 8791-8802 (2006)
102. Hallay, H., Locker, N., Ayadi, L., Ropers, D., Guittet, E. & Branlant, C.: Biochemical and NMR study on the competition between proteins SC35, SRp40, and heterogeneous nuclear ribonucleoprotein A1 at the HIV-1 Tat exon 2 splicing site. *J Biol Chem* 281, 37159-37174 (2006)
103. House, A. E. & Lynch, K. W.: An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat Struct Mol Biol* 13, 937-944 (2006)
104. Rothrock, C. R., House, A. E. & Lynch, K. W.: HnRNP L represses exon splicing via a regulated exonic splicing silencer. *EMBO J* 24, 2792-2802 (2005)
105. Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T. & Ast, G.: Comparative analysis identifies exonic splicing regulatory sequences-The complex definition of enhancers and silencers. *Mol Cell* 22, 769-781 (2006)
106. Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J. & Darnell, R. B.: An RNA map predicting Nova-dependent splicing regulation. *Nature* 444, 580-586 (2006)
107. Chen, C. D., Kobayashi, R. & Helfman, D. M.: Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev* 13, 593-606 (1999)
108. Dauksaite, V. & Akusjarvi, G.: Human splicing factor ASF/SF2 encodes for a repressor domain required for its inhibitory activity on pre-mRNA splicing. *J Biol Chem* 277, 12579-12586 (2002)
109. Ibrahim, E. C., Schaal, T. D., Hertel, K. J., Reed, R. & Maniatis, T.: Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci U S A* 102, 5002-5007 (2005)
110. Kanopka, A., Muhlemann, O. & Akusjarvi, G.: Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* 381, 535-538 (1996)
111. Schaub, M. C., Lopez, S. R. & Caputi, M.: Members of the Heterogeneous Nuclear Ribonucleoprotein H Family Activate Splicing of an HIV-1 Splicing Substrate by Promoting Formation of ATP-dependent Spliceosomal Complexes. *J Biol Chem* 282, 13617-13626 (2007)
112. Auweter, S. D., Oberstrass, F. C. & Allain, F. H.: Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* 34, 4943-4959 (2006)
113. Braddock, D. T., Louis, J. M., Baber, J. L., Levens, D. & Clore, G. M.: Structure and dynamics of KH domains from FBP bound to single-stranded DNA. *Nature* 415, 1051-1056 (2002)
114. Braddock, D. T., Baber, J. L., Levens, D. & Clore, G. M.: Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: solution structure of a complex between hnRNP KKH3 and single-stranded DNA. *EMBO J* 21, 3476-3485 (2002)
115. Jokan, L., Dong, A. P., Mayeda, A., Krainer, A. R. & Xu, R. M.: Crystallization and preliminary X-ray diffraction studies of UP1, the two-RRM domain of hnRNP A1. *Acta Crystallogr D Biol Crystallogr* 53, 615-618 (1997)
116. Lewis, H. A., Chen, H., Edo, C., Buckanovich, R. J., Yang, Y. Y., Musunuru, K., Zhong, R., Darnell, R. B. & Burley, S. K.: Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure* 7, 191-203 (1999)
117. Chandler, S. D., Mayeda, A., Yeakley, J. M., Krainer, A. R. & Fu, X. D.: RNA splicing specificity determined by the coordinated action of RNA recognition motifs in SR proteins. *Proc Natl Acad Sci U S A* 94, 3596-3601 (1997)
118. Singh, R. & Valcarcel, J.: Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* 12, 645-653 (2005)
119. Kanopka, A., Muhlemann, O., Petersen-Mahrt, S., Estmer, C., Ohrmalm, C. & Akusjarvi, G.: Regulation of adenovirus alternative RNA splicing by dephosphorylation of SR proteins. *Nature* 393, 185-187 (1998)
120. Tacke, R., Chen, Y. & Manley, J. L.: Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proc Natl Acad Sci U S A* 94, 1148-1153 (1997)
121. Graveley, B. R.: Sorting out the complexity of SR protein functions. *RNA* 6, 1197-1211 (2000)
122. Lee, C. G., Zamore, P. D., Green, M. R. & Hurwitz, J.: RNA annealing activity is intrinsically associated with U2AF. *J Biol Chem* 268, 13472-13478 (1993)
123. Shen, H. & Green, M. R.: RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes Dev* 20, 1755-1765 (2006)
124. Eddy, S. R., Mitchison, G. & Durbin, R.: Maximum discrimination hidden Markov models of sequence consensus. *J Comput Biol* 2, 9-23 (1995)
125. Burge, C. & Karlin, S.: Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268, 78-94 (1997)
126. Burge, C. B. & Karlin, S.: Finding the genes in genomic DNA. *Curr Opin Struct Biol* 8, 346-354 (1998)
127. Stanke, M. & Morgenstern, B.: AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33, W465-W467 (2005)
128. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B.: AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 32, W309-W312 (2004)
129. Stanke, M., Tzvetkova, A. & Morgenstern, B.: AUGUSTUS at EGASP: using EST, protein and genomic

alignments for improved gene prediction in the human genome. *Genome Biol* 7 Suppl 1, S11-S18 (2006)
130. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S.: Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7, 62-2006)
131. Stanke, M. & Waack, S.: Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 Suppl 2, ii215-ii225 (2003)
132. Thanaraj, T. A. & Robinson, A. J.: Prediction of exact boundaries of exons. *Brief Bioinform* 1, 343-356 (2000)
133. Buratti, E., Chivers, M., Kralovicova, J., Romano, M., Baralle, M., Krainer, A. R. & Vorechovsky, I.: Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res* 35, 4250-4263 (2007)
134. Vorechovsky, I.: Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res* 34, 4630-4641 (2006)
135. Aalberts, D. P., Daub, E. G. & Dill, J. W.: Quantifying optimal accuracy of local primary sequence bioinformatics methods. *Bioinformatics* 21, 3347-3351 (2005)
136. Shapiro, M. B. & Senapathy, P.: RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* 15, 7155-7174 (1987)
137. Lear, A. L., Eperon, L. P., Wheatley, I. M. & Eperon, I. C.: Hierarchy for 5' splice site preference determined *in vivo*. *J Mol Biol* 211, 103-115 (1990)
138. Salzberg, S. L.: A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput Appl Biosci* 13, 365-376 (1997)
139. Zhang, M. Q. & Marr, T. G.: A weight array method for splicing signal analysis. *Comput Appl Biosci* 9, 499-509 (1993)
140. Yeo, G. & Burge, C. B.: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11, 377-394 (2004)
141. Brunak, S., Engelbrecht, J. & Knudsen, S.: Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res* 18, 4797-4801 (1990)
142. Brunak, S., Engelbrecht, J. & Knudsen, S.: Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* 220, 49-65 (1991)
143. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D.: Improved splice site detection in Genie. *J Comput Biol* 4, 311-323 (1997)
144. Degroeve, S., De Baets, B., Van de, P. Y. & Rouze, P.: Feature subset selection for splice site prediction. *Bioinformatics* 18 Suppl 2, S75-S83 (2002)
145. Saeys, Y., Degroeve, S., Aeyels, D., Van de, P. Y. & Rouze, P.: Fast feature selection using a simple estimation of distribution algorithm: a case study on splice site prediction. *Bioinformatics* 19 Suppl 2, ii179-ii188 (2003)
146. Saeys, Y., Degroeve, S., Aeyels, D., Rouze, P. & Van de, P. Y.: Feature selection for splice site prediction: a new method using EDA-based feature ranking. *BMC Bioinformatics* 5, 64-2004)
147. Zhang, X. H., Heller, K. A., Hefter, I., Leslie, C. S. & Chasin, L. A.: Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res* 13, 2637-2650 (2003)
148. Bi, J., Xia, H., Li, F., Zhang, X. & Li, Y.: The effect of U1 snRNA binding free energy on the selection of 5' splice sites. *Biochem Biophys Res Commun* 333, 64-69 (2005)
149. Eperon, L. P., Estibeiro, J. P. & Eperon, I. C.: The role of nucleotide sequences in splice site selection in eukaryotic pre-messenger RNA. *Nature* 324, 280-282 (1986)
150. Serra, M. J. & Turner, D. H.: Predicting thermodynamic properties of RNA. *Methods Enzymol* 259, 242-261 (1995)
151. Reddy, R., Henning, D. & Busch, H.: Pseudouridine residues in the 5'-terminus of uridine-rich nuclear RNA I (U1 RNA). *Biochem Biophys Res Commun* 98, 1076-1083 (1981)
152. Roca, X., Sachidanandam, R. & Krainer, A. R.: Determinants of the inherent strength of human 5' splice sites. *RNA* 11, 683-698 (2005)
153. Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D. & Ast, G.: Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell* 14, 221-231 (2004)
154. Mathews, D. H. & Turner, D. H.: Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* 317, 191-203 (2002)
155. Bommarito, S., Peyret, N. & SantaLucia, J., Jr.: Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res* 28, 1929-1934 (2000)
156. Hodas, N. O. & Aalberts, D. P.: Efficient computation of optimal oligo-RNA binding. *Nucleic Acids Res* 32, 6636-6642 (2004)
157. Lund, M. & Kjems, J.: Defining a 5' splice site by functional selection in the presence and absence of U1 snRNA 5' end. *RNA* 8, 166-179 (2002)
158. Carmel, I., Tal, S., Vig, I. & Ast, G.: Comparative analysis detects dependencies among the 5' splice-site positions. *RNA* 10, 828-840 (2004)
159. Roca, X., Sachidanandam, R. & Krainer, A. R.: Intrinsic differences between authentic and cryptic 5' splice sites. *Nucleic Acids Res* 31, 6321-6333 (2003)
160. Freund, M., Hicks, M. J., Konermann, C., Otte, M., Hertel, K. J. & Schaal, H.: Extended base pair complementarity between U1 snRNA and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. *Nucleic Acids Res* 33, 5112-5119 (2005)
161. Rogan, P. K. & Schneider, T. D.: Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum Mutat* 6, 74-76 (1995)
162. Rogan, P. K., Faux, B. M. & Schneider, T. D.: Information analysis of human splice site mutations. *Hum Mutat* 12, 153-171 (1998)
163. Kol, G., Lev-Maor, G. & Ast, G.: Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet* 14, 1559-1568 (2005)
164. Burge, C. B., Tuschl, T., & Sharp, P. A.: Splicing of Precursors mRNAs by the Spliceosomes. In: The RNA

World (2nd). Eds: R.F.Gesteland and J.F.Atkins, New York, 525-560 (1999)
165. Norton, P. A.: Polypyrimidine tract sequences direct selection of alternative branch sites and influence protein binding. *Nucleic Acids Res* 22, 3854-3860 (1994)
166. Kralovicova, J., Christensen, M. B. & Vorechovsky, I.: Biased exon/intron distribution of cryptic and de novo 3' splice sites. *Nucleic Acids Res* 33, 4882-4898 (2005)
167. Stormo, G. D.: DNA binding sites: representation and discovery. *Bioinformatics* 16, 16-23 (2000)
168. Stadler, M. B., Shomron, N., Yeo, G. W., Schneider, A., Xiao, X. & Burge, C. B.: Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet* 2, e191 (2006)
169. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262, 208-214 (1993)
170. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D.: Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31, 3497-3500 (2003)
171. Schaal, T. D. & Maniatis, T.: Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol* 19, 1705-1719 (1999)
172. Coulter, L. R., Landree, M. A. & Cooper, T. A.: Identification of a new class of exonic splicing enhancers by *in vivo* selection. *Mol Cell Biol* 17, 2143-2150 (1997)
173. Smith, P. J., Zhang, C., Wang, J., Chew, S. L., Zhang, M. Q. & Krainer, A. R.: An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* 15, 2490-2508 (2006)
174. Liu, H. X., Cartegni, L., Zhang, M. Q. & Krainer, A. R.: A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* 27, 55-58 (2001)
175. Gabut, M., Mine, M., Marsac, C., Brivet, M., Tazi, J. & Soret, J.: The SR protein SC35 is responsible for aberrant splicing of the E1alpha pyruvate dehydrogenase mRNA in a case of mental retardation with lactic acidosis. *Mol Cell Biol* 25, 3286-3294 (2005)
176. Singh, N. N., Singh, R. N. & Androphy, E. J.: Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Research* 35, 371-389 (2007)
177. Singh, R. N.: Unfolding the mystery of alternative splicing through a unique method of *in vivo* selection. *Front Biosci* 12, 3263-3272 (2007)
178. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B.: Predictive identification of exonic splicing enhancers in human genes. *Science* 297, 1007-1013 (2002)
179. Yeo, G., Hoon, S., Venkatesh, B. & Burge, C. B.: Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* 101, 15700-15705 (2004)
180. Zhang, X. H. & Chasin, L. A.: Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 18, 1241-1250 (2004)
181. Zhang, X. H., Leslie, C. S. & Chasin, L. A.: Computational searches for splicing signals. *Methods* 37, 292-305 (2005)
182. Yeo, G. W., Nostrand, E. L. & Liang, T. Y.: Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet* 3, e85 (2007)
183. Sorek, R. & Ast, G.: Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res* 13, 1631-1637 (2003)
184. Brudno, M., Gelfand, M. S., Spengler, S., Zorn, M., Dubchak, I. & Conboy, J. G.: Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res* 29, 2338-2348 (2001)
185. Das, D., Clark, T. A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J. E. & Conboy, J. G.: A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res* 35, 4845-4857 (2007)

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