REVIEW | Genetics inMedicine

In silico tools for splicing defect prediction: a survey from the viewpoint of end users

Xueqiu Jian, MPH[1], Eric Boerwinkle, PhD[1,2] and Xiaoming Liu, PhD[1]

RNA splicing is the process during which introns are excised and exons are spliced. The precise recognition of splicing signals is critical to this process, and mutations affecting splicing comprise a considerable proportion of genetic disease etiology. Analysis of RNA samples from the patient is the most straightforward and reliable method to detect splicing defects. However, currently, the technical limitation prohibits its use in routine clinical practice. In silico tools that predict potential consequences of splicing mutations may be useful in daily diagnostic activities. In this review, we provide medical geneticists with some basic insights into some of the most popular in silico tools for splicing defect prediction, from the viewpoint of end users. Bioinformaticians in relevant areas who are working on huge data sets may also benefit from this review. Specifically, we focus on those tools whose primary goal is to predict the impact of mutations within the 5′ and 3′ splicing consensus regions: the algorithms used by different tools as well as their major advantages and disadvantages are briefly introduced; the formats of their input and output are summarized; and the interpretation, evaluation, and prospection are also discussed.

Genet Med advance online publication 21 November 2013

Key Words: bioinformatics; end user; in silico prediction tool; medical genetics; splicing consensus region; splicing mutation

INTRODUCTION TO PRE-mRNA SPLICING AND MUTATIONS AFFECTING SPLICING

Sixty years ago, the milestone discovery of the double-helix structure of the DNA molecule opened a door for scientists to uncover the secret of life. For long periods after this discovery, it was widely accepted that similar to prokaryotes, the genetic information manifested by proteins in eukaryotes was also carried by continuous DNA sequences. This specious assumption was proven wrong by a comparison between an mRNA sequence of adenovirus and the DNA from which it was transcribed, leading to the discovery of split genes and RNA splicing.[1,2] Generally speaking, DNA sequences coding for proteins (exons) are interrupted by noncoding sequences (introns); both exons and introns are transcribed to pre-mRNAs; before they are translated to proteins, introns are excised and discrete exons are spliced, resulting in mature mRNAs (**Figure 1**). Based on this new discovery, the molecular basis of RNA splicing was gradually revealed.

The completion and regulation of splicing lean on the complicated biochemical reactions between the nucleotide sequences (*cis*-acting elements) and different proteins binding to them (*trans*-acting elements). *Cis*-acting elements contain the 5′ splice site (junction between an exon and an intron), the 3′ splice site (junction between an intron and an exon), the branch point (tens of nucleotides upstream of the 3′ splice site), exonic splicing enhancers (ESEs), intronic splicing enhancers, exonic splicing silencers, and intronic splicing silencers. *Trans*-acting elements include the spliceosome that is made up of five small nuclear ribonucleoproteins and more than 150 proteins, serine/arginine-rich (SR) proteins, heterogeneous nuclear ribonucleoproteins, and the regulatory complex (**Figure 1**). During this process, the key step is to localize the exon–intron boundaries by capturing the splicing signals embedded in the pre-mRNA sequence by the spliceosome. Extensive comparisons of sequences at different exon–intron boundaries suggested not only the presence of almost invariant GT-AG sites (the respective first and last two sites of an intron) but also weaker conservation in the vicinity of these boundaries, named 5′ and 3′ splicing consensus sequences, respectively, which function as key splicing signals.[3] However, the so-called consensus sequence does not yet have a consensus definition. For example, one study used more than 1,400 5′ and 3′ splice sites from a variety of organisms to derive the consensus sequence from positions −3 to +6 at the 5′ splice site and from positions −14 to +1 at the 3′ splice site,[4] whereas another study used an alignment of conserved sequences from 1,683 human introns to yield the 5′ consensus sequence from positions −3 to +8 and the 3′ consensus sequence from positions −12 to +2.[5] Throughout the paper, we will loosely refer to the "splicing consensus region" as a few to tens of nucleotides in the vicinity of a 5′ or 3′ splice site.

For a certain gene, the final product of splicing may vary in different conditions as a result of alternative splicing that produces different protein sequences without deleterious effects on its functions. The consequence of alternative splicing can be the skipping of an exon (exon skipping), use of different 5′ or 3′ splice sites (alternative 5′ or 3′ splice sites, respectively),

[1]Division of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA; [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. Correspondence: Xiaoming Liu (Xiaoming.Liu@uth.tmc.edu)

Submitted 23 August 2013; accepted 9 October 2013; advance online publication 21 November 2013. doi:10.1038/gim.2013.176

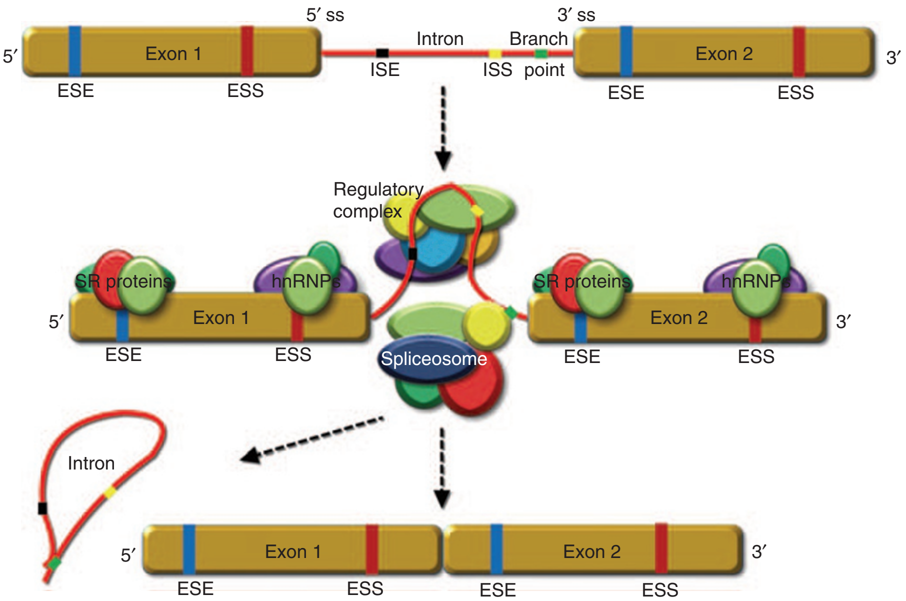

Figure 1 Schematic illustration of pre-mRNA splicing. 5′ Splice site and 3′ splice site are recognized by the spliceosome, and the intron is excised, and exons are spliced. The whole process is regulated by *trans*-acting elements such as SR proteins, heterogeneous nuclear ribonucleoproteins, and the regulatory complex. ESE, exonic splicing enhancer; ESS, exonic splicing silencer; ISE, intronic splicing enhancer; ISS, intronic splicing silencer; ss, splice site.

retaining one of the two exons but not both (mutually exclusive exons), or the retention of an intron (intron retention).[6] Alternative splicing diversifies gene expression (e.g., in different tissues or in different developmental stages) and is very common in human genes. Mutations at splice sites may also modify patterns of splicing in a deleterious way. For instance, a mutated splice site may disrupt an authentic exon–intron boundary and thus change the binding site of the spliceosome, which results in an aberrant splicing. For example, a G to T substitution at position 1 in intron 25 of the *DFNA1* gene can disrupt the canonical splice donor sequence and lead to a four-base insertion in the transcript, which further results in a frameshift and a premature termination that truncates 32 amino acids of the protein. This splicing mutation has been found to cause nonsyndromic deafness in humans.[7] Another example is the fact that a C to T point mutation at position 6 in exon 7 of the *SMN2* gene in individuals who already have deletions of the *SMN1* gene does not change the codon; instead, 80% of the time it inactivates an ESE and creates an exonic splicing silencer. This leads to exon 7 skipping and a truncated protein, thereby causing spinal muscular atrophy.[8–10] In addition to disrupting the primary linear sequence at splice sites, mutations may also have impact on other aspects of splicing, e.g., modification of the secondary structure of the region that hinders the binding of *trans*-acting elements.[11] Besides the causal role as shown in the above examples, splicing mutations can also act as a modifier of disease susceptibility and severity, which has been extensively reviewed elsewhere.[12]

Mutations affecting RNA splicing are not negligible in the population. For example, the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/ac/index.php) collects known human mutations responsible for inherited diseases. As of its Professional Version 2013.2, a total of 13,030 (9.2%) out of 141,161 disease-causing mutations have consequences for mRNA splicing. A widely cited paper estimated that among all human genetic diseases caused by point mutations, up to 15% are the results of splicing defects.[13] However, this estimate seems still conservative because mutations in coding regions are usually considered as missense, nonsense, or silent, which may have resulted in misclassification of splicing mutations and underestimation of the number of splicing mutations.[6] Recently, Lim et al.[14] and Sterne-Weiler et al.[15] provided similar estimates of the proportion of variants within exons that affect splicing but were originally classified as missense (missense or nonsense by Sterne-Weiler et al.[15]) mutations in the Human Gene Mutation Database using independent methods (22 and 25%, respectively). These statistics indicate that mutations affecting splicing comprise a considerable proportion of genetic disease etiology.

A PROBLEM OF SPLICING MUTATION DETECTION IN MEDICAL GENETICS

Disease diagnosis and treatment owe a great deal to our understanding of disease etiology and relevant laboratory techniques. With the importance of mutations affecting splicing being unraveled, their potential role in genetic diseases is increasingly attracting the attention of medical geneticists in their clinical practice. Analysis of RNA from the patient is the most straightforward and reliable method to detect splicing defects. Some other widely used laboratory techniques include in vitro splicing assay and minigene splicing assay.[16] However, our current knowledge of splicing is yet to be implemented in clinical practice on a routine basis due to RNA sample availability (especially specific tissue samples) and limitations in the use of these laboratory techniques.[16] Clinical genetic testing still relies largely on the DNA extracted from blood samples. Moreover, searching for particular splicing variants responsible for particular diseases in the genome may be akin to looking for a needle in a haystack. Since laboratory testing for all splicing variants is expensive and time consuming, medical geneticists are seeking a more economical and quicker way of screening thousands of variants without losing much accuracy so that limited medical resources can be used to serve as many patients as possible. One alternative is to use in silico prediction tools to filter out those variants with little odds of being deleterious and thus to narrow down the search to fewer candidate variants for further experimental validation. After decades of efforts, a number of in silico prediction tools have been developed to assess the effect of DNA sequence variations on splicing. Even so, medical geneticists may be uncertain as to which of the many prediction tools to choose when they have their patients' DNA sequences in hand. Most of the tools were initially designed and developed primarily for research purposes, making them much less useful in clinical practice. Therefore, in this review, we try to provide medical geneticists with some basic insights into some of the most popular in silico tools for splicing defect prediction. Although currently available prediction tools can cover almost all *cis*-acting elements, e.g., ESEfinder, a program that identifies putative ESEs responsive to SR proteins,[17] has successfully predicted the loss of a putative ESE motif in the *SMN2* gene in the previous example,[10] we restrict our review to those with the primary goal of predicting the impact of mutations within the 5′ and 3′ consensus regions for the following reasons: (i) the consensus regions are the prominent *cis*-acting elements; (ii) they have been understood and modeled much better than other elements; and (iii) prediction tools for mutations within consensus regions are better developed with more potential to be utilized in medical genetics. We focus on the application aspect of these tools and employ a user-oriented way to organize the logic of the text. This review may also be useful for bioinformaticians in relevant areas who are working on huge data sets such as whole-genome sequencing data. We anticipate that the information presented here will produce an intuitive picture of current progress in this field, from which readers may benefit when using these tools in their daily practice.

OVERVIEW OF IN SILICO PREDICTION TOOLS FOR 5′ AND 3′ SPLICE SITE MUTATION

The main purpose for using splice site prediction tools has shifted from the identification of possible exon–intron boundaries before the Human Genome Project was completed in 2003 to the prediction of the transcriptional impact of mutations at known splice sites and their vicinity regions in the post-Human Genome Project era. This transition reflects the need to understand human variation on splicing and its effect on human diseases, which is of most interest for medical geneticists. From the viewpoint of end users, the first interface presented when most of the tools are opened will be the input page, which asks the user to type or load the data they want to predict. Most tools require the input of one or more sequences with or without specifying exon–intron boundaries. In the former condition, users need to fix the length of the sequence, whereas in the latter condition, the computer program automatically searches for potential splice sites through the whole length of the input sequence. The shared feature of both formats is to provide the tool with a sequence around the splice site (either manually or automatically by the program), indicating that the prediction of a given mutation relies only on the sequence context itself, regardless of which tool is chosen. In fact, the major differences among tools are the consensus sequences that they used for the comparison with the input sequences, the statistical models applied to this comparison, and the training methods implemented in machine-learning approaches, which will be introduced later in this section. Although a number of in silico tools have been developed, the ideas behind them are not so diverse. Tools with the same backbone mainly differ in the extent to which the local sequence context is taken into account. Oftentimes, a new tool was introduced when certain components of the algorithm it stemmed from were improved. Although medical geneticists usually have more concern about the application aspect of these tools, which will be discussed later, a brief description of the principles is helpful for their understanding in the advantages and disadvantages of different tools.

The basic position weight matrix (PWM) model proposed by Shapiro and Senapathy[4] is to score and rank a sequence using appropriate weights for each nucleotide position based on the information from its aligned consensus sequence (**Table 1**), and it was used by the web interface Splice-Site Analyzer Tool (http://ibis.tau.ac.il/ssat/SpliceSiteFrame.htm). The PWM model is simple, easy to understand, and widely used for representing different patterns of sequences; however, it is overly simplified as it assumes independency (or no correlation) among all positions. That is, a PWM score of a sequence is the summation of position-specific scores for each of its bases (A, T, C, and G), and change of one score at a position has no impact on calculating the score at other positions. SpliceView (http://zeus2.itb.cnr.it/~webgene/wwwspliceview.html) improved the PWM

Table 1 A hypothetical example of a position weight matrix

	Site					
Nucleotide	**1**	**2**	**3**	**4**	**5**	**6**
A	0.73	0.17	0.00	0.00	0.05	0.62
T	0.05	0.26	0.00	1.00	0.49	0.00
C	0.15	0.57	0.00	0.00	0.16	0.22
G	0.07	0.00	1.00	0.00	0.30	0.16

For each site, the frequencies of different nucleotides observed in a set of aligned sequences are calculated to construct the position weight matrix (PWM), which is used to score and rank a sequence. For example, a sequence ACGTTA is most likely to be observed in the population and has the highest score, while a sequence TGACAT is one of the most unlikely and has the lowest score. The formula used to calculate the score varies between 5′ and 3′ splice sites and between different PWM algorithms.

model by considering mutual dependency between nucleotides in different positions.[18] A more general probabilistic model called the maximal dependence decomposition model, which is a decision tree method, captures potential strong dependencies between signal positions (adjacent and nonadjacent) by dividing the data set into subsets based on pairwise dependency between positions and modeling each subset separately.[19] The maximal dependence decomposition model was incorporated in the computer program GENSCAN (http://genes.mit.edu/GENSCAN.html). Pertea et al.[20] further enhanced the maximal dependence decomposition model by adding Markov models, which capture additional dependencies among adjacent positions. The source code for this method, called GeneSplicer, is downloadable at http://ccb.jhu.edu/software/genesplicer/.

In the previous examples, the features used for distinguishing true splice sites from decoy ones are selected by hand, e.g., using appropriate weights in the PWM model, which might not be optimized and introduce bias. To overcome this problem, machine-learning techniques such as artificial neural networks have been applied to the classification of splice sites. By training on the true-positive and true-negative data sets, a neural network automatically optimizes a criterion (e.g., a hyperplane) that separates the two classes. For instance, NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/) was developed using neural networks in which the threshold is also controlled by the exon signal,[21] and NNSplice (http://www.fruitfly.org/seq_tools/splice.html) was trained only on examples with consensus splice sites, and it also accounts for strong correlations in neighboring positions.[22] The support vector machine is another type of machine-learning technique. SplicePort (http://spliceport.cbcb.umd.edu/) used the feature generation algorithm that automatically identifies sequence-based features important for sequence classification as input for the support vector machine.[23] Although the machine-learning approach is highly automatic, the drawback is as obvious as its advantage, which is overfitting. If a classifier fits the training data "too well," e.g., too many parameters relative to the number of examples, the generalizability will likely be poor. That means, when using an overfitted neural network or support vector machine to predict unknown splice sites, the optimized criterion might not be appropriate any longer. One of the common ways to minimize overfitting is to use Bayesian models. One attempt by Brendel et al.[24] used three variables for splice site prediction, and the model was implemented by the web server SplicePredictor (http://bioservices.usd.edu/splicepredictor/).

To date, the most unbiased approximation for modeling short sequence motifs is to use the maximum entropy distribution (MED). Compared with other methods, the only assumption of MED is the consistency with the features of the empirical distribution estimated from available data.[25] MED also considers dependencies between both nonadjacent and adjacent positions. Rather than a single model, MED is a framework with much flexibility for generating different models by simply changing the sets of constraints. The approach has been utilized by the tool MaxEntScan (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html). Users can either use the default models or build their own. The model has been successfully applied to the prediction of splicing mutations in the *ATM* gene responsible for the neurological disorder ataxia telangiectasia, in which three apparently nonsense, missense, or silent exonic mutations were correctly interpreted as disrupting normal splice sites and creating new ones by using MED that had been confirmed by cDNA analysis.[26] MaxEntScan can also output results using other algorithms such as the PWM, maximal dependence decomposition model, and Markov models for easy comparison.

In addition to the approaches described above, various other methods used for splicing defect prediction have been proposed. Examples with user-friendly web interfaces include HBond (http://www.uni-duesseldorf.de/rna/html/hbond_score.php): hydrogen bond model describing the interaction of U1 small nuclear RNA and its binding sites;[27] automated splice site analyses (http://splice.uwo.ca/, free registration required): information theory-based models by which changes in the affinity of potential splice and regulatory sites caused by mutation are calculated;[28] CRYP-SKIP (http://cryp-skip.img.cas.cz/): multiple logistic regression model which distinguishes exons that are skipped and that activate cryptic splice sites as a result of splicing mutations;[29] and Spliceman (http://fairbrother.biomed.brown.edu/spliceman/index.cgi): prediction of how likely distant mutations around annotated splice sites disrupt splicing by clustering hexamers into distinct groups based on positional distributions.[30]

Some tools also incorporate multiple algorithms for the sake of user convenience. Human Splicing Finder (http://www.umd.be/HSF/) outputs splicing defect predictions based on the PWM and MED models as well as the predictions of branch points, ESEs, and exonic splicing silencers.[31] SROOGLE (http://sroogle.tau.ac.il/) is a comprehensive platform that combines nine different prediction algorithms to score four main splicing signals, in which 5′ and 3′ splice sites are predicted by both the PWM and MED models.[32] Automatic Analysis of SNP sites (AASsites; http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar/) is a new analysis pipeline that predicts splicing pattern change caused by single-nucleotide polymorphisms using outputs from five gene prediction programs.[33]

Information about the input and output of these tools is listed in **Table 2**. Unfortunately, at least one sequence is required as the input for almost all tools, thereby making the application of these tools in clinical practice much less convenient. One exception is automated splice site analyses, which has the option not to input sequence information because automated splice site analyses can also localize the variant based on the user-provided gene name, mRNA accession number, or dbSNP rs number. This may be more useful for medical geneticists, as it is better to have a simple format of input that prevents their focus from being distracted by technical concerns. Another common drawback of these tools is their limitation on the length of the sequence being analyzed (**Table 2**).

Table 2 Summary of input, output, and interpretation of prediction scores for selected currently available in silico tools for 5′ and 3′ splice site prediction with user-friendly web interface

Tool	Input	Output	Interpretation
Splice-Site Analyzer Tool	Single/multiple sequences (5′: 9 bp (–3 to +6); 3′: 15 bp (–14 to +1))	S & S score (0–100)	Higher score implies a more similar ss sequence with the consensus sequence
NetGene2	Single sequence (200 bp < length < 80,000 bp)	Confidence score (0–1)	Higher score implies a higher confidence of true site
NNSplice	Single/multiple sequences	Score (0–1)	Higher score implies greater potential for splice site
GENSCAN	Single sequence ≤1 million bp	Probability score (0–1)	Higher score implies a higher probability of correct exon
SpliceView	Single sequence ≤31,000 bp	S & S score (0–100)	Higher score implies a more similar ss sequence with the consensus sequence
Hbond	Single/multiple 11 bp sequences (–3 to +8) containing GT in +1/+2 or one genomic sequence	Hbond score	Higher score implies a stronger capability of forming H-bonds with U1 small nuclear RNA
MaxEntScan	Single/multiple sequences (5′: 9 bp (–3 to +6); 3′: 23 bp (–20 to +3))	Maximum entropy score (log odds ratio)	Higher score implies a higher probability of the sequence being a true splice site
SplicePredictor	Single/multiple sequences	*-Value (3–15) determined by *P*, ρ, and γ values	Higher value implies greater reliability of the predicted splice site
Automated splice site analyses	Mutation to be analyzed and the reference sequence	Information contents Ri	Color coded by direction and type of change in Ri
SplicePort	Single/multiple sequences ≤30,000 bp	Feature generation algorithm score	Higher score implies a more precise prediction of splice site
Human Splicing Finder	Single sequence ≤5,000 bp	S & S score (0–100)	Higher score implies greater potential for splice site
CRYP-SKIP	Single/multiple sequences ≤4,000 bp containing one exon in upper case and flanking intronic sequence ≥4 bp in lower case	Probability of cryptic ss activation (0–1)	Higher value implies a higher probability of cryptic ss activation as opposed to exon skipping
SROOGLE	Target exon along with two flanking introns	Different scores with their percentile scores (0–1)	Higher percentile score implies a higher ranking of the ss within precalculated distributions
AASsites	Single sequence containing the SNP(s) and the Ensembl gene ID to which the SNP(s) belong(s)	Classification of the probability for a change in splicing	Probable, likely, or unlikely
Spliceman	Single/multiple sequences with one mutation and ≥5 bp in each side of the mutation	L1 distance and percentile rank	Higher percentile rank implies a higher likelihood the point mutation is to disrupt splicing

SNP, single-nucleotide polymorphism; ss, splice site.

INTERPRETATION, EVALUATION, AND PROSPECTION

A simple, clear, but informative, interpretation of the output of prediction tools is extremely important for their application in clinical practice. Most tools output a score as a numerical measure of the strength of the splicing signal. Although the range varies, a higher score always indicates a higher degree of similarity to the consensus sequence or a higher probability or confidence of a site being a true splice site. A common misinterpretation of the score by end users is to treat the score as a measure of the effect size. Since the score is a reflection of how likely the variant is deleterious, it is by no means appropriate to consider a variant with a lower score as more deleterious than that with a higher score. Furthermore, the score itself is meaningless because there is no recognized threshold distinguishing positive sites from negative ones. This might be partially due to the fact that other factors besides splicing signals have an impact on splicing. A common way to interpret the scores and facilitate the comparison between different methods is to use score variation by comparing the mutant score with the reference score.[34,35] Users should use a criterion, usually a cutoff value, to determine whether the mutation is causing splicing defects. However, setting this value is usually arbitrary across different tools in different studies. Since the choice of the threshold might not be optimized, the apparently poor performance of a tool is probably due to human errors rather than the algorithm itself, and this will lead to the incomparability of different tools, thus impeding the development of interpretation guidelines.

Lack of interpretation guidelines for splicing defect prediction is also attributable to the small-scale nature of published studies (**Table 3**). As a recent example, Houdayer et al.[35] systematically evaluated several in silico prediction tools using 272 variants of unknown significance in *BRCA1* and *BRCA2*

Table 3 Selected recent publications whose primary goal (or one of the goals) was to evaluate in silico tools for splicing defect prediction

Number of variants	Gene(s)	Prediction tools evaluated	Year (reference)
39	*RB1*	NNSplice, PWM, MaxEntScan, ASSA, ESEfinder, RESCUE-ESE[a]	2008 (ref. 38)
18	*LDLR*	MaxEntScan, NNSplice, NetGene2	2009 (ref. 39)
29	*BRCA1/BRCA2*	NNSplice, NetGene2, PWM, ASSA, MaxEntScan, HSF	2009 (ref. 40)
623	Multiple	GENSCAN, GeneSplicer, HSF, MaxEntScan, NNSplice, SplicePort, SplicePredictor, SpliceView, SROOGLE	2010 (ref. 34)
53	*BRCA1/BRCA2*	PWM, GeneSplicer, NNSplice, MaxEntScan, HSF	2011 (ref. 41)
272	*BRCA1/BRCA2*	NNSplice, PWM, MaxEntScan, ESEfinder, RESCUE-ESE, HSF	2012 (ref. 35)
24	*BRCA1/BRCA2*	PWM, MaxEntScan, NNSplice, GeneSplicer, HSF, NetGene2, SpliceView, SplicePredictor, ASSA	2013 (ref. 42)

[a]ESEfinder and RESCUE-ESE are web tools that predict ESEs.

ASSA, automated splice site analyses; ESE, exonic splicing enhancer; HSF, Human Splicing Finder; PWM, position weight matrix.

genes. These variants were analyzed in vitro and in silico; the receiver operating characteristic curve was used to identify the optimized cutoff value for each tool and to compare their predictive performance for variants in 5′ and 3′ consensus regions (excluding GT-AG sites because all mutations at GT-AG sites affect splicing and were successfully predicted by all tools). They found that the combination of MaxEntScan with a 15% cutoff value and the PWM model with a 5% cutoff value led to an optimized sensitivity of 96% and specificity of 83%. Although the number of variants investigated is still relatively small and only from two genes, we consider it an encouraging step in the right direction. From this study, more interesting findings other than the result itself are the opportunities it provides for improvement. (i) The currently available prediction tools perform perfectly for mutations at invariant GT-AT sites. The real difficulty is to predict their vicinity regions (consensus sequences) and more distant sites. (ii) If the consensus sequence has a higher score, the prediction is more reliable. Ideally, a good algorithm should give the maximum score to the consensus sequence, but it relies on accurate and representative population sequence information to build consensus sequences. The high-throughput next-generation sequencing technologies provide amazing tools to rapidly resequence the whole genome and transcriptome; thus, a better definition of splicing consensus sequence is possible. (iii) The guidelines proposed based on a single study might only apply to a specific data set. The generalizability to other genes and populations is still unknown. As more and more whole-genome and transcriptome sequences are available, real large-scale splicing analyses are expected, which are not limited to certain genes in certain populations. Guidelines for splicing defect prediction based on more general data will be more reliable and generic. (iv) From the epidemiological point of view, all existing evaluations are retrospective, which inevitably suffer from a series of selection biases. Though expensive and time consuming, establishing large cohorts is still preferred to avoid the impact of these biases introduced in a retrospective study that cannot be fully controlled by any analytical method. For example, to associate a disease with its possible causal mutations, investigators often choose to retrospectively compare the nucleotide difference between the cases and the controls because the comparison can be accomplished quickly and economically. However, it might be more convincing to use a random cohort sequenced at baseline (exposure) and observe whether the disease emerges (outcome) prospectively to eliminate the impact of nonrandom selection of cases and controls. This can probably be achieved by using resequencing DNA samples collected at baseline from existing well-established cohorts, such as the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium.[36] Besides this, recently, the National Institutes of Health funded a 5-year research program that will explore the use of genomic sequencing in newborn screening.[37] This provides the opportunity to establish new large cohorts from the very beginning of life and has the potential for studying germline mutations and Mendelian diseases, especially for those with early age of onset.

Besides a standard interpretation guideline, ease of use is another important concern for medical geneticists. As previously mentioned, almost all currently available web tools are not convenient to use in clinical practice. In addition, computational efficiency determines the waiting time of end users and whether the tool has a local standalone version influences its usefulness when end users encounter internet outage. A commercial software package called Alamut (Interactive Biosoftware, Rouen, France) integrates multiple reliable, regularly updated data sources and multiple prediction algorithms (for splicing signal detection, PWM, MaxEntScan, NNSplice, GeneSplicer, and Human Splicing Finder are included). By entering only the variant and specifying its coordinates, users can easily obtain all results at the same time without worrying about the sequence context. This should be the future of in silico prediction tools, and it is expected that more and more such software with user-friendly interfaces will be developed and launched. For bioinformaticians who usually have large quantities of variants to annotate and predict, a tool that can conduct "batch" analysis is preferable (e.g., NetGene2, GENSCAN, GeneSplicer, MaxEntScan, and SplicePredictor have this option). The high-throughput version of the Alamut software, Alamut-HT, can handle ×1,000 variants using its server option (Windows and Linux) or standalone option (Linux only).

In summary, in silico tools for splicing defect prediction (especially for 5′ and 3′ splice sites) have potential value in disease diagnosis in view of the infeasibility of laboratory testing of large number of variants in daily clinical practice. There seems to be no simpler way other than relying on the currently available prediction algorithms until we have a more in-depth understanding of splicing mechanism. Reliable and straightforward interpretation guidelines for the results and an easy-to-use interface will accelerate the popularization of in silico tools among medical geneticists.

ACKNOWLEDGMENTS

This work was supported by the grants Genomic Approaches to Common Chronic Disease (National Institute of General Medical Sciences/P50GM065509) and Building on GWAS for NHLBI-disease: the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) consortium (National Heart, Lung, and Blood Institute/RC2-HL102419). We thank Huiqi Qu for his careful review and valuable comments on the manuscript. We also thank Sara Barton for her help on polishing the writing.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 1977;74:3171–3175.
2. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 1977;12:1–8.
3. Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem* 1981;50:349–383.
4. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* 1987;15:7155–7174.
5. Burge CB, Tuschl T, Sharp PA. Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland RF, Cech TR, Atkins JF (eds). *The RNA World*. 2nd edn. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, 1999:525–560.
6. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002;3:285–298.
7. Lynch ED, Lee MK, Morrow JE, Welcsh PL, León PE, King MC. Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the Drosophila gene diaphanous. *Science* 1997;278:1315–1318.
8. Lefebvre S, Bürglen L, Reboullet S, et al. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 1995;80:155–165.
9. Cartegni L, Krainer AR. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet* 2002;30:377–384.
10. Cartegni L, Hastings ML, Calarco JA, de Stanchina E, Krainer AR. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am J Hum Genet* 2006;78:63–77.
11. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 2007;8:749–761.
12. Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell* 2009;136:777–793.
13. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* 1992;90:41–54.
14. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci USA* 2011;108:11093–11098.
15. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res* 2011;21:1563–1571.
16. Baralle D, Lucassen A, Buratti E. Missed threads. The impact of pre-mRNA splicing defects on clinical practice. *EMBO Rep* 2009;10:810–816.
17. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* 2003;31:3568–3571.
18. Rogozin IB, Milanesi L. Analysis of donor splice sites in different eukaryotic organisms. *J Mol Evol* 1997;45:50–59.
19. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997;268:78–94.
20. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* 2001;29:1185–1190.
21. Brunak S, Engelbrecht J, Knudsen S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* 1991;220:49–65.
22. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol* 1997;4:311–323.
23. Dogan RI, Getoor L, Wilbur WJ, Mount SM. SplicePort–an interactive splice-site analysis tool. *Nucleic Acids Res* 2007;35:W285–W291.
24. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics* 2004;20:1157–1169.
25. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004;11:377–394.
26. Eng L, Coutinho G, Nahas S, et al. Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. *Hum Mutat* 2004;23:67–76.
27. Freund M, Asang C, Kammler S, et al. A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res* 2003;31:6963–6975.
28. Nalla VK, Rogan PK. Automated splicing mutation analysis by information theory. *Hum Mutat* 2005;25:334–342.
29. Divina P, Kvitkovicova A, Buratti E, Vorechovsky I. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur J Hum Genet* 2009;17:759–765.
30. Lim KH, Fairbrother WG. Spliceman–a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics* 2012;28:1031–1032.
31. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009;37:e67.
32. Schwartz S, Hall E, Ast G. SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Res* 2009;37(Web Server issue):W189–W192.
33. Faber K, Glatting KH, Mueller PJ, Risch A, Hotz-Wagenblatt A. Genome-wide prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASsites. *BMC Bioinformatics* 2011;12(suppl 4):S2.
34. Desmet FO, Hamroun D, Collod-Beroud G, Claustres M, Beroud C. Bioinformatics identification of splice site signals and prediction of mutation effects. In: Mohan RM (ed). *Research Advances in Nucleic Acids Research*. Global Research Network: Kerala, 2010:1–14.
35. Houdayer C, Caux-Moncoutier V, Krieger S, et al. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum Mutat* 2012;33:1228–1238.
36. Psaty BM, O'Donnell CJ, Gudnason V, et al.; CHARGE Consortium. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* 2009;2:73–80.
37. NIH program explores the use of genomic sequencing in newborn healthcare. 2013. http://www.nih.gov/news/health/sep2013/nhgri-04.htm. Accessed 4 September 2013.
38. Houdayer C, Dehainault C, Mattler C, et al. Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum Mutat* 2008;29:975–982.
39. Holla ØL, Nakken S, Mattingsdal M, et al. Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: Comparison of wet-lab and bioinformatics analyses. *Mol Genet Metab* 2009;96:245–252.
40. Vreeswijk MP, Kraan JN, van der Klift HM, et al. Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Hum Mutat* 2009;30:107–114.
41. Théry JC, Krieger S, Gaildrat P, et al. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *Eur J Hum Genet* 2011;19:1052–1058.
42. Colombo M, De Vecchi G, Caleca L, et al. Comparative *in vitro* and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS ONE* 2013;8:e57173.