Bioinformatics identification of splice site signals and prediction of mutation effects

François-Olivier Desmet, Dalil Hamroun, Gwenaëlle Collod-Béroud, Mireille Claustres, Christophe Béroud

Bioinformatics identification of splice site signals and prediction of mutation effects

François-Olivier Desmet[1,2], Dalil Hamroun[3,2], Gwenaëlle Collod-Béroud[1,2], Mireille Claustres[1,2,3] and Christophe Béroud[1,2,3]

1: INSERM, U827, Montpellier, F-34000 France.

2: Université Montpellier1, UFR Médecine, Montpellier, F-34000 France.

3: CHU Montpellier, Hôpital Arnaud de Villeneuve, Laboratoire de Génétique Moléculaire, Montpellier, F-34000 France.

Correspondence/Reprint request:

Dr Christophe Béroud, CHU de Montpellier, Hôpital Arnaud de Villeneuve, Laboratoire de Génétique Moléculaire, 641 Av du Doyen G. Giraud, 34093 Montpellier Cedex 5, France.

Tel: 33 4 67 41 53 60 ; Fax: 33 4 67 41 53 65 ; e-mail: christophe.beroud@inserm.fr

ABSTRACT: Maturation of mRNA in eukaryotes is a very complex process that includes exon recognition through specific elements such as branch point motifs, 5' and 3' splice sites and splicing regulators. Mutations may affect these signals either directly by disrupting constitutive splice sites or indirectly by creating cryptic splice sites. We thus evaluated the prediction efficiency of nine software programs based on Position Weight Matrices, Markov Models, Maximal Dependence Decomposition, Neural Networks or Feature Generation Algorithms using 623 mutations for which the consequence on mRNA has been demonstrated *in vitro*. Position Weight Matrix-based tools were the most efficient in predicting the impact of a given mutation. Overall, at least one system correctly predicted 100% of mutations affecting invariant positions as well as -1, +3 and +5 positions of the 5'ss. Deep intronic mutations resulting in the activation of cryptic exons were almost all correctly predicted (92.31%), while other intronic mutations were less efficiently (70-80%). Exonic mutations that create cryptic splice sites were also efficiently detected (70%). Because of the prediction heterogeneity and specificity, a single tool could not be used for all predictions. Moreover, since these systems are all based on text analysis an in vitro validation step is still required.

INTRODUCTION

With the recent breakthroughs in sequencing and genotyping, millions of single nucleotide polymorphisms (SNPs) have been characterized in the human genome. Among the 7,736,157 SNPs reported in genes at NCBI (dbSNP build 130; http://www.ncbi.nlm.nih.gov/SNP/), 86.20 % are localized in introns, 2.74% in exons and the others at 5' and 3' regions (1). Among the exonic variations, 52.19% are missense and 36.86% synonymous substitutions and may represent either neutral variations or pathogenic mutations. A review of data available from the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/ac/index.php) confirms these findings with 61.10% of disease-causing mutations being missense mutations (2). Concomitantly, 1610 mutations (2.55%) have also been reported that affect the invariant positions +1, +2, -1 or -2 of donor (5'ss) and acceptor (3'ss) splice sites, which delineate the limits of exons. These data are nevertheless far from reality as it is now recognized that many missense variations (up to 50% of disease-causing single nucleotide variations) could indeed have a critical impact on mRNA maturation through disruption or creation of splice sites. It has also been shown that these variations can affect other splicing

motifs such as auxiliary sequences also known as splicing regulators (Exonic Splicing Enhancers (ESE) or Exonic Splicing Silencer (ESS)) (3-5). In parallel, other intronic variations affect donor and acceptor splice sites although they do not concern the invariant positions. The prediction of the consequences of a nucleotide variation on mRNA is thus of major interest for molecular diagnosis not only of intronic mutations but also of exonic mutations. As the majority of disease-causing mutations that affect mRNA maturation modify donor or acceptor splices sites, many tools have been designed to predict the consequence of nucleotide substitutions on these motifs.
Two types of introns have been described (6,7). U2 snRNP-dependent introns represent more than 99.9% of all introns (7) and are excised by a spliceosome containing the U1, U2, U4, U5 and U6 snRNPs. U12 snRNP-dependent introns are the minor class of introns and are excised by a spliceosome containing U11, U12, U4atac, U6atac and U5 snRNPs (10) and are usually shorter than U2 introns. Both U2 and U12 snRNP-dependent introns are recognized by the cellular machinery through consensus motifs called donor (5'ss) and acceptor (3'ss) splice sites (11). These motifs are characterized by their terminal dinucleotides GT-AG, GC-AG and AT-AC, the vast majority (>98%) being of the GT-AG subtype (12). If these 4 bases are very conserved among genes, adjacent positions around the splice sites may vary, leading to short degenerated consensus sequences: usually 9 bp for the 5'ss and 14 bp for the 3'ss. Since these motifs are small, the probability to find them by chance within an intron is high thus creating pseudo-exons or decoy sites. Pseudo-exons are intronic sequences that match the exon requirements, but are not selected as exons by the spliceosome. This is usually related to the splice site motifs themselves but also to the RNA structure that impacts the motif accessibility (13) and/or other splicing signals. Indeed, three types of sequences are involved in splicing (14): constitutive splice sites, branch point sequences (15) and splicing regulators (ESE and ESS), which are important for discriminating real exons from pseudo-exons (4,16,17). Today, research for detecting real exons and splice sites also takes into account the link between chromatin and gene structure (18,19) as well as hydrophobicity profiles (10). If the splicing process is complex and much work has to be done yet to fully understand it, various software and web applications are already available on Internet to detect splicing signals, at the forefront of which stands donor and acceptor splice site predictions (20).
To detect 5'ss and 3'ss, six methods have been developed (21) and implemented in various software (22). The most frequently used relies on Position Weighted Matrices (PWM) and Position Specific Score Matrices (PSSM). Senapathy and Shapiro (23,24) defined PWM, which are based upon sequence alignments (25,26). PWM have subsequently been refined and the weight of each position has been modulated to underline their respective biological importance (27). Application software based on PWM processes a given sequence into short sequences (whose length is equal to that of the 5' or the 3' splice site). A weight is then attributed to each nucleotide according to its position within the PWM. If the sum of the consecutive nucleotides weights is superior or equal to a specific threshold (28), a motif is consequently detected. In order to increase the efficiency of 5'ss detection, Carmel *et al.* proposed to combine a Delta-G (DG), which predicts the variation of free energy that comes from the base-pairing of the 5'ss and U1 snRNP (29).
The Maximal Dependence Decomposition (MDD) tool relies on the assumption that splice site positions are not independent and that conditional probabilities can be determined between adjacent and non-adjacent positions. First, the algorithm calculates the amount of dependence between the consensus indicator variable and a nucleotide indicator at the position i. This value is then used to separate the dataset (constituted of aligned sequences)

in two subsets: sequences with and sequences without consensus nucleotide at the given position i. This is repeated for all positions to create a tree in which each subset will be used to generate separate weight matrix models that are then combined in a composite model (30).
The third approach uses Markov models (MM). The main advantage of this technique is that it takes into account bases dependencies. A Markov model (31-33) is defined as a suite of states (for example exon, 5' splice site, intron, 3' splice site, exon). Each state is associated to probabilities to evolve (stay in the current state or move to another one). After sequence reading, the algorithm computes the most likely suite of states (and corresponding sequences) that fulfills the various probabilities.
Support Vector Machines (SVM) are classifiers that help to take a decision (34). These classifiers need first to be trained using a set that contains both positive and negative data to efficiently assign the SVM parameters. Once the SVM is trained, it can be used to classify sequences. The efficiency of such algorithms is directly linked to the quality and the number of sequences used as positive and negative controls. The Feature Generation Algorithm (FGA) is an SVM derivative with an initial step to generate features. These features are groups of nucleotides (adjacent or not) that share a particular position or a range of positions. The obtained features are then used as input for a learning algorithm similar to that of SVM. Donor and acceptor sites use separate classifiers. Once the training is finished, each feature is weighted. When a sequence is processed for splice site prediction, the classifier checks if the sequence contains any feature according to the considered candidate (acceptor or donor) (35). Neural Networks (NN) use artificial neuron networks that mimic the function of real neuron networks (36). As for SVM, a training (learning) period to set up thresholds and weights is required for the network to work efficiently. Each artificial neuron will input new pieces of information (coming from external sources or other neurons) and weight them. Once a defined threshold is reached the sequence is considered as being identified and the output will be set to 1 (0 otherwise).
The last approach relies on Hybrid models. It has been shown that, in many cases, before processing data using an SVM-based approach, it could be more efficient to first sift data with Markov Models. This has been applied to splice site recognitions by Ho and Rajapakse (37), who created an hybrid approach consisting of two first- and two second-order Markov chain models followed by a three-layer neural network.
Because of the strong needs to efficiently predict the effects of sequence variations on splice site recognition, not only for molecular diagnosis of human genetic diseases but also for various therapeutic approaches, a comparison of the various methods and associated algorithms/ software was necessary. In this review we evaluated various available tools/methods to predict the impact of mutations on splice site recognition using a large set of validated mutations.

MATERIAL & METHODS

Method

We mined the literature for mutations affecting splicing and selected 623 mutations, using data from UMD locus-specific databases (38-41), the Human Gene Mutation Database (2) and other previously used datasets (42,43). Mutations were divided in four datasets (Figure 1): i) 72 mutations that affect the four invariant positions of 5' (n=49) and 3' splice sites (n=23) (positions +1, +2 and -2, -1,) (positive control); ii) 178 mutations that include intronic mutations either localized at splice sites in non-canonical positions (n=148), distant intronic mutations also known as "deep" intronic mutations (n=13) and short distance mutations (n=17); iii) 288 exonic mutations reported to affect splicing, including 10 exonic mutations that activate a cryptic splice site; and iv) 85 mutations that do not affect splice sites (negative control). These mutations have been re-

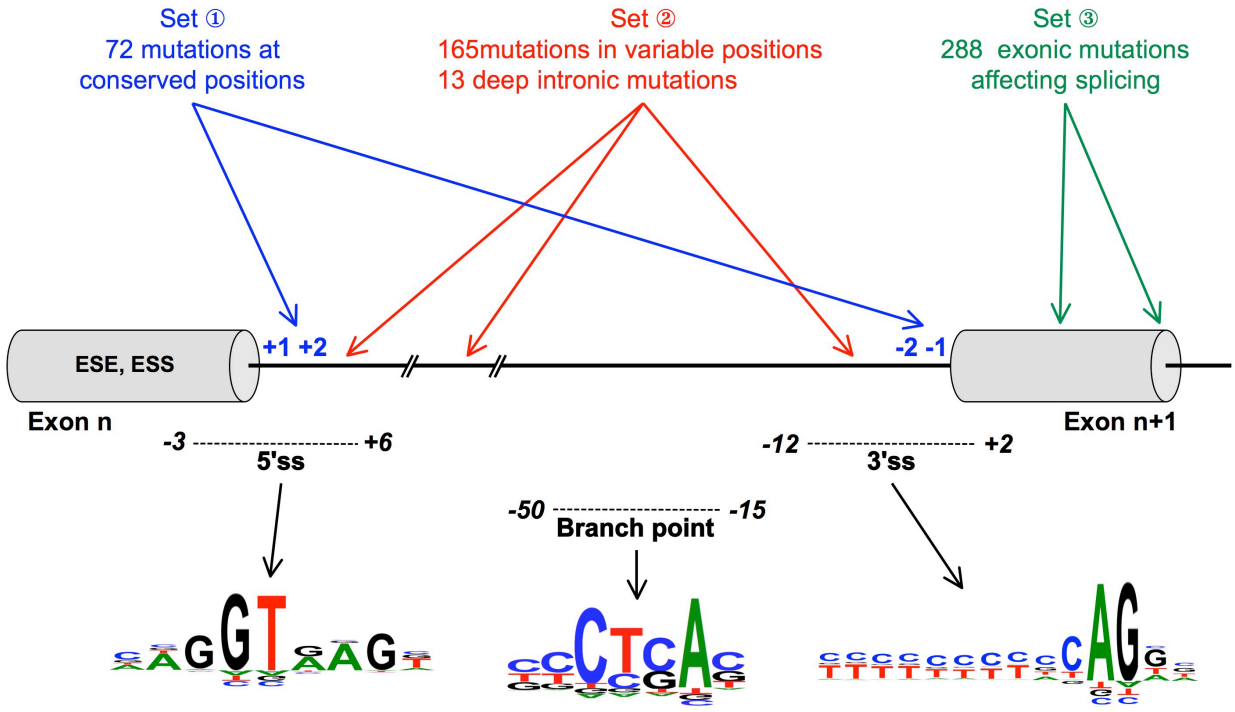

Figure 1: Distribution of mutation sets according to splicing motifs: 5'ss (donor splice site); 3'ss (acceptor splice site); branch point and splicing regulators (ESE, ESS).

ported to modify branch points (9 mutations), splicing regulators (72 mutations) or have been identified as polymorphisms in splice sites (4 mutations) during diagnosis procedures.

For each set, both reference and mutant sequences were extracted from the Ensembl database version 44 (44) using Human Splicing Finder (27). As some prediction programs require long intronic sequences, 200 nucleotides-long intronic sequences were added to each exon extremity. In order to facilitate data processing all sequences were stored as FASTA file (one file per set, cf. Supplementary Materials).

When possible, stand-alone versions of programs were preferred because they allowed the automation of the various prediction steps. As most tools have been designed to predict the position of splice sites within a sequence mainly to detect exons, they do not allow an easy comparison of a set of mutant sequences with the corresponding reference sequences. We thus developed a Java application that combines all results in a single file and outlines the differences between sequences (reference vs. mutant) for each mutation. When thresholds had to be set, we chose to use default thresholds and parameters except for the MaxEntScan software for which the analysis was processed with two different thresholds: 0 (default threshold) and 2 as suggested by Coutinho *et al.* (45). As the data obtained with the second threshold were more accurate (data not shown), these were subsequently used for all analyses. Note that data from MaxEntScan were extracted from the Human Splicing Finder website and not directly from MaxEntScan. Finally, when the downloaded software needed training, we used the included training sets.

Selected software programs

In order to evaluate all prediction methods, we selected the nine most widely used and freely available software programs: GenScan (30), GeneSplicer (46), Human Splicing Finder (HSF) (27), MaxEntScan (47), NNSplice (48), SplicePort (49), SplicePredictor (50), SpliceView (51) and Sroogle (52). Three use only PWM (Table 1), two combine PWM with MM or DG, two apply a MDD approach, in one case combined with MM, one NN and the

Software	Method	URL
GenScan	MDD	http://genes.mit.edu
GeneSplicer	MDD, MM	http://www.cbcb.umd.edu/software/GeneSplicer/
Human Splicing Finder	PWM	http://www.umd.be/HSF/
MaxEntScan	MM, PWM	http://genes.mit.edu/burgelab/maxent/
NNSplice	NN	http://www.fruitfly.org/seq_tools/splice.html
SplicePort	FGA	http://spliceport.cs.umd.edu/
SplicePredictor	PWM	http://deepc2.psi.iastate.edu/cgi-bin/sp.cgi
SpliceView	PWM	http://zeus2.itb.cnr.it/~webgene/wwwspliceview_ex.html
SROOGLE	PWM, DG	http://sroogle.tau.ac.il/

Table 1: Selected software programs to predict splice sites. PWM = Position Weight Matrix; MDD = Maximal Dependence Decomposition; MM = Markov Model; NN = Neural Network; FGA = Feature Generation Algorithm; DG = Delta-G

last uses an original approach called Feature Generation Algorithm (FGA).

Data management

In order to simplify interpretation of data from the three sets of mutations known to disrupt a wild type splice site, we converted all numerical information into a simple binary status: “broken” or not. The “broken” status was assigned when the software predicted that the mutation disrupted the natural splice site. This could be achieved either when the mutant splice site value was below the specific detection threshold or when the absolute variation between the reference and the mutant values was 10% higher than a specific variation that was set as previously described for HSF (27). Similarly, for mutations known to create a cryptic splice site (datasets 2 and 3), we converted all numerical information into a simple binary state: “new” or not. The “new” status was assigned when the program predicted that the mutation created a splice site. This new splice site could either be near the natural splice site itself or in a deeper intronic region of the sequence, depending on the mutation localization.

For dataset 4 (negative controls), both “broken” and “new” status were evaluated when a mutation was localized within a splice site or elsewhere.

RESULTS

The 72 mutations affecting the 3'ss and 5'ss invariant positions (Figure 1) were predicted to impact splice site motifs with a high accuracy (>95%) by four programs: HSF (100%), MaxEntScan (100%), SpliceView (100%), Sroogle (100%) and NNSplice (97.22%). Unexpectedly two software programs gave poor results: GenScan (45.83%) and GeneSplicer (63.89%). Overall results were more efficient for 3’ss than for 5’ss (93.72% vs. 83.90%) as reported in Table 2.

Among the 148 mutations from the second dataset localized in splice site motifs, 24 affected the +3 position and 55 the +5 position, which are frequently involved in human diseases. Because these positions

Software	Global	3’ss	5’ss
GeneSplicer	46 (63.89%)	19 (82.61%)	27 (55.10%)
GenScan	33 (45.83%)	19 (82.61%)	15 (30.61%)
Human Splicing Finder	72 (100%)	23 (100%)	49 (100%)
MaxEntScan	72 (100%)	23 (100%)	49 (100%)
NNSplice	70 (97.22%)	23 (100%)	47 (95.92%)
SplicePort	63 (87.50%)	19 (82.61%)	44 (89.80%)
SplicePredictor	63 (87.50%)	22 (95.65%)	41 (83.67%)
SpliceView	72 (100%)	23 (100%)	49 (100%)
Sroogle	72 (100%)	23 (100%)	49 (100%)
Average	***86.88%***	***93.72%***	***83.90%***

Table 2: Prediction of mutations affecting the invariant positions -2, -1, +1 and +2 of 3’ss (n=23) and 5’ss (n=49).

Software	Last base	+3	+5	5'ss	3'ss
GeneSplicer	9 (42.86%)	14 (58.33%)	29 (52.73%)	11 (44.00%)	29 (65.91%)
GenScan	8 (38.10%)	5 (20.83%)	8 (14.55%)	3 (12.00%)	11 (25.00%)
Human Splicing Finder	21 (100 %)	20 (83.33%)	55 (100%)	14 (56.00%)	20 (45.45%)
MaxEntScan	21 (100%)	24 (100%)	55 (100%)	19 (76.00%)	34 (77.27%)
NNSplice	20 (95.24%)	22 (91.67%)	54 (98.18%)	19 (76.00%)	20 (45.45%)
SplicePort	21 (100%)	22 (91.67%)	52 (94.55%)	17 (68.00%)	34 (77.27%)
SplicePredictor	10 (47.62%)	16 (66.67%)	36 (65.45%)	10 (40.00%)	15 (34.09%)
SpliceView	12 (57.14%)	11 (45.83%)	37 (67.27%)	6 (24.00%)	13 (29.55%)
Sroogle	21 (100%)	13 (54.17%)	55 (100%)	12 (48.00%)	2 (4.55%)
Average	***75.66%***	***68.06%***	***76.97%***	***49.33%***	***44.95%***

Table 3: Prediction of mutations affecting the last base of the exons as well as positions +3 and +5 of the 5'ss and of mutations affecting other positions of 3'ss and 5'ss.

are degenerated and dependent on the sequence context, splice site prediction is more difficult. This was confirmed by the overall lower prediction efficiency (76.97% for the +5 position and 68.06% for the +3 position) and its heterogeneity (from 20.83% to 100% for the +3 position and from 14.55% to 100% for the +5 position) (Table 3). In addition to these two intronic key positions, position -1 (last base of the exon) also plays a major role in 5'ss recognition. Once again, the overall prediction efficiency was around 75% with high heterogeneity (from 38.10% to 100%) among tools. Only four software programs gave homogeneous results with efficiency higher than 80% for these 3 positions (HSF, MaxEntScan, NNSplice and SplicePort).

For the mutations involving the other 3'ss (n=44) and 5'ss (n=25) positions, overall predictions were poorer (49.33% for 5'ss and 44.95% for 3'ss) (Table 3) and individual prediction efficiencies were even more scattered (from 4.55% to 77.27%). MaxEntScan and SplicePort gave the most accurate predictions.

Then 17 intronic mutations localized at a short distance (<100 bp) and 13 deep intronic mutations (>100 bp) from the second dataset were tested. Since these mutations create cryptic splice sites, only data from the "New" status were taken into account. For the short distance mutations, results were unsatisfactory with a mean prediction efficiency of 27.45% and only Human Splicing Finder showed a detection rate higher than 70% (Table 4). With deep intronic mutations, which activate cryptic splice sites leading to cryptic exon inclusions, the efficiency increased to 60.68% and three prediction tools (Human Splicing Finder, MaxEntScan and SpliceView) displayed efficiencies higher than 80%.

Concomitantly to these intronic mutations localized at a distance from constitutive splice sites, we also evaluated exonic mutations that do not directly affect constitutive splice sites but rather activate a cryptic splice site. As only few of such mutations have been studied experimentally (53-57), we collected a small set of 10 mutations belonging to this category. Human Splicing Finder, MaxEntScan and SplicePort predicted the activation of cryptic splice sites with good accuracy (70%), while the overall prediction was poor with only 44.44% accuracy (Table 5).

Software	Short distance (<100 bp)	Long distance (> 100 bp)
GeneSplicer	7 (41.18%)	6 (46.15%)
GenScan	2 (11.76%)	0 (0.00%)
Human Splicing Finder	12 (70.59%)	12 (92.31%)
MaxEntScan	4 (23.53%)	12 (92.31%)
NNSplice	1 (5.88%)	9 (69.23%)
SplicePort	6 (35.29%)	8 (61.54%)
SplicePredictor	4 (23.53%)	9 (69.23%)
SpliceView	3 (17.65%)	11 (84.62%)
Sroogle	3 (17.65%)	4 (30.77%)
Average	***27.45%***	***60.68%***

Table 4: Prediction of intronic mutations localized at a distance from splice sites.

Finally, to evaluate the proportion of exonic mutations that can affect splicing through the activation of a cryptic splice site, we selected a set of 267 mutations for which an effect on splicing has been demonstrated experimentally but without clues about the involved mechanisms. In this context, the average prediction was 23.01% with a heterogeneity ranging from 7.49% to 50.94% (Table 5).

To evaluate the specificity of the various tools, we collected a set of 85 negative controls. Among them 4 polymorphisms were localized within constitutive splice sites at not highly conserved positions (-12 and +6). Five tools (GenScan, Human Slicing Finder, SplicePredictor, SpliceView and Sroogle) did not detect any impact on the corresponding splice sites as expected, while the others reported false positive predictions for 1 to 4 of these sites (Table 5). The analysis of the other 81 negative controls led to prediction of cryptic splice site activation in 17.86% of the cases with a heterogeneity ranging from 4.94% to 30.86% (Table 6).

DISCUSSION

Splice sites are key elements for exon recognition and therefore mutations leading to alteration of these signals have a strong impact on mRNA maturation and protein synthesis. It is now recognized that a wide range of mutations, which can be localized in introns but also in exons, may affect these signals either directly (disruption of constitutive splice sites) or indirectly (creation of cryptic splice sites). In order to evaluate the efficiency of prediction tools, we only evaluated the impact of mutations on splice sites discarding any effect on splicing regulators and branch points. For this purpose, we selected three sets of mutations for which the consequence on mRNA has been demonstrated in vitro and that encompass all situations:

Software	Cryptic ss activation	Unknown mechanism
GeneSplicer	3 (30%)	70 (26.22%)
GenScan	1 (10%)	20 (7.49%)
Human Slicing Finder	7 (70%)	82 (30.71%)
MaxEntScan	7 (70%)	67 (25.09%)
NNSplice	6 (60%)	34 (12.73%)
SplicePort	7 (70%)	136 (50.94%)
SplicePredictor	4 (40%)	51 (19.10%)
SpliceView	5 (50%)	58 (21.72%)
Sroogle	0 (0%)	35 (13.11%)
Average	***44.44%***	***23.01%***

Table 5: Exonic mutations known to result in splice defects.

Software	ss polymorphism	Other positions
GeneSplicer	2 (50%)	23 (28.40%)
GenScan	0 (0%)	5 (6.17%)
Human Slicing Finder	0 (0%)	19 (23.46%)
MaxEntScan	2 (50%)	22 (27.16%)
NNSplice	1 (25%)	11 (13.58%)
SplicePort	4 (100%)	25 (30.86%)
SplicePredictor	0 (0%)	10 (12.35%)
SpliceView	0 (0%)	13 (16.05%)
Sroogle	0 (0%)	4 (4.94%)
Average	***10%***	***17.86%***

Table 6: Prediction of intronic and exonic mutations that do not affect splicing. ss polymorphism = mutations localized in constitutive splice sites; other positions = intronic or exonic mutations not localized in constitutive splice sites.

disruption of constitutive splice sites by affecting invariant or variable positions, activation of a cryptic exon by intronic mutations and exonic mutations.

We anticipated that all prediction tools and methods could efficiently predict the effect of mutations involving the four invariant positions. Overall predictions were in agreement with these expectations and four software programs (Human Splicing Finder, MaxEntScan, SpliceView and Sroogle) showed an accuracy of 100%. Unexpectedly, GeneSplicer and Genscan gave poor results especially for 5'ss.

Three other positions play an important role in 5'ss recognition: the -1, +3 and +5 positions. It has been recently demonstrated that the +3 position is intrinsically linked to sequence contexts leading to difficult predictions (58), while the +5 position is less influenced (59). Interestingly, despite these limitations MaxEntScan could accurately predict all pathogenic mutations affecting these positions. HSF, NNSplice and SplicePort tools also gave very accurate results. Like for the "invariant" positions, GenScan gave poor predictions while GeneSplicer, SplicePredictor and SpliceView showed intermediate accuracy. Sroogle was very efficient for the -1 and +5 position (100%), but performed less efficiently for the +3 position (54.17%). This is probably due to the use of a PWM that does not take into account the sequence context, this limit not being compensated by the addition of the DG algorithm.

Besides these mutations that disrupt a constitutive splice site, other mutations can activate a cryptic splice site that will either be in competition with the constitutive site for recognition by the cellular machinery or lead to the inclusion of a pseudo exon (60,61). Since these sites are localized in introns at short or long distance from the constitutive splice sites, they are usually harder to predict, as pseudo exons do not match criteria for exon recognition by Markov Models and Neural Networks. On the other hand, PWM-based tools should accurately identify them because they analyze only splice sites motifs. As expected PWM tools gave the best results especially for the deep intronic mutations (average of 73.85%), but performed less well in the identification of intronic mutations localized at a short distance from constitutive splice sites (average 30.59%). Only HSF gave accurate results in both situations (92.31% and 70.59%). Programs based on Markov Models and Neural Networks were less efficient with a detection rate ranging from 0 to 69.23% with poor results for short distance cryptic sites (5.88 to 41.18%).

Exonic mutations can also result in cryptic splice site activation. Because these sites are localized in a favorable context for recognition by Markov Models and Neural Networks, we expected them to be as efficient as PWM. Indeed, the best NN- and

PWM-based tools gave similar results with a detection rate of 70% (HSF, MaxEntScan and SplicePort) while the other three PWM-based tools gave unexpectedly low predictions, especially Sroogle (0%). MDD-based tools were less efficient as for previous situations.

In order to evaluate the proportion of exonic mutations that affect splicing and are associated with splice site modifications (disruption of the constitutive splice site or creation of a cryptic splice site) we selected a series of 267 exonic mutations belonging to this category. If we consider that only 70% of exonic mutations that result in the activation of a cryptic exon are correctly predicted by the most efficient tools (HSF, MaxEntScan and SplicePort) and we remove tools for which a false positive prediction rate is high (SplicePort, MaxEntScan) we can predict that between 31 to 44% of exonic mutations may alter splicing by involving a splice site.

Finally, to evaluate the specificity of predictions we selected 85 intronic and exonic mutations for which it has been demonstrated that they do not affect constitutive or cryptic splice sites (negative control). Four of them were localized in constitutive splice sites and therefore directly reflected false positive predictions for the inactivation of these sites. Five tools (GenScan, HSF, SplicePredictor, SpliceView and Sroogle) did not predict any consequence for these mutations. Conversely, NNSplice predicted the disruption of one of the constitutive splices sites, GeneSplicer and MaxEntScan of two and SplicePort of all. If these results are not significant due to the small number of mutations, they question the specificity of these last three tools. The evaluation of predictions for the other exonic and intronic mutations revealed a low rate (17.86%) of positive predictions. Because the creation of a cryptic splice site is not sufficient to activate a cryptic exon, these results could not be considered as false positive. In fact, two main situations could be encountered. The cryptic splice site is in the vicinity of the constitutive splice site and therefore is in competition for recognition by the cellular machinery. The strength of this new site, its localization in relation to the branch point for 3'ss as well as the sequence context (splice regulators) should all be taken into account before predicting if this splice site will be recognized by the cellular machinery. To our knowledge, today no tool can handle such information and a manual analysis has to be performed. In the second situation, the cryptic splice site is at a distance from an exon. In this context a cryptic exon can be recognized only if the complementary splice site (3'ss for a 5' cryptic ss and vice versa) is present at a short distance. In addition, other criteria should also be satisfied: presence of a branch point and a favorable splice regulator context. Once again, the tools tested in this study can not handle such information. GeneSplicer and GenScan were expected to perform better but cryptic exons are usually different from natural exons (no selection for codons and presence of stop codons), reducing their ability to recognize the cryptic exons.

Predictions limitations

Several limitations should be taken into account when interpreting the predictions of the tools used for this analysis. The first is related to the intron type. Although two types of introns have been described (6,7), only the GT-AG subtype of U2 snRNP-dependent intron model is used for predictions as it accounts for more than 98% of the cases. Consequently, a specific attention should be paid to wild type splice sites to confirm that they belong to this category before performing prediction analyses. The second limit is associated with the algorithms. They are all based on a text analysis (search or comparison) without addition of other parameters such as accessibility of the motif based on 2D or 3D structures as well as DNA-protein interactions that could mask a site. Thus strong splice sites could be predicted but not used *in vivo* because they are not accessible to the spliceosome. Moreover, the splicing process

is complex and involves many signals (branch point, splicing signals, splice regulators) and proteins, which are not taken into account for predictions (8). The software programs we tested in this study use different methods to achieve splice site recognition. They all rely on the quality of their respective training/building sets. With the completion of the Human Genome Project, many exons have been identified and are available to build positive sets of 5' and 3'ss. This is particularly useful for PWM-based tools that only require a positive control set. On the contrary, for SVM methods, both positive and negative control sets are required. In addition, some software programs need to know where the exon-intron boundaries are while others make *ab initio* prediction. In this review, we tested both splice site detectors and gene/exon annotation programs. The best results were obtained with PWM-based software. Gene/exon annotation programs (GenScan, GeneSplicer) have a global approach and need a longer sequence to detect splice sites (ranging from 80 to 200 intronic nucleotides) and evaluate splice regulators to efficiently detect true exons. This additional step is critical and thus led to a reduced efficiency of such tools in this analysis. For instance, GenScan had first to detect exon boundaries, which was not possible for all reference sequences. This does not impair their ability to predict exons in a large-scale genomic analysis.

Which software for which situation?

To decide which is the most appropriate software, the user has to consider not only its efficiency, but also additional characteristics such as its availability, amount of data to be processed (batch analysis or multiple query option), interface, etc. In this review we have tested the nine most frequently used tools. Only Human Splicing Finder was created to compare a mutant sequence with a reference sequence, while the other tools were designed to identify splice sites within a sequence (MaxEntScan, Sroogle, SplicePort, SplicePredictor and SpliceView) or to detect exons (GeneSplicer and GenScan). Because of these different designs and underlying algorithms, they displayed different efficiencies. To evaluate a mutation involving invariant positions most tools may be used with a preference for Human Splicing Finder, MaxEntScan, NNSplice, SpliceView and Sroogle, the last two being less efficient when positions -1, +3 and +5 of 5'ss need to be evaluated. When other less conserved positions have to be investigated, users may prefer MaxEntScan, NNSplice or SplicePort. Nevertheless, it is also critical to evaluate the specificity of the different prediction tools as MaxEntScan, NNSplice or SplicePort seem to have a reduced specificity compared to Human Splicing Finder, SplicePredictor, SpliceView and Sroogle. When searching for a cryptic splice site, Human Splicing Finder is the most efficient, but MaxEntScan and SpliceView also are very efficient for deep intronic mutations. Finally, for exonic mutations Human Splicing Finder, MaxEntScan, SplicePort and NNSplice are the most efficient.

In this study, we focused on splice site detection, one of the many features of the complex mechanism of splicing. In diagnostic and research situations, geneticists have to address all aspects of splicing defects including alteration of branch points and the effect of splicing regulators since they modulate splicing by allowing fixation of proteins/complexes on the pre-mRNA, thus enhancing weak splice site signals or decreasing strong splice site signals. The user should thus combine the tools presented here with other systems/approaches to evaluate these many situations (3,5,27,62-66). Although Human Splicing Finder is not the most efficient tool for all situations, it is the first that can address all splicing aspects (branch points, splice sites and splicing regulators). Users can thus include in their toolbox Human Splicing Finder, MaxEntScan and Sroogle. Finally, since all *in silico* predictions are based only on statistics, an *in vitro* validation step is still required.

FUNDING
The research leading to these results has received funding from the European Community Seventh Framework Program (FP7/2007-2013) under grant agreement number 200754 – GEN2PHEN project. The European Community Sixth Framework Program (FP6) under grant agreement number 036825 – TREAT-NMD Network of Excellence, also funded this work.

REFERENCES
1. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308-311.
2. Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S. and Cooper, D.N. (2009) The Human Gene Mutation Database: 2008 update. *Genome Med*, **1**, 13.
3. Zhang, C., Li, W.H., Krainer, A.R. and Zhang, M.Q. (2008) RNA landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci U S A*, **105**, 5797-5802.
4. Chasin, L.A. (2007) Searching for splicing motifs. *Adv Exp Med Biol*, **623**, 85-106.
5. Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G.P., Bresolin, N., Giorda, R. and Pozzoli, U. (2004) Silencer elements as possible inhibitors of pseudo-exon splicing. *Nucleic Acids Res*, **32**, 1783-1791.
6. Burge, C.B., Padgett, R.A. and Sharp, P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol Cell*, **2**, 773-785.
7. Sharp, P.A. and Burge, C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875-879.
8. Nilsen, T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147-1149.
9. Zhou, Z., Licklider, L.J., Gygi, S.P. and Reed, R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature*, **419**, 182-185.
10. Boldina, G., Ivashchenko, A. and Regnier, M. (2009) Using profiles based on nucleotide hydrophobicity to define essential regions for splicing. *Int J Biol Sci*, **5**, 13-19.
11. Jacob, M. and Gallinaro, H. (1989) The 5' splice site: phylogenetic evolution and variable geometry of association with U1RNA. *Nucleic Acids Res*, **17**, 2159-2180.
12. Burset, M., Seledtsov, I.A. and Solovyev, V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*, **28**, 4364-4375.
13. Buratti, E. and Baralle, F.E. (2004) Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol*, **24**, 10505-10514.
14. Wang, Z. and Burge, C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna*, **14**, 802-813.
15. Gao, K., Masuda, A., Matsuura, T. and Ohno, K. (2008) Human branch point consensus sequence is yUnAy. *Nucleic Acids Res*, **36**, 2257-2267.
16. Blencowe, B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci*, **25**, 106-110.
17. Cartegni, L., Chew, S.L. and Krainer, A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, **3**, 285-298.
18. Schwartz, S., Meshorer, E. and Ast, G. (2009) Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, **16**, 990-995.
19. Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J. and Guigo, R. (2009) Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, **16**, 996-1001.
20. Houdayer, C., Dehainault, C., Mattler, C., Michaux, D., Caux-Moncoutier, V., Pages-Berhouet, S., d'Enghien, C.D., Lauge, A., Castera, L., Gauthier-Villars, M. *et al.* (2008) Evaluation of in silico splice

tools for decision-making in molecular diagnosis. *Hum Mutat*, **29**, 975-982.
21. Zhang, X.H., Leslie, C.S. and Chasin, L.A. (2005) Computational searches for splicing signals. *Methods*, **37**, 292-305.
22. Brent, M.R. (2007) How does eukaryotic gene prediction work? *Nat Biotechnol*, **25**, 883-885.
23. Senapathy, P., Shapiro, M.B. and Harris, N.L. (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol*, **183**, 252-278.
24. Shapiro, M.B. and Senapathy, P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res*, **15**, 7155-7174.
25. D'Haeseleer, P. (2006) How does DNA sequence motif discovery work? *Nat Biotechnol*, **24**, 959-961.
26. Zhang, M.Q. and Marr, T.G. (1993) A weight array method for splicing signal analysis. *Comput Appl Biosci*, **9**, 499-509.
27. Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M. and Beroud, C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*, **37**, e67.
28. Touzet, H. and Varre, J.S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol*, **2**, 15.
29. Carmel, I., Tal, S., Vig, I. and Ast, G. (2004) Comparative analysis detects dependencies among the 5' splice-site positions. *RNA*, **10**, 828-840.
30. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, **268**, 78-94.
31. Eddy, S.R. (1996) Hidden Markov models. *Curr Opin Struct Biol*, **6**, 361-365.
32. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.
33. Eddy, S.R. (2004) What is a hidden Markov model? *Nat Biotechnol*, **22**, 1315-1316.
34. Noble, W.S. (2006) What is a support vector machine? *Nat Biotechnol*, **24**, 1565-1567.
35. Dogan, R.I., Getoor, L., Wilbur, W.J. and Mount, S.M. (2007) Features generated for computational splice-site prediction correspond to functional elements. *BMC Bioinformatics*, **8**, 410.
36. Krogh, A. (2008) What are artificial neural networks? *Nat Biotechnol*, **26**, 195-197.
37. Ho, L.S. and Rajapakse, J.C. (2003) Splice site detection with a higher-order markov model implemented on a neural network. *Genome Inform*, **14**, 64-72.
38. Collod-Beroud, G., Beroud, C., Ades, L., Black, C., Boxer, M., Brock, D.J., Holman, K.J., de Paepe, A., Francke, U., Grau, U. *et al.* (1998) Marfan Database (third edition): new mutations and new routines for the software. *Nucleic Acids Res*, **26**, 229-223.
39. Beroud, C., Hamroun, D., Collod-Beroud, G., Boileau, C., Soussi, T. and Claustres, M. (2005) UMD (Universal Mutation Database): 2005 update. *Hum Mutat*, **26**, 184-191.
40. Beroud, C., Collod-Beroud, G., Boileau, C., Soussi, T. and Junien, C. (2000) UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. *Hum Mutat*, **15**, 86-94.
41. Frederic, M.Y., Monino, C., Marschall, C., Hamroun, D., Faivre, L., Jondeau, G., Klein, H.G., Neumann, L., Gautier, E., Binquet, C. *et al.* (2009) The FBN2 gene: new mutations, locus-specific database (Universal Mutation Database FBN2), and genotype-phenotype correlations. *Hum Mutat*, **30**, 181-190.
42. Rogan, P.K. and Schneider, T.D. (1995) Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum Mutat*, **6**, 74-76.

43. Rogan, P.K., Faux, B.M. and Schneider, T.D. (1998) Information analysis of human splice site mutations. *Hum Mutat*, **12**, 153-171.
44. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res*, **35**, D610-617.
45. Eng, L., Coutinho, G., Nahas, S., Yeo, G., Tanouye, R., Babaei, M., Dork, T., Burge, C. and Gatti, R.A. (2004) Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. *Hum Mutat*, **23**, 67-76.
46. Pertea, M., Lin, X. and Salzberg, S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*, **29**, 1185-1190.
47. Yeo, G. and Burge, C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, **11**, 377-394.
48. Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997) Improved splice site detection in Genie. *J Comput Biol*, **4**, 311-323.
49. Dogan, R.I., Getoor, L., Wilbur, W.J. and Mount, S.M. (2007) SplicePort--an interactive splice-site analysis tool. *Nucleic Acids Res*, **35**, W285-291.
50. Brendel, V., Xing, L. and Zhu, W. (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, **20**, 1157-1169.
51. Rogozin, I.B. and Milanesi, L. (1997) Analysis of donor splice sites in different eukaryotic organisms. *J Mol Evol*, **45**, 50-59.
52. Schwartz, S., Hall, E. and Ast, G. (2009) SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Res*, **37**, W189-192.
53. Chen, W., Kubota, S., Teramoto, T., Nishimura, Y., Yonemoto, K. and Seyama, Y. (1998) Silent nucleotide substitution in the sterol 27-hydroxylase gene (CYP 27) leads to alternative pre-mRNA splicing by activating a cryptic 5' splice site at the mutant codon in cerebrotendinous xanthomatosis patients. *Biochemistry*, **37**, 4420-4428.
54. Keeratichamroen, S., Ketudat Cairns, J.R., Wattanasirichaigoon, D., Wasant, P., Ngiwsara, L., Suwannarat, P., Pangkanon, S., Kuptanon, J., Tanpaiboon, P., Rujirawat, T. *et al.* (2008) Molecular analysis of the iduronate-2-sulfatase gene in Thai patients with Hunter syndrome. *J Inherit Metab Dis*.
55. Lualdi, S., Pittis, M.G., Regis, S., Parini, R., Allegri, A.E., Furlan, F., Bembi, B. and Filocamo, M. (2006) Multiple cryptic splice sites can be activated by IDS point mutations generating misspliced transcripts. *J Mol Med*, **84**, 692-700.
56. Pomponio, R.J., Reynolds, T.R., Mandel, H., Admoni, O., Melone, P.D., Buck, G.A. and Wolf, B. (1997) Profound biotinidase deficiency caused by a point mutation that creates a downstream cryptic 3' splice acceptor site within an exon of the human biotinidase gene. *Hum Mol Genet*, **6**, 739-745.
57. Su, C.C., Yang, J.J., Shieh, J.C., Su, M.C. and Li, S.Y. (2007) Identification of novel mutations in the KCNQ4 gene of patients with nonsyndromic deafness from Taiwan. *Audiol Neurootol*, **12**, 20-26.
58. Guedard-Mereuze, S.L., Vache, C., Molinari, N., Vaudaine, J., Claustres, M., Roux, A.F. and Tuffery-Giraud, S. (2009) Sequence contexts that determine the pathogenicity of base substitutions at position +3 of donor splice-sites. *Hum Mutat*, **30**, 1329-1339.
59. Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J. and Cooper, D.N. (2007) Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat*, **28**, 150-158.
60. Tuffery-Giraud, S., Saquet, C., Chambert, S. and Claustres, M. (2003) Pseudoexon activation in the DMD gene as a novel mechanism for Becker muscular dystrophy. *Hum Mutat*, **21**, 608-614.
61. Beroud, C., Carrie, A., Beldjord, C., Deburgrave, N., Llense, S., Carelle, N., Peccate, C., Cuisset, J.M., Pandit, F.,

Carre-Pigeon, F. *et al.* (2004) Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. *Neuromuscul Disord*, **14**, 10-18.
62. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q. and Krainer, A.R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, **31**, 3568-3571.
63. Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T. and Ast, G. (2006) Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol Cell*, **22**, 769-781.
64. Zhang, X.H. and Chasin, L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, **18**, 1241-1250.
65. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M. and Burge, C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831-845.
66. Fairbrother, W.G., Yeo, G.W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P.A. and Burge, C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res*, **32**, W187-190.