

AU1 ▶

A G-to-T Transversion at the Splice Acceptor Site of Dystrophin Exon 14 Shows Multiple Splicing Outcomes That Are Not Exemplified by Transition Mutations

Mitsunori Ota, Yasuhiro Takeshima, Atsushi Nishida, Hiroyuki Awano, Tomoko Lee, Mariko Yagi, and Masafumi Matsuo

Mutations at splicing consensus sequences have been shown to induce splicing errors such as exon skipping or cryptic splice site activation. Here, we identified eight splicing products caused by a G-to-T transversion mutation at the splice acceptor site of exon 14 of the dystrophin gene (c.1603–1G>T). Unexpectedly, the most abundant product showed skipping of the two consecutive exons 14 and 15, and exon 14 skipping was observed as the second most abundant product. To examine the cause of this splicing multiplicity, minigenes containing dystrophin exons 14 and 15 with their flanking introns were constructed and subjected to *in vitro* splicing. Minigenes with the wild-type sequence or a G>A transition at position c.1603–1 produced only the mature mRNA. On the other hand, the minigenes with a G>T or G>C transversion mutation produced multiple splicing products. A time-course analysis of the *in vitro* splicing revealed that splicing of the middle intron, intron 14, was the first step in transcript maturation for all four minigene constructs. The identity of the mutant nucleotide, but not its position, is a factor leading to multiple splicing outcomes. Our results suggest that exon skipping therapy for Duchenne's muscular dystrophy should be carefully monitored for their splicing outcomes.

Introduction

SPLICING IS A highly regulated process that removes intron sequences from pre-mRNA. The process proceeds from the 5' to the 3' end of the pre-mRNA to assemble the ordered array of exons in mature transcripts (Keren *et al.*, 2010; Licatalosi and Darnell, 2010). Splicing depends on the correct identification of exons, which must be recognized within the pre-mRNA despite being relatively short compared with introns. It is known that the presence of well-defined cis-elements, namely the splice donor and acceptor sites and the branch point, are necessary but not sufficient to define intron-exon boundaries (Senapathy *et al.*, 1990). A critical step in pre-mRNA splicing is the recognition and pairing of splice donor and acceptor sites, where GT and AG dinucleotides are strictly conserved sequences, respectively. Mutations at these consensus sequences result in splicing errors such as exon skipping and cryptic splice site activation and are responsible for nearly 15% of all genetic diseases (Hertel, 2008).

Duchenne's and Becker muscular dystrophy (DMD/BMD) are the most common forms of inherited myopathy and are caused by mutations in the dystrophin gene. DMD is a rapidly progressive disease that usually results in the death of patients in their twenties, whereas BMD is a clinically less severe

form of the disease and often has only a slightly debilitating effect. The difference between DMD and BMD can be explained by the reading frame rule: frame-shift mutations that generate premature stop codons in the coding sequence of dystrophin mRNA usually result in the DMD phenotype, whereas mutations that maintain the original reading frame cause the milder BMD phenotype (Monaco *et al.*, 1988). Therefore, it is important to consider how splice site mutations could alter the splicing products of the dystrophin gene, and thus whether they will lead to either the DMD or the BMD phenotype. For this reason, dystrophinopathy is one of the most extensively studied diseases related to splice site mutations and their resultant products (Matsuo *et al.*, 1991; Nishiyama *et al.*, 2008).

The dystrophin gene is the largest human gene, spanning 2500 kb on the X chromosome. It has a complex structure, including a large number of exons, long introns, and several alternative promoters (Den Dunnen *et al.*, 1989; Ahn and Kunkel, 1993; Nishio *et al.*, 1994). Dystrophin pre-mRNA splicing would seem to require extraordinarily strict regulation. However, single-nucleotide mutations in splice acceptor or donor sites have been reported to cause splicing alterations, including exon skipping, cryptic splice site activation, or both (Hagiwara *et al.*, 1994; Tuffery-Giraud *et al.*, 1999; Adachi *et al.*,

2003; Thi Tran *et al.*, 2005; Habara *et al.*, 2009). Despite the identification of a large number of splice site mutations in many diseases, there is still no established way to predict what splicing patterns these mutations will produce (Wimmer *et al.*, 2007; Hertel, 2008).

In this study, we identified multiple splicing outcomes caused by a transversion mutation at the splice acceptor site of dystrophin exon 14. In *in vitro* splicing assays, transversion but not transition mutations induced multiple splicing outcomes. We suggest that the manipulation of splicing with antisense oligonucleotides (AOs) should be carefully evaluated in DMD treatment.

Materials and Methods

Genomic DNA analysis

Genomic DNA was extracted from a BMD male (KUCG557) and a control as previously described (Matsuo *et al.*, 1990). A mutation in the dystrophin gene was sought by direct sequencing (Nishiyama *et al.*, 2008; Takeshima *et al.*, 2010).

Dystrophin mRNA analysis

Dystrophin mRNA in skeletal muscle was analyzed as previously described (Roberts *et al.*, 1991; Surono *et al.*, 1999). A fragment spanning exons 13 to 18 was amplified by reverse transcription-polymerase chain reaction (RT-PCR) using a set of primers (c13f; 5'-GCTGCTTTGGAAGAACAACCTT-3' and 1F; 5'-CTTCTGAGCGAGTAATCCAGCT-3') using conditions as previously described (Habara *et al.*, 2009). The amplified products were sequenced after subcloning, as previously described (Thi Tran *et al.*, 2005).

In vitro splicing analysis

In vitro splicing analysis was conducted using the preconstructed minigene of the expression vector H492 that encodes two cassette exons (exons A and B) and a multicloning site in the intervening sequence (Habara *et al.*, 2008).

A region spanning intron 13, exon 14, intron 14, exon 15, and intron 15 was PCR amplified from genomic DNA of the index case and a control using the primer set: 5'-GCGGCTAGCTCAGAAAGAGTGCCCTTCCAA-3' and 5'-GCGG GATCCCACCTTAATTCAGAAAAGTAGCAA-3'. Amplified products were digested with *NheI* and *BamHI* restriction enzymes (New England Biolabs, Ipswich, MA) and inserted into the H492 vector that was predigested with the same enzymes. Two mutant constructs, c.1603-1G>A and c.1603-1G>C, were prepared by means of overlap extension PCR and were inserted into the H492 vector. After checking their sequences, the resultant plasmids were transfected into HeLa cells for splicing assays as previously described (Habara *et al.*, 2009). The cells were harvested after 24 h of incubation for extraction of total RNA.

Time-course analysis

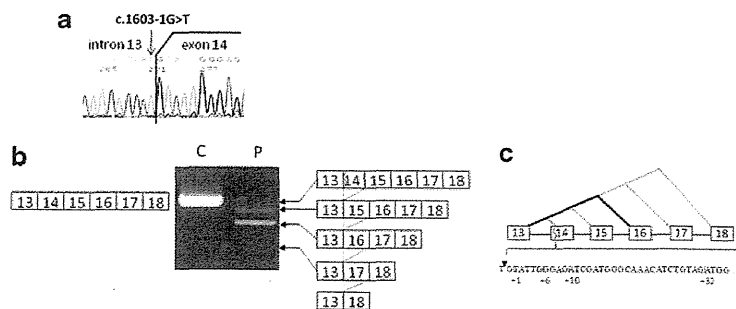
To determine the splicing order of the introns, *in vitro* splicing was stopped after 1, 2, 3, or 4 h by extracting total RNA. The resulting splicing products were analyzed as previously described. RT-PCR products were semiquantified using a DNA 1000 LabChip kit on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) and their sequences were determined by subcloning and sequencing.

Results

A G-to-T transversion mutation was identified at the 3' end of intron 13 of the dystrophin gene (c.1603-1G>T) in a Japanese BMD patient (Fig. 1). Although the mutation was located within an intron, the mutation changed the highly conserved AG dinucleotides at the splice acceptor site to AT, thereby decreasing the splice site strength (Ri) of the splice acceptor site from 12.3 to 3.5. It was strongly expected that skipping of the downstream exon 14 would be the splicing outcome of this mutation. However, dystrophin mRNA expressed in the patient's skeletal muscle was found to contain multiple splicing products. When a region spanning exons 13 to 18 was amplified by RT-PCR, four amplified bands were visualized

◀ F1

FIG. 1. Genomic and mRNA analysis of the dystrophin gene. (a) Sequence of the intron 13/exon 14 junction. A genomic PCR product from the index case (KUCG557), a Becker muscular dystrophy patient, was directly sequenced. The terminal nucleotide of intron 13 (c.1603-1) was identified to be T, indicating a G-to-T transversion mutation (c.1603-1G>T; indicated by the arrow). (b) RT-PCR products encompassing exons 13 to 18. Amplified products obtained from the control (C) and patient (P) are shown. The identity of the bands is indicated on either side of the gel image. Boxes and numbers within the boxes indicate exons and exon numbers, respectively. The vertical wavy line indicates activated cryptic splice sites. (c) Schematic of potential splicing patterns. The approximate frequency of each identified splicing pattern is distinguished by the lines joining the boxes (bold line, high frequency; dashed line, low frequency; regular line, intermediate frequency). Details of cryptic splice sites detected around the 5' end of exon 14 are shown as a magnified image (bottom). Four vertical dashed lines located at nucleotides +1, +6, +10, and +32 within exon 14 indicate cryptic splice sites that were activated by the mutation. The inverted triangle indicates the junction between intron 13 and exon 14. RT-PCR, reverse transcription-polymerase chain reaction.



SPLICE MUTATION SHOWING MULTIPLE SPLICING OUTCOMES

on agarose gels, with different densities from clearly visible to barely visible (Fig. 1). Unexpectedly, subcloning and sequencing revealed that the most abundant product lacked not only exon 14 but also exon 15, indicating the skipping of two consecutive exons (Fig. 1). The second most abundant product corresponded to the expected product that lacked only exon 14. As these two abundant mRNAs maintained the original dystrophin reading frame, the index case was diagnosed as BMD at the mRNA level. This matched with his clinical phenotype.

The third faintly visible product, which corresponded to the wild-type dystrophin size, maintained all six exons from exons 13 to 18. However, subcloning and sequencing of this band revealed four different clones, containing a deletion of 1, 6, 10, or 32 nucleotides at the 5' end of exon 14 (Fig. 1). It was unclear whether these were genuine splicing products or artifacts. Examination of the genomic sequences around these unexpected deletions revealed that the ends of two of the deletions, at the 10th and 32nd nucleotides, comprise AG dinucleotides, whereas the ends of the other two, at the 1st and 6th nucleotides, comprise TG dinucleotides (Fig. 1). Because all these four sites scored as potential splice acceptor sites, it was concluded that all four clones were derived from alternative splicing products, activating a cryptic splice acceptor site located +1, +6, +10, or +32 nucleotides from the intron 13/exon 14 boundary (Fig. 1). The smallest, barely visible band lacked exons 14, 15, and 16 (Fig. 1) and clones lacking exons 14, 15, 16, and 17 were also identified. Eight splicing patterns had been identified in the patient's muscle RNA: four patterns of exon skipping and four patterns of cryptic splice site activation (Fig. 1).

Multiple splicing outcomes caused by a single-nucleotide change at the splice acceptor site of the mammoth dystrophin gene were unexpected. This suggested that the mutation identified in the index case exerts a strong influence on the splicing machinery. To demonstrate this, an *in vitro* splicing system using the H492 minigene was employed. Fragments encompassing exons 14–15 and their flanking introns, with wild-type or mutant (c.1603–1G>T) sequence, were amplified from genomic DNA by PCR and inserted into the multicloning site within the intron between exons A and B of the H492 minigene (Fig. 2a). Constructed minigenes were transfected into HeLa cells and the resulting splicing products were analyzed by RT-PCR after 24 h. From the wild-type construct, a single amplified product was obtained and was confirmed to be the expected mature mRNA consisting of exons A, 14, 15, and B (Fig. 2). This *in vitro* system was therefore confirmed to effect accurate splicing of dystrophin exons 14 and 15.

For the mutant minigene harboring the c.1603–1G>T transversion, multiple RT-PCR amplification products were identified: one major band, two minor bands, and one weak band (Fig. 2). Sequencing of the major band revealed an intermediate splicing product that lacked both introns 14 and 15 but retained intron A unspliced. This indicated that the splicing machinery recognized the two downstream introns, but not the upstream intron A in the presence of the c.1603–1G>T change. This outcome was compatible with the disruption of the splice acceptor site of intron A by the mutation. By subcloning and sequencing of the two minor products and the one weak product, four clones were identified, which differed in their exon 14 5' end sequences in the mature mRNA. Each had a 1-, 4-, 10-, or 32-nucleotide deletion from the 5' end

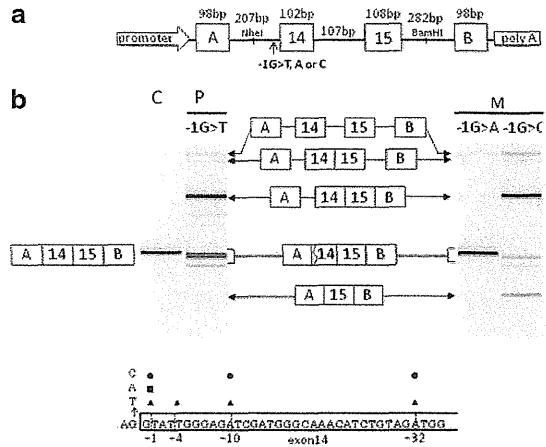


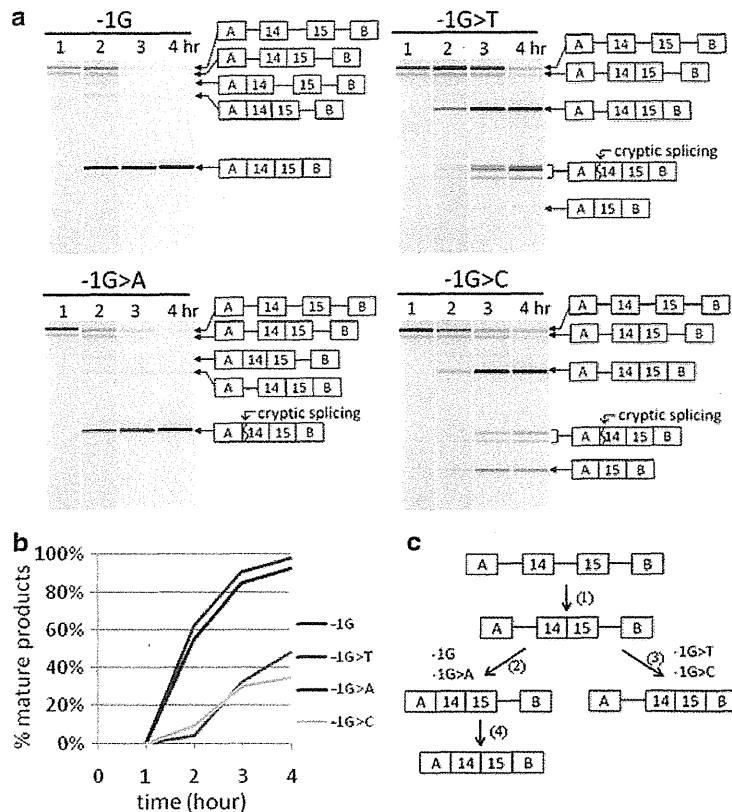
FIG. 2. *In vitro* splicing analysis. (a) Schematic showing the structure of the minigenes. The genomic regions spanning dystrophin exons 14, 15, and their respective adjacent introns, containing wild-type or mutant (c.1603–1G>T, G>A, or G>C) sequence, were inserted into the *NheI* and *BamHI* sites between exons A and B in the H492 vector. Boxes and lines indicate exons and introns, respectively. At the 5' and 3' end, respectively, there is a promoter and poly(A) addition signal. The lengths of the exons and introns are indicated. (b) RT-PCR products of minigene transcripts. Capillary gel electrophoresis patterns of RT-PCR-amplified products from the minigenes carrying the wild-type (c.1603–1G) (C) or the mutant (c.1603–1G>T or G>C) (M) sequence are shown. The identity of the splicing products is indicated on either side. Boxes and numbers within the boxes indicate exons and exon numbers, respectively. The vertical wavy line indicates the position of the cryptic splice sites. Details of the cryptic splice sites (vertical dashed lines) detected around the 5' end of exon 14 are shown as a magnified image (bottom). The numbers indicate the nucleotide number within exon 14. The activated cryptic splice sites are marked by black triangles, black squares, and black circles in the minigenes with c.1603–1G>T, G>A, and G>C, respectively.

F2 ▶

of exon 14 (Fig. 2); the 10-bp deletion was the most common. Although activation of four cryptic splice acceptor sites within exon 14 was shown *in vitro* as well as *in vivo*, only three were common to both. *In vivo*, the sixth nucleotide was activated as a cryptic splice site (TG), whereas *in vitro*, the 4th nucleotide was activated (AT). Remarkably, pre-mRNAs maintaining all three introns and mRNA lacking only intron 14 were identified. Because the pre-mRNA was not detected for the wild-type minigene, the G>T mutation was considered to affect the splicing reaction fundamentally. However, skipping of exons 14 and 15, or the single exon 14, was not observed *in vitro*. It was unclear why the expected exon skipping did not occur, even though multiple splicing outcomes caused by the G>T transversion mutation were exemplified *in vitro* (Fig. 2).

We questioned whether nucleotide changes other than the G>T transversion located at an identical position would also produce multiple splicing outcomes. Minigenes with substitution of G to A or C were constructed and subjected to the same *in vitro* splicing assays. Unexpectedly, the minigene with the G>A transition mutation produced one clear band

FIG. 3. Time-course analysis of minigene splicing reactions. (a) Capillary electrophoresis patterns of RT-PCR products from minigenes with G, T, A, or C at position c.1603-1 are shown. The identity of the splicing products is indicated on the right-hand side. (b) The percentage of mature mRNA among all RT-PCR products is shown according to the reaction time. Mature mRNA identified at each time point was expressed as the percentage of the total amplified product. Mature mRNA rapidly increased and, after 4 h, constituted >90% of the total for the wild-type and G>A mutant minigenes. In contrast, at the same time point, mature mRNA constituted less than half of the total for the mutant minigenes with transversion mutations (G>T or G>C). (c) Schematic showing the main splicing pathways of the minigene transcripts. For all four minigenes, splicing of intron 14 occurred first (1), producing a splicing intermediate. For the wild-type minigene or the G>A transition mutant, the splicing of intron A then proceeded, producing another intermediate (2), whereas for the G>T or G>C construct, splicing of intron 15 occurs, producing another intermediate (3). Finally, splicing of intron 15 proceeds to produce mature mRNA in the wild-type or G>A mutant minigene (4).



(Fig. 2). Interestingly, this product corresponded to the mature mRNA but had a deletion of a G nucleotide at the 5' end of exon 14. This G nucleotide was probably recognized as the last nucleotide of intron A, resulting in a simple splicing outcome. This minigene also produced a faint band corresponding to pre-mRNA; this was probably due to disturbance of the entire splicing reaction (Fig. 2).

Remarkably, the minigene with the G>C transversion mutation produced many products: one major band, four additional weak bands, and one barely discernible band. The most abundant band corresponded to a splicing intermediate that removed two downstream introns (introns 14 and 15) but retained intron A (Fig. 2). The second abundant, smallest product had the exon structure A-15-B, indicating exon 14 skipping. In addition, three transcripts were identified with 1-, 10-, or 32-nucleotide deletions at the 5' end of exon 14 of the mature mRNA (Fig. 2), indicating three cryptic splice site activations. The G>C transversion mutation induced multiple splicing outcomes. These results indicated that a single-nucleotide change at the splice acceptor site changed the splicing pathways, but the splicing outcomes *per se* were dependent on the particular substituted nucleotide.

To clarify the splicing process in transversion or transition mutations at c.1603-1, the time course of the splicing reactions in the minigenes was investigated 1-4 h after transfection. After 1 h, an intermediate product that lacked intron 14 was detected for all four minigenes, in addition to the pre-mRNA

(Fig. 3). This indicated that the first step of splicing occurs at the middle intron (intron 14), even in the presence of a normal splice acceptor site for intron A. After 2 h, the mature mRNA constituted >60% of the total mRNA for the wild-type minigene and 55% for the G>A mutant minigene, indicating an uneventful splicing reaction. In contrast, mature mRNA constituted only 10%-20% for minigenes with the G>T or G>C transversions. This indicated that splicing of intron A in the context of a G>T or G>C transversion was strongly inhibited. After 3 h, >80% of the total transcripts were mature mRNA for the wild-type minigene or for the minigene with the G>A transition. However, mature mRNA with activated cryptic splice sites within intron A occupied <40% for the minigenes with the G>T or G>C transversion. This indicated that the last step, the splicing of intron A, was strongly hampered in the presence of a transversion mutation.

After 4 h, the splicing reaction was almost complete, with mature mRNA constituting >90% of the total transcripts for the wild-type minigene or the minigene with the G>A mutation. For minigenes with the G>T or G>C transversion, mature mRNA comprised less than half of the total. We considered this to be due to severe disturbance of the splicing of intron A in the presence of a transversion mutation.

In summary, we found that transversion mutations led to multiple splicing outcomes by suppressing splicing reactions. Our data suggest that the identity of the mutant nucleotide can dramatically influence splicing outcome.

◀F3

Discussion

In this study, a transversion mutation from G to T (c.1603-1G>T) at the splice acceptor site of exon 14 of the dystrophin gene was shown to induce multiple splicing outcomes in a Japanese BMD case. Altogether, eight splicing patterns for dystrophin mRNA were identified in the patient's muscle (Fig. 1). We questioned why this wide variety of splicing patterns was produced from a single-nucleotide change.

In the dystrophin gene, 80 donor and 137 acceptor splice site mutations have been reported (www.dmd.nl/nmdb/home.php). Most of these showed skipping of an adjacent exon as the outcome of the mutation, and some resulted in cryptic splice site activation (Bartolo *et al.*, 1996; Adachi *et al.*, 2003; Thi Tran *et al.*, 2005). Skipping of multiple exons has been reported only rarely: among the reported splice site mutations of the dystrophin gene, only six have been described to induce skipping of multiple exons (Tuffery-Giraud *et al.*, 2004, 2005; Deburgrave *et al.*, 2007; Taylor *et al.*, 2007; Takeshima *et al.*, 2010). Of these six, three were splice acceptor site mutations including the present case. Three were splice donor site mutations, but two of these were an identical mutation at the splice donor site of exon 29 (Deburgrave *et al.*, 2007). It was supposed that splice acceptor site mutations have a greater tendency to promote multiple exon skipping. It is remarkable that mutations at the splice donor site of exon 29 and at the splice acceptor site of exon 13 have each been reported twice to cause multiple exon skipping. This suggests that these two exons predispose to multiple exon skipping in the presence of a splice site mutation.

In the index mutation, we supposed that the double exon skipping was due to the short length of intron 14 (107 bp). However, *in vitro* splicing of the minigene with an intact intron 14 failed to produce double exon skipping (Fig. 2). This indicated that the short length of intron 14 is not the sole determinant of multiple exon skipping. Remarkably, the time-course analysis of minigene splicing revealed that splicing of the middle intron, intron 14, occurred first for all four minigenes (Fig. 3). Because the structure of the intron and neighboring exons was accurate in the minigenes, we presume that intron 14 is spliced out at an early stage also *in vivo*. It is supposed that, in the patient's muscle, the removal of intron 14 is the major splicing event in the early stages, and then the splicing machinery identifies the splice acceptor site of exon 16 and the donor site of intron 13, thereby promoting the skipping of the two consecutive exons 14 and 15. As a minor pathway, the splicing machinery identifies the splice acceptor site of exon 15 and the donor site of intron 13, leading to exon 14 skipping.

Multiple splicing outcomes were exemplified for minigenes with transversion mutations but not for the minigene with a transition mutation (Fig. 2). Although a multiplicity of splicing outcomes was demonstrated *in vitro* as *in vivo*, the actual splicing patterns were different: a splicing intermediate retaining intron A was a major product *in vitro* (Fig. 2), whereas exon-skipped mature mRNAs were the main products *in vivo* (Fig. 1). Among the four *in vitro* splicing reactions, only the minigene with the G>C mutation produced exon 14-skipped mRNA. Even though the two transversion mutations both exerted drastic changes (Fig. 2), each produced different splicing outcomes. It is difficult to explain this difference. We

conclude that both transversion mutations cause drastic changes in the splicing machinery and multiple splicing outcomes, but the identity of the substituted nucleotide induces the splicing pattern differences.

Recently, the manipulation of dystrophin pre-mRNA splicing is attracting much attention as a way to treat DMD (Matsuo, 1996; van Deutekom and van Ommen, 2003). AOs against splicing regulatory elements have been shown to induce exon skipping, thereby enhancing dystrophin expression in DMD patients (Takeshima *et al.*, 2006; van Deutekom *et al.*, 2007). Our results flag the possibility of multiple splicing outcomes after manipulation of splicing with AOs. Particular caution should be exercised when inducing skipping of the exons that have been reported to induce multiple exon skipping in the presence of a splice site mutation (Tuffery-Giraud *et al.*, 2004, 2005; Deburgrave *et al.*, 2007; Taylor *et al.*, 2007; Takeshima *et al.*, 2010). Considering that even a single-nucleotide change within exon sequence has been shown to induce splicing error (Shiga *et al.*, 1997; Nishiyama *et al.*, 2008), AO is supposed to induce unwanted splicing outcomes, such as activation of cryptic splice site created by new junction. In establishing exon skipping therapy with AO, therefore, splicing outcomes should be carefully monitored.

Acknowledgments

The authors thank Professor Peter K. Rogan for his encouraging support and Ms. Kanako Yokoyama for her secretarial help. This work was supported in part by a Grant-in-Aid for Scientific Research (B) and Grant-in-Aid for Exploratory Research from the Japan Society for the Promotion of Science; a Health and Labor Sciences Research Grant for Research on Psychiatric and Neurological Diseases and Mental Health; and a research grant for Nervous and Mental Disorders from the Ministry of Health, Labor, and Welfare, Japan.

Disclosure Statement

No competing financial interests exist.

References

- Adachi K, Takeshima Y, Wada H, *et al.* (2003) Heterologous dystrophin mRNAs produced by a novel splice acceptor site mutation in intermediate dystrophinopathy. *Pediatr Res* 53: 125-131.
- Ahn AH, Kunkel LM (1993) The structural and functional diversity of dystrophin. *Nat Genet* 3:283-291.
- Bartolo C, Papp AC, Snyder PJ, *et al.* (1996) A novel splice site mutation in a Becker muscular dystrophy patient. *J Med Genet* 33:324-327.
- Deburgrave N, Daoud F, Llense S, *et al.* (2007) Protein- and mRNA-based phenotype-genotype correlations in DMD/BMD with point mutations and molecular basis for BMD with nonsense and frameshift mutations in the DMD gene. *Hum Mutat* 28:183-195.
- Den Dunnen JT, Grootsholten PM, Bakker E, *et al.* (1989) Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. *Am J Hum Genet* 45:835-847.
- Habara Y, Doshita M, Hirozawa S, *et al.* (2008) A strong exonic splicing enhancer in dystrophin exon 19 achieve proper splicing without an upstream polypyrimidine tract. *J Biochem* 143:303-310.

- Habara Y, Takeshima Y, Awano H, *et al.* (2009) *In vitro* splicing analysis showed that availability of a cryptic splice site is not a determinant for alternative splicing patterns caused by +1G→A mutations in introns of the dystrophin gene. *J Med Genet* 46:542–547.
- Hagiwara Y, Nishio H, Kitoh Y, *et al.* (1994) A novel point mutation (G⁻¹ to T) in a 5' splice donor site of intron 13 of the dystrophin gene results in exon skipping and is responsible for Becker muscular dystrophy. *Am J Hum Genet* 54:53–61.
- Hertel K (2008) Combinatorial control of exon recognition. *J Biol Chem* 283:1211–1215.
- Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11:345–355.
- Licalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 11:75–87.
- Matsuo M (1996) Duchenne/Becker muscular dystrophy: from molecular diagnosis to gene therapy. *Brain Dev* 18:167–172.
- Matsuo M, Masumura T, Nakajima T, *et al.* (1990) A very small frame-shifting deletion within exon 19 of the Duchenne muscular dystrophy gene. *Biochem Biophys Res Commun* 170:963–967.
- Matsuo M, Masumura T, Nishio H, *et al.* (1991) Exon skipping during splicing of dystrophin mRNA precursor due to an intron deletion in the dystrophin gene of Duchenne muscular dystrophy Kobe. *J Clin Invest* 87:2127–2131.
- Monaco AP, Bertelson CJ, Liechti-Gallati S, *et al.* (1988) An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. *Genomics* 2:90–95.
- Nishio H, Takeshima Y, Narita N, *et al.* (1994) Identification of a novel first exon in the human dystrophin gene and of a new promoter located more than 500 kb upstream of the nearest known promoter. *J Clin Invest* 94:1037–1042.
- Nishiyama A, Takeshima Y, Zhang Z, *et al.* (2008) Dystrophin nonsense mutations can generate alternative rescue transcripts in lymphocytes. *Ann Hum Genet* 72:717–724.
- Roberts RG, Barby TF, Manners E, *et al.* (1991) Direct detection of dystrophin gene rearrangements by analysis of dystrophin mRNA in peripheral blood lymphocytes. *Am J Hum Genet* 49:298–310.
- Senapathy P, Shapiro MB, Harris NL (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol* 183:252–278.
- Shiga N, Takeshima Y, Sakamoto H, *et al.* (1997) Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. *J Clin Invest* 100:2204–2210.
- Surono A, Takeshima Y, Wibawa T, *et al.* (1999) Circular dystrophin RNAs consisting of exons that were skipped by alternative splicing. *Hum Mol Genet* 8:493–500.
- Takeshima Y, Yagi M, Okizuka Y, *et al.* (2010) Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. *J Hum Genet* 55:379–388.
- Takeshima Y, Yagi M, Wada H, *et al.* (2006) Intravenous infusion of an antisense oligonucleotide results in exon skipping in muscle dystrophin mRNA of Duchenne muscular dystrophy. *Pediatr Res* 59:690–694.
- Taylor PJ, Maroulis S, Mullan GL, *et al.* (2007) Measurement of the clinical utility of a combined mutation detection protocol in carriers of Duchenne and Becker muscular dystrophy. *J Med Genet* 44:368–372.
- Thi Tran HT, Takeshima Y, Surono A, *et al.* (2005) A G-to-A transition at the fifth position of intron 32 of the dystrophin gene inactivates a splice donor site both *in vivo* and *in vitro*. *Mol Genet Metab* 85:213–219.
- Tuffery-Giraud S, Chambert S, Demaille J, *et al.* (1999) Point mutations in the dystrophin gene: evidence for frequent use of cryptic splice sites as a result of splicing defects. *Hum Mutat* 14:359–368.
- Tuffery-Giraud S, Saquet C, Chambert S, *et al.* (2004) The role of muscle biopsy in analysis of the dystrophin gene in Duchenne muscular dystrophy: experience of a national referral centre. *Neuromuscul Disord* 14:650–658.
- Tuffery-Giraud S, Saquet C, Thorel D, *et al.* (2005) Mutation spectrum leading to an attenuated phenotype in dystrophinopathies. *Eur J Hum Genet* 13:1254–1260.
- van Deutekom J, Janson A, Girjaar I, *et al.* (2007) Local dystrophin restoration with antisense oligonucleotide PRO051. *N Engl J Med* 357:2677–2686.
- van Deutekom JC, van Ommen GJ (2003) Advances in Duchenne muscular dystrophy gene therapy. *Nat Rev Genet* 4:774–783.
- Wimmer K, Roca X, Beiglböck H, *et al.* (2007) Extensive *in silico* analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5' splice-site disruption. *Hum Mutat* 28:599–612.

Address correspondence to:
 Masafumi Matsuo, M.D., Ph.D.
 Department of Pediatrics
 Graduate School of Medicine
 Kobe University
 7-5-1 Kusumokicho
 Chuo-ku
 Kobe 650-0017
 Japan
 E-mail: matsuo@kobe-u.ac.jp

AUTHOR QUERY FOR GTMB-2010-0276-VER9-OTA_1P

AU1: Please italicize all the gene symbols mentioned in the article.

RESEARCH ARTICLE

Open Access

Categorization of 77 *dystrophin* exons into 5 groups by a decision tree using indexes of splicing regulatory factors as decision markers

Rusdy Ghazali Malueka¹, Yutaka Takaoka², Mariko Yagi¹, Hiroyuki Awano¹, Tomoko Lee¹, Ery Kus Dwianingsih¹, Atsushi Nishida^{1,3}, Yasuhiro Takeshima¹ and Masafumi Matsuo^{1,4*}

Abstract

Background: Duchenne muscular dystrophy, a fatal muscle-wasting disease, is characterized by dystrophin deficiency caused by mutations in the *dystrophin* gene. Skipping of a target *dystrophin* exon during splicing with antisense oligonucleotides is attracting much attention as the most plausible way to express dystrophin in DMD. Antisense oligonucleotides have been designed against splicing regulatory sequences such as splicing enhancer sequences of target exons. Recently, we reported that a chemical kinase inhibitor specifically enhances the skipping of mutated *dystrophin* exon 31, indicating the existence of exon-specific splicing regulatory systems. However, the basis for such individual regulatory systems is largely unknown. Here, we categorized the *dystrophin* exons in terms of their splicing regulatory factors.

Results: Using a computer-based machine learning system, we first constructed a decision tree separating 77 authentic from 14 known cryptic exons using 25 indexes of splicing regulatory factors as decision markers. We evaluated the classification accuracy of a novel cryptic exon (exon 11a) identified in this study. However, the tree mislabeled exon 11a as a true exon. Therefore, we re-constructed the decision tree to separate all 15 cryptic exons. The revised decision tree categorized the 77 authentic exons into five groups. Furthermore, all nine disease-associated novel exons were successfully categorized as exons, validating the decision tree. One group, consisting of 30 exons, was characterized by a high density of exonic splicing enhancer sequences. This suggests that AOs targeting splicing enhancer sequences would efficiently induce skipping of exons belonging to this group.

Conclusions: The decision tree categorized the 77 authentic exons into five groups. Our classification may help to establish the strategy for exon skipping therapy for Duchenne muscular dystrophy.

Keywords: Splicing, *Dystrophin*, Exon, Splicing enhancer, Decision tree

Background

Duchenne muscular dystrophy (DMD), a fatal muscle-wasting disease, is the most common inherited muscle disease, affecting one in every 3500 male births. DMD is characterized by dystrophin deficiency caused by mutations in the *dystrophin* gene, the largest human gene that spans over 2500 kb on the X-chromosome. For the treatment of DMD, antisense oligonucleotides (AOs) against splicing regulatory sequences have been proposed to

produce in-frame *dystrophin* mRNA from the out-of-frame mRNA by inducing exon skipping during splicing [1]. The newly generated in-frame *dystrophin* mRNA is expected to produce semi-functional, internally deleted dystrophin protein. Currently, induction of exon skipping with AOs is considered one of the most promising treatments for DMD [2,3].

The *dystrophin* gene encodes a 14-kb mRNA consisting of 79 exons and is characterized by its huge intron size; the largest, intron 44, is 248 kb long. In addition, eight alternative promoters that are activated in a tissue-specific manner have been identified. Each tissue-specific exon 1 under the control of a cryptic promoter is spliced correctly

* Correspondence: mmatsuo@reha.kobegakuin.ac.jp

¹Department of Pediatrics, Graduate School of Medicine, Kobe University, Chuo, Kobe 6500017, Japan

Full list of author information is available at the end of the article

to one of the downstream authentic *dystrophin* exons, producing a tissue-specific dystrophin isoform [4,5]. Furthermore, alternative splicings of some exons lead to the production of additional isoforms of the tissue-specific transcripts [6-8]. Remarkably, 14 cryptic exons that resemble authentic exons in terms of length and splice site strength, but are very rarely if ever spliced, have been reported within the huge introns [9,10].

Splicing is the process that removes introns from pre-mRNA, and is performed in the spliceosome, a ribonucleoprotein assembly. The spliceosome is one of the most complex cellular machineries, comprising approximately 150 proteins and five small nuclear RNAs (snRNAs U1, U2, U4, U5, and U6) [11]. Three sites, the splice donor site (5'ss), the splice acceptor site (3'ss), and the branch point sequences are the core splice site signals that are present in every intron. Despite the high potential for errors, the splicing process appears to occur with high fidelity, implying the widespread involvement of additional transcript features. These exonic elements are conventionally classified as exonic splicing enhancers (ESEs) or silencers (ESSs) and they function to promote or inhibit inclusion of the exon in which they reside, respectively. These splicing regulatory elements function by recruiting *trans*-acting splicing factors that activate or suppress splice site recognition or spliceosome assembly by various mechanisms [12]. AOs that induce skipping of *dystrophin* exons have been mainly designed against ESEs of target exons to hamper recruitment of splicing factors [2,3].

Recently, we reported that a small chemical enhances skipping of mutated *dystrophin* exon 31 in a sequence-specific manner, not altering the splicing of other *dystrophin* exons [13]. It was strongly suggested that *dystrophin* exons have their own splicing regulatory systems. However, the characteristics of the splicing regulatory systems of individual *dystrophin* exons are not well understood.

Decision trees are classifiers that predict class labels for data items and can make very accurate predictions [14,15]. They have been used to establish an integrated method that is one of the best available ways to find genes in the human genome [15].

Here, we identified a novel cryptic exon in intron 11 of the *dystrophin* gene in a DMD patient, and constructed decision trees to discriminate authentic exons from cryptic exons. Finally, we categorized 77 authentic exons into five groups based on indexes of their splicing regulatory factors. From the decision tree, we suppose that one group of exons showing high density for ESE sequences are a good target for exon skipping therapy.

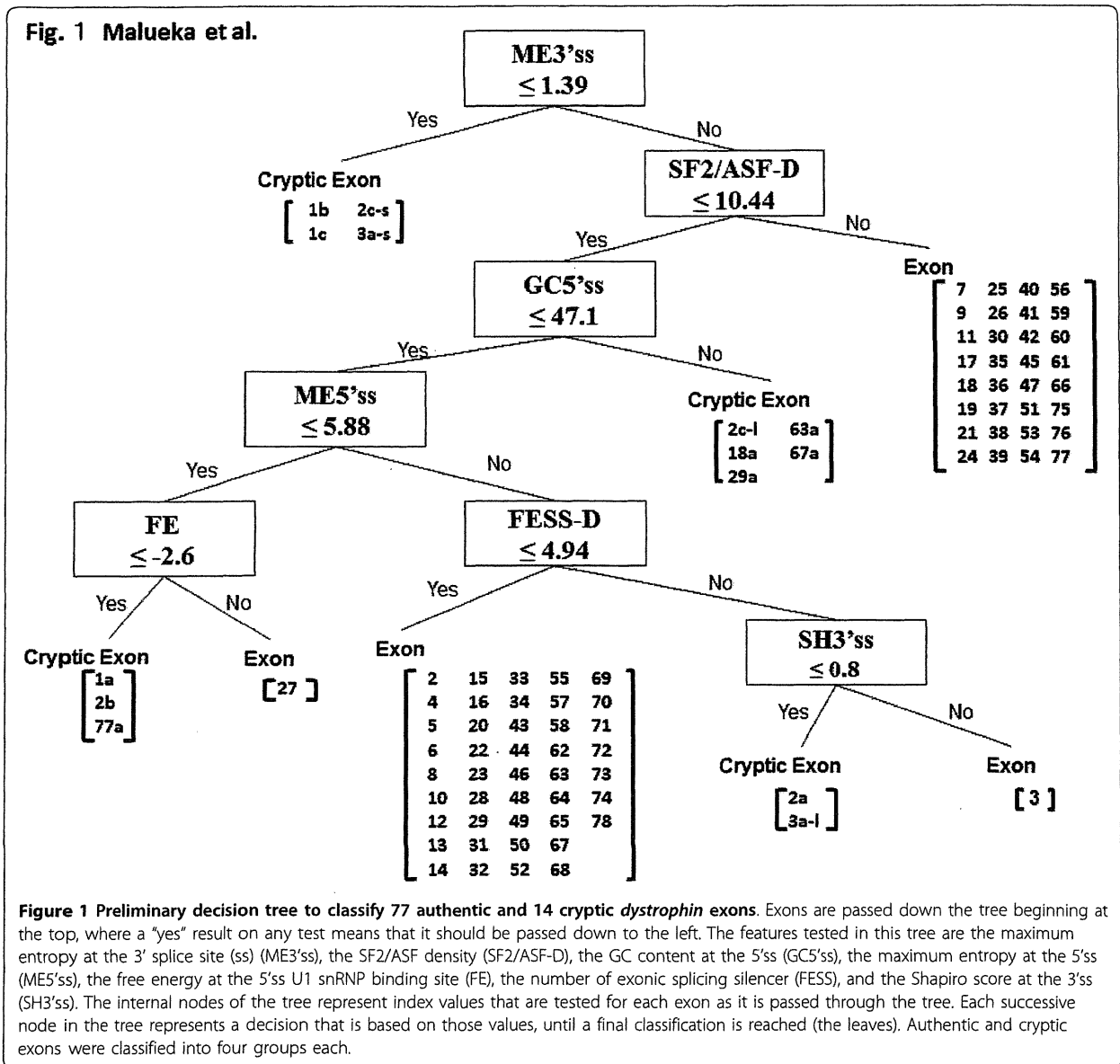
Results

To examine the splicing regulatory factors that characterize particular exons, we constructed decision trees classifying authentic from cryptic exons using indexes of

splicing regulatory factors as decision markers. Cryptic exons within the *dystrophin* gene resemble authentic exons in terms of length and splice site strength, but are very rarely if ever spliced [10]. Therefore, analyzing the exon recognition parameters of these exons compared to the authentic *dystrophin* exons can give insight into which splicing regulatory elements actually play a critical role in the splicing of *dystrophin* exons. The goal of the decision tree was to determine the critical parameters that provided the most accurate categorization of authentic exons and cryptic exons. A preliminary decision tree was constructed to discriminate 77 authentic exons from 14 known cryptic exons. To classify these exons, we used 26 indexes that have been reported as important in proper splicing (see Methods). The decision tree system output a simple data structure (Figure 1). The decision tree revealed that the strength of the 3'ss calculated by maximum entropy (ME3'ss) was the first splitting variable, with a cut-off point of 1.39. At this node, four cryptic exons were classified into one group. At the next node, SF2/ASF-D was used as the splitting variable, with a cut-off point of less than 10.44, generating a group of 32 authentic exons. In this way, seven nodes were used to separate clearly the 77 authentic exons from the 14 cryptic exons. The authentic exons were categorized into four groups, comprising 43, 32, 1, and 1 exon, respectively; similarly, the cryptic exons were also categorized into four groups.

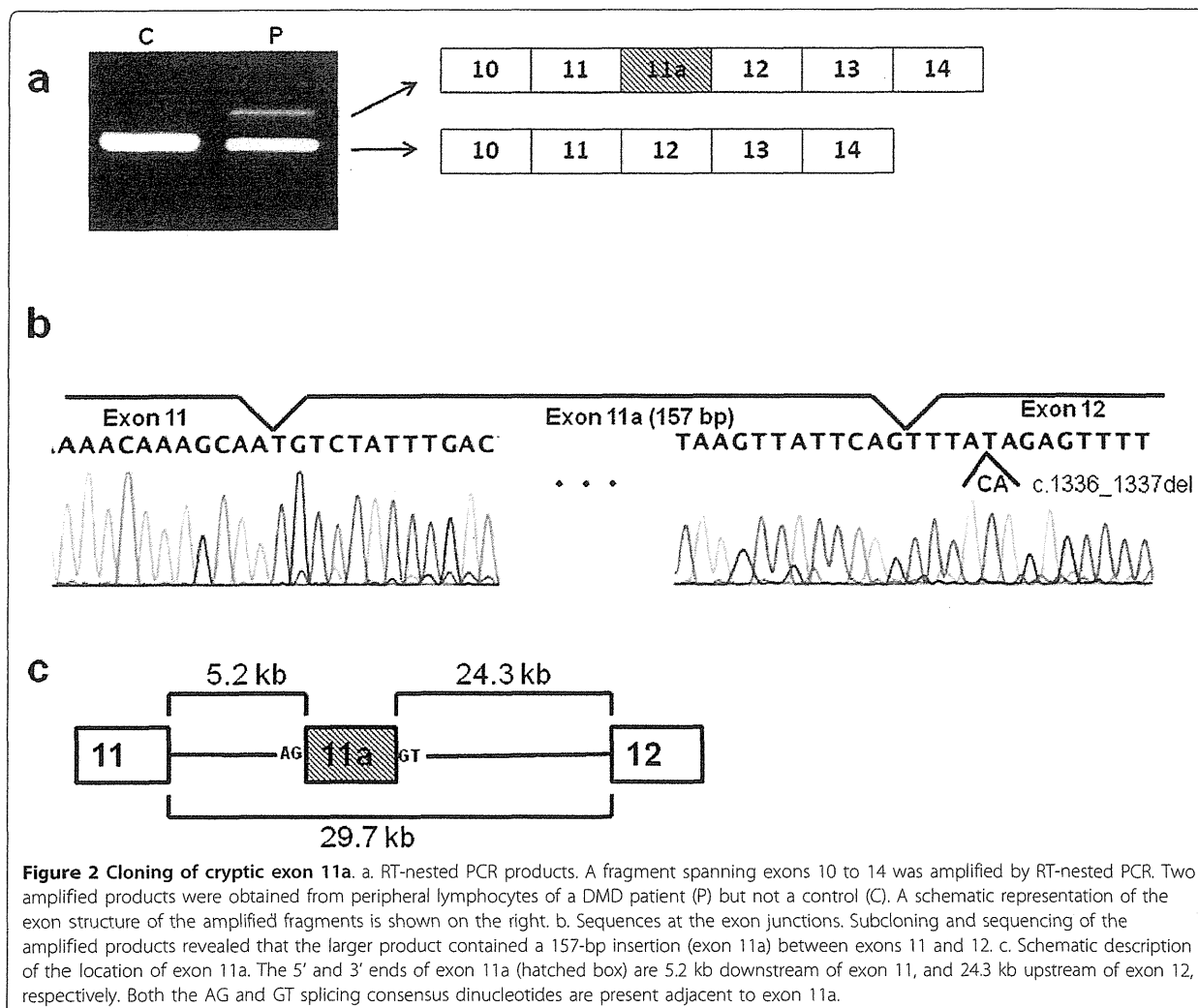
We evaluated the decision tree by analyzing a novel cryptic exon 11a that was identified in this study. Exon 11a was found inserted into *dystrophin* mRNA in one DMD case who had a two-nucleotide (CA) deletion at the 5th and 6th nucleotides of exon 12 (c.1336_1337del). RT-nested PCR amplification of a fragment spanning exons 10 to 14 from this individual revealed two products: one corresponding to the normal size and the other larger than expected (Figure 2). Sequencing of the two products revealed that the normal-size band had the predicted exon content with the two-nucleotide deletion, and the larger product contained a 157-bp unknown insertion between exons 11 and 12. The inserted 157-bp sequence was identical to part of the 29.7-kb-long intron 11. The corresponding intronic sequence was present 5244 nt downstream of the 3' end of exon 11 and 24,278 nt upstream of exon 12 (Figure 2c). The inserted sequence maintained the AG and GT dinucleotide consensus sequences for splicing acceptor and donor sites at either end. Although the sequences flanking the inserted sequence were examined in this individual, no nucleotide change was found. The inserted sequence was named exon 11a.

When we tested exon 11a on our decision tree, it was classified not as a cryptic exon but as a real exon. Therefore, we decided that this tree was not suitable to classify



exon 11a. We reconstructed the decision tree, including exon 11a as an additional cryptic exon using the same 26 indexes that were used to construct the first decision tree. The final tree, which had eight nodes, was able to separate all 15 cryptic exons from all 77 authentic exons with 100% accuracy (Figure 3). ME3'ss was the first splitting variable, with a cut-off point of 1.39. At this node, four cryptic exons were classified into one group (group a; Figure 3). SF2/ASF-D was the splitting variable at the second node, with a cut-off point of 10.53. A group of 32 exons (group A) was categorized on its "no" branch. Group A, therefore, was characterized by a high-density of ESEs recognized by SF2/ASF. At the third node,

ME5'ss with a cut-off point of 5.58 was used to divide the data into two subnodes. On the "yes" branch, FE with a cut-off point of -2.6 split the group: one exon group consisting of five cryptic exons (group b) and one exon group containing only exon 27 (group B). On the other, "no" branch, FESS-D with a cut-off point of 4.45 divided the group into two subnodes. On the "yes" branch, GC5'ss with a cut-off point of 46.8 was used at the fifth node. One big exon group (group C), comprising 42 exons, formed the "yes" branch. On the other, "no" branch, SH3'ss with a cut-off point of 0.79 was used at the sixth node, generating a cryptic exon group (group c) and a group containing two authentic exons (exons 70



and 78; group E). On the “no” branch at the fourth node, SIZE with a cut-off point of 144 was next used, categorizing the final data into one exon group consisting of exons 2 and 3 (group D) and one group containing four cryptic exons (group d).

We evaluated our re-constructed tree using nine disease-causing exons that have been identified in the *dystrophin* mRNAs of dystrophinopathy patients. These pathological exons are created by the effect of deep-intron single nucleotide mutations. These mutations confer the characteristics of an authentic exon on a portion of intronic sequence, and thus the intron segment becomes recognized as an exon during the splicing process. Therefore, the decision tree should categorize these pathological exons as authentic exons. Remarkably, all nine pathological exons were correctly categorized into one of the four exon groups: four in group A, three in group C, and one in each of groups D and E.

Therefore, we consider our decision tree a suitable classifier of pathological *dystrophin* exons.

Seven indexes were used in common by both the preliminary and re-constructed trees. However, their values in each tree were different, except for one (≤ 1.39 ME3'ss score at the first node). When we compared the two trees, we found that three exons (2, 70, and 78) and two exons (40 and 56) were re-categorized into small groups (groups E and D) and the largest group (group C), respectively. This indicated slight changes in the categorizing pathways between the two trees. The cryptic exons were ultimately categorized into four groups, and exon 11a belonged to group d, which comprises exons 1a, 2a, and 3a-l. It was interesting that group D was categorized by a SIZE score of ≤ 114 (Figure 3). This indicates that a small exon can be real if the splicing silencer density is high.

Among the 26 indexes used as candidates for the decision markers, only eight were used in the tree. Furthermore, five

Fig. 3 Malueka et al.

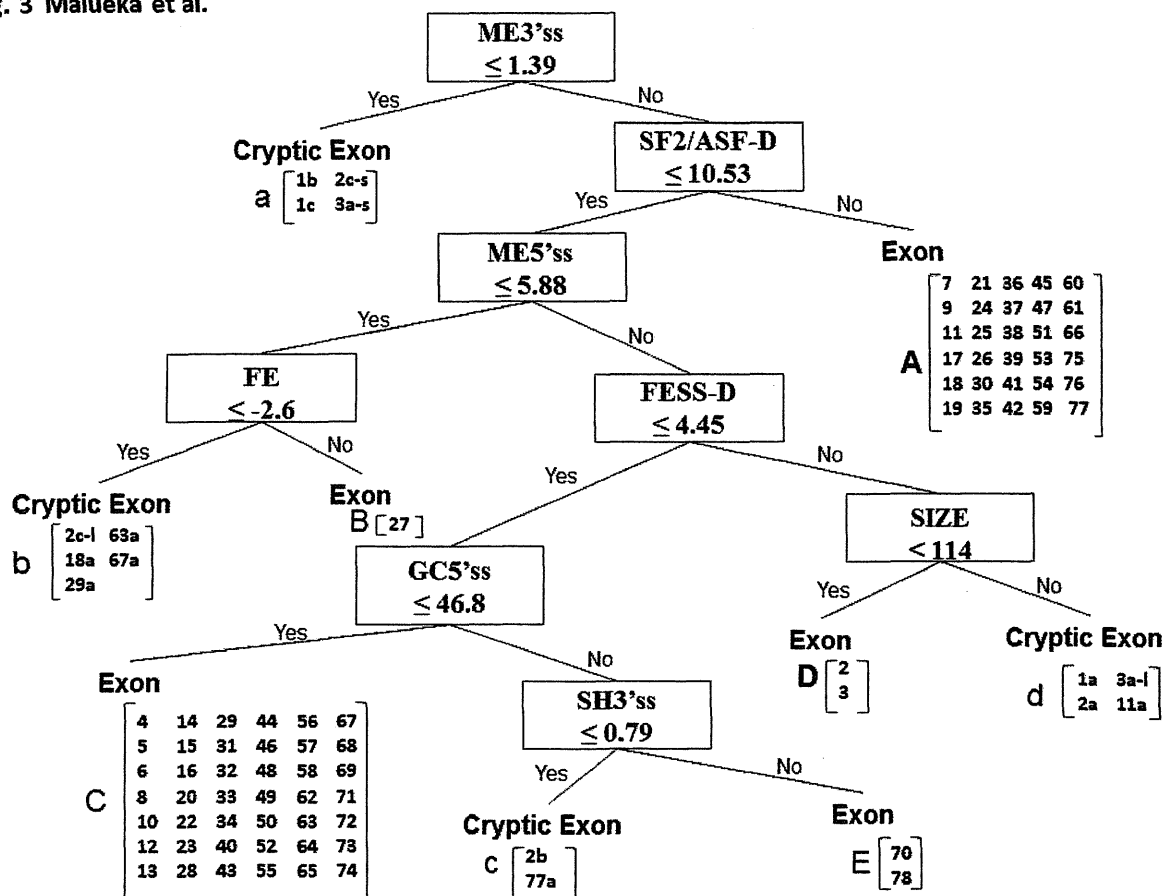


Figure 3 Final decision tree to classify 77 authentic and 15 cryptic *dystrophin* exons. The structure of the tree is as described for Figure 1. The features tested in this tree include seven indexes used in the preliminary tree (Figure 1) and one additional index, exon size (SIZE). This tree classified the authentic *dystrophin* exons into five groups (groups A to E), containing 30, 1, 42, 2, and 2 exons, respectively. The cryptic exons were classified into four groups (groups a to d).

of these eight were those determining the strength of splice sites (Table 1). This indicates that recognition of the splicing sites is very important for proper splicing. Group E (exons 70 and 78) was categorized through six nodes; the final one discriminating it from group c (cryptic exons 2b and 77a). For this categorization, the strength of the 3'ss was tested twice using different indexes (ME3'ss and SH3'ss). This indicates the importance of the 3'ss strength relative to the 5'ss strength in *dystrophin* splicing. Indexes of ESE density (SF2/ASF-D) and ESS density (FESS-D) were also used in the tree. This indicates that both the ESE and ESS densities are important in the splicing process.

Next, we used the tree to characterize exons that are subjected to specific splicing patterns. First, we marked exons known to be alternatively spliced (data not shown). However, they were found in all five groups. This indicates that no particular *cis*-elements predispose to alternative splicing. Second, we examined exons that were prone to

splicing errors caused by intra-exon mutation. We have previously reported nonsense mutation-induced exon skipping in seven exons (exons 14, 15, 17, 27, 41, 42, and 70) [16]. Among these, only four (exons 17, 27, 41, and 42) cause disruption of an ESE. Remarkably, exons 17, 41, and 42 were categorized in group A, while all those that do not disrupt an ESE were categorized in other groups. This indicates that, for exons in group A, nonsense mutation-induced exon skipping is caused by disruption of an ESE, while for exons in other groups, nonsense mutation-induced exon skipping is caused by other factors. Third, we examined exons that splice to tissue-specific exon 1 (exons 2, 30, 45, 56, and 63). We did not identify any particular categorization characteristics for these exons.

Discussion

The *dystrophin* gene produces many mRNAs because it is so complex: it contains 79 exons, huge introns, eight

Table 1 Summary of categorization of exons by the decision tree

Group	Number of Exon		Indexes							
			ME3'ss	SF2/ASF-D	ME5'SS	FE	FESS-D	GC5'ss	SH3'ss	SIZE
A	30		H	H	-	-	-	-	-	-
B	1		H	L	L	H	-	-	-	-
C	42		H	L	H	-	L	L	-	-
D	2		H	L	H	-	H	-	-	L
E	2		H	L	H	-	L	H	H	-

Results of categorization of 77 *dystrophin* exons are tabulated. Indexes determining the strength of splice sites are marked bold. H and L represent high and low at each node, respectively. Bars indicate not applicable.

tissue-specific promoters, 15 cryptic exons, and generates many alternatively spliced products. It has been reported that there is a gradient in exon and intron definition at the level of pre-mRNA splicing [17]. For example, efficient use of intrinsically weak cryptic splice sites in exons is facilitated by a higher-than-average density of ESSs and a high density of SF2/ASF ESE motifs [17]. This suggests that each exon has a specific splicing regulatory mechanism in which particular splicing factors could compensate for the lack of other splicing factors. It is important to understand more about the splicing regulatory mechanisms for each exon of the *dystrophin* gene because treatment with exon skipping is a promising therapeutic approach.

Decision trees have been used in many kinds of classifications, such as for the management of Parkinson's disease, disease severity profiling, large-scale proteomic studies, and microarray data classification [14,18]. Here, we classified 77 authentic and 15 cryptic *dystrophin* exons into five groups based on their splicing regulatory factor characteristics. In this study, we needed to modify our preliminary decision tree (Figure 1) based on data from the novel exon 11a (Figure 2). This suggested a possibility that the tree needs further modification when another cryptic exon is identified in the future construction. However, we believe any modifications, if necessary, will be minimal as exon 11a was additionally analyzed (Figure 3). Moreover, our trees were constructed using indexes obtained only from *cis*-elements; further modification incorporating information from *trans*-elements may further improve the reliability of the tree.

We used our decision tree to determine the parameters that are most useful for the discrimination of authentic and cryptic *dystrophin* exons. We separated the two categories completely using eight variables: (1) strength of the 3'ss measured by maximum entropy score; (2) density of ESEs particularly SF2/ASF; (3) strength of the 5'ss measured by maximum entropy score; (4) free energy in U1-snRNP binding to the 5'ss; (5) density of ESSs as identified by FESS-D; (6) GC content at the 5'ss; (7) exon size; and (8) strength of the 3'ss as measured by the Shapiro score. In other words, the strength of the 3'ss consensus

sequence as indicated by maximum entropy score is the most critical, as all authentic *dystrophin* exons must have a maximum entropy score of more than 1.39 at the 3'ss. The next most important deciding factor is ESE density, particularly for SF2/ASF; almost half (30) of *dystrophin* exons can be classified based on this parameter and 3'ss strength only. These findings are in line with those of a previous report, which showed that exons that are skipped because of splice site mutations have a weaker 3'ss and a lower-than-average density of ESEs [17].

Our exon categorization suggested that there are at least five different patterns of splicing regulatory mechanisms for *dystrophin* pre-mRNAs (Table 1). At the extreme, one group contained only one exon (exon 27; group B), implying that this exon has a unique splicing regulatory mechanism. In fact, alteration of exon 27 splicing has been reported for two nonsense mutations, resulting in a mild phenotype [19]. Exon 27 may be particularly vulnerable to splicing alterations caused by intra-exonic mutations.

On the other hand, two groups (A and C) contained 72 out of 77 exons, suggesting that most *dystrophin* exons are under similar splicing control. Group A, consisting of 32 exons, was characterized by a high density of ESEs, particularly SF2/ASF (more than 10.53) (Figure 3) (Table 1), indicating that ESE density plays a critical role in the splicing of these exons. Disruption of ESEs for exons in this group is more likely to cause exon skipping compared with disruption of ESEs in the other exon groups. Nonsense mutations within exons 17, 41, and 42, all belonging to group A, have been shown to induce exon skipping [16]. Accordingly, the nonsense mutations in these three exons disrupt an ESE sequence and thus cause exon skipping, while nonsense mutations that induce skipping of exons in other groups do not disrupt an ESE [16]. However, not all ESE-disrupting nonsense mutations identified in these exons induce exon skipping [16]. It may be necessary to consider *trans*-elements to explain these differences, along with the strength of splicing factor binding to the ESEs and their positions in the pre-mRNA secondary structure.

Intronic pseudo-exons that look like perfect exons, maintaining splicing consensus sequences, are now under intensive investigation [10,20-22]. There is

evidence that inclusion of many of these sequences in mRNAs is actively inhibited because of the presence of intrinsic defects, the presence of silencer elements, or the formation of an inhibitory RNA secondary structure [23]. We categorized the cryptic exons into four groups (Figure 3), indicating a heterogeneous contribution of splicing factors required for their activation.

Group C contained 42 exons and was separated at the fifth node by a score of less than 46.8% GC content at the 5'ss. More than half of the *dystrophin* exons belonged to this group and are therefore presumed to be subject to similar splicing regulation. From the decision tree we can see that this group is characterized also by an ESS density ≤ 4.45 (FESS-D), indicating that splicing silencer signals play a critical role in the splicing of these exons. Indeed, we showed previously that nonsense mutation-induced exon skipping of exon 31 was caused by the creation of a splicing silencer-binding site for hnRNP A1 [13]. It has been reported that negative elements play important roles in distinguishing real splicing signals from the vast number of false-positive splicing signals [9].

Because the *dystrophin* gene is so complex, it has the potential to produce many alternatively spliced transcripts that are translated into protein variants [24]. However, studies on alternative splicing are limited [6-8]. In one study, Sironi et al. identified 16 alternative transcripts and examined splicing regulatory factors such as the 3' and 5' consensus values and exonic splicing enhancer scoring matrices; however, no reasonable explanation of the alternative splicing was identified [7]. This is consistent with our findings that the cryptic exons fell into five different groups. It is possible that alternative splicing does not rely completely on specific sequence elements and is regulated by *trans*-acting factors.

Interest in the splicing regulation of *dystrophin* pre-mRNA was first aroused when exon skipping caused by a small intra-exonic deletion was identified in a DMD patient [25]. Subsequently it was revealed that the deleted region was an ESE sequence within exon 19 [1] and that AOs against this ESE successfully induced the skipping of exon 19 [26]. This has led to the establishment of exon-skipping therapy [27]. In this study, we demonstrated that exon 19 was categorized into group A, characterized by a high ESE density; hence it is reasonable that AOs against the exon 19 ESE would induce exon skipping efficiently. Currently exon skipping is recognized as the most promising way to express dystrophin in DMD. The main targets for exon skipping therapy are exons 44, 45, 51, and 53 [28] and AOs against exon 51 are now in phase II or III clinical trials [2,3]. Because exons 45, 51, and 53 belong to group A, we would expect AOs against the ESEs in these exons to work well.

Conclusions

The decision tree categorized the 77 authentic *dystrophin* exons into five groups. Our classification may help to establish the strategy for exon skipping therapy for DMD.

Methods

Indexes of splicing regulatory factors

The sequences of 14 known cryptic exons, and nine disease-causing mutations were obtained from our previous report [10] and the literature [29,30]. Twenty six indexes of splicing regulatory factors of each exon were obtained as described below (Table 2).

Splice site strength

Splice site strength was determined in three ways: Shapiro's splicing probability matrix scores at the 5'ss and the 3'ss were calculated using published formulae [17,31] (SH5'ss and SH3'ss, respectively). Maximum entropy scores at the 5'ss and the 3'ss were obtained using online tools available at http://genes.mit.edu/burgelab/maxent/Xmaxentcan_scoreseq.html (ME5'ss and ME3'ss, respectively) [32]. Information content values at the 5'ss and the 3'ss were obtained using the Delila server at <https://splice.uwo.ca/> (Ri5'ss and Ri3'ss, respectively) [33].

Free energy of U1 snRNA binding to the 5'ss

U1 snRNA binding to constitutive splice sites has lower free energy than that of its binding to alternatively spliced exon splice sites [34]. Free energy was analyzed using the SROOGLE server at <http://sroogle.tau.ac.il/> (FE) [35].

Numbers and densities of ESEs and ESSs

The number of ESEs in each exon was calculated using the prediction algorithm at <http://genes.mit.edu/burgelab/rescue-ese/> (RESE). The number of ESSs was calculated using two algorithms: at <http://genes.mit.edu/fas-ess/> (FESS) and <http://cubweb.biology.columbia.edu/pesx/> (PESS) [17,36-38]. To calculate the densities of ESSs/ESEs, the RESE, FESS, and PESS numbers were divided by the sequence length (in nucleotides) and this figure was multiplied by 100 [17] to give the RESE-D, FESS-D, and PESS-D scores.

The numbers of binding sites for the four SR proteins (SF2/ASF, SRp40, SC35, and SRp55) were obtained using ESEfinder (v. 3.0) available at <http://rulai.cshl.edu/cgi-bin/tools/ESE3/ese finder.cgi?process=home> (SF2/ASF-N, SRp40-N, SC35-N, and SRp55-N) [39]. The density of SR protein-binding sites (SF2/ASF-D, SC35-D, SRp40-D, SRp55-D) was obtained by dividing each number by its nucleotide length then multiplying it by 100 [17].

Minimum free energy value of pre-mRNA secondary structure and GC content around splice sites

Minimum free energy values of the pre-mRNA secondary structure at the 5'ss and 3'ss were obtained using the free energy minimization program RNAfold <http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi> using the 70

Table 2 Indexes of splicing regulatory factors

No	Features	Symbol	Reference
1	5' splice site strength (Shapiro score)	SH5'ss	[31]
2	3' splice site strength (Shapiro score)	SH3'ss	[31]
3	5' splice site strength (maximum entropy)	ME5'ss	[32]
4	3' splice site strength (maximum entropy)	ME3'ss	[32]
5	5' splice site strength (information content/Ri)	Ri5'ss	[33]
6	3' splice site strength (information content/Ri)	Ri3'ss	[33]
7	U1 SnRNA binding free energy	FE	[35]
8	ESE density (RESCUE ESE/RESE)	RESE-D	[36]
9	ESE density (PESE)	PESE-D	[38]
10	ESS density (FAS-ESS/FESS)	FESS-D	[37]
11	ESS density (PESS)	PESS-D	[38]
12	SF2/ASF number	SF2/ASF-N	[39]
13	SF2/ASF (IgM/BRCA1) number	SF2/ASF (IgM, BRCA1)-N	[39]
14	SRp40 number	SRp40-N	[39]
15	SC35 number	SC35-N	[39]
16	SRp55 number	SRp55-N	[39]
17	SF2/ASF score density	SF2/ASF-D	[39]
18	SF2/ASF (IgM/BRCA1) score density	SF2/ASF (IgM-BRCA1)-D	[39]
19	SC35 score density	SC35-D	[39]
20	SRp40 score density	SRp40-D	[39]
21	SRp55 score density	SRp55-D	[39]
22	5' splice site pre-mRNA secondary structure free energy	RSS5'ss	[40]
23	3' splice site pre-mRNA secondary structure free energy	RSS3'ss	[40]
24	5' splice site GC content	GC5'ss	[40]
25	3' splice site GC content	GC3'ss	[40]
26	The number of nucleotides in exon	SIZE	

nucleotides both up- and down-stream of each splice site (RSS5'ss and RSS3'ss, respectively) [40,41].

GC content around splice sites has been reported to affect splicing [40]. Thus, the percentage of GC content for each 70 nucleotides both up- and down-stream of the 5'ss and the 3'ss was calculated (GC5'ss and GC3'ss, respectively) [40].

Exon size

The number of nucleotides in each exon was also taken into account (SIZE).

Construction of decision trees

The C4.5 algorithm was used to construct decision trees. The C4.5 decision tree algorithm is an approach for pattern recognition and data mining [42,43] and was developed by Quinlan [44,45]. The algorithm uses information gain, which is an entropy-based criterion in information theory for decision tree construction. In this study, the conditions were as follows: (1) at least one object was to be contained in each branch, (2) the pruning confidence level was set to 100%, and (3) the iterative mode was used to avoid a local optimum.

We used 92 out of 94 data points (79 exons and 15 cryptic exons) for decision tree construction. We excluded two exons, exons 1 and 79, because they lacked some feature values.

Dystrophin mRNA analysis

A 1-year-old Japanese boy (KUCG 952) without any family history of neuromuscular disorders was referred to Kobe University Hospital (Kobe, Japan) because of a high serum creatine kinase level (37,110 IU/l). DMD was tentatively diagnosed, and mutation in the *dystrophin* gene was analyzed after obtaining informed consent from his parents. *Dystrophin* mRNA expressed in his peripheral lymphocytes was analyzed as described previously [10,25]. A fragment spanning exons 10 to 14 was amplified by reverse transcription (RT)-nested PCR using two sets of primers. The outer primer set comprised a forward primer, 1A 5'-TTTTTATCGCTGCCTTGATATAACA-3' and a reverse primer, 1B 5'-ACTCTGCAACACAGCTTCTGAG-3'; the inner primer set comprised a forward primer, 1E 5'-TTGCAAGCACAAGGAGAGATT-3' and a reverse primer, c14r 5'-ACGTTGCCATTTGAGAAGGAT-3'. The

amplified fragments were resolved by agarose gel electrophoresis. Sequencing of the amplified products was performed by subcloning sequencing as described previously [46].

The mutation study was approved by our ethical committee and mutation analysis of the *dystrophin* gene was done after obtaining the informed consent from the parents of the patient.

Abbreviations

DMD: Duchenne muscular dystrophy; AO: Antisense oligonucleotide; ESE: Exonic splicing enhancer; ESS: Exonic splicing silencer; ME3'ss: Maximum entropy of 3' splice site; SF2/ASF-D: SF2/ASF score density; ME5'SS: Maximum entropy of 5' splice site; FE: Free energy of U1 SnRNA binding; FE55-D: Fluorescence-activated screen-for exonic splicing silencers density; GC5'ss: GC content of 5' splice site; SH3'ss: Shapiro score of 3' splice site; SIZE: Size of exon.

Acknowledgements

This work was supported by a Grant-in-Aid for Scientific Research (B), and a Grant-in-Aid for Exploratory Research from the Japan Society for the Promotion of Science, a Health and Labour Sciences Research Grant for Research on Psychiatric and Neurological Diseases and Mental Health, and a research grant for Nervous and Mental Disorders from the Ministry of Health, Labour and Welfare, Japan.

Author details

¹Department of Pediatrics, Graduate School of Medicine, Kobe University, Chuo, Kobe 6500017, Japan. ²Division of Medical Informatics and Bioinformatics, Kobe University Hospital, Chuo, Kobe 6500017, Japan. ³Department of Clinical Pharmacy, Kobe Pharmaceutical University, Higashinada, Kobe 6588558, Japan. ⁴Department of Medical Rehabilitation, Faculty of Rehabilitation, Kobegakuin University, 518 Arise, Ikawadani, Nishi, Kobe 651-2180, Japan.

Authors' contributions

RGM carried out the molecular genetic studies, participated in the bioinformatics analysis and drafted the manuscript. YT carried out the bioinformatics analysis. MY, HA, TL and EKD participated in the clinical and genetic analysis. YT participated in the design of the study. MM conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 18 January 2012 Accepted: 31 March 2012
Published: 31 March 2012

References

1. Takeshima Y, Nishio H, Sakamoto H, Nakamura H, Matsuo M: Modulation of *in vitro* splicing of the upstream intron by modifying an intra-exon sequence which is deleted from the *dystrophin* gene in *dystrophin* Kobe. *J Clin Invest* 1995, **95**:515-520.
2. Aartsma-Rus A: Antisense-mediated modulation of splicing: Therapeutic implications for duchenne muscular dystrophy. *RNA Biol* 2010, **7**(4):453-461.
3. Lu QL, Yokota T, Takeda S, Garcia L, Muntoni F, Partridge T: The Status of Exon Skipping as a Therapeutic Approach to Duchenne Muscular Dystrophy. *Mol Ther* 2011, **19**:9-15.
4. Ahn AH, Kunkel LM: The structural and functional diversity of *dystrophin*. *Nat Genet* 1993, **3**:283-291.
5. Nishio H, Takeshima Y, Narita N, Yanagawa H, Suzuki Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M: Identification of a novel first exon in the human *dystrophin* gene and of a new promoter located more than 500 kb upstream of the nearest known promoter. *J Clin Invest* 1994, **94**:1037-1042.
6. Feener CA, Koenig M, Kunkel LM: Alternative splicing of human *dystrophin* mRNA generates isoforms at the carboxy terminus. *Nature* 1989, **338**:509-511.
7. Sironi M, Cagliani R, Pozzoli U, Bardoni A, Comi GP, Giorda R, Bresolin N: The *dystrophin* gene is alternatively spliced throughout its coding sequence. *FEBS Lett* 2002, **517**:163-166.
8. Kubokawa I, Takeshima Y, Ota M, Enomoto M, Okizuka Y, Mori T, Nishimura N, Awano H, Yagi M, Matsuo M: Molecular characterization of the 5'-UTR of retinal *dystrophin* reveals a cryptic intron that regulates translational activity. *Mol Vis* 2010, **16**:2590-2597.
9. Sun H, Chasin LA: Multiple splicing defects in an intronic false exon. *Mol Cell Biol* 2000, **20**:6414-6425.
10. Zhang Z, Habara Y, Nishiyama A, Oyazato Y, Yagi M, Takeshima Y, Matsuo M: Identification of seven novel cryptic exons embedded in the *dystrophin* gene and characterization of 14 cryptic *dystrophin* exons. *J Hum Genet* 2007, **52**:607-617.
11. Valadkhan S: Role of the snRNAs in spliceosomal active site. *RNA Biol* 2010, **7**:345-353.
12. Matlin AJ, Clark F, Smith CW: Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* 2005, **6**:386-398.
13. Nishida A, Kataoka N, Takeshima Y, Yagi M, Awano H, Ota M, Itoh K, Hagiwara M, Matsuo M: Chemical treatment enhances skipping of a mutated exon in the *dystrophin* gene. *Nat Commun* 2011, **2**:308.
14. Kingsford C, Salzberg SL: What are decision trees? *Nat Biotechnol* 2008, **26**:1011-1013.
15. Allen JE, Majoros WH, Pertea M, Salzberg SL: JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biol* 2006, **7**(Suppl 1:S9):1-13.
16. Nishiyama A, Takeshima Y, Zhang Z, Habara Y, Tran TH, Yagi M, Matsuo M: *Dystrophin* nonsense mutations can generate alternative rescue transcripts in lymphocytes. *Ann Hum Genet* 2008, **72**:717-724.
17. Královicová J, Vorechovsky I: Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Res* 2007, **35**:6399-6413.
18. David MP, Concepcion GP, Padlan EA: Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. *BMC Bioinforma* 2010, **11**:79.
19. Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M: Disruption of the splicing enhancer sequence within exon 27 of the *dystrophin* gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. *J Clin Invest* 1997, **100**:2204-2210.
20. Buratti E, Chivers M, Královicová J, Romano M, Baralle M, Krainer A, 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* 2007, **35**:4250-4263.
21. Dhir A, Buratti E: Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *FEBS J* 2010, **277**:841-855.
22. Vorechovsky I: Transposable elements in disease-associated cryptic exons. *Hum Genet* 2010, **127**:135-154.
23. Dhir A, Buratti E, van Santen MA, Luhrmann R, Baralle FE: The intronic splicing code: multiple factors involved in ATM pseudoexon definition. *EMBO J* 2010, **29**:749-760.
24. Wei B, Jin JP: Troponin T isoforms and posttranscriptional modifications: Evolution, regulation and function. *Arch Biochem Biophys* 2011, **505**:144-154.
25. Matsuo M, Masumura T, Nishio H, Nakajima T, Kitoh Y, Takumi T, Koga J, Nakamura H: Exon skipping during splicing of *dystrophin* mRNA precursor due to an intraexon deletion in the *dystrophin* gene of Duchenne muscular dystrophy Kobe. *J Clin Invest* 1991, **87**:2127-2131.
26. Pramono ZA, Takeshima Y, Alimsardjono H, Ishii A, Takeda S, Matsuo M: Induction of exon skipping of the *dystrophin* transcript in lymphoblastoid cells by transfecting an antisense oligodeoxynucleotide complementary to an exon recognition sequence. *Biochem Biophys Res Commun* 1996, **226**:445-449.
27. Takeshima Y, Yagi M, Wada H, Ishibashi K, Nishiyama A, Kakumoto M, Sakaeda T, Saura R, Okumura K, Matsuo M: Intravenous infusion of an antisense oligonucleotide results in exon skipping in muscle *dystrophin* mRNA of Duchenne muscular dystrophy. *Pediatr Res* 2006, **59**:690-694.
28. Takeshima Y, Yagi M, Okizuka Y, Awano H, Zhang Z, Yamauchi Y, Nishio H, Matsuo M: Mutation spectrum of the *dystrophin* gene in 442 Duchenne/

- Becker muscular dystrophy cases from one Japanese referral center. *J Hum Genet* 2010, **55**:379-388.
29. Yagi M, Takeshima Y, Wada H, Nakamura H, Matsuo M: Two alternative exons can result from activation of the cryptic splice acceptor site deep within intron 2 of the *dystrophin* gene in a patient with as yet asymptomatic dystrophinopathy. *Hum Genet* 2003, **112**:164-170.
 30. Beroud C, Carrie A, Beljord C, Deburgrave N, Llense S, Carelle N, Peccate C, Cuisset JM, Pandit F, Carre-Pigeon F, et al: Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. *Neuromuscul Disord* 2004, **14**:10-18.
 31. 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.
 32. Yeo G, Burge CB: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2003, **11**:377-394.
 33. Nalla VK, Rogan PK: Automated splicing mutation analysis by information theory. *Hum Mutat* 2005, **25**:334-342.
 34. Ast G: How did alternative splicing evolve? *Nat Rev Genet* 2004, **5**:773-782.
 35. Schwartz S, Hall E, Ast G: SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Res* 2009, **37**(Web Server):W189-192.
 36. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: Predictive identification of exonic splicing enhancers in human genes. *Science* 2002, **297**:1007-1013.
 37. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB: Systematic identification and analysis of exonic splicing silencers. *Cell* 2004, **119**:831-845.
 38. Zhang XH, Chasin LA: Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 2004, **18**:1241-1250.
 39. 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.
 40. Zhang J, Kuo CC, Chen L: GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* 2011, **12**:90.
 41. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL: The Vienna RNA websuite. *Nucleic Acids Res* 2008, **36**(Web Server):W70-74.
 42. Han J, Kamber M, Pei J: *Data Mining: Concepts and Techniques* Waltham, Massachusetts: Morgan Kaufmann Publishers; 2011.
 43. Witten IH, Frank E: *Practical Machine Learning Tools and Techniques* Waltham, Massachusetts: Morgan Kaufmann Publishers; 2005.
 44. Quinlan JR: *C4.5: Programs for Machine Learning* Waltham, Massachusetts: Morgan Kaufmann Publishers; 1993.
 45. Quinlan JR: Improved Use of Continuous Attributes in C. *J Artif Intell Res* 1996, **4**:77-90.
 46. Surono A, Takeshima Y, Wibawa T, Ikezawa M, Nonaka I, Matsuo M: Circular *dystrophin* RNAs consisting of exons that were skipped by alternative splicing. *Hum Mol Genet* 1999, **8**:493-500.

doi:10.1186/1471-2156-13-23

Cite this article as: Malueka et al.: Categorization of 77 *dystrophin* exons into 5 groups by a decision tree using indexes of splicing regulatory factors as decision markers. *BMC Genetics* 2012 **13**:23.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



