Biochemical and Biophysical Research Communications **267,** 321–328 (2000)
doi:10.1006/bbrc.1999.1962, available online at http://www.idealibrary.com on IDEAL®

A Novel Cryptic Exon in Intron 2 of the Human Dystrophin Gene Evolved from an Intron by Acquiring Consensus Sequences for Splicing at Different Stages of Anthropoid Evolution

Zacharias Aloysius Dwi Pramono,* Yasuhiro Takeshima,† Agus Surono,* Takafumi Ishida,‡ and Masafumi Matsuo*,[1]

**Division of Genetics, International Center for Medical Research, Kobe University School of Medicine, Kobe 650-0017, Japan; †Department of Pediatrics, Kobe University School of Medicine, Kobe 650-0017, Japan; and ‡Department of Anthropology, Graduate School of Science, University of Tokyo, Tokyo 113-0033 Japan*

Received August 22, 1999

The dystrophin gene, which is mutated in Duchenne muscular dystrophy, is thus the largest human gene. A full spectrum of splicing of the dystrophin transcript has not been elucidated yet, though more than 10 alternative splicings have been identified in the 5′ region of the dystrophin gene. In this study, two novel dystrophin transcripts containing a 140-nucleotide insertion precisely between exons 2 and 8 or between exons 2 and 18 were identified in skeletal muscle. The genomic region corresponding to and surrounding this 140-nucleotide sequence was sequenced to reveal that the insertion possessed a branch point and both acceptor and donor splice site consensus sequences perfectly. Therefore, the 140-bp insertion sequence was considered to be a novel exon. The novel exon was mapped to intron 2 and was designated exon 2a. Reverse-transcription PCR screening for transcripts containing exon 2a in 12 human tissues revealed its presence in 3 of them, including skeletal muscle. Phylogenetic studies disclosed that exon 2a evolved from intron DNA by the progressive acquisition of nucleotide substitutions in ancestral hominoids.

The human dystrophin gene, which is defective in patients with Duchenne or Becker muscular dystrophy (DMD/BMD), spans approximately 3000 kb of the X-chromosome and encodes a 14-kb transcript consisting of 79 exons (1, 2). Consequently, more than 99% of the gene sequence is composed of introns. The extraordinary nature of the dystrophin gene with respect to size and the number of exons points to several potential problems in the production and processing of dystrophin transcripts (3, 4).

Pre-mRNA processing to remove introns utilizes conserved sequence at the 5′ and 3′ splice sites, and branch consensus site of the intron (5). However, these conserved sequences alone are not sufficient to explain the splicing process which are preferentially involved in tissue specific alternative splicing and it has been suggested that other *cis*- or *trans*-acting elements work together to determine splicing sites. Several dystrophin isoforms produced by alternative splicing have been reported to be produced in a tissue- or development- specific manner (6–10). Alternative splicing of the 5′ region of the dystrophin gene was identified during a molecular study of dystrophinopathy (8, 10, 11) and the possibility that alternative splicing could modify the clinical phenotype of DMD/BMD by editing translational reading frame has been discussed (11–13).

Two exon sequences that were transcribed from intron of the dystrophin gene have been reported; Exon X, which is inserted into between exons 1 and 2, has been identified in intron 1 of the dystrophin gene (14). Though exon X is detectable in half of the ectopic dystrophin transcripts expressed in peripheral blood lymphocytes, it is not detectable in skeletal muscle. The physiological role of exon X remains unknown, since it is in an ectopic transcript and contains an in-frame nonsense codon. Recently, a sequence derived from intron 11 was identified in dystrophin cDNA from human cardiac and skeletal muscle tissues. Since this transcript maintains the open reading frame, it was proposed to be part of the unknown gene that overlaps the human dystrophin gene (15). These data suggested

[1] To whom correspondence should be addressed at Division of Genetics, International Center for Medical Research, Kobe University School of Medicine, 7-5-1 Kusunokicho, Chuo, Kobe 650-0017, Japan. Fax: +81-78-382-5719. E-mail: matsuo@kobe-u.ac.jp.

0006-291X/00 $35.00

TABLE 1

Sequences of Primers Used in This Study

Target region	Forward primer	Reverse primer
	cDNA	
Exon 1–exon 2a	1A (TTTTTATCGCTGCCTTGATTATACA)	R1 (TTCAGAAGGATGAATAATGTGC)
Exon 1–exon 5	1A	C5R (TGCCAGTGGAGGATTATATTCCAA)
5′RACE		
1st PCR	(Gibco BRL Life Technologies)	R1
2nd PCR	(Gibco BRL Life Technologies)	R2 (CTTCTTTACTTGAACAGGAGTGA)
	Genomic DNA	
5′ flanking intron of exon 2a		
1st PCR	C1 (Takara Shuzo Co.)	R1
2nd PCR	C2 (Takara Shuzo Co.)	R2
3′ flanking intron of exon 2a		
1st PCR	F1 (TTGGTGGAGCCTCACTCCTGTT)	C1
2nd PCR	F2 (CTGTTCAAGTAAAGAAGACACAT)	C2
Exon 2a	g2aF (TAGAGTTATCCTAGAGAGGTGG)	g2aR (TCAGTGCATCATCCAGCAAT)
Exon 2–exon 2a	g2F (AGATGAAAGAGAAGATGTTCAAAAG)	g2aR
Exon 2a–exon 3	g2aF	g3R (CAGGCGGTAGAGTATGCCAAATGAAAATCA)

that cryptic exons that are expressed in a tissue- or development specific manner might be located within other introns of the dystrophin gene.

We recently identified six novel alternative splicing patterns in the 5′ region of the dystrophin gene (3) in addition to the six already known ones, indicating that exon usage in the 5′ region of the dystrophin gene is highly complicated. During study on alternative splicing, an unknown sequence was identified to be inserted into dystrophin transcript. Here we report the identification of a novel exon located in intron 2 of the dystrophin gene that is activated in a limited number of alternative splicing reactions. Phylogenetic analysis showed that the novel exon evolved from an intron in the ancestral hominoid by acquiring consensus sequences necessary for splicing.

METHODS

PCR and DNA sequencing. PCR was performed essentially as described previously (16). To amplify genomic DNA fragments stretching from exon 2 to exon 2a or from exon 2a to exon 3 of the dystrophin gene, an LA PCR kit (Takara Shuzo Co., Kyoto, Japan) was used. Primer sequences used in this study are listed in Table 1.

For DNA sequencing, amplified products were separated by electrophoresis in low-melting agarose gels. Bands of amplified products were cut out and the DNA was purified. The purified DNA was subcloned into pT7 vector (Novagen, Inc., Madison, WI) and the inserted DNA was sequenced using an automatic DNA sequencer (model 373A, Perkin–Elmer Applied Biosystem Inc., Norwalk, CT) as described previously (17).

Analysis of dystrophin transcript. Total RNA was isolated from human skeletal muscle as previously described (16). A muscle sample was obtained at the time of an orthopedic operation after obtaining informed consent. Fragments encompassing exons 1 to 11 or 18 of dystrophin mRNA were analyzed by reverse-transcription (RT) PCR as described previously (3).

To examine the efficiency of exon 2a activation in different tissues, fragments stretching from exon 1 to exon 2a and from exon 1 to exon 5 (Fig. 1a) were amplified from cDNA prepared from total RNA of twelve human tissues (Sawady Co. Tokyo, Japan). cDNA corresponding to 0.125 mg of each of the RNA samples was subjected to PCR amplification. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) cDNA was amplified as a control (18) and the yield of each of the amplified products was measured by densitometry (Densitograph; ATTO Co., Tokyo, Japan).

To analyze sequences flanking the inserted sequence in transcripts, 5′ RACE for rapid amplification of cDNA ends (5′ RACE) (CLONTECH Laboratories, Inc., Palo Alto, CA) was performed as suggested by the supplier, using two primers complementary to novel exon (Table 1).

Analysis of genomic DNA. Human genomic DNAs were isolated by the previously-described method (16) from lymphocytes of a normal individual and three DMD patients. Non-human anthropoids DNA samples from cotton-top tamarin (*Saguinus oedipus*), night monkey (*Aotus trivirgatus*), Japanese macaque (*Macaca fuscata*), gibbon (*Hylobates lar*), orangutan (*Pongo pygmaeus*), gorilla (*Gorilla gorilla*) and chimpanzee (*Pan troglodytes*), and mouse (JcL:ICR strain) were obtained from blood samples or cell lines (19).

Nested PCR was employed to clone the genomic regions flanking the inserted sequence using a commercially-available cassette (20). The Sau3A1 cassette (Takara Shuzo Co., Kyoto, Japan) was ligated to Sau3A1- digested genomic DNA according to the protocol suggested by the supplier. Then, the 5′ and 3′ flanking regions were cloned by PCR amplification using cassette primers C1 or C2 (Takara Shuzo Co., Kyoto, Japan) and primers on or complementary to exon 2a (Table 1). The amplified products were sequenced.

Once the intron sequences that flank exon 2a had been sequenced, a set of primers (g2aF, corresponding to the intron region upstream of exon 2a and g2aR complementary to the intron region downstream of exon 2a (Table 1) were designed to amplify a 304 bp fragment of genomic DNA (Fig. 1b). A region encompassing exon 2a was amplified from three DMD patients (cases A, B, and C) in order to map exon 2a in the dystrophin gene. Cases A, B and C had deletions stretching from the 5′ end to exon 2, from exon 3 to exon 7 and from exon 5 to exon 11, respectively. Genomic DNA fragments encompassing exon 2 and exon 3 were amplified as previously described (21).

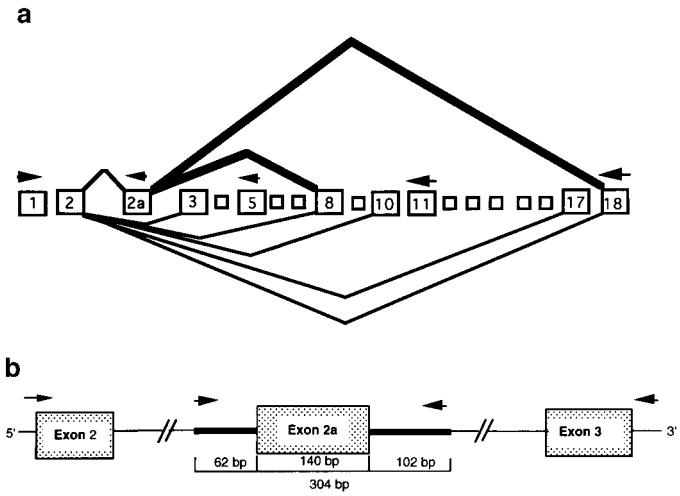

FIG. 1. Schematic representation of the examined regions. (a) Exons and splicing patterns starting from exon 2 of the dystrophin gene. Five patterns of splicing consisting of one canonical and 4 alternative splicing reactions are indicated as diagonal lines originating from exon 2 beneath the schematic representation of the gene (3). The novel splicing patterns that contain exon 2a are presented above the gene as bold diagonal lines. Boxes and numbers in boxes indicate exons and exon numbers, respectively. Arrows indicate the location and orientation of primers that were used to examine the dystrophin transcript. (b) Genomic region flanking the novel exon 2a. Novel 140 bp exon 2a was identified between exons 2 and 3. 62 and 103 bp of upstream and downstream introns, respectively, were analyzed. Open boxes indicate exons, horizontal lines indicate introns and bold lines indicate intron sequences that were clarified in this study by cassette PCR. Horizontal arrows indicate the hybridization sites and orientations of primers that were used to amplify the corresponding genomic region of the dystrophin gene. Brackets indicate the size of each segment of gene. The figure is not drawn to scale.

RESULTS

During studies on alternative splicing in the 5′ region of the dystrophin gene, we detected two unidentified or unexpected transcripts among subcloned fragments of RT-PCR products. First; when a fragment stretching from exon 1 to exon 11 was amplified from the dystrophin transcript in skeletal muscle (Fig. 1a), amplified products in addition to that of the expected size (1230 bp) were obtained (Fig. 2a). Sequencing of a subcloned 674-bp amplified fragment revealed an unidentified sequence inserted precisely between exon 2 and exon 8 (Fig. 3a), in addition to skipping of exon 9 (22). Second; additional product of 319 bp was obtained when the region spreading from exon 1 to 18 was amplified (Fig. 2b). And sequence analysis revealed an unidentified sequence of 140 bp inserted precisely between exons 2 and 18 (Figs. 3b). The sequences of these two unidentified 140-bp insertions were found to be same (Fig. 3). A BLAST search failed to reveal significant homologue of this 140-bp DNA insert.

The fact that it was inserted precisely between exon 2 and two different exons (8 and 18) led us to speculate that the 140 nucleotide sequence could be present downstream of exon 2 in dystrophin genomic DNA. To test the genomic origin of the 140-bp sequence, we attempted to amplify it from normal genomic DNA using primers specific for the 140-nucleotide sequence. A product of the expected size was obtained (data not shown), indicating that the normal human genome contains the sequence. However, this result did not reveal whether or where the sequence is located in the dystrophin gene.

At this point, we considered two mechanisms that could have given rise to the novel insertion: (i) use of an alternative splice donor site downstream of intron 2 and (ii) use of a cryptic novel exon present in one of introns between exons 2 and 8. The first possibility can be discounted because of the lack of homology between the novel sequence and the known sequence of 5′ end of intron 2 (GenBank Accession No. Wo2). Therefore, the second possibility was examined by characterizing the sequences flanking the 140 bp segment. Cloning and sequencing of both upstream and downstream regions of the inserted sequence revealed novel 62- and 102-bp long sequences, respectively (Fig. 1b). As a result a total of 304 bp of genomic region was clarified (Fig. 4) (GenBank Accession No. AB016196). Remarkably, the AG and GT dinucleotides that are absolutely conserved at splice acceptor and donor sites of all introns, respectively, were identified immediately adjacent to the 5′ and 3′ ends of the 140-bp sequence. Shapiro's scores for splice acceptor and donor sites were 99 and 79, respectively (23). Furthermore, the sequence CTTTGAC, with a perfect match to the branch point consensus sequence (YNYURAC, CY = C or T; R = G or A; N = any base), was identified 35 nt upstream of the novel

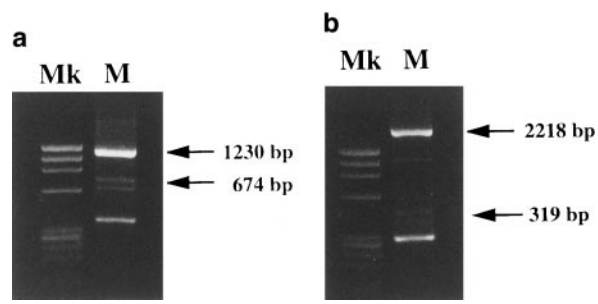

FIG. 2. Amplification of cDNA fragments stretching from exon 1 to exon 11 or from exon 1 to exon 18 of the dystrophin transcript. (a) Amplification of the region encompassing exons 1 to 11 led to the appearance of several small products in addition to that of the expected size (1230 bp). All small products except a 674-bp product were from alternative splicing products consisting of authentic exons (3). (b) Amplification of the region encompassing exons 1 to 18 also resulted in the appearance of several small products in addition to the expected product (2218 bp). All small products except a 319-bp product were from alternative splicing products consisting of authentic exons (3). M refers to skeletal muscle cDNA used a template for the PCRs. Mk represents a size marker (*Hinc*II-digested ϕ X174 phage DNA; Toyobo Co., Osaka, Japan).

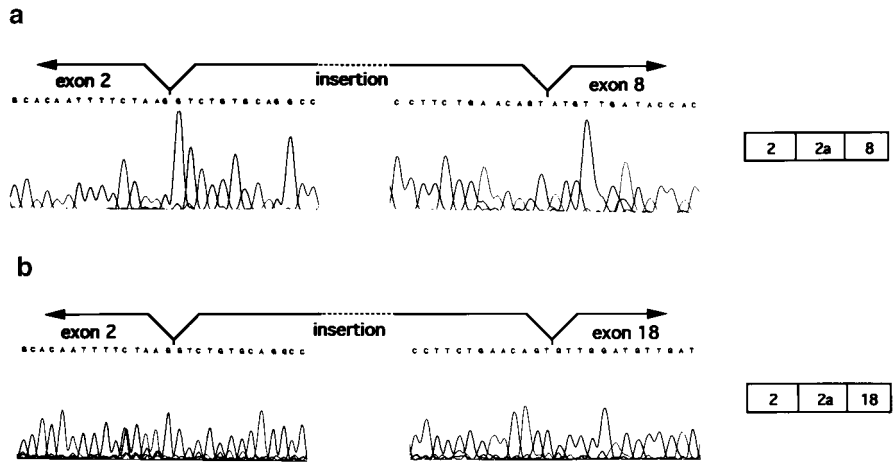

FIG. 3. Sequence of the insertion located downstream of exon 2. The sequences at the boundaries of the 140 bp long inserted sequence and its flanking exons are shown. A schematic representation of the organization of the exons in the amplified fragments is shown at the right of the sequence. (a) Sequence analysis of the 674-bp product revealed that 8 of 10 subcloned fragments matched with those of alternative splicing product (3), but two clones had insertion of 140 bp that did not correspond to a known dystrophin gene exon. The 3′ terminal sequence of exon 2 (CTAAG) is joined precisely to the inserted sequence (GTCTG), and at its 3′ end (ACAGT) the unknown sequence joined precisely to the 5′ end of the sequence of exon 8 (ATGTT). (b) The sequence of 7 clone of 10 subcloned 319-bp fragment matched with those of alternative splicing product (3) and three clones contained the same unidentified 140-bp segment as the 674-bp fragment amplified in a. The 3′ terminal sequence of exon 2 (CTAAG) is joined precisely to the inserted sequence (GTCTG) and the 3′ end of the insertion (ACAGT) is joined precisely to the 5′ end sequence of exon 18 (GTTGG).

sequence (Fig. 4). A polypyrimidine tract was also identified between the putative branch point and the putative splice acceptor site. Since the 140-bp inserted sequence exhibited all of the characteristics typical of a genomic exon and was inserted between dystrophin exons, we refer it as the novel exon.

Since the novel exon was inserted between either exons 2 and 8 or exons 2 and 18, we suspected that it must be located within one of the introns between 2 and 7 of the dystrophin gene. To map the novel exon further, we performed a deletion analysis by using genomic DNA prepared from DMD patients with different deletion mutations (Fig. 5). The 304-bp fragment encompassing the novel exon could be amplified from DNA samples of cases B or C who had deletions of exons from 3 to 7 or from 5 to 11, respectively (Fig. 5), indicating that the novel exon is located outside of the region flanked by exons 3 to 11. However, the 304-bp fragment could not be amplified from the DNA of case A with a deletion stretching from the 5′ end of the dystrophin gene to exon 2 (Fig. 5), indicating that the novel exon is located within this region of the dystrophin gene. These results led us to conclude that the novel exon is really located in intron 2 and named as exon 2a. The location of exon 2a was further examined in the 150 kb long intron 2 (24, 25) by amplifying genomic regions spanning from exon 2 to exon 2a or from exon 2a to exon 3 (Fig. 1b). However, no product was obtained in either case (data not shown), suggesting that exon 2a is located too far from either of these exons to be PCR amplified.

The protein coding capacity of exon 2a was examined by analyzing translational reading frame of a transcript retaining exon 2a between either exons 2 and 8 or exons 2 and 18. Neither transcripts maintained the open reading frame of the dystrophin gene. Furthermore, since exon 2a does not contain an in-frame ATG codon after the last termination codon, it is unlikely that transcripts containing exon 2a would direct translation of a novel product. Therefore, these transcripts would only be expected to allow reinitiation of translation at a downstream ATG codon (26).

To analyze the tissue-specific usage of exon 2a, a fragment extending from exon 1 to exon 2a was amplified from cDNA prepared from total RNA of 12 different tissues. A product containing exon 2a was obtained from 3 tissues such as skeletal muscle, small intestine and colon tissues (Fig. 6). A fragment stretching from exon 1 to exon 5 of the dystrophin transcript was also amplified from these 3 tissues. Parallel studies failed to reveal any significant differences in the yield of amplified reference cDNA (GAPDH) from any of these tissues (Fig. 6). These results indicated the tissue specificity of incorporation of exon 2a.

We also considered the possibility that exon 2a could be a part of a gene embedded in the dystrophin gene, a situation that would be similar to that reported in the factor VIII gene (27). To examine this possibility, the 5′ region of the transcript containing exon 2a was cloned as described in the methods section and fragments were obtained from human muscle cDNA. Sequencing of the product disclosed that the fragment consisted of

```
                          1                                  *******                                  50
Human                :    tagagttatc     ctagagaggt     ggctttgacc     tgttagaagc     tatgttagtg
Chimpanzee           :    a---------     ----------     ----------     ----------     ----------
Gorilla              :    ----------     ----------     ----------     ----------     ----------
Orangutan            :    ----------     ----------     ----------     ----------     ----------
Gibbon               :    ----------     ----------     ----------     ----------     ----------
Japanese macaque     :    ----------     ----------     ----------     -------t--     ----------
Night monkey         :    ----------     ----------     --------ta     --------tg     -----c----
Cotton-top tamarin:       ----------     ----------     --------t-     -------gt-     -----c----
                          51                                                                  101
Human                :    ttttttgttc     agGTCTGTGC     AGGCCAAGGT     TATGGCCTAG     CTGAGAAGAG
Chimpanzee           :    ----------     ----------     ----------     ----------     ----------
Gorilla              :    ----------     ----------     ----------     ----------     ----------
Orangutan            :    ----------     --------A-     ----------     ----------     ----------
Gibbon               :    ----------     --------A-     ----------     ----------     ------G---
Japanese macaque     :    ------a---     -a------A-     --------A-     -G--------     ---------A
Night monkey         :    ----------     -a-C--T-A-     ----------     -G--------     ----------
Cotton-top tamarin:       ----------     -a----T-A-     ----------     -GC------A     ------CA--
                          101                                                                 150
Human                :    GGCTCAGAGG     AGCCTAGCTA     AAGTTTGGTC     AAGGGGAGCG     TCTTGGTGGA
Chimpanzee           :    ----------     ----------     ----------     ----A-----     ----------
Gorilla              :    ----------     ----------     ----------     ----A-----     ----------
Orangutan            :    ----------     ----------     ----------     ----A-----     ----------
Gibbon               :    ----------     ----------     ----------     ----A-----     ----------
Japanese macaque     :    ------C---     -----G----     -C--------     ----A---A-     ----------
Night monkey         :    ----------     --TG-G----     ----------     ---AA--CA-     ----------
Cotton-top tamarin:       ---C------     -----G----     ----------     ---AA--CA-     ----------
                          151                                                                 200
Human                :    GCCTCACTCC     TGTTCAAGTA     AAGAAGACAC     ATTATTCATC     CTTCTGAACA
Chimpanzee           :    --------AT     ----------     ----------     ----------     ----------
Gorilla              :    ---------T     ----------     ----------     ---G------     ----------
Orangutan            :    ---------T     ----------     ----------     ----------     ----------
Gibbon               :    ---------T     ----------     ----------     ----------     ----------
Japanese macaque     :    ---------T     C---------     ----------     -------T--     ----------
Night monkey         :    ---------T     ----------     ----------     -----A-T--     ---G----A-
Cotton-top tamarin:       ---------T     ----------     ----------     ---G---T--     ---G------
                          201                                                                 250
Human                :    GTgtaagtcc     atttgtcatt     cgattgcctt     ttgtttacta     aggactagct
Chimpanzee           :    ----------     ----------     ----------     ----------     ----------
Gorilla              :    ----------     ----------     ----------     ----------     ----------
Orangutan            :    ----------     ----------     ----------     ----------     ----g-----
Gibbon               :    ----------     ----------     ----------     ------g---     ----g-----
Japanese macaque     :    --------g-     ----------     ------g---     --------c-     ----------
Night monkey         :    ACa-------     ----------     ----------     -------t--     ----------
Cotton-top tamarin:       CAt-------     ----------     ----------     -------t--     ----------
                          251                                                                 300
Human                :    aaatcatcgg     ttgagacttg     ctagtagagc     atagttgctg     gatgatgcac
Chimpanzee           :    -------t--     ----------     ----------     ----------     ----------
Gorilla              :    --------t-     ----------     ----------     ----------     ----------
Orangutan            :    --------t-     ----------     ----------     ----------     ----------
Gibbon               :    --------t-     ----------     ----------     ----------     ----------
Japanese macaque     :    --------t-     --c--g----     -c--------     ----------     ----------
Night monkey         :    --------t-     ----------     ----------     ----------     ----------
Cotton-top tamarin:       --------t-     ----------     ----------     ----------     ----------
                          301
Human                :    gtga
Chimpanzee           :    ----
Gorilla              :    ----
Orangutan            :    ----
Gibbon               :    ----
Japanese macaque     :    ----
Night monkey         :    ----
Cotton-top tamarin:       ----
```

FIG. 4. Genomic nucleotide sequence of the inserted sequence and its flanking introns and its alignment with sequences from anthropoids. The sequence of exon 2a encompassing of the human genome is shown at the top (GenBank Accession No. AB016196). The 140 nt of exon 2a are shown in uppercase letters, and the 62 and 102 nt upstream and downstream of the exon 2a sequence, respectively, are shown in lower case letters. Absolutely-conserved AG and GT dinucleotides are present at the boundaries between exon 2a and its flanking regions. Asterisks indicate the branch point consensus sequence. Sequence differences detected in amplified anthropoid DNA are indicated below the sequence determined for amplified human DNA. Sequences are arranged according to evolutionary lineage. Dots indicate identity to the human sequence.

exon 1, exon 2, and exon 2a (data not shown), indicating that exon 2a was incorporated into transcript transcribed from the muscle promoter of the dystrophin gene. This rules out the possibility that exon 2a is part of the transcript of another gene.

The phylogeny of exon 2a was analyzed by sequencing genomic DNA from anthropoids and mouse. We attempted to amplify the corresponding region from DNA samples obtained from 7 species of anthropoids (chimpanzee, gorilla, orangutan, gibbon, Japanese ma-

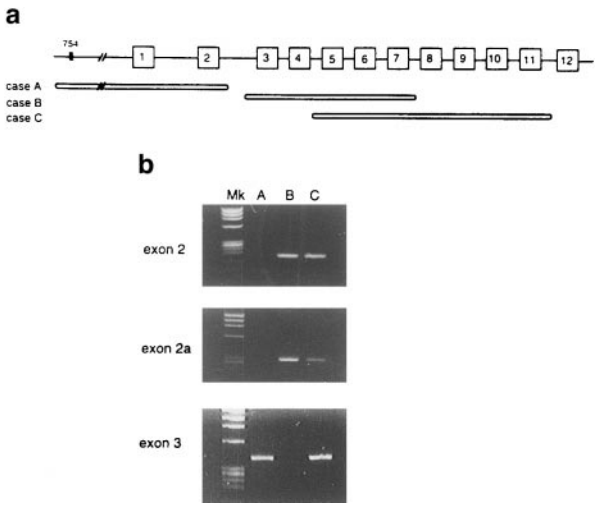

FIG. 5. Mapping of exon 2a in the dystrophin gene. The genomic region encompassing exon 2a was amplified from DMD patients who have deletions in the 5′ region of the dystrophin gene. (a) Schematic description of the positions of their deletions. Case A had a deletion from the 5′ end of the gene to exon 2. Cases B and C had deletions of exons from 3 to 7 and 5 to 11, respectively. Boxes with numbers indicate exons. Columns indicate deleted region in respective cases. Thin lines indicate introns. The figure is not drawn to scale. (b) Amplified products from genomic DNA were shown. Fragments encompassing exons 2, 2a, or 3 were examined by amplifying DNA from the three DMD cases. Case A had amplified no products from exons 2 and 2a but product from exon 3, while case B had product from exon 2 and 2a, but no product from exon 3. Case C had amplified products from all three examined exons. Mk represents a size marker (*Hinc*II-digested ϕ X174 phage DNA; Toyobo Co., Osaka, Japan).

caque, night monkey and cotton-top tamarin) and from mouse by using the same primer set as that used to amplify exon 2a of human genomic DNA. A product of the same size as that obtained from human DNA was obtained from the 7 anthropoids DNA, but no product was obtained from mouse DNA (data not shown). Therefore, it was concluded that the region encompassing exon 2a is conserved among anthropoids. All of the amplified products obtained from anthropoids were sequenced and compared with the sequence of the region amplified from human DNA (Fig. 4). Higher substitution frequencies were found in the samples amplified from night monkey and cotton-top tamarin DNA 26 (8.6%) and 27 (8.8%) nucleotide substitutions, respectively, compared to the human sequence. A phylogenetic tree of these sequences, constructed by the neighbor-joining method (28) using the CLUSTAL W program (29), corresponded to the accepted phylogeny of anthropoids (Fig. 7). No significant homologies were found when data banks were screened using the BLAST program with the sequence of the most ancient exon 2a-encompassing region, that of the cotton-top tamarin.

Interestingly, we could not identify all of the three consensus sequences in the DNA amplified from night monkey and cotton-top tamarin, which therefore correspond to purely intron DNA (Fig. 7). However, a perfect branch point consensus sequence was created in the region amplified from Japanese macaque DNA as a results of a C to T substitution at the 34th nucleotide from the splice site (Fig. 7). Similarly, a splice donor site consensus sequence was created in Japanese macaque DNA as a result of the introduction of a G at nucleotide +1, whereas the night monkey and cotton-top tamarin have A and T nucleotides, respectively, at +1 position (Figs. 4 and 7). Furthermore, a splice acceptor site consensus sequence was created in the gibbon DNA sequence as a result of the A to G substitution at nucleotide −1 (Figs. 4 and 7). These consensus sequences are conserved in gibbons, orangutans, gorillas, and chimpanzees as well as humans (Figs. 4 and 7). Therefore, we conclude that exon 2a evolved from intron DNA in ancestral hominoids as a result of the creation of consensus sequences for splicing and that it has been conserved in humans (Fig. 7).

DISCUSSION

In this report, a novel exon 2a was identified in intron 2 of the dystrophin gene (Fig. 3). This study reveals that consensus branch point and splice donor and acceptor sites in human exon 2a were created by

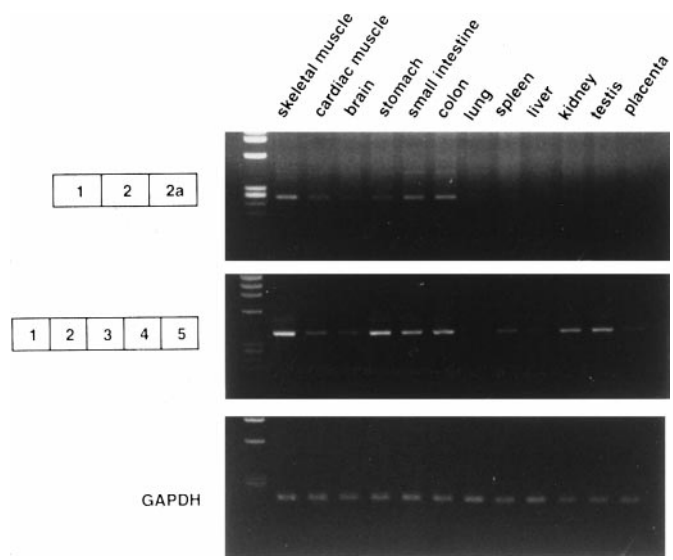

FIG. 6. RT-PCR products containing exon 2a. A cDNA fragment stretching from exon 1 to exon 2a was amplified from 12 different human tissues (upper panel) and, for comparison, cDNA fragments stretching from exon 1 to exon 5 of the dystrophin transcript (middle panel) and of GAPDH cDNA (lower panel) were also amplified. A product including exons 1 and 2a was obtained from skeletal muscle (lane 1), small intestine (lane 5) and colon (lane 6), as well as products corresponding to the dystrophin transcript from exon 1 to exon 5 (middle panel) and to the GAPDH transcript (lower panel). Mk represents a size marker (*Hinc*II-digested ϕ X174 phage DNA; Toyobo Co., Osaka, Japan).

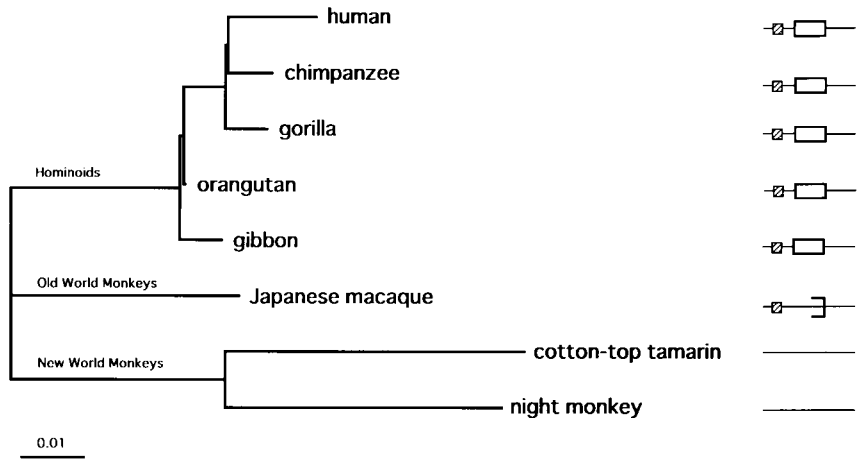

FIG. 7. Phylogenetic tree and evolution of exon 2a. A phylogenetic tree of the region encompassing exon 2a was constructed according to the genomic sequences of 8 anthropoids by the neighbor-joining method using the CLUSTAL W program. The evolution of exon 2a is depicted at the right. The shaded box indicates a branch point consensus sequence and the open box indicates exon 2a, while lines represent introns. The bracket indicates a consensus splice sequences for splice acceptor site.

nucleotide substitutions in intron 2 at different stages in anthropoid evolution (Figs. 4 and 7). In contrast to New World monkeys that have no consensus sequences for splicing, Old World monkeys and hominoids possess both branch point and splice donor site consensus sequences. This observation strongly suggests that these consensus sequences were created in the DNA of a common ancestor of Old World monkeys and hominoids. In addition, the splice acceptor consensus sequence evolved after divergence from the ancestral Old World monkeys (Fig. 7). As a result, exon 2a apparently acquired its consensus sequences for splicing early in hominoid evolution, before branching of the gibbon lineage occurred. Therefore, it was estimated that exon 2a evolved more than 17 million years ago (30). Our results support "introns early theory" since exon 2a was evolved from intron.

Exon 2a is the third example of part of a dystrophin gene intron sequence that is incorporated into mRNA. The first example is exon X in intron 1 (14) and another transcribed sequence has been identified in intron 11 (15). However, the physiological role of these novel transcripts remains unclear. Considering the huge size of the introns in the dystrophin gene, we suspect that more novel exons will be identified within them as analysis on the dystrophin gene is extended. It needs further study to confirm this supposition.

Exon 2a-containing transcripts occupy a very small proportion of dystrophin transcripts (Fig. 2). This is in contrast to exon X, which is incorporated between exons 1 and 2 of an authentic transcript that amounts to half of the dystrophin transcript produced in leukocytes (14). Even though Shapiro's probability scores for 5′ acceptor and 3′ donor splice sites are high and a perfect branch point sequence is present at the proper position in exon 2a, transcripts containing exons 2, 2a and 3 in tandem could not be detected (data not shown). It has been reported that different regulatory programs for splicing run concurrently within the same cell, suggesting that the production of different alternatively spliced pre-mRNAs is regulated by distinct programs that use different sets of cis-elements and trans-acting factors (31). Therefore, incorporation of exon 2a might be regulated in a very specific manner by a number of factors, including those that regulate alternative splicing of exon 2 to 8 or exon 2 to 18, since exon 2a was found to be inserted into these transcripts (Fig. 3).

Secondary or tertiary structures of mRNA precursors are among the factors thought to regulate splicing. However, analysis of the region encompassing exon 2a by Zuker's algorithm (32) failed to reveal a putative secondary structure that might affect splicing (33). A proof-reading mechanism has also been proposed to check the protein-coding capability of transcripts and to lead to a dramatic decrease in the levels of aberrant transcripts (34). And we assume that proof-reading would be expected to prevent the use of exon 2a because exon 2a does not contain an open reading frame.

ACKNOWLEDGMENTS

We thank Dr. A. Pugsley for advice and for critically reading the manuscript. We acknowledge Ms. N. Kageyama for secretarial help. This work was supported by grants from the Ministry of Education, Science, and Culture of Japan and Research Grant 8A-1 for Nervous and Mental Disorders from the Ministry of Health and Welfare of Japan.

REFERENCES

1. Ahn, A. H., and Kunkel, L. M. (1993) *Nature Genet.* **3,** 283–291.
2. Nishio, H., Takeshima, Y., Narita, N., Yanagawa, H., Suzuki, Y.,

Ishikawa, Y., Ishikawa, Y., Minami, R., Nakamura, H., and Matsuo, M. (1994) *J. Clin. Invest.* **94,** 1037–1042.

3. Surono, A., Takeshima, Y., Wibawa, T., Pramono, Z. A. D., and Matsuo, M. (1997) *Biochem. Biophys. Res. Commun.* **239,** 895–899.
4. Surono, A., Takeshima, Y., Wibawa, T., Ikezawa, M., Nonaka, I., and Matsuo, M. (1999) *Hum. Mol. Genet.* **8,** 493–500.
5. Green, M. R. (1986) *Annu. Rev. Genet.* **20,** 671–708.
6. Feener, C. A., Koenig, M., and Kunkel, L. M. (1989) *Nature* **338,** 509–511.
7. Chelly, J., Hamard, G., Koulakoff, A., Kaplan, J.-C., Kahn, A., and Berwald-Netter, Y. (1990) *Nature* **344,** 64–65.
8. Reiss, J., and Rininsland, F. (1994) *Hum. Mol. Genet.* **3,** 295–298.
9. Barbieri, A. M., Soriani, N., Ferlini, A., Michelato, A., Ferrari, M., and Carrera, P. (1996) *Eur. J. Hum. Genet.* **4,** 183–187.
10. Torelli, S., and Muntoni, F. (1996) *Hum. Genet.* **97,** 521–523.
11. Chelly, J., Gilgenkrantz, H., Hugnot, J., Hamard, G., Lambert, M., Recan, D., Akli, S., Cometto, M., Kahn, A., and Kaplan, J. (1991) *J. Clin. Invest.* **88,** 1161–1166.
12. Winnard, A. V., Klein, C. J., Coovert, D. D., Prior, T., Papp, A., Snyder, P., Bulman, D. E., Ray, P. N., McAndrew, P., King, W., Moxley, R. T., Mendell, J. R., and Burghes, A. H. M. (1993) *Hum. Mol. Genet.* **2,** 737–744.
13. Winnard, A. V., Mendell, J. R., Prior, T. W., Florence, J., and Burghes, H. M. (1995) *Am. J. Hum. Genet.* **56,** 158–166.
14. Roberts, R. G., Bentley, D. R., and Bobrow, M. (1993) *Hum. Mutat.* **2,** 293–299.
15. Ferlini, A., and Muntoni, F. (1998) *Biochem Biophys. Res. Commun.* **242,** 401–406.
16. Matsuo, M., Masumura, T., Nishio, H., Nakajima, T., Kitoh, Y., Takumi, T., Koga, J., and Nakamura, H. (1991) *J. Clin. Invest.* **87,** 2127–2131.
17. Hagiwara, Y., Nishio, H., Kitoh, Y., Takeshima, Y., Narita, N., Wada, H., Yokoyama, M., Nakamura, H., and Matsuo, M. (1994) *Am. J. Hum. Genet.* **54,** 53–61.
18. Dveksler, G. S., Basile, A. A., and Dieffenbach, C. W. (1992) *PCR Methods Appl.* **1,** 283–285.
19. Ishida, T., and Yamamoto, K. (1987) *J. Med. Primatol.* **16,** 359–371.
20. Shiga, N., Takeshima, Y., Sakamoto, H., Inoue, K., Yokota, Y., Yokoyama, M., and Matsuo, M. (1997) *J. Clin. Invest.* **100,** 2204–2210.
21. Beggs, A. H., Koenig, M., Boyce, F. M., and Kunkel, L. M. (1990) *Hum. Genet.* **86,** 45–48.
22. Roberts, R. G., Barby, T. F. M., Manners, E., Bobrow, M., and Bentley, D. R. (1991) *Am. J. Hum. Genet.* **49,** 298–310.
23. Shapiro, M. B., and Senapathy, P. (1987) *Nucleic Acids Res.* **15,** 7155–7174.
24. Coffey, A. J., Roberts, R. G., Green, E. D., Cole, C. G., Butler, R., Anand, R., Giannelli, F., and Bentley, D. R. (1992) *Genomics* **12,** 474–484.
25. Nobile, C., Marchi, J., Nigro, V., Roberts, R. G., and Danieli, G. A. (1997) *Genomics* **45,** 421–424.
26. Malhotra, S. B., Hart, K. A., Klamut, H. J., Thomas, N. S. T., Bodrug, E., Burghes, A. H. M., Borrow, M., Harper, P. S., Thompson, W., Ray, P. N., and Worton, R. G. (1988) *Science* **242,** 755–759.
27. Levinson, B., Kenwrick, S., Lakich, D., Hammonds, G., Jr., and Gitschier, J. (1990) *Genomics* **7,** 1–11.
28. Saitou, N., and Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
29. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
30. Pilbeam, D. (1986) *Am. Authropol.* **88,** 295–312.
31. Cooper, T. A., and Mattox, W. (1997) *Am. J. Hum. Genet.* **61,** 259–266.
32. Zuker, M., and Stiegler, P. (1981) *Nucleic Acids Res.* **9,** 133–148.
33. Matsuo, M., Nishio, H., Kitoh, Y., Francke, U., and Nakamura, H. (1992) *Biochem. Biophys. Res. Commun.* **182,** 495–500.
34. Dietz, H. C. (1997) *Am. J. Hum. Genet.* **60,** 729–730.