Quantification Analysis of 3′-Splice Signal Sequences in mRNA Precursors. Mutations and Exon Skipping in Rabbit β-Globin Gene

Yôichi IIDA
Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060
(Received February 18, 1994)

The signals which direct excision of introns from mRNA precursors in mammalian genes have been studied. The consensus sequence for the 3′-splice site is (T or C)$_{11}$N(C or T) AG/G, where N is A, G, C or T, and where the stroke (/) indicates the boundary between intron and exon. In the present research, the nucleotide sequence at the 3′-splice junction in rabbit β-globin gene was analyzed by a quantification method proposed previously. Using a sample score of 16-nucleotide sequence at the 3′-splice junction, we proposed the strength of the 3′-splice signal. This approach could explain not only the location of the authentic 3′-splice sites in rabbit β-globin pre-mRNA but also the experimental results of various point mutations introduced into the 3′-splice region of the second intron. Our analysis also explains a skipping mechanism of the second exon, in which the first exon is spliced directly to the third exon if the 5′-splice site of the second intron is destroyed.

Most eukaryotic genes are interrupted by introns, which are removed from mRNA precursors (pre-mRNAs) by the RNA splicing mechanism. As for the reaction of splicing and the spliceosome assembly, see reviews.[1,2] The sequences required for splicing in mammalian pre-mRNAs consist of conserved elements at the 5′- and 3′-splice sites and a less conserved branch point sequence at the site of lariat formation. Each of these sequences has multiple roles in the reaction of splicing, which involves two reaction steps (lariat formation and exon ligation, respectively). Earlier, the 3′-splice consensus sequence was determined as (T or C)$_{11}$N(C or T)AG/G, where N is A, G, C or T, and where the stroke (/) indicates the boundary between intron and exon.[3,4] The description of invariant AG dinucleotide preceded by short polypyrimidine stretch (ca. 12 nucleotides) only shows the qualitative features of the 3′-splice signal, and actual 3′-splice sequences differ from it to a greater or lesser degree.

At least two pathways to direct the branch site and the 3′-splice site have been proposed.[5] In the usual case of a short polypyrimidine stretch, however, the 16-nucleotide consensus sequence of the 3′-splice site plays an essential role in directing both the branch site and the 3′-splice site. One major problem of the consensus sequence is that the effect of mutations in the sequence on the splicing effeciencies depends greatly on the position and the species of nucleotides. Another unsolved problem involves splice site selection. Within exons and introns, there are a number of places at which the sequences resemble the 3′-splice consensus sequence but are normally inactive in splicing. However, they are activated if the authentic 3′-splice site is abolished. In order to solve those problems, we previously developed a quantification analysis, and proposed a quantitative measure for the 3′-splice signal.[6—10] To test the applicability of our approach, we studied the rabbit β-globin pre-mRNA and its various mutant pre-mRNAs, where Aebi et al.[11] introduced nucleotide changes around 3′-splice site of the second intron. In studying the strength of the 3′-splice signal, we further examined mutants showing exon skipping, where the first exon is spliced directly to the third exon if the 5′-splice site of the second intron is destroyed.

Quantification Analysis of 3′-Splice Signal in Rabbit β-Globin pre-mRNA

As was mentioned previously, the 16-nucleotide 3′-splice consensus sequence qualitatively shows the signal essential for directing both the branch point and 3′-splice site. In the present paper, the strength of 3′-splice signal was introduced by quantification method. Principles and procedure of this method were described previously,[6] so that details of the calculation are not given here.

As given in Table 1, we constructed two groups of 16-nucleotide sequence data. The first group ($r=1$) is composed of 185 sequences containing the 3′-splice signal. Such 16-nucleotide sequences (15-nucleotides in intron and one in exon) are taken from authentic 3′-splice sites of various mammalian pre-mRNAs.[12] We set $n_1=185$, which denotes the number of samples in the first group. The second group ($r=2$) is composed of sequences including no 3′-splice signal. In order to analyze rabbit β-globin pre-mRNA, sample sequences of the second group were taken in the following way. The pre-mRNA sequence is separated into three exons by two introns, and there are two positions (271/272 and 1067/1068) of authentic 3′-splice junctions.[12] Here, we first take 16-nucleotide sequence at the 5′-cap site. Next, we progress one nucleotide in the 3′-direction, and window the next 16-nucleotide sequence. In this way, we window a 16-nucleotide sequence at every position of the whole pre-mRNA. In those sequences, however, there lie two sequences due to the authentic 3′-splice sites, which belong to the first group. These two are excluded, and the remaining 1271 sequences are summarized in the second group. Some of those sequences are also shown in Table 1. The number of samples in the second group is then $n_2=1271$.

Table 1. The 16-Nucleotide Sequence Data of 3′-Splice Signal To Be Analyzed by Quantification Method

No. (ν)	Group (r)[a]	Sequence	Gene
1	1	TTTTCATTTTCTCAGG	Rabbit β-globin
2	1	TTCTTTTTCCTACAGC	
3	1	TGCTTCTCCCCGCAGG	Human α-globin
4	1	CTCTTCTCTGCACAGC	
:	:	:	
185	1	TCTCTTTCTGTGTAGG	Rat cytochrome C
1	2	ACACTTGCTTTTGACA	Rabbit β-globin
2	2	CACTTGCTTTTGACAC	
:	:	:	
1271	2	AATTTATTTTCATTGC	

a) Group 1 is composed of sequences at mammalian authentic 3′-splice sites, while group 2 comprises sequences including no 3′-splice signal. Sequences of the latter group are constructed by using rabbit β-globin pre-mRNA. See text for further details.

Next, we introduce a dummy variable, $x_{i(\alpha)}^{r(\nu)}$, which is defined by item (i=1, 2, ⋯, 16), category (α=1, 2, 3, 4), group (r=1, 2), and sample (ν=1, 2, ⋯, n_r). Sixteen items correspond to the positions of nucleotides in the 16-nucleotide sequences, i being given by the order from the 5′- to 3′-ends of the sequence. Four categories denote the kinds of nucleotides, where A, G, C, or T is specified by α=1, 2, 3, or 4 at every item, respectively. The parameter ν specifies each sample sequence belonging to the group (r=1 or 2). The dummy variable, $x_{i(\alpha)}^{r(\nu)}$, becomes 1, if the sample sequence (ν) of the group (r) has a nucleotide (α) at the position (i), and otherwise it is 0. Using this variable, we transform the sequence data of Table 1 into the item-category data composed of 0 or 1.

Quantification of each sequence can be done by calculating the sample score value,

$$y^{r(\nu)} = \sum_{i=1}^{16} \sum_{\alpha=1}^{4} x_{i(\alpha)}^{r(\nu)} a_{i(\alpha)}, \tag{1}$$

where r=1, 2 and ν=1, 2, ⋯, n_r. The coefficient of $a_{i(\alpha)}$ is a real number and is called the category weight. Our quantification method determines the $a_{i(\alpha)}$ and $y^{r(\nu)}$ values in such a way that the two groups of 3′-splice site sequences (r=1) and sequences other than 3′-splice sites (r=2) may be discriminated most distinctly. This optimization can be achieved by the following procedure. First, we calculate the mean value of sample scores within the group r, $\overline{y}^r$, and the mean value of the total samples, $\overline{y}$. The variance of the total samples, σ^2, and the variance between groups r=1 and 2, ${\sigma_B}^2$, are then given by

$$\sigma^2 = (1/N) \sum_{r=1}^{2} \sum_{\nu=1}^{n_r} (y^{r(\nu)} - \overline{y})^2, \tag{2}$$

$${\sigma_B}^2 = (1/N) \sum_{r=1}^{2} n_r (\overline{y}^r - \overline{y})^2, \tag{3}$$

where $N = n_1 + n_2$. In order to discriminate the sequences between the groups r=1 and 2 most distinctly, we maximize the ${\sigma_B}^2/\sigma^2$ value. This can be done by solving the eigen-value problem, and the procedure to estimate $a_{i(\alpha)}$ values at this optimum condition was described in our previous paper.[6] The $a_{i(\alpha)}$ values thus calculated are given in Table 2. Using these data, the sample score of any 16-nucleotide sequence in rabbit β-globin pre-mRNA is calculated by Eq. 1. Our analysis demonstrates that the higher the score of a sequence is, the stronger the 3′-splice signal that the sequence contains. In the next section, we study the sequences of the rabbit β-globin pre-mRNA and its mutant genes in terms of such sample scores.

Analysis of Nucleotide Substitutions in Rabbit β-Globin pre-mRNA *in Vitro*

So far, we have used the rabbit β-globin pre-mRNA

Table 2. The Optimum Category Weight Values of $a_{i(\alpha)}$ Calculated with Quantification Analysis of 3′-Splice Site Sequences[a]

Category (α) / Nucleotide / Item (i)	1 A	2 G	3 C	4 T
1	−0.460	−1.281	1.333	0.203
2	−1.389	0.260	−0.513	1.021
3	−0.440	−1.662	0.662	0.830
4	−1.881	−0.977	0.275	1.497
5	−0.551	−1.603	0.115	1.151
6	−1.495	−0.459	1.529	0.187
7	−1.197	−0.170	0.783	0.365
8	−2.203	−2.076	2.230	1.031
9	−1.074	−1.668	2.001	0.218
10	−1.482	−1.879	1.661	0.897
11	−0.964	−1.965	1.211	0.931
12	0.931	0.145	−0.565	−0.433
13	−3.384	−3.329	4.468	0.407
14	8.228	−3.747	−2.077	−5.621
15	−5.110	9.190	−3.582	−4.173
16	−0.211	2.444	−0.320	−1.655

a) Item number (i) specifies the position of nucleotide, while category number (α), the kind of nucleotide. For further details, see text.

sequence to construct 16-nucleotide sequences in Table 1. There lie two introns in this pre-mRNA, and thus, there are two positions of authentic 3′-splice junctions. The 16-nucleotide sequence at the 3′-splice site of the first intron is TTTTCATTTTCTCAG/G at position (271/272), while that of the second intron is TTCTTTTTCCTACAG/C at position (1067/1068), where the stroke (/) indicates the boundary between intron and exon. Using the $a_{i(\alpha)}$ values in Table 2, the sample score value for TTTTCATTTTCTCAG/G at 271/272 is calculated as 29.79, while that for TTCTTTTTCCTACAG/C at 1067/1068 is 33.21. Among the 16-nucleotide sequences of the entire rabbit β-globin pre-mRNA, TTCTTTTTCCTACAG/C shows the highest score, and TTTTCATTTTCTCAG/G, the next highest. These quantification results are consistent with the finding that those two are the authentic 3′-splice site sequences of the second and first introns, respectively.

Sequences other than the authentic 3′-splice sequences are not recognized as 3′-splice site in the normal gene. However, nucleotide changes within a 3′-splice sequence may decrease its sample score and abolish the authentic site, resulting in abnormal splicing. In this respect, the experimental results of Aebi et al.[11] are very interesting; they reported on the effect of various point mutations in the 3′-splice region of the second intron of rabbit β-globin pre-mRNA on splicing activity *in vivo* and *in vitro*. First, they substituted the three nucleotides preceding the 3′-splice site individually. Those mutant pre-mRNAs were spliced in HeLa cell nuclear extract, and full-length pre-mRNAs thus formed *in vitro* were tested. Mutations changing the invariant AG at positions 1066 and 1067 (−2 and −1 relative to the 3′-splice site of the second intron, respectively) to either TG, CG, AT, or AA showed neither the RNA species from which the second intron had been excised, nor the excised lariat of the second intron (denoted by L). Only after long exposure of the autoradiogram was the lariat intermediate L-exon 3 detected, indicating that cleavage at the 5′-splice site was not totally abolished; however, no exon joining took place between exon 2 and exon 3. To estimate the reaction more quantitatively, Aebi et al. measured the relative molecular yields of the products of (exon 1)–(exon 2)–(exon 3) and (exon 1)–(exon 2)–L–(exon 3) in the wild-type and mutant pre-mRNAs. They are given by the values of E1–E2–E3 and E1–E2–L–E3 in Fig. 1 of Ref. 11, respectively. The former molecular species corresponds to a correctly spliced product, where the 3′-splice signal of the second intron functions normally. By contrast, the latter species retaining L is the unspliced product. Relative molecular yields of E1–E2–E3 and E1–E2–L–E3 in wild-type pre-mRNA were 1.0 and 1.0, respectively, which were quantitated from the 90 min sample. This means that, after 90 min reaction of *in vitro* splicing of the wild pre-mRNA, half of the pre-mRNAs were processed, while the remaining half were unprocessed. Under the same condition, however, the pre-mRNA with AG→TG mutation gives the yields of E1–E2–E3 and E1–E2–L–E3 as <0.02 and 3.5, respectively. In this case, the correctly spliced products were virtually not observable, while all of the second introns in the pre-mRNAs were unspliced, because the 3′-splice signal of the second intron was destroyed completely. In spite of such a defect, the first intron was excised normally, indicating that the *in vitro* system worked functionally. Similarly, in all of the pre-mRNAs with AG→CG, AG→AT and AG→AA mutations, the correctly spliced products were also unobservable, and all of the second introns in the pre-mRNAs were unspliced (Fig. 1 of Ref. 11). These results clearly indicate that point mutations within the invariant AG dinucleotide at the 3′-splice site greatly reduce cleavage at the 5′-splice site and abolish 3′-cleavage and splicing *in vitro*.

Another example of mutation around the 3′-splice site of the second intron is the C→A transversion at position 1065 (position −3 of the second intron), where the nucleotide change occurs out of the invariant AG dinucleotide but apparently severely affects 3′-splicing. As is shown in Fig. 1 of Ref. 11, the mutation did not always abolish the 3′-splice site, but did reduce splicing of the second intron by 70%. That is, after 90 min reaction of *in vitro* splicing of the pre-mRNA with C→A mutation, the relative molecular yields of E1–E2–E3 and E1–E2–L–E3 were 0.31 and 2.7, respectively. These values should be compared with those of the wild-type pre-mRNA (both 1.0) under the same condition.

We analyze those *in vitro* mutational results in terms of our sample scoring scheme. The AG→TG mutation changes the wild-type 16-nucleotide sequence at the 3′-splice site of the second intron TTCTTTTTC-CTACAG/C into TTCTTTTTCCTACTG/C. As was estimated earlier, the sample score of the wild-type sequence is 33.21, the highest value among the whole 1273 sets of 16-nucleotide sequences in the entire rabbit β-globin pre-mRNA. However, if we use our $a_{i(\alpha)}$ data in Table 2, the AG→TG mutation dramatically decreases its score to 19.36. Since the magnitude of such a score shows the extent to which the altered sequence contains the 3′-splice signal, the lower the sample score, the lower the relative splicing efficiency. Therefore, point mutation within the invariant AG dinucleotide destroys the 3′-splice signal completely. This quantification result is consistent with the experimental finding that correctly spliced products are virtually not observable in this pre-mRNA. Similarly, the mutation changing the AG dinucleotide to CG, AT or AA is found to decrease the score of the wild-type sequence into 22.90, 19.84 or 18.91, respectively. Again in these mutant pre-mRNAs, the decrease in the sample score is so drastic that the 3′-splice signal of the second intron may be lost completely.

Next, we consider another example of C→A transversion at position 1065. This mutation changes the

wild-type 16-nucleotide sequence TTCTTTTTCC-TACAG/C into TTCTTTTTCCTAAAG/C. Then, such a mutation reduces the sample score 33.21 of the wild-type sequence to 25.35 of the mutant sequence. The 25.35 value is greater than 19.36 of the AG→TG mutant discussed previously, but is smaller than 33.21 of the normal sequence. Therefore, in the C→A mutation, 3′-splice signal is considerably decreased, but is not so severe as that of the AG→TG mutant. Such a situation leads to an inefficient splicing of the second intron, giving some amount of normally spliced product of E1–E2–E3. Then, we can understand the previous experimental results of decreased amount (70% decrease) of E1–E2–E3 reported by Aebi et al.[11]

The nucleotides out of the invariant AG dinucleotide are also functionally important. As described by Mount,[3] the consensus sequence preceding the 3′-splice site consists of a stretch of 11 or more pyrimidines, followed by N(C or T)AG. The 3′-terminal AG is strictly conserved in all functional introns, whereas the nucleotide preceding it is a C residue in 65% of the cases and a T in 31% of the cases. Experimental results demonstrated that substitution of any of the AG dinucleotide led to total reduction in the rate of splicing measured *in vitro*, while substitution of the less conserved C residue at the −3 position by A residue led to partial reduction. These tendencies are quantitatively explained in terms of our sample scoring approach. If we refer to the category weight values of $a_{i(\alpha)}$ in Table 2, subsitution of pyrimidine to purine in the 11-nucleotide region of polypyrimidine stretch may reduce the splicing efficiency to a lesser extent than that of the C→A mutant. However, no such experimental data are given by Aebi et al.

Analysis of Nucleotide Substitutions *in Vivo*

Aebi et al.[11] also examined the effects of the 3′-AG mutations described above *in vivo*. They were observed in transiently transfected HeLa cells. As is shown in Fig. 4A of Ref. 11, none of the 3′-AG mutants gave the correctly spliced transcripts; instead, a major component of protected RNA fragments of around 84 nucleotides was observed, which had about 40—80% of the intensity of the 133-nucleotide signal in the wild-type control. Although Aebi et al. did not report so explicitly, another minor component of 70-nucleotide signal was also detected. The fragment in the wild-type control comes from use of the normal 3′-splice site of the second intron (TTCTTTTTCCTACAG/C at position 1067/1068). The signal of 84 nucleotides in the mutant is attibutable to use of a cryptic 3′-splice site (ATCATTTTGGCAAAG/A at 1116/1117 in the third exon), which corresponds to the closest AG dinucleotide downstream of the mutated 3′-splice site of the second intron. The fragment of the 70 nucleotides arises from use of another cryptic 3′-splice site (GAATTCACTCCTCAG/G at 1130/1131), which is the second closest AG dinucleotide downstream of the mutated 3′-splice site.

Mutations of the invariant 3′-AG dinucleotide gave unspliced product *in vitro*, while *in vivo* the 3′-AG mutations activated cryptic 3′-splice sites. This is probably because the splicing system *in vitro* lacks certain factors which manage to find cryptic 3′-splice sites to complete splicing *in vivo*. The experimental results *in vivo* can be explained in terms of our scoring scheme. Sample scores of the above cryptic 3′-splice sequences are calculated as 14.50 (ATCATTTTGGCAAAG/A) and 29.09 (GAATTCACTCCTCAG/G). Note that the latter value is the second largest in the region downstream of the 5′-splice site of the second intron. If a mutation occurs within the AG dinucleotide of the authentic 3′-splice site of the second intron, the largest score (33.21) of the authentic sequence (TTCTTTTTCCTACAG/C) decreases dramatically. For example, as was discussed previously, the AG→TG mutation decreases the score to 19.36, which becomes much smaller than the second largest score (29.09) of the cryptic sequence (GAATTCACTCCTCAG/G), but which is still greater than the score (14.50) of the cryptic sequence (ATCATTTTGGCAAAG/A) nearest the mutated 3′-splice site of the second intron. Therefore, it is reasonable to consider that the AG→TG mutation destroys the normal 3′-splice site of the second intron completely, but that splicing occurs at the cryptic site with the second largest score (GAATTCACTCCTCAG/G at 1130/1131). However, it is puzzling that the RNA product spliced at 1130/1131 is the minor component, while most of the product is spliced with the 14-nucleotide upstream site (ATCATTTTGGCAAAG/A at 1116/1117), whose score (14.50) is much smaller than 29.09 at 1130/1131. This problem was solved previously with two examples of human β-globin thalassemia pre-mRNAs.[10] Prior to cleavage and splicing of 3′-splice site (the second step of the splicing reaction), the 3′-splice signal sequence possessing high sample score directs branch point and lariat formation in the first step of the reaction. In our case, the sequence GAATTCACTC-CTCAG/G with the second largest score (29.09) plays this role. Once the branch point is selected by such a sequence possessing high sample score, the 3′-splice site itself is recognized in the second step of the reaction, and the cleavage usually occurs at the first AG downstream of the branch site by a distance-independent scanning process.[13,14] The 1116/1117 and 1130/1131 sites lie so close to each other (only 14-nucleotides apart) that they may use a common branch point; usually, the branch point is located 18-40 nucleotides upstream of the 3′-splice site. Then, major 3′-splicing takes place at position 1116/1117 nearest the branch point, but minor splicing occurs at 1130/1131. This may be a reason why the upstream sequence with rather low score (14.50) is used preferentially over the downstream sequence with high score (29.09).

Mutations in the 5′-Splice Region and Exon Skipping (Aberrant Joining of Exon 1 to Exon 3) *in Vitro*

So far, we have examined various examples of mutations around 3′-splice site of the second intron of rabbit β-globin pre-mRNA. In connection with sample score and 3′-splice signal intensity, we study a mechanism of exon skipping, where mutations around 5′-splice region of the second intron cause aberrant joining of exon 1 to exon 3. Aebi et al.[11] found that mutations of the invariant 5′-GT impaired splicing at that position. For example, two such mutations, 495 G→A and 496 T→A (positions +1 and +2 of the 5′-splice site of the second intron, respectively), gave a drastic change in the pattern of products *in vitro*; correctly spliced RNA E1–E2–E3 was replaced by the 213-nucleotide aberrant product E1–E3. Moreover, the lariat of the second intron was replaced by another lariat, in which the 5′-end of the first intron was linked to the branch point of the second intron. These experimental results can be explained in terms of sample scores of 3′-splice sites calculated above.

In the normal pre-mRNA, two introns are spliced correctly, giving E1–E2–E3. Here, two pieces of spliceosome bind 5′-splice sites of the first and second introns, and find 3′-splice sites of the first and second introns, respectively. However, if the invariant GT within the 5′-splice site of the second intron is mutated, the 5′-splice signal of the second intron is destroyed completely. This was demonstrated previously by our quantification analysis,[8] where sample score of the 9-nucleotide sequence (AGG/GTGAGT) decreased dramatically by the mutations. In such a case, the recognition factor of 5′-splice site in spliceosome, U1 small nuclear ribonucleoprotein particle (snRNP), can no longer bind at the authentic 5′-splice site of the second intron. This gives rise to U1 snRNP of single spliceosome bound at the authentic 5′-splice site of the first intron, and there are two possible candidates of the 3′-splice sites (the authentic sites of the first and second introns) for the single spliceosome. Choice of one 3′-splice site out of the two should be determined by the intensity of their 3′-splice signals. If such intensity is given by a sample score of 16-nucleotide sequence in the consensus region, the score 33.21 of the authentic 3′-splice site of the second intron (TTCTTTTTCCTACAG/C at 1067/1068) is the greatest and is larger than 29.79 of the first intron (TTTTCATTTTCTCAG/G at 271/272). In this way, the 3′-splice site of the first intron is no longer recognized after competition, but only that of the second intron is selected. This explains a reason why the first exon is spliced directly to the third exon (exon skipping). This also explains the above experimental results that the lariat of the second intron was replaced by another lariat, in which the 5′-end of the first intron was linked to the branch point of the second intron.

The mechanism responsible for inclusion or skipping of exon is sensitive to the binding of U1 snRNP to the 5′-splice site of an intron. Usually, in addition to the U1 snRNP binding at the upstream intron, another U1 snRNP interacts with the 5′-splice site of the downstream intron. Then, two such U1 snRNPs define the boundaries of a functional splicing unit of the upstream intron.[15] Our scoring scheme shows that, in such a functional unit, the 3′-splice sequence possessing the highest sample score is selected as the authentic 3′-splice site. Whether the downstream 5′-splice site is bound by U1 snRNP or not determines not only the range of the functional unit of the upstream U1 snRNP but also inclusion or skipping of exon. Binding of U1 snRNP to the downstream 5′-splice site may be inhibited if the 5′-splice signal is destroyed by mutation (the present study) or if the 5′-splice site is masked by certain factors. The latter may lead to tissue-specific alternative splicings (see discussion in Concluding Remarks).

Analogous Exon Skippings as Revealed by Other Genes

So far, we have examined exon skipping caused by site-directed mutagenesis of 5′-splice site. Analogous exon skippings have also been observed in several naturally-ocurring genes, where genetic diseases have been known to come from exon skippings and defects in 5′-splice sites. Next, we demonstrate two such examples.

One example is a human β-globin mutant gene of thalassemia. As is similar to the previous rabbit β-globin, normal human β-globin pre-mRNA consists of three exons divided by two introns. Splicing in the normal pre-mRNA is illustrated in Fig. 1a, where two introns are spliced correctly. A pre-mRNA of β°-thalassemia mutant contains GT→AT transition at the 5′-end of the second intron.[16] Two abnormally spliced products were detected. The predominant mRNA differs from normal mRNA by the insertion of the first 47 nucleotides of the second intron sequence between the second and third exons (Fig. 1b). The less-abundant mRNA is composed of the normal first exon spliced directly to the third exon, skipping the second exon sequence (Fig. 1c). In

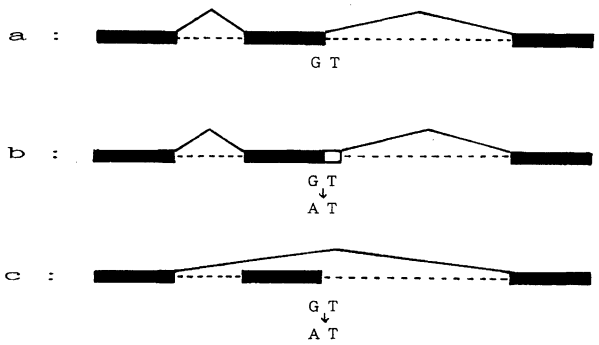

Fig. 1. (a) Normal splicing in human β-globin pre-mRNA; (b) cryptic splicing and (c) exon skipping in β°-thalassemia pre-mRNA. See text and Ref. 16 for further details.

this β°-thalassemia, the GT→AT mutation at the 5′-end of the second intron destroys its 5′-splice site completely, giving cryptic splicing or exon skipping. In the predominant mRNA, the first intron is spliced normally, while the extra 47 nucleotide insertion arises from splicing from the cryptic 5′-splice site at ATG/GTTAAG (position 542/543 within the second intron) to the authentic 3′-splice site of the second intron. In this mutant pre-mRNA, U1 snRNP (recognition factor for 5′-splice site) binds at the cryptic site, but its 5′-cryptic splice signal is not strong enough to bind tightly with U1 snRNP. Then, some portions of U1 snRNP can neither bind at the cryptic site of the second intron nor form a spliceosome assembly. In the less-abundant mRNA, single U1 snRNP binds at the authentic 5′-splice site of the first intron. Then, two possible candidates of 3′-splice sites (the authentic splice sites of the first and second introns) arise for the single spliceosome. As in the rabbit β-globin case, splice site selection should be determined by sample scores of their 3′-splice sites. For the human β-globin pre-mRNA, we calculated such scores of 16-nucleotide sequences of the 3′-splice sites previously.[7,10] The score 34.03 of the authentic site of the second intron (ATCTTCCTCCCACAG/C at 1345/1346) was much larger than 27.09 of the first intron (TTTTCCCACCCTTAG/G at 272/273). Also in this case, only the 3′-splice site of the second intron is selected, and we can explain the reason why the first exon is spliced directly to the third exon in the less-abundant mRNA.

Another example involves a human phenylalanine hydroxylase (PAH) gene. Classical phenylketonuria is an autosomal recessive human genetic disorder caused by a deficiency of hepatic PAH. Marvit et al.[17] found a certain mutant PAH gene, in which internal 116 nucleotides were deleted compared to the normal PAH cDNA. The sequence of the 116-nucleotides corresponds precisely to the 12-th exon sequence of the PAH gene. Such deletion leads to a lack of the C-terminal 52 amino acids, producing inactive PAH protein. In the mutant gene, a GT→AT substitution was found at the 5′-splice site of the 12-th intron. It appears that this mutation abolishes the 5′-splice signal completely and causes skipping of the preceding 12-th exon during pre-mRNA splicing. This situation is very similar to those discussed in the rabbit β-globin mutant and human β°-thalassemia. Normal human PAH gene contains thirteen exons spanning approximately 90 kb of DNA. In the mutant gene, we only pay attention to the splicings of the 11-th and 12-th introns, since abnormal splicing is limited to these introns and the other introns are spliced normally. If there is no appropriate 5′-cryptic site in the region between the 3′-splice site of the 11-th and 12-th introns, single U1 snRNP binds to the 5′-splice site of the 11-th intron, but no U1 snRNP binds to the mutated 5′-splice site of the 12-th intron. Then, for the single spliceosome bound at the 5′-end of the 11-th intron, there arise two possible candidates of 3′-splice sites (the authentic 3′-splice sites of the 11-th and 12-th introns). Unfortunately, whole sequence data of the PAH pre-mRNA are not available to calculate sample scores of their sites. However, when we compare 16-nucleotide sequence of the 3′-splice site of the 11-th intron (TGGTTTTGGTCTTAG/G) with that of the 12-th intron (TGTTTTTCTTTGTAG/G),[18] only five nucleotides differ from each other, and the remaining nucleotides are identical. If category weight values of the previous human β-globin pre-mRNA (Table 2 of Ref. 7) are available for comparison, the sample score of the 3′-splice sequence of the 12-th intron is much larger than that of the 11-th intron, by 10.05. This clearly indicates that the spliceosome bound at the 5′-splice site of the 11-th intron should preferably recognize the authentic 3′-splice site of the 12-th intron, skipping the entire 12-th exon. Our sample scoring approach explains the above experimental results of the phenylketonuria mutant.[17]

Concluding Remarks

The sample score of 16-nucleotide sequence at the 3′-splice site was calculated with category weight values of $a_{i(\alpha)}$, which show the relative importance of nucleotides at each position. The data of Table 2 were obtained by taking the sequences of the second group (sequences other than authentic 3′-splice sites) from rabbit β-globin pre-mRNA. We examined whether or not the values of $a_{i(\alpha)}$ would be unaltered, and compared the values with those of human β-globin pre-mRNA.[7,10] Relative magnitudes of $a_{i(\alpha)}$ were practically the same between the two systems, so that the data of Table 2 appear to be applicable to characterize 3′-splice signals in various genes.

In connection with the abnormal splicings in the present study, we note the evolutionary role in alternative splicings of pre-mRNA. If splice sites are mutated artificially or naturally, abnormal splicing such as unsplicing, cryptic splicing or exon skipping occurs, generating various mRNAs from a single gene. This may produce, in addition to the normal protein, various other proteins, which may give a chance of getting a new function in the mutant (see a recent review[19]). Alternative splicings also play important roles in tissue-specific gene expression of higher eukaryotes, where different mRNAs and proteins are often generated from a single gene.

References

1) T. Maniatis and R. Reed, *Nature*, **325**, 673 (1987).
2) M. M. Konarska and P. A. Sharp, *Cell*, **49**, 763 (1987).
3) S. M. Mount, *Nucl. Acids Res.*, **10**, 459 (1982).
4) M. B. Shapiro and P. Senapathy, *Nucl. Acids Res.*, **15**, 7155 (1987).

5) R. Reed and T. Maniatis, *Genes Dev.*, **2**, 1268 (1988).
6) Y. Iida, *Comput. Appl. Biosci.*, **3**, 93 (1987).
7) Y. Iida, *J. Theor. Biol.*, **135**, 109 (1988).
8) Y. Iida, *Biochim. Biophys. Acta* (*Gene Structure and Expression*), **1007**, 270 (1989).
9) Y. Iida, *J. Theor. Biol.*, **145**, 523 (1990).
10) Y. Iida, *J. Biochem.* (*Tokyo*), **108**, 934 (1990).
11) M. Aebi, H. Hornig, R. A. Padgett, J. Reiser, and C. Weissmann, *Cell*, **47**, 555 (1986).
12) GenBank, "Genetic Sequence Data Bank," Release 68.0, USA (1991).
13) R. Reed, *Genes Dev.*, **3**, 2113 (1989).
14) C. W. J. Smith, E. B. Porro, J. G. Patton, and B. Nadal-Ginard, *Nature*, **342**, 243 (1989).
15) B. L. Robberson, G. J. Cote, and S. M. Berget, *Mol. Cell. Biol.*, **10**, 84 (1990).
16) R. Treisman, N. J. Proudfoot, M. Shander, and T. Maniatis, *Cell*, **29**, 903 (1982).
17) J. Marvit, A. G. DiLella, K. Brayton, F. D. Ledley, K. J. M. Robson, and S. L. C. Woo, *Nucl. Acids Res.*, **15**, 5613 (1987).
18) A. G. DiLella, S. C. M. Kwok, F. D. Ledley, J. Marvit, and S. L. C. Woo, *Biochemistry*, **25**, 743 (1986).
19) H. Kuo, F. H. Nasim, and P. J. Grabowski, *Science*, **251**, 1045 (1991).