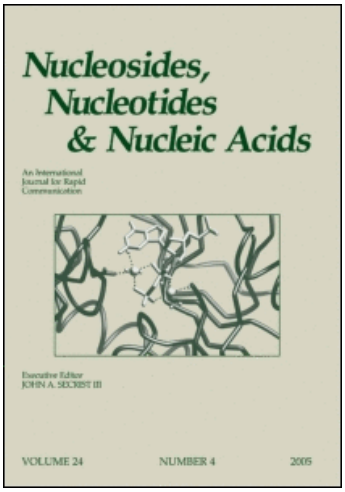

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

Quantification Analysis of Human α- and δ-Globin Genes. Mutations in 5'-Splice Junction Sequence and α- and δ- Thalassemias

Yôichi Iida[a]
[a] Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo, Japan

To cite this Article Iida, Yôichi(1992) 'Quantification Analysis of Human α- and δ-Globin Genes. Mutations in 5'-Splice Junction Sequence and α- and δ- Thalassemias', Nucleosides, Nucleotides and Nucleic Acids, 11: 2, 311 — 317
To link to this Article: DOI: 10.1080/07328319208021705
URL: http://dx.doi.org/10.1080/07328319208021705

NUCLEOSIDES & NUCLEOTIDES, 11(2-4), 311-317 (1992)

QUANTIFICATION ANALYSIS OF HUMAN α- AND δ-GLOBIN GENES. MUTATIONS IN 5'-SPLICE JUNCTION SEQUENCE AND α- AND δ-THALASSEMIAS

Yôichi Iida
Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060, Japan

Abstract: Nucleotide sequences of the exon-intron junction in human α- and δ-globin genes were analyzed by the quantification method proposed previously. We further studied several mutants of α- and δ-thalassemias, where mutational changes occur around the 5'-splice junction of the first intron. These changes abolish the normal 5'-splice site completely, but activate a cryptic site lying in the first exon. Such behaviours were well explained in terms of our quantification analysis.

Most of the mammalian genes are interrupted by introns, which are removed from mRNA precursors (pre-mRNAs) by the RNA splicing mechanism. A number of genetic diseases have been known to come from defects in RNA splicing, by which abnormal mRNAs and proteins are produced. Thalassemia is one of genetic diseases showing defects in human globin chain synthesis. [1,2] The discovery that globin genes consist of three exons divided by two introns led to the suggestion that some forms of thalassemia might arise from incorrect splicing of pre-mRNA. As for signals about where to splice, there are consensus sequences around the 5'- and 3'-splice sites of the intron and around the branch point. For example, the 5'-splice signal has been given by a 9-nucleotide consensus sequence, (C or A)AG/GT(A or G)AGT, where the stroke (/) indicates

---

This paper is dedicated to the memory of Professor Tohru Ueda (Hokkaido University).

the boundary between exon and intron.[3] One major problem in thalassemia is why certain mutations in the splice signal sequence abolish the normal splice sites completely. These mutations lead to another problem concerning splice site selection. Within exons and introns, there are several places at which sequences resemble the consensus splice sites but are normally inactive in splicing. However, they are activated if the authentic splice site is abolished. In this report, we analyzed 5'-splice sequences of the human α- and δ-globin genes, and found that the above problems actually take place in several α- and δ-thalassemia mutants. Using our quantification method, we explained why abnormal splicing occurs in such thalassemia genes.

Method of Quantification Analysis

Quantification method (categorical discriminant analysis) is a mathematical analysis of DNA sequences to measure the 5'-splice signal quantitatively. The method was reported previously,[4-6] so that its detail is not mentioned here. Taking human α-globin gene as an example, we show the procedure of calculation briefly. The analysis demonstrated that a 9-nucleotide sequence in the consensus region (three nucleotides in exon and six in intron) is almost sufficient to specify the 5'-splice site in pre-mRNA splicing.[6] Then, we constructed two groups of sequence data, as given by TABLE 1. The first group is composed of 155 sets of 9-nucleotide sequences, which are taken from authentic 5'-splice sites in various mammalian genes. Sequences of the second group were taken from human α-globin pre-mRNA in the following way. First, we start with the 9-nucleotide sequence at the cap site. Next, we progress one nucleotide in the 3'-direction, and take the next 9-nucleotide sequence. In this way, we window 9-nucleotide sequences at every position of the whole pre-mRNA. In those sequences, however, there lie two sequences due to the authentic 5'-splice sites, which belong to the first group. These two are excluded, and the remaining 822 sequences are summarized in the second group (see TABLE 1). Quantification analysis discriminates sequences between the two groups most distinctly by a mathematical optimization technique. In this process, we can estimate sample score values for each 9-nucleotide sequence of the pre-mRNA, where the larger the score, the stronger the 5'-splice signal such a sequence has (for detail of the calculation, see Refs. 4-6).

Analysis of Human α-Globin Pre-mRNA and α-Thalassemia

As was mentioned, the human α-globin pre-mRNA is composed of

TABLE 1. The 9-nucleotide sequence data to be analyzed by quantification method

No. (ν)	Group (r) [a]	Sequence	Gene
1	1	GAGGTGAGG	Human α-globin
2	1	AAGGTGAGC	
3	1	CAGGTTGGT	Human β-globin
4	1	AGGGTGAGT	
⋮	⋮	⋮	
155	1	AGGGTGAGC	Dog insulin
1	2	ACTCTTCTG	Human α-globin
2	2	CTCTTCTGG	
⋮	⋮	⋮	
822	2	GTGGGCGGC	

a Group 1 is composed of 5'-splice site sequences, while group 2 comprises sequences other than 5'-splice sites. See text and Refs. 4-6 for further details.

three exons and two introns, and there are two positions of a 5'-splice site. Hereafter, positions of the junctions are specified by numbering them from the 5'-end of the pre-mRNA. Quantification analysis of the data in TABLE 1 demonstrated that, among all of the 9-nucleotide sequences in the entire pre-mRNA, AAG/GTGAGC (position 454/455), GAG/GTGAGG (132/133), GGG/GTAAGG (83/84) and AAG/GTCGGC (88/89) give the highest sample scores of 40.8717, 34.0718, 32.8155 and 27.5589, respectively (See TABLE 2). The former two sequences coincide well with the authentic 5'-splice site sequences of the second and first introns, respectively. The third and fourth sequences, GGG/GTAAGG (83/84) and AAG/GTCGGC (88/89), resemble the consensus sequence, but the normal gene does not undergo splicing at these positions. This is because the scores of 32.8155 and 27.5589 of the third and fourth sequences are smaller than 34.0718 of GAG/GTGAGG (132/133) at the authentic 5'-splice site of the first intron. Since those three sequences lie rather close to each other and share a common 3'-splice site of the first intron at (249/250), it appears that the GAG/GTGAGG sequence gets the 5'-splice signal after competition. This view is strongly supported by the following experiments with the α-thalassemia gene. Felber et al. reported the α-thalassemia mutant gene, [7] where a pentanucleotide deletion occurs at the underlined position of GAG/GTGAGG (132/133) of the

TABLE 2. The results of quantification analysis and sample scores of 9-nucleotide sequences in human α- and δ-globin genes and their thalassemia genes [a]

Gene	Position	Sequence	Sample Score
α-Globin	83/84	GGG/GTAAGG	32.8155
	88/89	AAG/GTCGGC	27.5589
	132/133	GAG/GTGAGG	34.0718
	454/455	AAG/GTGAGC	40.8717
α-Thalassemia	132/133	GAG/GCTCCC	-1.2616
δ-Globin	126/127	GTG/GTGAGG	27.7026
	142/143	CAG/GTTGGT	31.5675
	493/494	AGG/GTGAGT	36.3472
δ-Thalassemia	142/143	CAG/GCTGGT	22.9328
δ-Thalassemia	126/127	GTG/GTGAGT	32.3851

a See text for further details.

authentic 5'-splice sequence of the first intron. Expression of this mutant gene revealed that the deletion abolishes the normal 5'-splice site completely but that alternative splicing occurs at the cryptic splice site of GGG/GTAAGG (83/84). Such a splicing led to abnormal mRNA, giving severe α-thalassemia phenotype. We can explain this abnormal splicing in terms of our sample scoring scheme. The pentanucleotide deletion changes the authentic 5'-splice site sequence, GAG/GTGAGG, into GAG/GCTCCC. This change decreases dramatically the score of 34.0718 of the authentic signal into - 1.2616 (see TABLE 2). Then, the altered score is much smaller than 32.8155 of the cryptic site of GGG/GTAAGG at position (83/84). However, the latter score is larger than the next largest score 27.5589 of the sequence AAG/GTCGGC at position (88/89). Therefore, these sample scores can well explain why the authentic site is abolished and the cryptic site at (83/84) is activated as the alternative 5'-splice site. Note that, in both of the normal and abnormal splicings, the 5'-splice signal sequence overcomes, in the magnitude of sample score, the other sequences which resemble the consensus sequence.

Analysis of Human δ-Globin Pre-mRNA and δ-Thalassemia

The human δ-globin is the non-α chain of HbA_2 ($\alpha_2\delta_2$), the

minor component of adult human hemoglobin. Also in the δ-globin gene,[1,8] the pre-mRNA is composed of three exons and two introns, and there are two positions of 5'-splice site. The 9-nucleotide sequence at the 5'-splice site of the first intron is CAG/GTTGGT at position (142/143), while that of the second intron is AGG/GTGAGT at (493/494). As is similar to TABLE 1, we constructed two groups of the sequence data, where sequences of the seond group are taken from the δ-globin pre-mRNA, and 1631 sets of 9-nucelotide sequences are summarized in this group. Quantification analysis showed that, among all of the 9-nucleotide sequences in the entire pre-mRNA, those sequences at positions (493/494) and (142/143) give the highest sample scores of 36.3472 and 31.5675, respecively (see TABLE 2). This is consistent with the finding that they are authentic 5'-splice sequences of the two introns.

On the other hand, at position (126/127), there lies a cryptic sequence GTG/GTGAGG, which resembles the consensus 5'-splice sequence and shows the third highest sample score, 27.7026. However, it is not recognized as the 5'-splice site in the normal gene. This is also explained by our scoring scheme. For the common 3'-splice site of the first intron at (270/271), there lie two potential candidates for the 5'-splice site, the authentic sequence CAG/GTTGGT at (142/143) and the cryptic sequence GTG/GTGAGG at (126/127). In view of their sample scores, the stronger sequence at the authentic site gets the 5'-splice signal after competition in the normal gene. This view is strongly supported by abnormal splicings observed with two δ-thalassemia mutant genes.

The first example is a single nucleotide substitution (T → C) at position 144 within the first intron.[8] This change leads to complete loss of splicing at the authentic 5'-splice site and thus to the absence of normal δ-globin mRNA, explaining the phenotype of δ^{o}-thalassemia. Such a substitution changes the authentic 9-nucleotide sequence of the 5'-splice site of the first intron CAG /GTTGGT into CAG/GCTGGT. In terms of our scoring, the mutation results in a dramatic decrease of the sample score of the authetic site, 31.5675 into 22.9328, whose value is found to be much smaller than the 27.7026 value of the cryptic sequence GTG/GTGAGG (see TABLE 2). This means that the intensity of the 5'-splice signal of the altered sequence is much weakened, compared to that of the cryptic sequence, and that normal splicing is completely abolished in the mutant. Instead of this, the cryptic splicing may occur at position (126/127).

The second example of δ-thalassemia shows a G → T substitution at position (132).[8] This place corresponds to the first nucleotide of codon 27, which produces an amino acid change (Ala → Ser). Moreover, the mutation occurs within the 9-nucleotide sequence of the above cryptic site, and changes the wild-type sequence GTG/GTGAGG to GTG/GTGAGT. Then, the sample score increases from 27.7026 to 32.3851, whose value becomes larger than the 31.5675 value of CAG/GTTGGT at the authentic 5'-splice site of the first intron (see TABLE 2). Therefore, in this mutant, the GTG/GTGAGT sequence gets the 5'-splice signal after competition, and a fraction of RNA transcript is abnormally spliced from this site to the normal 3'-splice site of the first intron. Although the score 32.3851 (GTGGTGAGT) is larger than 31.5676 (CAGGTTGGT), the difference between them (0.8175) is rather small. This implies that splicing occurs not only at (126/127) but also at the authentic splice site of (142/143). This mutant produces some normal mRNA, resulting in δ^+-thalassemia. It should be noted that, in case of α-globin gene, the difference between the above-mentioned sample scores for the authentic 5'-splice site (132/133) and the cryptic site (83/84) is 1.2563 (= 34.0718 - 32.8155), which is slightly larger than the difference (0.8175) observed here in δ^+-thalassemia. This seems to explain why normal α-globin pre-mRNA can be spliced exclusively at the authentic site (132/133), and not at the site (83/84).

In view of the above examples, a quite high sample score is required for a sequence to get the signal, but the sequence is not always selected as 5'-splice site. If two or more such potential sequences of 5'-splice site share a common 3'-splice site, they compete with each other, and the sequence possessing the greatest score gets the 5'-splice signal. If it is destroyed by mutation, the next strongest sequence gets the signal. In this way, whether a potential sequence may get or lose the signal depends on the surrounding sequences.

REFERENCES

1. Nienhuis, A. W.; Anagnou, N. P.; Ley, T. J. Blood, 1984, 63, 738.
2. Higgs, D. R.; Vickers, M. A.; Wilkie, A. O. M.; Pretorius, I. - M.; Jarman, A. P.; Weatherall, D. J. Blood, 1989, 73, 1081.
3. Mount, S. M. Nucleic Acids Res., 1982, 10, 459.
4. Iida, Y. Comput. Appl. Biosci., 1987, 3, 93.

5. Iida, Y. Biochim. Biophys. Acta (Gene Structure and Expression), 1989, 1007, 270.
6. Iida, Y. J. Theor. Biol., 1990, 145, 523.
7. Felber, B. K.; Orkin, S. H.; Hamer, D. H. Cell, 1982, 29, 895.
8. Moi, P.; Paglietti, E.; Sanna, A.; Brancati, C.; Tagarelli, A.; Galanello, R.; Cao, A.; Pirastu, M. Blood, 1988, 72, 530.

Received 8/9/91
Accepted 12/4/91