  Bull. Chem. Soc. Jpn., 60, 2977—2981 (1987)

Nucleotide Sequences of DNA and Their Structural Analysis. Categorical Discrimination of 5′-Splice Site Sequences in mRNA Precursors

Yôichi IIDA
Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060
(Received February 28, 1987)

The signals which direct excision of introns from mRNA precursors in higher eukaryotes' genes are not well-understood. Although a consensus sequence, $^{C}_{A}$AG/GT$^{A}_{G}$AGT, has been proposed with the 5′-splice site, actual 5′-splice site sequences differ from it to a greater or lesser degree. In the present paper, structural analysis of 5′-splice site sequences was done by using quantification theory (class II). Nucleotide sequences were transformed into categorical data, and categorical discriminant analysis was carried out between 5′-splice site sequences and sequences other than 5′-splice site. Relative importance of each position of nucleotide sequence was then estimated by calculating partial correlation coefficient on each position.

Many eukaryotic genes are interrupted by introns, which are removed from mRNA precursors (pre-mRNAs) by the RNA splicing mechanism.[1-5] The splicing of pre-mRNA includes, as a first step, cleavage of the 5′-splice junction and formation of a lariat involving the 5′-end of the intron joined by a 2′—5′ phosphodiester bond to an A residue upstream of the 3′-splice junction. This is then followed by cleavage at the 3′-splice junction together with ligation of the 5′ and 3′ exons. Although both steps are related to several reactions, the machinery for splicing (macromolecular ribonucleoproteins containing small nuclear RNAs) carries out this process in a concerted manner.[3-5] The compilation of sequences surrounding intron boundaries demonstrates consensus sequences at both the 5′ (donor) and the 3′ (acceptor) splice sites.[1,2] The 5′-splice site consensus sequence is described by 5′-(exon)-$^{C}_{A}$AG/GT$^{A}_{G}$AGT-(intron)-3′, where the stroke (/) indicates boundary between exon and intron. Such a sequence plays an important role in defining the 5′-splice site, because a number of naturally occurring and in vitro derived mutations within the consensus regions result in splice site inactivation.[6,7] However, the consensus sequence has been ambiguous in what degree of matching between observed junction sequences and the consensus sequence is necessary to specify the exact splice site. That is, actual splice site sequences differ from the consensus sequence to a greater or lesser degree. It is also impossible for the consensus sequence to account for relative importances of each of the nucleotides in the sequence. In order to study such situation more quantitatively, Staden used the weight matrix method.[8] On the other hand, we applied previously multivariate statistical analysis to 5′-splice site sequences of pre-mRNAs, and performed categorical discriminant analysis, where nucleotide sequences were transformed into categorical data.[9] In the present paper, discriminating 5′-splice site sequences and sequences other than 5′-splice site, we calculate a value for partial correlation coefficient on each position of the sequence. This approach enables us to know relative importance of each position of nucleotide sequence.

Categorical Discriminant Analysis

Categorical Data. The formulation of categorical data was described in the previous paper.[9] In the following, we summarize the procedure briefly. Our data assume two groups of nucleotide sequences. The first group is composed of 5′-splice site sequences as recognized by the machinery for splicing. They are taken from authentic 5′-splice site sequences in various mammalian genes containing introns, such as globin genes, insulin genes, etc.[10] We take into consideration the consensus sequence, $^{C}_{A}$AG/GT$^{A}_{G}$AGT, for 5′-splice site, and take 9-nucleotide sequence composed of three nucleotides at the 3′-end of exon and six nucleotides at the 5′-end of intron. It is assumed that essential pattern to define 5′-splice site may lie within such 9-nucleotide sequence. In this way, we summarized 155 sequences into the first group. Some of the sequence data are shown in Table 1.

The second group is composed of sequences other than 5′-splice site. They are taken from human β-globin gene in the following way. The β-globin gene sequence is separated into three exons by two introns,

Table 1. Some of the 9-Nucleotide Sequences Belonging to Group 1 and Group 2[a)]

No.	Group	Sequence	Gene
1	1	GAGGTGAGG	Human alpha-Globin
2	1	AAGGTGAGC	
3	1	CAGGTTGGT	Human beta-Globin
4	1	AGGGTGAGT	
:	:	:	
155	1	AGGGTGAGC	Dog Insulin
156	2	ACATTTGCT	Human beta-Globin
157	2	CATTTGCTT	
:	:	:	
1751	2	TTTCATTGC	

a) Group 1 is composed of 5′-splice site sequences, while Group 2, of sequences other than 5′-splice sites. For further details, see text.

and there are two positions of 5′-splice junctions.[11] In this gene, we first take the 9-nucleotide sequence at the 5′-cap site. Next, we progress one nucleotide in the 3′ direction, and take the next 9-nucleotide sequence. In this way, we truncate 9-nucleotide sequence at every position of the whole pre-mRNA. In those sequences, however, there lie two sequences of the authentic 5′-splice sites, belonging to the first group. These two are excluded, and the remaining 1596 sequences are summarized into the second group, as shown in Table 1.

In order to transform sequence data into categorical data, we introduce a dummy variable $x_{i(\alpha)}^{r(\nu)}$ and parameters of r, ν, i, and α. The parameter, r, indicates the groups, where r=1 or 2 corresponds to the first or second group, respectively. We denote n_r as sample size of each group, and take ν=1, 2, $\cdots$, n_r, where n_1=155 and n_2=1596. Next, we consider item and category. In our case, there are nine items with i=1, 2, $\cdots$, 9. They correspond to the positions of nucleotides in 9-nucleotide sequence, and are defined by the order from the 5′- to 3′-ends of the sequence. Category denotes the kind of nucleotide, where we specify nucleotide A, G, C, or T by α=1, 2, 3, or 4 at every item, respectively. Using these parameters, dummy variable of $x_{i(\alpha)}^{r(\nu)}$ is expressed by

$$x_{i(\alpha)}^{r(\nu)} = \begin{cases} 1: \text{if the sample sequence } (\nu) \text{ of the group } (r) \\ \quad \text{has a nucleotide } (\alpha) \text{ at the position } (i), \\ 0: \text{otherwise.} \end{cases} \tag{1}$$

In this way, we transformed the sequence data of Table 1 into the categorical data. Here, we note the redundant condition for any sample sequence $r(\nu)$,

$$\sum_{\alpha=1}^{4} x_{i(\alpha)}^{r(\nu)} = 1, \quad (i=1, 2, \cdots, 9). \tag{2}$$

Categorical Discriminant Analysis. It was made according to Hayashi's quantification analysis (class II).[12,13] Quantification of the categorical data leads to the sample score value

$$y^{r(\nu)} = \sum_{i=1}^{9} \sum_{\alpha=1}^{4} x_{i(\alpha)}^{r(\nu)} a_{i(\alpha)}, \tag{3}$$

where r=1, 2 and ν=1, 2, $\cdots$, n_r. Coefficient of $a_{i(\alpha)}$ is the real number and is called category weight. Our analysis estimates the $a_{i(\alpha)}$ and $y^{r(\nu)}$ values in such a way that the two groups of 5′-splice site sequences (r=1) and sequences other than 5′-splice site (r=2) may be discriminated most distinctly. For this purpose, we calculate the mean value of sample scores within the group r as

$$\bar{y}^r = \frac{1}{n_r} \sum_{\nu=1}^{n_r} y^{r(\nu)} = \sum_{i=1}^{9} \sum_{\alpha=1}^{4} \bar{x}_{i(\alpha)}^{r} a_{i(\alpha)}, \tag{4}$$

where

$$\bar{x}_{i(\alpha)}^{r} = \frac{1}{n_r} \sum_{\nu=1}^{n_r} x_{i(\alpha)}^{r(\nu)}, \quad (r=1, 2). \tag{5}$$

On the other hand, the mean value of the total samples is

$$\bar{y} = \frac{1}{N} \sum_{r=1}^{2} n_r \bar{y}^r, \tag{6}$$

where $N=n_1+n_2=1751$ is the number of total samples. Using these values, variance of the total samples, σ^2, and variance between groups 1 and 2, σ_B^2, are given by

$$\sigma^2 = \frac{1}{N} \sum_{r=1}^{2} \sum_{\nu=1}^{n_r} (y^{r(\nu)} - \bar{y})^2, \tag{7}$$

$$\sigma_B^2 = \frac{1}{N} \sum_{r=1}^{2} n_r (\bar{y}^r - \bar{y})^2. \tag{8}$$

In order to discriminate the groups 1 and 2 most distinctly, we maximize the following η^2 value,

$$\eta^2 = \frac{\sigma_B^2}{\sigma^2}. \tag{9}$$

The procedure to maximize η^2 and to obtain $a_{i(\alpha)}$ values at the optimum condition was briefly summarized in the previous paper.[9]

Partial Correlation Coefficient. Using the optimum values of category weight $\hat{a}_{i(\alpha)}$, (i=1, 2, $\cdots$, 9; α=1, 2, 3, 4), we calculate the optimum values of sample scores $\hat{y}^{r(\nu)}$ by Eq 3. Then, we obtain from Eq. 4

$$\hat{y}^r = \frac{1}{n_r} \sum_{\nu=1}^{n_r} \hat{y}^{r(\nu)}, \quad (r=1, 2). \tag{10}$$

We define $x_i^{r(\nu)}$ by

$$x_i^{r(\nu)} = \sum_{\alpha=1}^{4} \hat{a}_{i(\alpha)} x_{i(\alpha)}^{r(\nu)}, \quad (\nu=1, 2, \cdots, n_r;\ i=1, 2, \cdots, 9). \tag{11}$$

Using these parameters, we calculate the following quantities;

$$\sigma_{ij} = \frac{1}{N} \sum_{r=1}^{2} \sum_{\nu=1}^{n_r} (x_i^{r(\nu)} - \bar{x}_i)(x_j^{r(\nu)} - \bar{x}_j), \quad (i, j=1, 2, \cdots, 9), \tag{12}$$

$$\sigma_{iy} = \frac{1}{N} \sum_{r=1}^{2} \sum_{\nu=1}^{n_r} (x_i^{r(\nu)} - \bar{x}_i)(\hat{y}^r - \hat{y}), \quad (i=1, 2, \cdots, 9), \tag{13}$$

$$\sigma_{yy} = \frac{1}{N} \sum_{r=1}^{2} n_r (\hat{y}^r - \hat{y})^2, \tag{14}$$

where

$$\bar{x}_i = \frac{1}{N} \sum_{r=1}^{2} \sum_{\nu=1}^{n_r} x_i^{r(\nu)}, \quad (i=1, 2, \cdots, 9), \tag{15}$$

$$\hat{y} = \frac{1}{N} \sum_{r=1}^{2} n_r \hat{y}^r. \tag{16}$$

Next, we introduce correlation coefficients between

items, r_{ij}, and between item and sample sequence, r_{iy}, by

$$r_{ij}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}},\ (i,j=1,2,\cdots,9), \tag{17}$$

$$r_{iy}=\frac{\sigma_{iy}}{\sqrt{\sigma_{ii}\sigma_{yy}}}=r_{yi},\ (i=1,2,\cdots,9,y). \tag{18}$$

We construct the 10-dimensional matrix, **R**, composed of those coefficients by

$$\mathbf{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{19} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{29} & r_{2y} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{91} & r_{92} & \cdots & r_{99} & r_{9y} \\ r_{y1} & r_{y2} & \cdots & r_{y9} & r_{yy} \end{bmatrix}. \tag{19}$$

This matrix leads to the inverse matrix, $\mathbf{R}^{-1}$, whose matrix elements are

$$\mathbf{R}^{-1}=\begin{bmatrix} r^{11} & r^{12} & \cdots & r^{19} & r^{1y} \\ r^{21} & r^{22} & \cdots & r^{29} & r^{2y} \\ \vdots & \vdots & & \vdots & \vdots \\ r^{91} & r^{92} & \cdots & r^{99} & r^{9y} \\ r^{y1} & r^{y2} & \cdots & r^{y9} & r^{yy} \end{bmatrix}. \tag{20}$$

By the use of these matrix elements, partial correlation coefficient between samples and the i-th item is given by

$$r_{iy\cdot 1,2,\cdots,i-1,i+1,\cdots,9}=\frac{-r^{iy}}{\sqrt{r^{ii}r^{yy}}},\ (i=1,2,\cdots,9). \tag{21}$$

where $r_{iy.1,2,\cdots,i-1,i+1,\cdots,9}$ indicates a correlation coefficient between i and y, removing influences of the other items of $1,2,\cdots,i-1,i+1,\cdots,9$. The magnitude of this coefficient tells us to what extent the i-th item may contribute to the discrimination of sample sequences between the first and second groups.

Results

We applied the above method to the $N=1751$ sample sequences, and discriminated 5′-splice site sequences ($r=1$) from sequences other than 5′-splice sites ($r=2$) most distinctly. The optimum values for category weights $\hat{a}_{i(\alpha)}$, ($i=1,2,\cdots,9$; $\alpha=1,2,3,4$), were estimated as shown in Table 2. Using these values together with Eqs. 10—18, we constructed the **R** matrix. Its matrix elements are symmetrical and the lower half of them is only given in Table 3. After transforming **R** matrix into $\mathbf{R}^{-1}$ matrix, partial correlation coefficients in Eq. 21 are calculated as 0.100, 0.162, 0.264, 0.424, 0.329, 0.175, 0.206, 0.351, and 0.157 for items $i=1,2,3,4,5,6,7,8$, and 9, respectively.

Discussion

So far, the splice junction signals have been summarized as "consensus" sequences. In the case of 5′-splice site consensus sequence, $\begin{matrix}C\\A\end{matrix}AG/GT\begin{matrix}A\\G\end{matrix}AGT$ has been proposed by Mount.[2] Attempts have been widely made to look for sequences relevant to the consensus sequence in the whole pre-mRNA. However, its technique is rather unsatisfactory, because actual splice site sequences differ from the consensus sequence to a greater or lesser degree. The consensus sequence is also ambiguous in which nucleotides are important to specify the 5′-splice site and which nucleotides are less important. The optimum values of category weight $\hat{a}_{i(\alpha)}$, ($i=1,2,\cdots,9$; $\alpha=1,2,3,4$), may answer this problem. In this case, the item, i, corresponds to the position of nucleotide in 9-nucleotide sequence, while the category, α, denotes the kind of nucleotide at the i-th position. Relative importance of the kind of nucleotide at each position was previously discussed in terms of the $\hat{a}_{i(\alpha)}$ value.[9]

Next, we consider relative importance of item (position of nucleotide) in 9-nucleotide sequence. There

Table 2. Optimum Category Weight Values of $\hat{a}_{i(\alpha)}$'s Calculated with Categorical Discriminant Analysis of 5′-Splice Site Sequences[a]

Item No. (i)	Category Name (α)	$\hat{a}_{i(\alpha)}$
1	A-1	0.5300
	G-2	−0.5163
	C-3	1.9291
	T-4	−1.4366
2	A-1	3.1444
	G-2	−1.9297
	C-3	−0.3227
	T-4	−1.3873
3	A-1	−1.9442
	G-2	5.6465
	C-3	0.2679
	T-4	−3.1345
4	A-1	−3.6933
	G-2	9.8117
	C-3	−4.9457
	T-4	−2.9199
5	A-1	−3.1491
	G-2	−4.8216
	C-3	−2.7129
	T-4	5.2915
6	A-1	2.7188
	G-2	1.4351
	C-3	−2.7449
	T-4	−1.8026
7	A-1	4.0432
	G-2	−1.7958
	C-3	−1.1522
	T-4	−1.9755
8	A-1	−3.4019
	G-2	7.9204
	C-3	−3.6093
	T-4	−1.8206
9	A-1	−1.1933
	G-2	−2.8199
	C-3	0.0108
	T-4	2.3215

a) For further details, see the text.

Table 3. The **R** Matrix and Its Matrix Elements r_{ij}, ($i, j=1, 2, \cdots, 9, y$), Shown by Eq. 19 in the Text[a)]

Item No.	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	0.171	1.000								
3	0.120	0.034	1.000							
4	0.103	0.129	0.197	1.000						
5	0.054	0.050	0.093	0.107	1.000					
6	0.036	0.037	0.113	0.140	0.094	1.000				
7	0.055	0.032	0.147	0.166	0.100	0.098	1.000			
8	0.099	0.051	0.123	0.175	0.150	0.133	0.104	1.000		
9	0.018	−0.009	−0.035	0.053	0.104	−0.025	−0.013	−0.026	1.000	
10	0.197	0.206	0.352	0.506	0.377	0.264	0.296	0.416	0.131	1.000

a) The tenth item in this table is read as $i, j=y$. Since $r_{ij}=r_{ji}$, the lower half of the matrix elements is only given.

have been known a few approaches to measure quantitatively contribution of item to the discrimination between two groups of 5′-splice site sequences and sequences other than 5′-splice sites. One approach is to calculate partial correlation coefficient between sample sequences and the i-th item under the optimum condition of discrimination. This was done in the previous sections. Here, the larger the value of such a partial correlation coefficient, the greater the contribution of the i-th item to the discrimination between the two groups. As is shown in Results, the largest value of the partial correlation coefficient was found to be 0.424 at the fourth position. The next largest values range as 0.351 at the eighth position, 0.329 at the fifth position, and 0.264 at the third position, while the smallest values were found to be 0.100 at the first position and 0.157 at the ninth position. Therefore, the fourth, eighth and fifth positions are more important to specify 5′-splice site sequences, whereas the first, ninth and second positions are less important. This prediction is consistent with the experimental results that the GT dinucleotide at the fourth and fifth positions and the G nucleotide at the eighth position are almost invariant in every sequence of 5′-splice junction.[2)] It is also found that nucleotides at the other positions are rather variable, differing from junction to junction. The small values of partial correlation coefficients at the other positions may correspond to this phenomenon.

Another approach to measure relative importance of item (position of nucleotide) is to estimate the range of category weights, R_i, which is defined by

$$R_i = \max_{\alpha}(\hat{a}_{i(\alpha)}) - \min_{\alpha}(\hat{a}_{i(\alpha)}). \quad (22)$$

This parameter implies that the larger the R_i value, the greater the influence of $\hat{a}_{i(\alpha)}$'s at the i-th item on the total sample score $y^{r(\nu)}$. Using the data of $\hat{a}_{i(\alpha)}$ in Table 2, the R_i values are calculated as R_1=3.3657, R_2=5.0741, R_3=8.7810, R_4=14.7574, R_5=10.1131, R_6=5.4637, R_7=6.0187, R_8=11.5297, and R_9=5.1414. These results show that the degree of importance of items aligns in the order of i=4, 8, 5, 3, 7, 6, 9, 2, and 1. It appears that this order of importance is in good accordance with that of the previous partial correlation coefficients given in the previous section of Results.

Several other evidences which support our conclusion are provided by experimental results on mutational changes of nucleotides around 5′-splice junctions. One example is the study of Treisman et al.,[6)] who reported three mutants of human β-globin gene. The β-globin gene sequence is separated into three exons by two introns, and there are two positions of 5′-splice junctions.[11)] In the normal gene, the 9-nucleotide sequence at the 5′-splice site of the first intron is ··· CAG/GTTGGT ···. All of the above three mutant genes (i)—(iii) contain single nucleotide changes within the 9-nucleotide sequence, causing β-thalassemia phenotype. The β-thalassemia comprises a group of diseases in which the synthesis of normal β-globin polypeptide is either absent or reduced. In the (i) mutant, a G→A transition took place at the fourth position of the 9-nucleotide sequence, resulting in ··· CAGATTGGT ···. The (ii) mutant underwent a G→C transversion at the eighth position, and has a sequence ··· CAGGTTGCT ···. In the (iii) mutant, a T→C transition took place at the ninth position, resulting in ··· CAGGTTGGC ···. Treisman et al. analyzed the RNA synthesis and expression of the mutant gene after its introduction into cultured mammalian cells on a suitable vector.[6)] In all of the (i)—(iii) genes, the mutational changes inactivated the authentic 5′-splice site signal at the first intron, failing in normal β-globin mRNA synthesis or reducing the synthesis. The mutation at the fourth position completely inactivated the 5′-splice site of the first intron, giving no production of normal β-globin mRNA. On the other hand, the mutation at the eighth position resulted in partial inactivation of the 5′-splice site, while the mutation at the ninth position showed a less dramatic decrease in the production level of normal mRNA. Although the latter two mutations led to β^+-thalassemia, the eighth position mutation caused a more pronounced deficiency of β-globin synthesis than did the ninth position mutation. These experimental results may correspond well to our categorical discriminant analysis. Since the relative importance of items (position of nucleotide) is greatest at the fourth

position, mutation at this position will influence the 5′-splice site signal most strongly. This may be a reason why the G→A change at the fourth position of the (i) mutant inactivated completely the 5′-splice site. According to our discriminant analysis, the next important item corresponds to the eighth position, and the least important item, to the ninth position. In accordance with this view, both of the eighth and ninth position mutations caused partial inactivation of the 5′-splice site, but the deficiency of mRNA synthesis is more severe in the eighth position mutant than in the ninth position mutant.

Another example of mutations which supports our conclusion is the work of Wieringa et al.,[7] who reported mutagenic inactivation of an authentic 5′-splice site of the second intron in rabbit β-globin gene. The 9-nucleotide sequence of the normal gene at this site is ···AGG/GTGAGT···. Six mutant genes were prepared by site-directed mutagenesis involving purine transitions; (i) A→G transition at the first position, (ii) G→A at the second position, (iii) G→A at the third position, (iv) G→A at the fourth position, (v) G→A at the sixth position and (vi) A→G at the seventh position. Expression of these genes in HeLa cells revealed that mutations of (i)—(iii), (v) and (vi) affected neither the quantity nor the structure of correct β-globin mRNA. However, the mutation of (iv) inactivated completely the 5′-splice site, giving no correctly spliced β-globin mRNA. These experimental results seem to be consistent with the previous discriminant analysis, because the 5′-splice site sequence is influenced most strongly by the mutational change at the fourth position, and because the effects of the other positions are much weaker than that of the fourth position.

References

1) M. R. Lerner, J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz, *Nature*, **283**, 220 (1980).
2) S. M. Mount, *Nucl. Acids Res.*, **10**, 459 (1982).
3) B. Ruskin, A. R. Krainer, T. Maniatis, and M. R. Green, *Cell*, **38**, 317 (1984).
4) R. Reed and T. Maniatis, *Cell*, **41**, 95 (1985).
5) S. M. Mount, I. Petterson, M. Hinterbergen, A. Karmas, and J. A. Steitz, *Cell*, **33**, 509 (1983).
6) R. Treisman, S. H. Orkin, and T. Maniatis, *Nature*, **302**, 591 (1983).
7) B. Wieringa, F. Meyer, J. Reiser, and C. Weissmann, *Nature*, **301**, 38 (1983).
8) R. Staden, *Nucl. Acids Res.*, **12**, 505 (1984).
9) Y. Iida, *Compt. Appl. Biosci.*, **3**, 93 (1987).
10) GenBank, Genetic Sequence Data Bank, Release 40.0, BBN Laboratories, U. S. A. (1986).
11) R. M. Lawn, A. Efstratiadis, C. O'Connell, and T. Maniatis, *Cell*, **21**, 647 (1980).
12) C. Hayashi, *Ann. Inst. Statist. Math.*, **2**, 35 (1950).
13) C. Hayashi, *Ann. Inst. Statist. Math.*, **3**, 69 (1952).