  Bull. Chem. Soc. Jpn., 61, 3801—3805 (1988)

Classification of DNA Sequences by Hayashi's Quantification (III) Method. Sequences Involving 5′-Splice Signals in Pre-mRNA of Higher Eukaryotes' Genes

Yôichi IIDA
Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060
(Received April 15, 1988)

Concerning the signals which direct the excision of introns from mRNA precursors in higher eukaryotes' genes, a consensus sequence, $\overset{C}{\underset{A}{}}$AG/GTGAGT, has been proposed with a 5′-splice site, but actual 5′-splice site sequences differ from it to a greater or lesser degree. In the present paper, we took a set of 1751 sequences, which are composed of 9-nucleotide sequences including 5′-splice signals and of other sequences including no such signals. Hayashi's quantification analysis (Class III) was applied to such a system. Assuming no external criterion for 5′-splice signals, we examined how the 1751 sequences show responses to the positions and species of nucleotides. The optimum classification led to a fair separation of the 5′-splice signal sequences from sequences including no such signals. The results were compared with those derived previously by the use of the quantification (Class II) method.

Many eukaryotic genes are interrupted by introns, which have been removed from the mRNA precursors (pre-mRNAs) by the use of the RNA splicing mechanism. It became clear that the splicing pathway could be divided into two stages. In the first stage, the pre-mRNA is cleaved at the 5′-splice site of the intron, generating the first exon and a species composed of the intron and the following exon. The latter intermediate is present in a lariat form, where the 5′-end of the intron is joind by a 2′–5′ phosphodiester linkage to an A residue upstream of the 3′-splice junction. The sequence encompassing the A residue is called "the branch-point sequence". The second stage involves cleavage at the 3′-splice site and ligation of the exons, with a concomitant release of the intron lariat. Both stages are related to several reactions which occur in a large multicomponent complex termed a "spliceosome". Small nuclear ribonucleoprotein particles (snRNPs) including U1, U2, U4—6 are important components of the spliceosome (for reviews, see Refs. 1—3).

Nucleotide sequences defining 5′- and 3′-splice junctions in pre-mRNAs have been found to share considerable homology in higher eukaryotes.[4] In the case of 5′-splice sites, a consensus sequence, 5′-(exon)-$\overset{C}{\underset{A}{}}$AG/GTGAGT-(intron)-3′, has been known, where the stroke (/) indicates the boundary between exon and intron. However, this is only a qualitative description of the splice signal. The consensus sequence cannot tell what degree of matching between the observed sequence and the consensus sequence is necessary to specify the exact splice site. Another major problem involves the relative importance of each of the nucleotides in the sequence. That is, mutational studies have demonstrated that the importance of nucleotide matching to the consensus sequence differs from position to position. For example, a few positions, such as the G at the 5′- end of intron, are absolutely necessary, while mutations in the other positions, though they reduce the efficiency of correct splicing, sometimes do not affect correct splicing seriously.[5,6] In order to study such splice signals more quantitatively and to ascertain the relative importance of each of the nucleotides, we previously studied multivariate statistical analysis and then performed Hayashi's quantification analysis (Class II).[7,8] For this purpose, we took 1751 samples of the 9-nucleotide sequences belonging to either of two groups. Group 1 is composed of sequences including 5′-splice signals, while sequences of Group 2 include no such signals. In the Class II analysis, using external criterion for the 5′-splice signals, we searched for the optimum condition under which sequences of the two groups could be discriminated most distinctly; we also determined the optimum weights for the position and species of nucleotides.

In the present paper, we utilizes another approach of Hayashi's quantification analysis (Class III)[9,10] to the same set of 1751 sequences. In contrast to the Class II method, the Class III method distributes the total of 1751 sequences most distinctly and determines the optimum weights for the position and species of nucleotides, assuming an absence of external criterion as to whether a sequence may belong to the 5′-splice sequence or not. It was found that the classification by this technique could well separate the sequences of Group 1 from those of Group 2. This approach enables us, to a considerable extent, to characterize 5′-splice site sequences in terms of position and species of nucleotides, even if there is no external criterion for 5′-splice signals.

Analysis by Quantification (III) Method

The nucleotide sequence data used in the present paper are the same as those used by quantification (Class II) in a previous paper.[7] In Table 1, some of the 9-nucleotide sequences are reproduced. Here, sequences belonging to Group 1 include 5′-splice signals, where nine nucleotides (composed of three at the 3′-end of the exon and six at the 5′-end of the intron) are

truncated in each sequence. In the group, we summarized 155 sequences (i_1=1, 2, ···, 155), which were taken from authentic 5′-splice site sequences in various mammalian genes containing introns: globin genes, insulin genes, etc.[11] On the other hand, Group 2 is composed of sequences other than the 5′-splice site. Those sequences were taken from the human β-globin gene, as was described in the previous paper.[7] Excluding the two sequences of the authentic 5′-splice sites in the β-globin, we gave 1596 sequences (i_1=156, 157, ···, 1751) composed of 9-nucleotides in Group 2. Therefore, a total of N_1=1751 sample sequences (i_1=1, 2, ···, N_1) are summarized in Table 1.

In constrast to the Class II analysis, the Class III method does not use the criterion of Groups 1 and 2, but examines how uniquely the 1751 sequences respond to items and categories.[9,10] This method is similar to the dual scaling approach of Maung and Guttman.[12,13] In our case, there are nine items, identified k=1, 2, ···, 9. They correspond to the positions of nucleotides in the 9-nucleotide sequence, being defined in terms of their order from the 5′ to 3′ ends of the sequence. On the other hand, the category denotes the kind of nucleotide, where A, G, C, or T is specified by α=1, 2, 3, or 4 respectively at every item.

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 is composed of sequences other than 5′-splice sites. For further details, see text.

We introduce a running suffix $i_2=4(k-1)+\alpha$, (i_2=1, 2, ···, N_2; N_2=36), and transform the sequence data of Table 1 into the item-category data by using a dummy variable, $n(i_1, i_2)$. It takes a value of 1 if the sample sequence (i_1) has a nucleotide (α) at the position (k), but otherwise it takes 0. Such item-category data are shown in Table 2.

The Class III method rearranges sample sequences and item-categories in such a way that sample sequences showing similar responses to items and categories are gathered together. If we obtain the optimum condition for this, we can find what responses to items and categories are similar to each other. Moreover, we can classify not only items and categories, but also sample sequences in the most characteristic way. For this purpose, let us give the quantity $x(i_1)$ to the i_1-th sample sequence (i_1=1, 2, ···, N_1) and a quantity $y(i_2)$ to the i_2-th item-category ($i_2=4(k-1)+\alpha$; i_2=1, 2, ···, N_2). Then we introduce the correlation coefficient, $r(X, Y)$, by the use of

$$r(X,Y) = V(X,Y)/\{S(X)S(Y)\}, \tag{1}$$

$$V(X,Y) = (1/N)\sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2} x(i_1)y(i_2)n(i_1,i_2) - \bar{x}\bar{y}, \tag{2}$$

$$S(X) = \{(1/N)\sum_{i_1=1}^{N_1} x(i_1)^2 n_1(i_1) - (\bar{x})^2\}^{1/2}, \tag{3}$$

$$S(Y) = \{(1/N)\sum_{i_2=1}^{N_2} y(i_2)^2 n_2(i_2) - (\bar{y})^2\}^{1/2}, \tag{4}$$

where;

$$N = \sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2} n(i_1,i_2), \tag{5}$$

$$\bar{x} = (1/N)\sum_{i_1=1}^{N_1} x(i_1)n_1(i_1), \tag{6}$$

$$\bar{y} = (1/N)\sum_{i_2=1}^{N_2} y(i_2)n_2(i_2), \tag{7}$$

$$n_1(i_1) = \sum_{i_2=1}^{N_2} n(i_1,i_2), \tag{8}$$

$$n_2(i_2) = \sum_{i_1=1}^{N_1} n(i_1,i_2). \tag{9}$$

In order to classify sample sequences together with items and categories most distinctly, we maximize $|r(X, Y)|$. That is, if the correlation coefficient

Table 2. Categorical Data to be Used in Hayashi's Quantification Analysis (Class III)[a]

No.	Group	Item 1				Item 2				···	Item 9			
		Category				Category					Category			
		1	2	3	4	1	2	3	4		1	2	3	4
1	1	0	1	0	0	1	0	0	0	···	0	1	0	0
2	1	1	0	0	0	1	0	0	0	···	0	0	1	0
:	:			:				:					:	
155	1	1	0	0	0	0	1	0	0	···	0	0	1	0
156	2	1	0	0	0	0	0	1	0	···	0	0	0	1
:	:			:				:					:	
1751	2	0	0	0	1	0	0	0	1	···	0	0	1	0

a) For the 9-nucleotide sequences given in Table 1, the position of nucleotide (item) and the kind of nucleotide (category) are specified with a dummy variable (0 or 1). In the Class III analysis, the item number (k) and the category number (α), where k=1, 2, ···, 9 and α=1, 2, 3, 4, are further combined into a parameter: $i_2=4(k-1)+\alpha$, where i_2=1, 2, ···, 36. For further details, see text.

approaches 1, the arrangements of $n(i_1, i_2)=1$ versus i_1 and i_2 become linear. In this case, the (i_1) sequence may be well characterized by the item-category (i_2). The maximization of $|r(X, Y)|$ can be done by differentiating:

$$\partial r(X, Y)^2/\partial x(i_1) = 0, \quad (i_1 = 1, 2, \cdots, N_1), \tag{10}$$
$$\partial r(X, Y)^2/\partial y(i_2) = 0, \quad (i_2 = 1, 2, \cdots, N_2), \tag{11}$$

where we can put $\bar{x}=0$ and $\bar{y}=0$ without loss of generality. These equations lead to:

$$\sum_{i_2'=1}^{N_2}\sum_{i=1}^{N_1} n(i, i_2)n(i, i_2')y(i_2')/n_1(i) = r(X, Y)^2 n_2(i_2)y(i_2), \quad (i_2 = 1, 2, \cdots, N_2), \tag{12}$$

$$x(i_1) = \{1/r(X, Y)\}\sum_{j=1}^{N_2} n(i_1, j)y(j)/n_2(j), \quad (i_1 = 1, 2, \cdots, N_1). \tag{13}$$

Eq. 12 can be solved by eigen value problem with the aid of computer calculation. It has the maximum eigen value of 1. However, this value is a trivial solution and is excluded, because it gives an inadequate value of $\bar{y}=(1/N)\sum_{i_2=1}^{N_2} y(i_2)n_2(i_2)\neq 0$. Therefore, we take the second largest eigen value of $r(X, Y)^2$. This value, together with Eq. 12, leads to the eigen vector of $y_2=\{y(12), \cdots, y(i_2 2), \cdots, y(N_2 2)\}^t$, where the suffix 2 corresponds to the second largest eigen value and where the superscript t indicates the transposition of the vector. From y_2 together with Eq. 13, we can obtain the eigen vector of $x_2=\{x(12), \cdots, x(i_1 2), \cdots, x(N_1 2)\}^t$. If we arrange the components of x_2 and y_2 in the order of their magnitudes, we can classify not only items and categories, but also sample sequences most distinctly.

In a similar way, if we obtain the third and fourth largest eigen values of Eq. 12, we can calculate the corresponding eigen vectors: y_3, y_4, and x_3, x_4. If a sample sequence i_1, $(i_1=1, 2, \cdots, N_1)$, is plotted by the use of the coordinate of $\{x(i_1 2), x(i_1 3)\}$, we can map sample sequences on a two-dimensional graph most distinctly. It is also possible to distribute the items and categories i_2, $(i_2=1, 2, \cdots, N_2)$, in the most characteristic way by plotting $\{y(i_2 2), y(i_2 3)\}$.

Results and Discussion

For the item-category data of Table 2, the second, third, and fourth largest eigen values of $r(X, Y)^2$ were calculated as 0.2155, 0.1571, and 0.1511 respectively. From Eq. 12, we estimated the components of y_2, y_3 and y_4 (normalized item-category scores), where the variance of the components was normalized to unity. In view of the magnitudes of $r(X, Y)^2$, the vector of y_2 distinguishes item-categories and sample sequences most strongly, while the vectors of y_3 and y_4 have less distinguishing power. The degree of distinction by y_3 is almost the same as that by y_4. Such normalized scores are summarized for item (k) and category (α), where $i_2=4(k-1)+\alpha$, $(i_2=1, 2, \cdots, 36)$. For example, we give in Fig. 1 a graph of normalized item-category scores obtained by plotting $\{y(i_2 2), y(i_2 3)\}$ with $i_2=1, 2, \cdots, 36$. This graph corresponds to the second and third largest eight values and shows the most distinct distribution among item-categories, where some form a cluster. For item-categories $(k-\alpha)$, (4–2), (3–2), (8–2), (7–1), (2–1), (6–2), (5–4), (6–1), (1–1), and (1–3) form a cluster and show relatively high scores in the direction of the second largest eigen value. These item-categories come from G at the fourth position, G at the third position, G at the eighth position, A at the seventh position, A at the second position, G at the sixth position, T at the fifth position, A at the sixth position, A at the first position, and C at the first position respectively in the 9-nucleotide sequence of Table 1. These nucleotides and positions coincide well with the consensus sequence $\overset{C}{A}AG/GT\overset{A}{G}AGT$ proposed by Mount,[4] possessing a strong power to distinguish the 5′-splice site sequence. From the magnitudes of the $y(i_2 2)$ values, G at the fourth position seems to be most effective. This result may correspond to the invariant G nucleotide at the 5′-end of every intron. As is shown in Fig. 1, the item-categories other than those due to the consensus

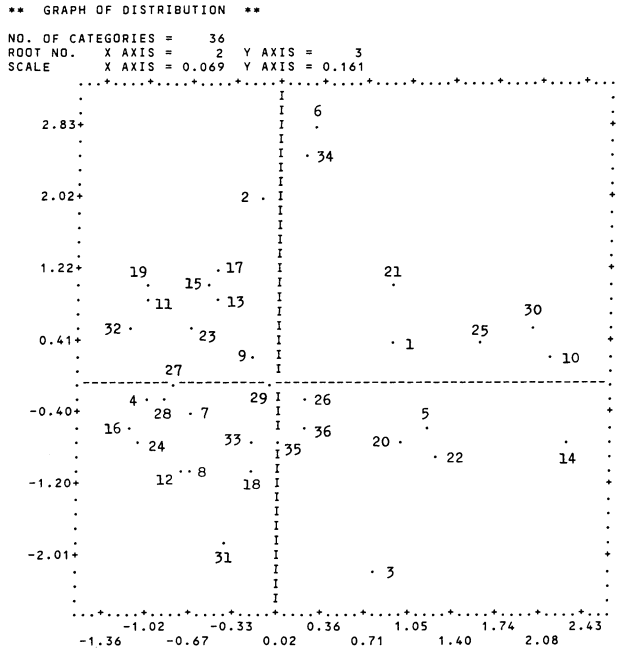

Fig. 1. Two-dimensional mapping of item-categories in Hayashi's quantification analysis (Class III) of 1751 members of 9-nucleotide sequences given in Tables 1 and 2. Abscissa shows normalized score of $y(i_2 2)$, which is calculated from the second largest eigen value of Eq. 12. Ordinate shows normalized score of $y(i_2 3)$ calculated from the third largest eigen value. Parameter i_2, $(i_2=1, 2, \cdots, 36)$, is related to item (k) and category (α) by $i_2=4(k-1)+\alpha$. Note that points of $i_2=1$, 3, 5, 10, 14, 20, 21, 22, 25, and 30 form a cluster. For further details, see text.

sequence form another cluster. Althogh some of them are located around the positive region of $y(i_2 2)$, most of the item-categories other than the consensus sequence lie in the negative region of $y(i_2 2)$. It appears that they make negative contribution to 5′-splice site signals.

Next let us examine how uniquely the 1751 sequences respond to the positions and species of nucleotides. For this purpose, we will calculate the sample scores of $x(i_1 2)$, $x(i_1 3)$ and $x(i_1 4)$, (i_1=1, 2, ···, N_1). As was mentioned above, the sample score of $x(i_1 2)$ can classify the 1751 sequences most distinctly. Therefore, we first arranged their $x(i_1 2)$ values in the order of their magnitudes, as is shown in Fig. 2. Here, we categorized the sample scores at appropriate equal intervals, and constructed a graph of the distribution and the frequency percentage versus the categorized sample score ($z=x(i_1 2)$) for the 1751 members of the 9-nucleotide sequences shown in Table 1. For the sake of comparison, the lower half of the graph illustrates a histogram of the sequences belonging to Group 1 (155 members of the 5′-splice site sequences), while the upper half illustrates those belonging to Group 2 (1596 members of sequences other than the 5′-splice sites). As has already been mentioned, the present Class III approach assumes no external criterion as to whether or not a sequence may belong to the 5′-splice site sequence. However, Fig. 2 clearly shows that the Class III approach can discriminate, to considerable extent, sequences of Group 1 from those of Group 2. In order to estimate the rate of discrimination in terms of the sample score z, we note the overlapping of the distribution curve of Group 1, $f_1(z)$, and that of Group 2, $f_2(z)$, around z=0.55. The optimum discrimination rate, P, at the critical sample score, z^*, may be determined by using this relation:

$$P=\int_{-\infty}^{z^*} f_2(z)\mathrm{d}z=\int_{z^*}^{\infty} f_1(z)\mathrm{d}z=1-\int_{-\infty}^{z^*} f_1(z)\mathrm{d}z. \quad (14)$$

Here, the values for $\int_{-\infty}^{z^*} f_1(z)\mathrm{d}z$ and $\int_{-\infty}^{z^*} f_2(z)\mathrm{d}z$ can be obtained by estimating the cumulative frequency percentage curves for the distributions of sample scores in Groups 1 and 2 respectively. If we then plot $\int_{-\infty}^{z} f_2(z)\mathrm{d}z$ and $1-\int_{-\infty}^{z} f_1(z)\mathrm{d}z$ versus z, the intersect of their curves will give the values for z^* and P. In this way, z^*=0.572 and P=96.8% were estimated from the cases given in Fig. 2. This result implies that any sequence with a sample score of $z>0.572$ may be classified into Group 1 with a 96.8% probability, while a sequence with a sample score of $z<0.572$ may be placed in Group 2 with a 96.8% probability.

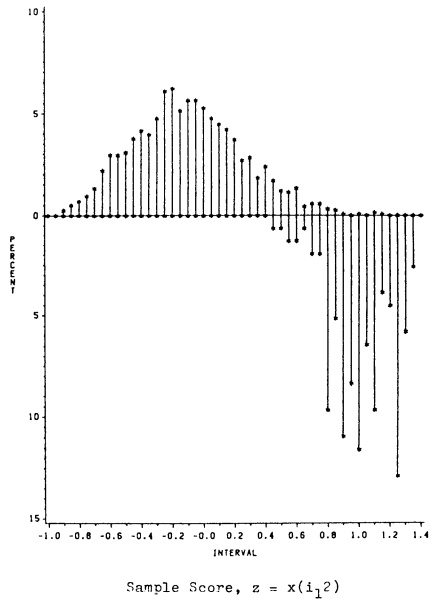

Fig. 2. Graph of distribution and frequency percentage versus categorized sample score ($z=x(i_1 2)$) for 1751 members of 9-nucleotide sequences shown in Table 1. Values of $x(i_1 2)$, (i_1=1, 2, ···, 1751), are calculated from the second largest eigen value of Eqs. 12 and 13. For purposes of comparison, the lower half of the graph shows a histogram of the sequences belonging to Group 1 (155 members of 5′-splice site sequences in Table 1), while the upper half, those belonging to Group 2 (1596 members of sequences other than 5′-splice sites). For further details, see text.

Let us compare the discrimination rate of the present Class III analysis with that of the Class II analysis reported previously.[7] The Class II analysis used an external criterion for Group 1 (5′-splice site sequences) and Group 2 (sequences other than 5′-splice sites), and discriminated the two groups most distinctly by maximizing the σ_B^2/σ^2 value, where σ^2 is the variance of the total samples and σ_B^2, the variance between Groups 1 and 2. Then, the optimum discrimination rate, P, was estimated to be as high as 98.2%. It is reasonable that this value is larger than the 96.8% value of the present Class III analysis. However, it is important to note that, even without any external criterion, Groups 1 and 2 can be discriminated as much as 96.8%. This implies that the sequences belonging to Group 1 show unique responses to the positions and species of nucleotides and that the 5′-splice signal may be composed of a particular pattern of nucleotide sequences.

In order to examine this situation more extensively, we map 1751 sequences on a two-dimensional graph by plotting the coordinates of $\{x(i_1 2), x(i_1 3)\}$, (i_1=1, 2, ···, N_1); as is shown in Fig. 3. Here, any sequence is represented by a dot (·), but multiple overlapping of dots is given by signs (*, +, etc.) (see the captions of Fig. 3). It is found that the 1751 sequences may be roughly classified into two clusters on the two-dimensional map. The first cluster lies around the origin of the coordinate: most of the members of this cluster belong to Group 2 (sequences other than 5′-splice sites). Those members are distributed almost symmetrically around the point of origin, indicating that their sequences exhibit no particular (random)

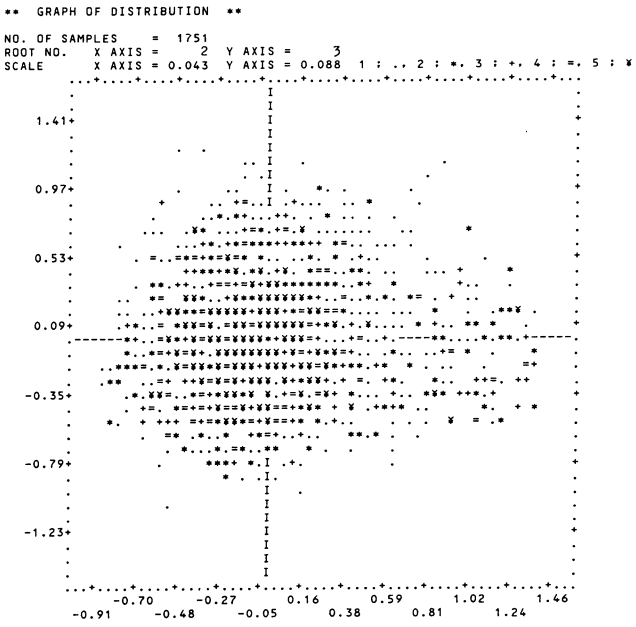

Fig. 3. Two-dimensional mapping of 1751 members of 9-nucleotide sequences (i_1=1, 2, ···, 1751) in Hayashi's quantification analysis (class III). Abscissa shows value of $x(i_12)$, which is calculated from the second largest eigen value of Eqs. 12 and 13. Ordinate shows value of $x(i_13)$ calculated from the third largest eigen value. Each sample sequence is given by a dot (·), but multiple overlapping of dots is indicated by signs (*, +, etc.). For example, the sign (*) shows double overlapping of dots. For further details, see text.

responses to the positions and species of nucleotides. On the other hand, the second cluster is out of the first cluster, deviating remarkably from the point of origin and lying around the positive region of $x(i_12)$=0.55—1.50. Most of the members of the second cluster belong to Group 1, as has been described previously. The distribution of those members versus the $x(i_13)$ axis is almost symmetrical, suggesting that the $x(i_13)$ value has no appreciable power to separate Group 1 (5′-splice site sequences) from Group 2. These observations are in good accordance with our previous calculation, where $x(i_13)$ comes from the third largest eigen value of Eq. 12 and has a lesser correlation coefficient than has $x(i_12)$. Even in the two-dimensional mapping of the 1751 sequences, the two clusters can be mostly separated by the use of the $x(i_12)$ parameter.

We further carried out two-dimensional mapping of the 1751 sequences (i_1=1, 2, ···, N_1) by plotting {$x(i_12)$, $x(i_14)$} or {$x(i_13)$, $x(i_14)$} (data not shown). In these cases also, the parameters of $x(i_13)$ and $x(i_14)$ are found to have no practical ability to discriminate: only the parameter of $x(i_12)$ is very effective.

Concluding Remarks

Hayashi's quantification analysis (Class III) is similar to the principal-component analysis for data of samples composed of continuous variables. However, the Class III analysis differs from the principal-component analysis in several points:

(1) The Class III analysis deals with samples composed of categorical data or item-category data.

(2) In the Class III analysis, we give appropriate quantities to both samples and item-categories in such a way that the correlation coefficient between samples and item-categories may be maximized (see Eq. 12).

(3) In this process, using the second largest eigen value, we can simultaneously classify not only item-categories, but also samples, in the most characteristic way. We can map them in two- or three-dimensional coordinates, using the third and fourth largest eigen values.

The Class III approach is applicable when samples have no external criterion. In the present investigation, the 1751 sample sequences composed of 5′-splice site sequences and sequences other than 5′-splice sites were analyzed with the Class III method, and a discrimination rate of up to 96.8% was attained. If there is an external criterion, the Class II method gives a higher rate of discrimination. However, if sample sequences exhibit unique responses to the positions and species of nucleotides, the Class III approach may give a satisfactory rate of discrimination even without any external criterion. In this respect, the Class III method is "an automatic classification technique" of unknown nucleotide sequences, and it may become generally useful in the analysis of nucleotide sequences other than those of splicing.

References

1) B. Ruskin and M. R. Green, *Science*, **229**, 135 (1985).
2) T. Maniatis and R. Reed, *Nature (London)*, **325** 673 (1987).
3) M. M. Konarska and P. A. Sharp, *Cell*, **49**, 763 (1987).
4) S. M. Mount, *Nucleic Acids Res.*, **10**, 459 (1982).
5) R. Treisman, S. H. Orkin, and T. Maniatis, *Nature (London)*, **302**, 591 (1983).
6) M. Aebi, H. Hornig, R. H. Padgett, J. Reiser, and C. Weissmann, *Cell*, **47**, 555 (1986).
7) Y. Iida, *Comput. Appl. Biosci.*, **3**, 93 (1987).
8) Y. Iida, *Bull. Chem. Soc. Jpn.*, **60**, 2977 (1987).
9) C. Hayashi, *Ann. Inst. Statist. Math.*, **3**, 69 (1952).
10) C. Hayashi, *Bull. Inst. Statist. Math.*, **4**, 19 (1956).
11) GenBank, Genetic Sequence Data Bank, Release 55.0, USA (1988).
12) L. Guttman, The Prediction of Personal Adjustment, ed by The Committee on Social Adjustment, New York: Social Science Research Council, 319 (1941).
13) K. Maung, *Ann. Eugenics*, **11**, 189 (1941).