ICOT Technical Report: TR-0876

TR-0876

Statistical Analysis of Human DNA Sequences in the Vicinity of POLY(A) SIGNAL

by
T. Yada (MRI), M. Ishikawa, Y. Totoki
& K. Okubo (Osaka Univ.)

May, 1994

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5

Institute for New Generation Computer Technology

STATISTICAL ANALYSIS OF HUMAN DNA SEQUENCES IN THE VICINITY OF POLY(A) SIGNAL

Tetsushi Yada, Masato Ishikawa †, Yasushi Totoki † and Kousaku Okubo ‡

Systems Science Department, Mitsubishi Research Institute, Inc. 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100, Japan

†*Institute for New Generation Computer Technology (ICOT)*

‡*Institute for Molecular and Cellular Biology, Osaka University*

Abstract

We have applied a statistical method to the analysis of human DNA sequences in the vicinity of the poly(A) signal, and have obtained new sequence patterns for 3'UTR. Although a large number of AATAAA subsequences exist in human DNA, some of them are identified as conserved subsequences of the eukaryote poly(A) signal. The main purpose of this study is to identify the environments in which AATAAA is meaningful as a conserved subsequence of the poly(A) signal. To distinguish the signal patterns possessing AATAAA from non-signal patterns, we have adopted a discriminant analysis based on class II quantification theory. According to the theory, we have assigned category weights to bases A, T, C, and G at each position in a DNA sequence. A set of category weight values is used for the discrimination of signal/non-signal patterns and the identification of conserved subsequences. The range of category weights at each position reflects the relative importance of the position in the discrimination of patterns. To examine features of the DNA sequence, we further evaluated correlations between range values at two distant positions in a DNA sequence and investigated the behavior of these correlations. We observed the following two features from the analysis. (1) In the case of human DNA sequences, a base C frequently appears on the upstream side of the AATAAA subsequence and a base T or C often appears downstream. (2) In the vicinity of the poly(A) signal, the correlation between the range values is at a maximum every 12 base pairs. (1) implies that CAATAAA$\begin{matrix}\mathrm{T}\\ \mathrm{C}\end{matrix}$ can be regarded as a conserved subsequence of the poly(A) signal. (2) indicates that the periodicity of 12 base pairs almost corresponds to the base pair numbers in one pitch of the DNA double-helix structure.

1 Introduction

As a result of recent advances in DNA sequencing and the instigation of a number of large genome sequencing projects, there has been an upsurge in the study of reliable gene identification methods [1]. Using these methods, the last few years has seen the development of a variety of computational tools for predicting gene structures in uncharacterized DNA

sequences [2, 3, 4, 5, 6, 7]. In general, the prediction of gene structures entails three steps. (1) Predicting each functional site, such as initiation codons, donor sites, acceptor sites and poly(A) signals. (2) Assembling a set of functional sites into a gene. (3) Evaluating and filtering the assembled gene based on information about the set of functional sites, as well as regularities in the coding and the intron regions. One of the main subjects of research into gene identification methods is the prediction of the functional sites with a satisfactory degree of accuracy. However, with our current level of knowledge of functional motifs, the methods are not sufficiently accurate for practical use. Therefore, a reliable method which is applicable for the recognition of any type of functional sites is desired earnestly.

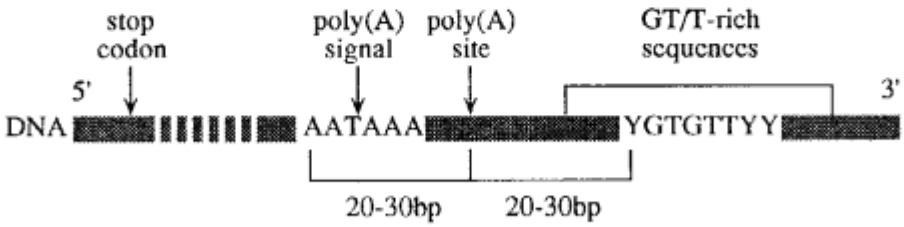

Figure 1: The characteristic features in 3'UTR of eukaryote genes

The AATAAA and YGTGTTYY subsequences locate 20 – 30 bp upstream and downstream of poly(A) site, respectively. Both of them are required for correct cleavage and polyadenylation of pre-mRNA. However, additional signals must be exist because of sequence diversity of 3'UTR in eukaryote genes.

The characteristic features in 3'UTR of eukaryote genes are shown in Figure 1. The poly(A) signals in eukaryote DNA sequences are located $20 - 30$ bp upstream of the poly(A) site. They are required to ensure the cleavage and polyadenylation of pre-mRNA [8]. Therefore, the poly(A) signals become guides to 3'end of the genes under the gene identification methods. In general, an AATAAA subsequence is known to be a consensus sequence of the poly(A) signals [8]. However, a large number of AATAAA subsequences other than poly(A) signals exist in the eukaryote DNA sequences. Therefore, it is not sufficient merely to recognize AATAAA subsequences as a termination signal for transcription. The characteristic feature of the region downstream from a poly(A) site is the presence of GT-rich and T-rich subsequences having various compositions and locations [9, 10]. The GT/T-rich subsequences are required for the correct polyadenylation of pre-mRNA [11]. Using a representative sample of mammalian pre-mRNAs, the presence of the consensus YGTGTTYY was found to be located $20 - 30$ bp downstream of the poly(A) site [12]. However, only 67% of the examined sites have this consensus subsequence. Consequently, the sequence patterns in the vicinity of the poly(A) signal can vary.

The main purpose of this study is to identify the environments in which AATAAA is meaningful as a conserved subsequence of the poly(A) signal in human DNA sequences. We have tried to obtain new sequence patterns in the vicinity of the poly(A) signal by applying a statistical method. From a viewpoint of reliance, a method which is efficient for the analysis of such variable 3'UTR can make great contribution to gene identification

methods.

2 Methods

To distinguish signal patterns possessing AATAAA subsequences from non-signal patterns, we have adopted discriminant analysis, based on class II quantification theory, as developed by Hayashi [14, 15]. The theory was applied to the structural analysis of a splice site [16, 17].

According to the theory, the quantification of categorical data produces a sample score. Let α_{ti} denote the score of the ith DNA sequence which is a member of group t. We assume the sample score to be given by the following equations:

$$\alpha_{ti} = \sum_{j=1}^{L} \sum_{k=1}^{4} \delta_i(jk) x_{jk} \tag{1}$$

where

$$\delta_i(jk) = \begin{cases} 1 & : \text{If the } i\text{th sample sequence of the group } t \text{ has a base } k \text{ at position } j \\ 0 & : \text{Otherwise} \end{cases} \tag{2}$$

t is the group number, and $t = 1, 2$ correspond to the positive and negative data sets, respectively. In our present study, group 1 is composed of DNA sequences in the vicinity of poly(A) signals, while group 2 consists of sequences other than poly(A) signals. i is the sample sequence number, and L is the length of the sequences. k is a base, and $k = 1, 2, 3, 4$ correspond to bases A, T, C and G, respectively. Coefficient x_{jk} is called the category weight, and we define it as being the weight of base k at position j in the sequences.

A set of category weight values is used to discriminate between signal/non-signal patterns and to identify of conserved subsequences. According to the theory, we estimate a set of x_{jk} and α_{ti}, to clearly discriminate between the two groups of the positive and the negative data sets. In general, the total variance is expressed by the sum of that within and between the group variance. Therefore, to discriminate between two groups most distinctly is equivalent to maximizing the ratio of the between group variance to the total variance. By letting η^2 denote the correlation ratio, this formulation can be expressed by the following equations:

$$\eta^2 = \frac{\sigma_b^2}{\sigma^2} \longrightarrow \max \tag{3}$$

where

$$\sigma^2 = \frac{1}{N} \sum_{t=1}^{2} \sum_{i=1}^{n_t} (\alpha_{ti} - \alpha)^2 \tag{4}$$

$$\sigma_b^2 = \frac{1}{N} \sum_{t=1}^{2} n_t (\bar{\alpha}_t - \bar{\alpha})^2 \tag{5}$$

$$N = \sum_{t=1}^{2} n_t \tag{6}$$

σ_b^2 and σ^2 are the between group and the total variances, respectively. N is the number of all sample sequences, and n_t is the number of sample sequences in group t. $\bar{\alpha}$ is the average of the sample scores of all sequences, and $\bar{\alpha}_t$ is the average of the sample scores of group t. $\bar{\alpha}$ and $\bar{\alpha}_t$ are given by the following equations:

$$\bar{\alpha} = \frac{1}{N}\sum_{t=1}^{2} n_t \bar{\alpha}_t \tag{7}$$

$$\bar{\alpha}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} \alpha_{ti} \tag{8}$$

The value of the correlation ratio is used as an index of the accuracy of the signal/non-signal pattern discrimination, and can vary between 0 and 1. The accuracy of the discrimination result increases with the value of the correlation ratio.

It is known that the procedure for maximizing the correlation ratio η^2 and obtaining a set of category weights x_{jk} at the optimum condition can be formulated as a latent equation [14, 15, 16]. The category weights are estimated as a latent vector for a maximum latent root. Since the degree of freedom is reduced, we calculated these weights, defining x_{j1} as 0, and normalized them to satisfy the conditions of $\sum_k^4 n_{jk}x_{jk} = 0$ for all j positions and $\sigma = 1$, where n_{jk} is the number of sample sequences possessing a base, k, at a position j. Then, we defined a positive value of these weights as a frequent appearance in the positive data set, because whether they are positive or negative is determined at will. The latent root corresponds to the correlation ratio. In the case of discrimination between two groups, it is known that the latent equation becomes simple, and is transformed into a linear equation system.

The range of category weights at each position reflects the relative importance of the position in the discrimination of patterns. The position in DNA sequences becomes important as the value of the range increases. Let s_j denote the range at position j in a DNA sequence, such that the range is given by the following equation :

$$s_j = \max_k(x_{jk}) - \min_k(x_{jk}) \tag{9}$$

To check the accuracy of the analysis, we cross-validated the category weights. In the following, we briefly summarize the cross-validation procedures. We randomly selected five DNA sequences from each of the positive and the negative data sets. Using the remaining sequences, we determined the category weight set. Based on these weights, we estimated the selected sequences. These procedures were repeated 501 times, so as to evaluate the accuracy of the analysis. The total accuracy percentage, d, is given by the following equation :

$$d = \frac{1}{2}\left(\frac{m_1}{M_1} + \frac{m_2}{M_2}\right) \times 100 \tag{10}$$

M_1 and M_2 are the numbers of the positive and negative test sequences, respectively. m_1 is the number of positive test sequences that are recognized correctly, while m_2 is the number of negative test sequences, that are recognized correctly.

3 Materials

According to the procedures described below, we created two data sets of human DNA sequences; a positive and a negative data set. To these data sets, we applied a type of statistical analysis called the class II quantification theory. All sequences were taken from GenBank release 76.0 [13].

1. From GenBank 76.0, we selected human entries which have a clear description of the poly(A) signal in the feature table. The poly(A) signal in the selected entries possesses the AATAAA pattern.

2. As a positive data set, we took DNA sequences in the vicinity of the poly(A) signal from the entries selected by applying procedure 1.

3. From GenBank 76.0, we selected human entries which include DNA fragment longer than 10 kbp. The reason for collecting such entries is to avoid sampling identical DNA sequences.

4. As a negative data set, we took the DNA sequences possessing the pseudo-poly(A) signal, which exists in the region from the 5'-end to the 3'-end of the gene, from the entries selected by applying procedure 3. Even though the pseudo-poly(A) signal includes the AATAAA pattern, it does not work as the transcription termination signal.

By applying the procedures described above, we collected 183 sequences as the positive data set, and 680 sequences as the negative data set. Some are shown in Table 1. We defined the location of the 5'-end of the AATAAA subsequence as being position 0. Each sequence consists of 128 bases corresponding to the region from -80 to $+47$, and possesses AATAAA subsequences in the region corresponding to positions 0 to 5. It is assumed that unknown sequence patterns in the vicinity of the poly(A) signal may lie within such a 128-nucleotide sequence.

4 Results

Using the positive and negative data sets collected by applying the above procedures, we obtained sets of category weights x_{jk} and ranges s_j. Figure 2 shows these values. The graph at the top shows a set of range values, while the other four graphs show the sets of category weights of bases A, T, C and G. The position in the DNA sequence becomes more important as the value of the range increases. As a category weight value x_{jk} of a base, k, at position j increases, the base frequently appears at a position in the positive data set. On the other hand, as the value decreases, the base appears more often in the negative data set. The values of the category weights and ranges corresponding to a region of AATAAA subsequences is 0, because the region does not contribute to discrimination between the two groups. The correlation ratio η^2, defined by equation 3, and which indicates the accuracy of the discrimination, is 0.576.

Table 1: Data sets of human DNA sequences

Groups 1 and 2 are composed of DNA sequences in the vicinity of the poly(A) signals and other than poly(A) signals, respectively. Each sequence consists of 128 bases, and possesses AATAAA subsequences in the region corresponding to positions 0 to 5.

No.	Group	Sequence	Gene
1	1	AAATGT···AT AATAAA TG···TGGTGA	Human NAT1 gene
2	1	AAATTA···GT AATAAA AT···ACACGC	Human Pit-1 gene
3	1	AACAGC···TC AATAAA TG···TTTCAA	Human interleukin 7 (IL7) gene
⋮	⋮	⋮	⋮
183	1	TTTTTT···AC AATAAA CA···AAAAAA	Human prothymosin-α pseudogene
184	2	AAAAAA···TA AATAAA CT···TTTCTC	Human tissue factor gene
185	2	AAAAAA···TT AATAAA AT···TAATTC	Human retinoblastoma susceptibility gene
⋮	⋮	⋮	⋮
863	2	TTTTTT···TG AATAAA AC···CAATCA	Human retinoblastoma susceptibility gene

Figure 3 shows the distribution of the sample scores α_{ti}, as estimated using equation 1. It was observed that the DNA sequences in the two groups are clearly discriminated by using the set of the calculated category weights x_{jk}. To the development of the 3'UTR recognition function, we applied a linear discriminant analysis. The optimum threshold, $\hat{\alpha}$, is 1.355. Using this threshold, we were able to correctly recognize 166 sequences in the 183 of the positive data set and 636 sequences in the 680 of the negative data set. The Mahalanobis distance, D_p^2, between the two groups is 8.108. The probability of a discriminant error is 7.75%.

Table 2 shows the results of cross-validation. Accuracies of 52.2% and 74.0 % were observed for the positive and the negative test sequences, respectively. The total accuracy percentage is estimated by equation 10. We could predict a poly(A) signal with an accuracy of 64.1% by using the category weights. This indicates that the category weight values determined by the analysis are an efficient means of signal discrimination.

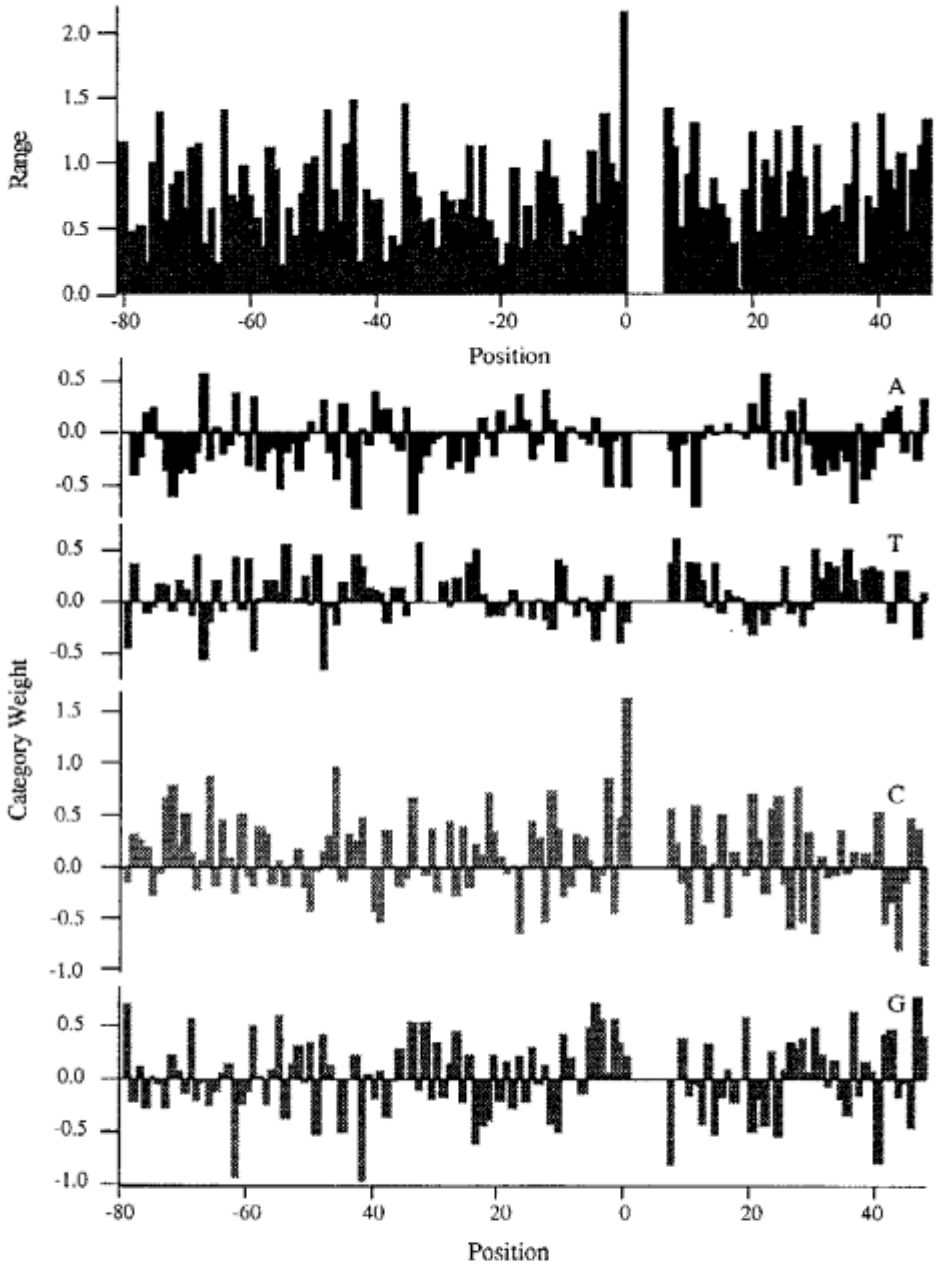

Figure 2: Category weights and ranges evaluated by class II quantification theory
The graph at the top shows a set of range values s_j. The position in the DNA sequence becomes important as the range value increases. The other four graphs show the sets of category weights x_{jk} of bases A, T, C and G. As the category weight value of a base at a position increases, the base frequently appears at a position in the positive data set. On the other hand, as the weight value decreases, the base tends to appear in the negative data set. The correlation ratio η^2 is 0.576.

Table 2: Results of cross-validation

The total accuracy percentage is estimated using equation 10. We were able to predict a poly(A) signal with an accuracy of 64.1% by using the category weights.

Accuracy		
Positive samples	Negative samples	Total
1358 / 2505 (54.2 %)	1854 / 2505 (74.0 %)	64.1 %

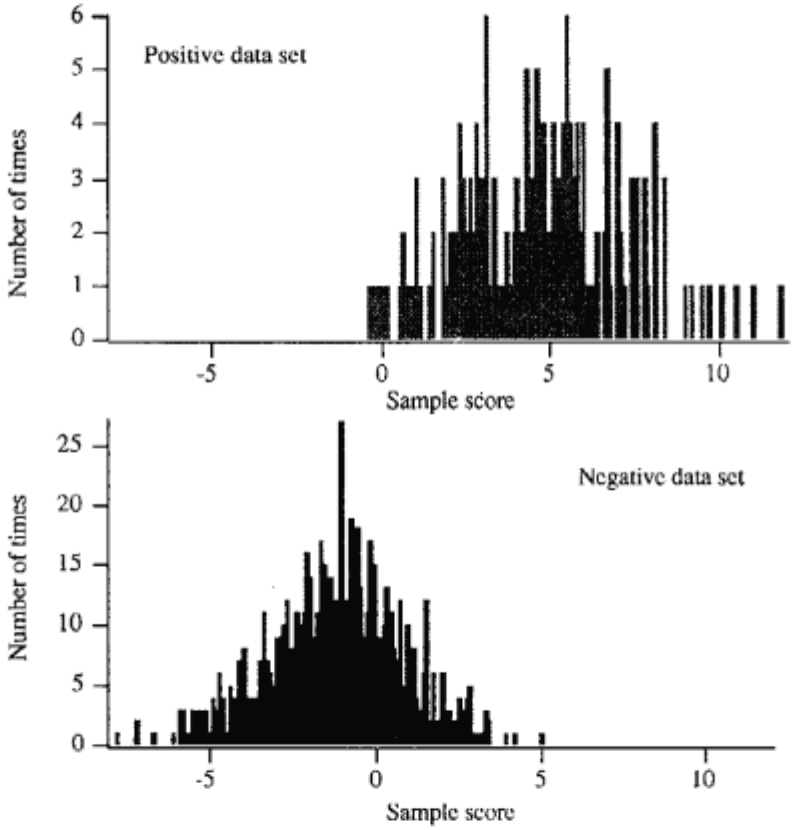

Figure 3: Distribution of sample scores

The upper and lower graphs show the distribution of the sample scores for the positive and the negative data sets, respectively. Based on linear discriminant analysis, the optimum threshold, $\hat{\alpha}$, for recognizing a DNA sequence as 3'UTR is 1.355. The Mahalanobis distance, D_p^2, between two groups is 8.108, and the probability of discriminant error is 7.75%.

5 Discussion

Since we took DNA sequences consisting of 128 bases into consideration, a large number of category weights were evaluated. Therefore, we tried to obtain the minimum set of these weights that is significant to the discrimination of 3'UTR. Based on a statistical test, we examined whether there is a significant difference in the discrimination ability resulting when all are used and when only some are used. Let p denote the number of variables, and q denote the number of some variables. The statistical test is performed using the following F-distribution:

$$F = \frac{n_1+n_2-p-1}{p-q}\frac{n_1n_2(D_p^2-D_q^2)}{(n_1+n_2)(n_1+n_2-2)+n_1n_2D_q^2} \tag{11}$$

where the number of degrees of freedom is $(p-q, n_1+n_2-p-1)$. $n_i(i=1,2)$ is the number of samples in group i. D_p^2 and D_q^2 are the Mahalanobis distances between the two groups using p and q variables, respectively. We excluded some category weights from all of them, while calculating the F value. In the lower range values, the corresponding category weights were taken away, four-by-four. Based on significance at the 5% level, we obtained the minimum set of category weights shown in Figure 4.

Based on the category weights and the ranges obtained by applying class II quantification theory, we were able to confirm the known characteristic features in 3'UTR. The category weights, ranked by position in Figure 4, indicate that the DNA sequences in the vicinity of the poly(A) signal have a tendency to possess bases G and T upstream of position 29. According to the distance from the poly(A) signal, this corresponds to the GT/T-rich sequence region.

We were also able to confirm the YGTGTTYY consensus sequences [12]. Since the sequence patterns in the vicinity of the poly(A) signal tend to vary, a glance at these weights does not show the conserved sequence clearly. Therefore, we applied an oligonucleotide-based analysis, similar to k-tuple analysis [18], to confirm the existence of the YGTGTTYY pattern. Let u_j denote the score of the pattern at position j in the DNA sequence. Using the category weights x_{jk} obtained by applying the theory, we assume the score, u_j, to be given by the following equations:

$$u_j = \sum_{i=j}^{j+l}\sum_{k=1}^{4}\delta(i-j+1,k)x_{ik} \tag{12}$$

where

$$\delta(i-j+1,k) = \begin{cases} 1 & : \text{If the } i-j+1\text{th base of the pattern} \\ & \quad \text{corresponds a } k \text{ base} \\ 0 & : \text{Otherwise} \end{cases} \tag{13}$$

l is the length of the pattern. Figure 5 shows the scores for the YGTGTTYY pattern in 3'UTR. The figure suggests the existence of a YGTGTTYY pattern in the vicinity of position 30.

Further, we revealed new patterns in the vicinity of the poly(A) signal by using the category weights and ranges. As can be clearly seen from Figures 3 and 4, the ranges of the positions close to the poly(A) signal are remarkably large. From the corresponding

Figure 4: Minimum set of category weights

The filled bars are located at positions corresponding to the minimum set of the category weights based on significance at the 5% level. Bases A, T, C, and G at each position are ranked in the order of higher category weight values by position.

category weights, base C frequently appears upstream of the AATAAA subsequence, and base T or C often appears downstream in the case of human DNA sequences. This implies that CAATAAA$\begin{matrix}\mathrm{T}\\ \mathrm{C}\end{matrix}$ can be regarded as being a conserved subsequence of the poly(A) signal.

Another new feature is the periodicity of the signal patterns. In Figures 3 and 4, the periodicity of the range values can be observed upstream of the poly(A) signal. To examine the features in the region, we evaluated the correlation between the range values at two distant positions in a DNA sequence, and investigated the behavior of these correlations. Figure 6 shows the normalized autocorrelation of the ranges upstream of the poly(A) signal. In the region, the correlation between the range values is a maximum every 12 base pairs. This indicates that the periodicity of the 12 base pairs almost corresponds to the base pair numbers in a single pitch of the DNA double-helix structure. It is quite interesting that the primary sequence related to the expression of the poly(A) signal is highly related to the tertiary DNA structure.

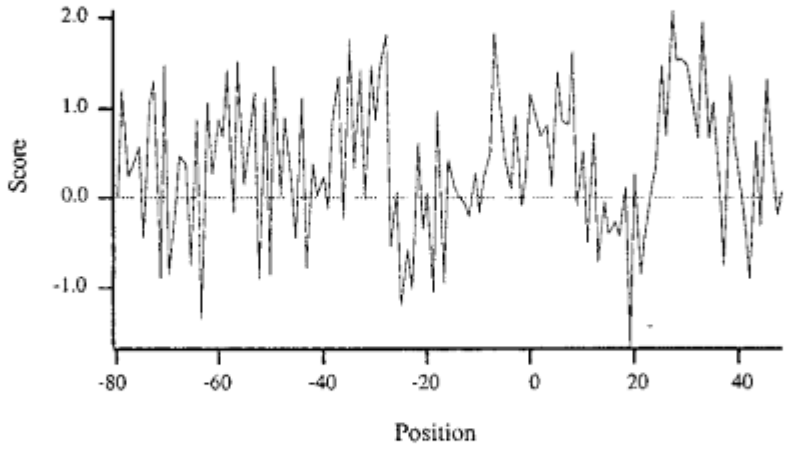

Figure 5: Scores of YGTGTTYY pattern in 3'UTR

The graph shows the average scores for eight variations of the YGTGTTYY pattern, such as the CGTGTTCC, CGTGTTCT, CGTGTTTC, CGTGTTTT, TGTGTTCC, TGTGTTCT, TGTGTTTC and TGTGTTTT patterns, and suggests the existence of a YGTGTTYY pattern in the vicinity of position 30.

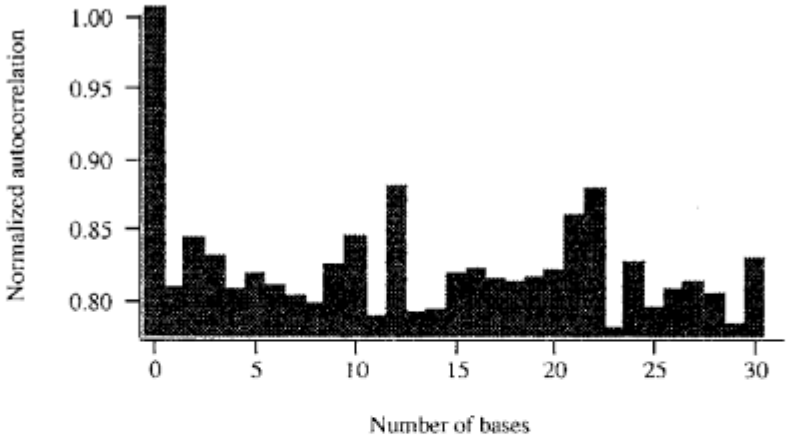

Figure 6: Autocorrelation of the ranges upstream of the poly(A) signal

Each value is normalized by the value at gap 0. Upstream side of the poly(A) signal, the correlation between the range values is a maximum every 12 base pairs.

6 Conclusion

We applied a statistical method, called class II quantification theory, to the analysis of human DNA sequences in the vicinity of the poly(A) signal. The category weights and ranges evaluated by the theory were analyzed by applying numerical methods including linear discriminant analysis, statistical test, oligonucleotide-based analysis and autocorrelation analysis. As a result, we have confirmed the known characteristic features of 3'UTR. Further, we have obtained new sequence features for 3'UTR, which are an extension of the consensus sequence, AATAAA, and the periodicity of the signal patterns. Since methods discussed in this paper dose not lose the generality, these methods is applicable for the recognition of any type of functional sites, and is able to make great contribution to gene identification methods.

We believe that combining quantification theory and the methods described above will provide an efficient means of analyzing the general functional sites in DNA sequences. Especially, a method combining theory with oligonucleotide-based analysis would reveal new conserved patterns in uncharacterized DNA sequences. Using such a combinational method, we plan to develop a computational tool for formulating signal patterns in DNA sequences.

Acknowledgments

The authors are grateful to Dr. K. Nakai, National Institute for Basic Biology of Japan, for his constant encouragement throughout this project.

References

[1] Fickett, J.W., Tung, C.S. : *Nucleic Acids Res.*, **24**, 6441-50 (1992)

[2] Hutchinson, G.B., Hayden, M.R. : *Nucleic Acids Res.*, **20**, 3453-62 (1992)

[3] Mural, R.J., Einstein, R., Guan, X., Mann, R.C., Uberbacher, E.C. : *Trends Biotech.*, **10**, 66-9 (1992)

[4] Guigo, R., Knudsen, S., Drake, N., Smith, T. : *J. Mol. Bio.*, **225**, 141-57 (1992)

[5] Fields, C.A., Soderlund, C.A. : *Comput. Applic. Biosci.*, **6**, 263-70 (1990)

[6] Milanesi, L., Kolchanov, N.A., Rogozin, I.B., Ischenko, I.N., Kel, A.E., Orlov, Y.L., Ponomarenko, M.P., Vezzoni, P. : In *Proceedings of the 2nd International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, (1993)

[7] Snyder, E.E., Stormo, G. : *Nucleic Acids Res.*, **21**, 607-13 (1993)

[8] Gil, A., Proudfoot, N.J. : *Cell*, **49**, 399-406 (1987)

[9] Taya, Y., Devos, R., Tavernier, J., Cheroutre, H., Engler, G., Fiers, W. : *EMBO J.*, **1**, 953-8 (1982)

[10] Brinstiel, M.L., Busslinger, M., Strub, K. : *Cell,* **41,** 349-59 (1985)

[11] Levitt, N., Briggs, D., Gil, A., Proufoot, N.J., : *G&D,* **7,** 647-59 (1993)

[12] McLauchlan, J., Gaffney, D., Whitton, J.L., Clements, J.B. : *Nucleic Acids Res.,* **13,** 1347-69 (1985)

[13] GenBank, Genetic Sequence Data Bank, Release 76.0, BBN Laboratories, U.S.A. (1993)

[14] Hayashi, C. : *Ann. Inst. Statist. Math.,* **2,** 35 (1950)

[15] Hayashi, C. : *Ann. Inst. Statist. Math.,* **3,** 69 (1952)

[16] Iida, Y. : *Bull. Chem. Sco. Jpn.,* **60,** 2977-81 (1987)

[17] Iida, Y. : *Comput. Applc. Biosci.,* **3,** 93-8 (1987)

[18] Claverie, J.M., Sauvaget, I., Bougueleret, L. : In *Methods in Enzymology*(ed. R.F.Doolittle), **183,** 237-52 (1990)