Finding Characteristic Patterns Embedded in Non-transcribed Region of Dictyostelium Discoideum by Computed Moiré

Toshiro ONITANI [1], Ikuo YOSHIHARA [2], Kunihito YAMAMORI [3], Moritoshi YASUNAGA [4]

Abstract

Finding transcription regulatory pattern of *Dictyostelium discoideum* (*D. discoideum*) is one of the most important tasks in genomics. Life cycle of *D. discoideum* is divided into vegetative stage and morphogenesis stage. The goal of this paper is to develop the technique for extracting the feature patterns in vegetative stage and morphogenesis stage. We calculate each score for candidate patterns ("*aaaaaaaa*"-"*tttttttt*") by matching these candidate patterns to DNA sequence of *D. discoideum.* The feature patterns of vegetative stage and morphogenesis stage are extracted by comparing scores of morphogenesis stage with those of vegetative stage.

Experiments reveal that "*cagcagca*", "*agcagcac*", "*gagagaga*", "*cagcatca*", and "*caccagca*" appear in vegetative stage more than morphogenesis stage, and "*cacaccc*", "*gtgtgtgt*", "*acgactac*", "*acacaccc*", and "*ctactact*" appear in morphogenesis stage more than vegetative stage.

Key Words:
Matching, *Dictyostelium Discoideum*, Vegetative Stage, Morphogenesis Stage

1. Introduction

In recent years, *Dictyostelium discoideum* has been researched all over the world, because it is based on the simple growth cycle. In the field of genome analysis, finding transcription regulatory pattern is one of the most important tasks. Life cycle of *D. discoideum* consists of four growth stages based on timing of gene expression. We assume that non-transcribed region of each growth stage have a specific transcription regulatory pattern.

The non-transcribed region data of *D. discoideum* consists of 2000 bases from the transcription start site (TSS). All the data used in this research belong to chromosome 2' among six chromosomes in *D. discoideum* [5].

We divided four growth stages into vegetative stage (V stage) and morphogenesis stage (non-V stage) by growth and development of *D. discoideum*, because the amount of non-transcribed region data is insufficient to experiment on each four growth stages. To search for transcription regulatory pattern in V stage and non-V stage, we develop a technique to extract feature patterns appearing in V stage and non-V stage.

2. Gene Expression and *Dictyostelium Discoideum*

2.1. Gene Expression

The process of composing protein by transcription of a part of DNA is called as gene expression. Fig.1 shows the process of composing protein. When gene expresses, the DNA sequence is copied to the RNA sequence. This process is called as transcription. A part of transcribed region composes protein. The starting position of transcription is called as TSS. Moreover, not transcribed region is called as non-transcribed region.

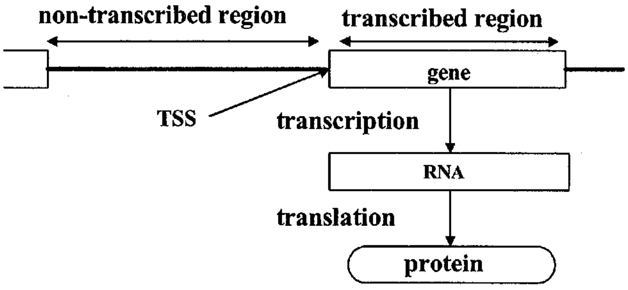

Fig.1 The Process of Composing Protein

2.2. *Dictyostelium Discoideum*

D. discoideum is a kind of primitive eukaryote. The total number of gene of *D. discoideum* is estimated at 8000-10000. Life cycle of *D. discoideum* consists of V stage with single cell and non-V stage (aggregating stage, slug stage, and culminating stage) with multicell. Fig.2 shows life cycle of *D. discoideum.*

In vegetative stage, the amoeba starts fission and breeding. In aggregating stage, the amoeba starts assembly, and pseudoplasmodium is formed. In slug stage, pseudoplasmodium begins to move. In culminating stage, pseudoplasmodium forms a fruiting body. Fruiting body has the spore, and amoebas appear by germination of spore.

[1] Postgraduate student, Graduate School of Engineering.
[2] Professor, Dept. of Computer Science and Systems Engineering.
[3] Associate Professor, Dept. of Computer Science and Systems Engineering.
[4] Professor, Inst. of Information and Electronics, Univ. of Tsukuba.

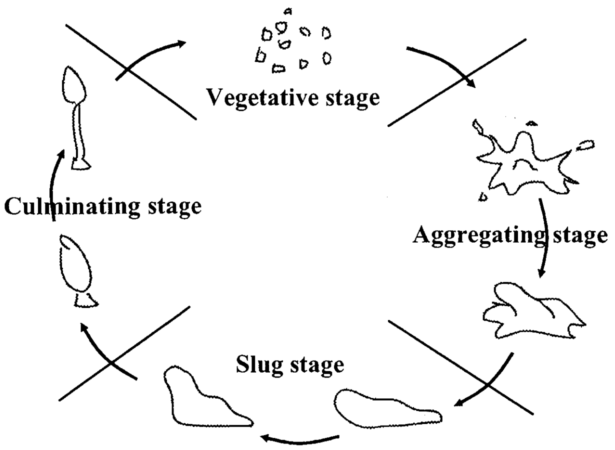

Fig.2 Life Cycle of *Dictyostelium Discoideum*

3. Score for A Candidate Pattern

As an example, we assume "*ACGTATGT*" is known transcription regulatory pattern in *D. discoideum*. We will show method to extract *"ACGTATGT"* from a DNA sequence and calculate score for it.

First, "*ACGTATGT*" is matched to a DNA sequence one by one position from the head of sequence, and "*ACGTATGT*" included in a DNA sequence are extracted. Then, we extracted 7 different significantly aligned "*ACGTATGT*": all matches, 1-gap, 1-missmatch, 2-gaps, 2-missmatch and 1-gap & 1-missmatch. Fig.3 shows extracted "*ACGTATGT*" from a DNA sequence. In Fig.3, under bar (_) means mismatch of a base, and (-) means a gap.

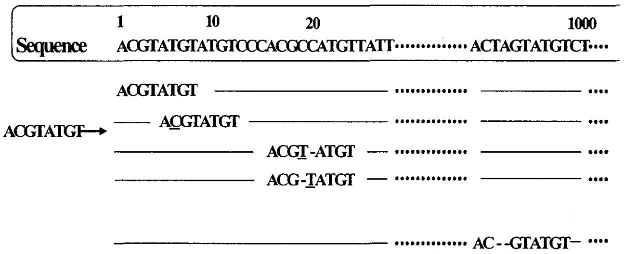

Fig.3 Extraction Aligned "*ACGTATGT*" from a DNA Sequence

Table.1 Similarity

	State	Similarity
(1)	All matches	8
(2)	1-gap	6
(3)	1-mismatch	6
(4)	2-gaps(consecutive)	5
(5)	2-gaps(non-consecutive)	4
(6)	2-mismatches	4
(7)	1-gap&1-mismatch	4

Next, the score for "*ACGTATGT*" is calculated. Aligned "*ACGTATGT*" are given similarity based on Table.1. A similarity for all matches is 8, and the penalty about mismatch is 2, the penalty about open gap is 2, and the penalty about extended gap is 1. Score for "*ACGTATGT*" is calculated one by one position from the head of a DNA sequence based on similarity. The score for "*ACGTATGT*" at the nth position from the head of sequence is sum of similarity of "*ACGTATGT*" extracted between n-50th and n+50th position from the head of sequence. However, when two or more extracted sequences appear at same position, only the highest similarity is used. This score shows the distribution of "*ACGTATGT*" in a DNA sequence.

4. Finding Feature Patterns Embedded in Dictyostelium Discoideum

4.1. Experimental Data

Non-transcribed region data of *D. discoideum* is used as experimental data. These data were created on experiments in Univ. of Tsukuba and genome database "cDNA project" [6][7]. These data are divided into V stage and non-V stage by growth and development of D. discoideum.

- V stage: 25data
- Non-V stage: 10data
  (Aggregating stage: 6data, Culminating stage: 3data, Slug stage: 1data)

Non transcribed region data of *D. discoideum* consist of 2000 bases from the TSS as shown in Fig.4. However, a few data is less than 2000 bases. In that case, the missing part is filled with the asterisk (*).

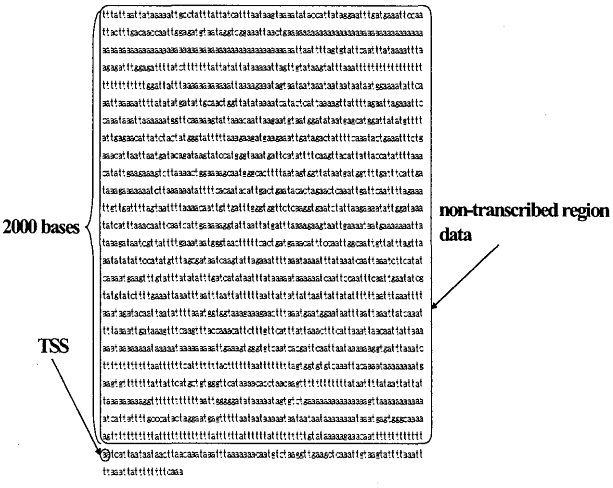
2000 bases
TSS
non-transcribed region data

Fig.4 An Example of Data

4.2. Comparing Scores of V Stage with those of non-V Stage

In this research, the length of candidate pattern is set to 8. We have 4^8 different candidate patterns from "*aaaaaaaa*" to "*tttttttt*". The scores for candidate patterns "*aaaaaaaa*"-"*tttttttt*" are calculated by matching one by one position from the TSS. Average of score for a candidate pattern in V stage (*Score_V*) is calculated. The *Score_V* is defined as follows,

$$Score\_V(Cp(j),k) = \frac{\sum_{i=1}^{i \le N(V)} Score(V(i), Cp(j), k)}{N(V)} \quad . (1)$$

V (i): i^{th} data of V stage
Cp (j): j^{th} candidate pattern ($0 \leq j < 4^8$)
k: k^{th} position from TSS ($-2000 \leq k < 0$)
N (V): the number of data about V stage

The average of scores in non-V stage (*Score_nonV*) is calculated in the same way.

Differences between *Score_V* and *Score_nonV* are represented by newly defined indices, *P_score_V* and P_score_nonV. *P_score_V* is defined as follows.

$$P\_score\_V(Cp(i), j) = \frac{Score\_V(Cp(i), j)}{Score\_V(Cp(i), j) + Score\_nonV(Cp(i), j)} , \quad (2)$$

where *Score_V* and *Score_nonV* must satisfy the following conditions.

$$\begin{cases} Score\_V(Cp(i), j) \geq 6.0 & (3) \\ Score\_nonV \geq \overline{Score\_nonV(Cp(i), j)} - \sigma(Score\_nonV(Cp(i), j)) & . (4) \end{cases}$$

Here, upper bar ($\overline{\qquad\quad}$) denotes the average score from j=2000 to j=0, and σ denotes standard deviation. If these condition are not satisfied, *P_score_V*=0. Condition (3) is to extract a candidate pattern with high score. Condition (4) prevents *P_score_V* from increasing by low *Score_nonV*. *Pmax_score_V*(*C*(*i*)) represents the maximum of *P_score_V*(Cp(i),j) at $-2000 \leq j < 0$ as follows,

$$P\text{max}\_score\_V(Cp(i)) = \max_{-2000 \leq j < 0} P\_score\_V(Cp(i), j) \quad . (5)$$

Pmax_score_nonV (*Cp* (i)) is calculated in the same way. The candidate patterns with the top five *Pmax_score_V* represent feature patterns in V stage. Feature patterns in non-V stage are represented in the same way.

4.3. Experimental Results and Discussions

4.3.1. V Stage

Experimental result on V stage is shown Table.2. Distribution of score for the discovered patterns is shown Fig.5-Fig 9. In Fig.5-Fig.9, ordinate axis shows the value of *Score_V* or *Score_nonV* for extracted pattern, and abscissa axis shows the position from TSS.

Table.2 Extracted Patterns in V Stage

Extracted pattern	Position	Score_V	Score_nonV	P_max_score_V
cagcagca	-1475	8.67	0.36	0.96
agcagcac	-1447	6.17	0.36	0.95
gagagaga	-1525	5.83	0.36	0.94
cagcatca	-1473	8.75	0.73	0.92
caccagca	-1472	8.17	0.73	0.92

In Fig.5, *P_score_V* for "*cagcagca*" is highest at position -1475. In Fig.6, *P_score_V* for "*agcagcac*" is highest at position -1447. In Fig.7, *P_score_V* for "*gagagaga*" is highest at position -1525. In Fig.8, *P_score_V* for "*cagcatca*" is highest at position -1437. In Fig.9, *P_score_V* for "*caccagca*" is highest at position -1472.

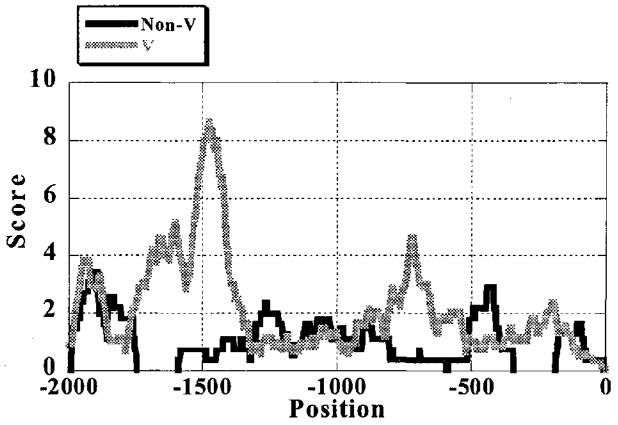

Fig.5 Distribution of *Score* for "*cagcagca*"

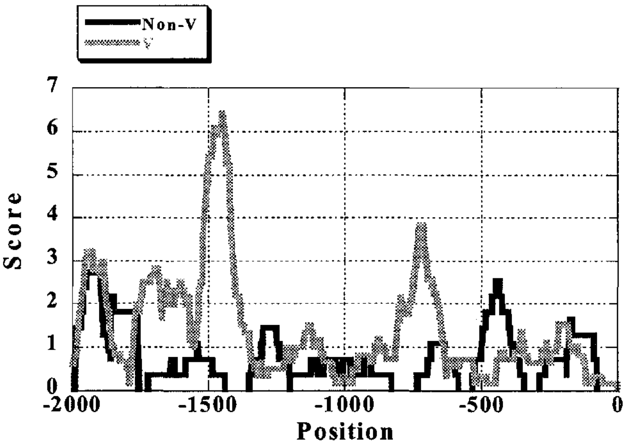

Fig.6 Distributions of Score for "*agcagcac*"

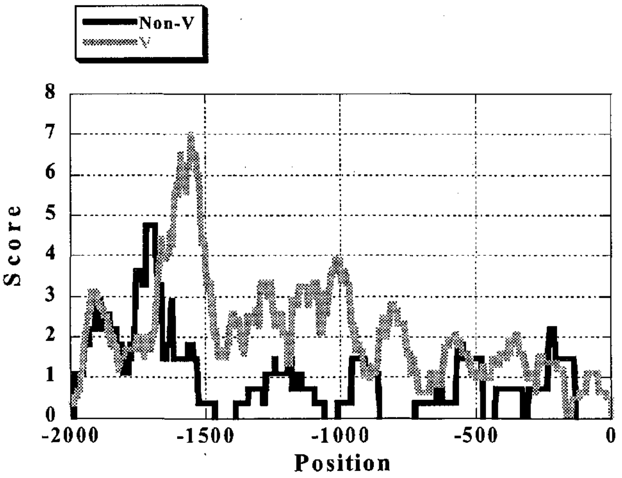

Fig.7 Distribution of Score for "*gagagaga*"

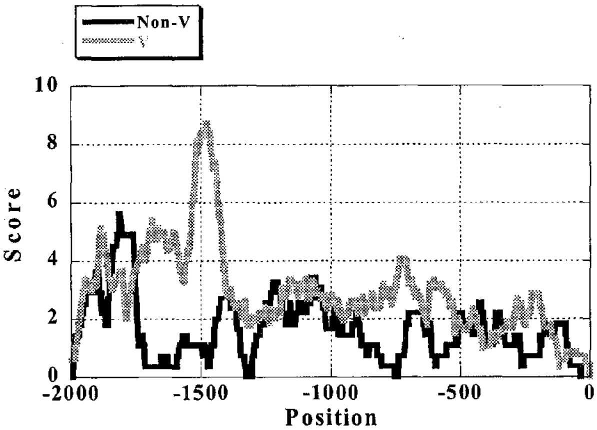

Fig.8 Distribution of Score for "*cagcatca*"

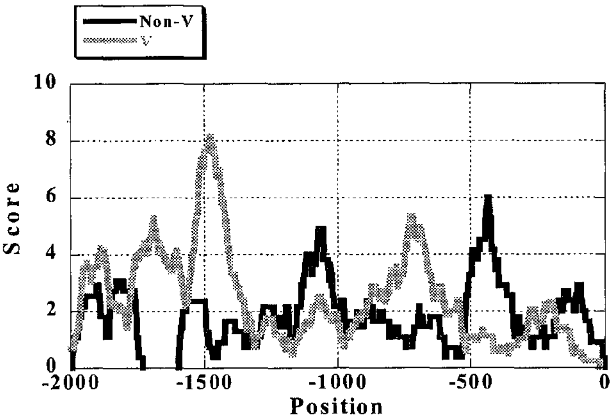

Fig.9 Distribution of Score for "*caccagca*"

"*cagcagca*", "*agcagcac*", "*gagagaga*", "*cagcatca*", and "*caccagca*" appear in V stage more than non-V stage. We found that "*cagca*" is often included in these extracted patterns. So we calculate *Score_V* and *Score_nonV* for candidate pattern "*cagca*" in the same way as shown Fig.10.

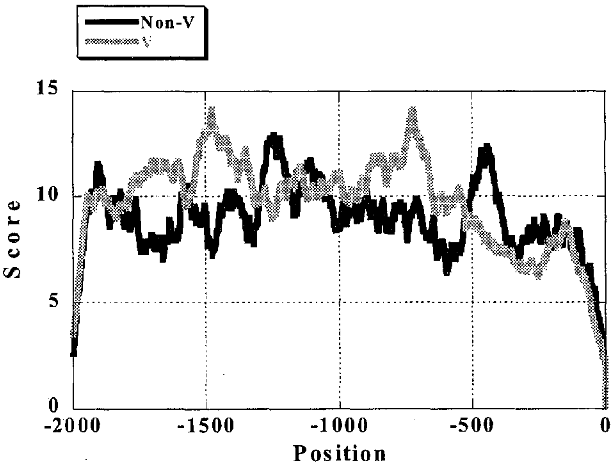

Fig.10 Distribution of *Score* for "*cagca*"

In Fig.10, "*cagca*" appears in V stag more than non-V stage at position between -1400 and 1500, and between -800 and -700.

4.3.2. Non-V Stage

Experimental result of non-V stage is shown Table.3. Distribution of score for the discovered pattern is shown Fig.11-Fig.15. In Fig.11-Fig.15, ordinate axis shows the *Score_V* and *Score_nonV* for extracted pattern, and abscissa axis shows the position from TSS.

Table.3 Extracted Patterns in non-V Stage

Extracted pattern	Position	Score_V	Score_nonV	P_max_score_nonV
cacacccc	-559	0.00	6.00	1.00
gtgtgtgt	-325	0.32	7.09	0.96
acgactac	-1378	0.50	9.45	0.95
acacaccc	-519	0.40	7.27	0.95
ctactact	-1207	0.50	8.18	0.94

In Fig.11, *P_score_nonV* for "*cacaccc*" is highest at position -559. In Fig.12, *P_score_nonV* for "*gtgtgtgt*" is highest at position -325. In Fig.13, *P_score_nonV* for "*acgactac*" is highest at position -1578. In Fig.14, *P_score_nonV* for "*acacaccc*" is highest at position -519. In Fig.15, *P_score_nonV* for "*ctactact*" is highest at position -1207.

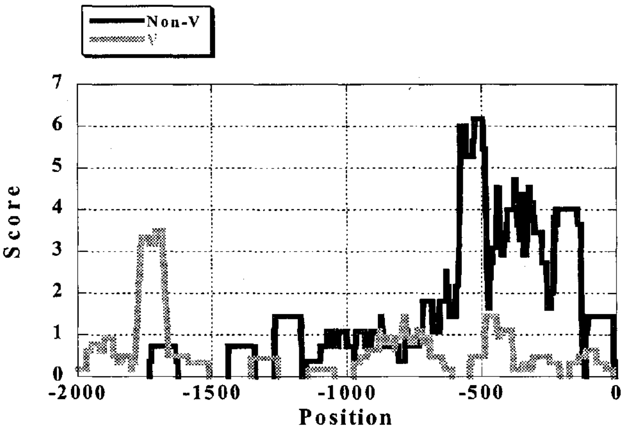

Fig.11 Distribution of *Score* for "*cacacccc*"

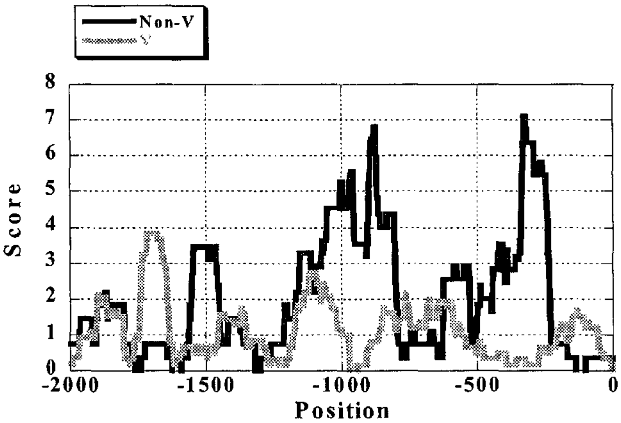

Fig.12 Distribution of *Score* for "*gtgtgtgt*"

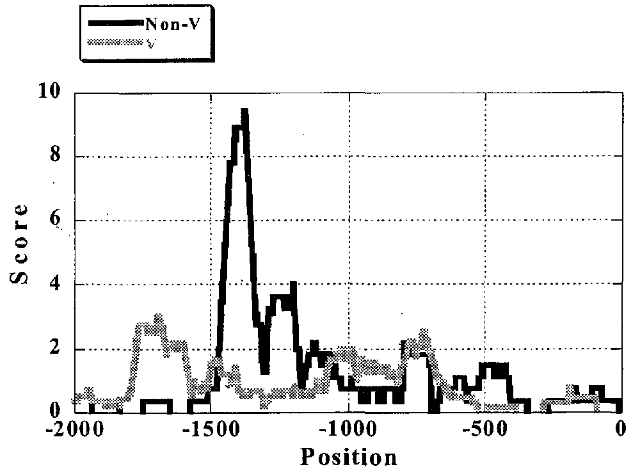

Fig.13 Distribution of *Score* for "*acgactac*"

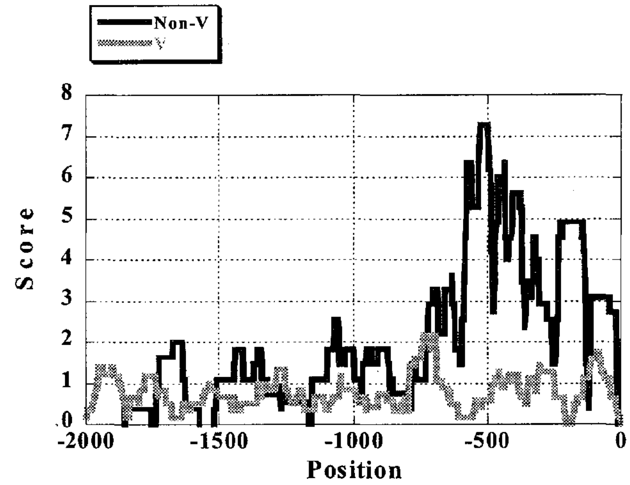

Fig.14 Distribution of *Score* for "*acacaccc*"

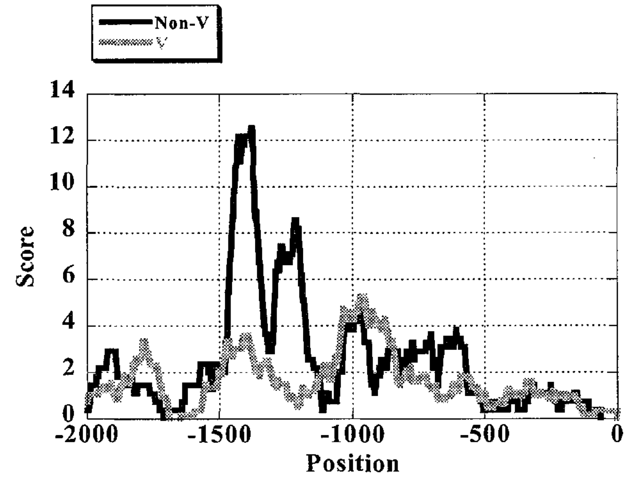

Fig.15 Distribution of *Score* for "*ctactact*"

"*cacaccc*", "*gtgtgtgt*", "*acgactac*", "*acacaccc*", and "*ctactact*" appear in non-V stage more than V stage. We found that "*actac*" or "*cacaccc*" is often included in these patterns. So we calculate *Score_V* and *Score_nonV* for candidate pattern "*actac*" and "*cacaccc*" in the same way as shown in Fig.16 and Fig17.

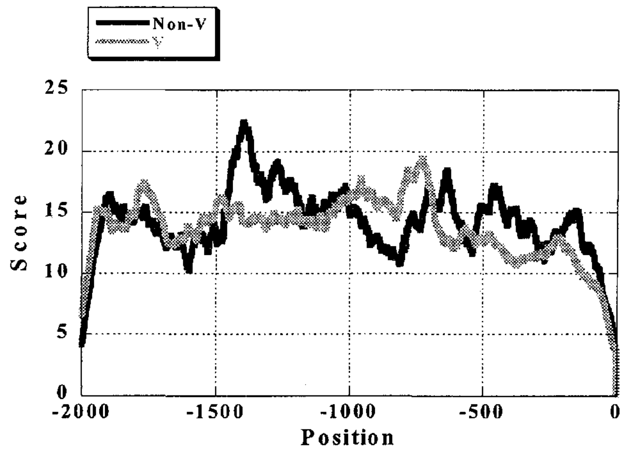

Fig.16 Distribution of *Score* for "*actac*"

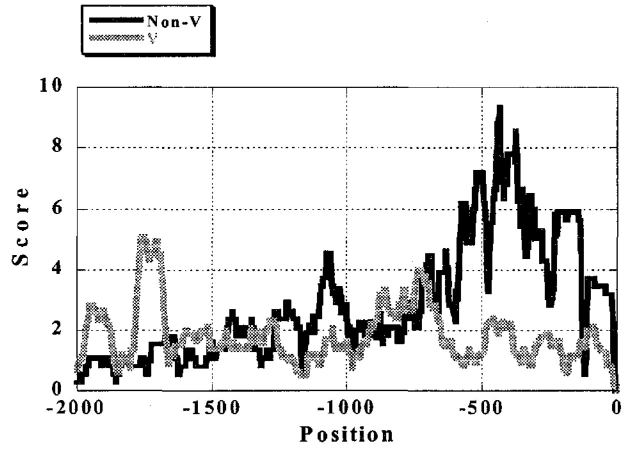

Fig.17 Distribution of *Score* for "*cacaccc*"

In Fig.16, "*actac*" appears in non-V stage more than V stage at position between -1400 and -1300. In Fig.17, "*cacaccc*" appears in non-V stage more than V stage at position between -400 and -100.

5. Conclusions

New indices *Score_V* and *Score_nonV* are proposed for extracting feature patterns in V stage and non-V stage.

Experiments based on the data compiled by Dr.Urushihara and Yasunaga in Univ. of Tsukuba lead us to the conclusion that "*cagcagca*", "*agcagcac*", "*gagagaga*", "*cagcatca*", and "*caccagca*" appear in V stage more than non-V stage, and "*cacaccc*", "*gtgtgtgt*", "*acgactac*", "*acacaccc*", and "*ctactact*" appear in non-V stage more than V stage. Additional experiments show that "*cagca*" is often included in extracted patterns in V stage, and "*actac*" and "*cacaccc*" are often included in extracted patterns in non-V stage.

Future works are refinement of scoring method, experiment on longer candidate patterns, and extraction feature patterns in each growth stage (Aggregating stage, Culminating stage, Slug stage, and Vegetative stage).

6. Acknowledgements

We are most grateful to Dr. Hideko Urushihara at

University of Tsukuba and the cDNA Project in Japan for providing us the information on cDNA sequences.

This research is partly supported by MEXT grant 16011204 in 2004.

7. References

[1] N. Miyake, M. Kanehisa, Project of Genome of Homo and Knowledge Information Processing System, *Baifukan Co., Ltd*, 1995 (in Japanese).

[2] N. Miyake, Y. Sakaki, Bioinformatics, *Tokyo Kagaku Dojin Co., Ltd*, 2003 (in Japanese).

[3] J. Tanida, "String Data Alignment by a Spatial Coding and Moiré Technique", Optics Letters, Vol.24, pp.1681-1683, 1999.

[4] T. Okamura, Y. Narita, DNA and RNA, *Natsume* Publishing Co., Ltd, 1999 (in Japanese).

[5] Glockner,G. et al., (2002) Sequence and analysis of chromosome 2 of Dictyostelium discoideum. Nature, 418, 79-85.

[6] Functional genomics of the social amoebae, Dictyostelium discoideum. H. Urushihara 2002. Mol. Cells 13(1) 1-4.

[7] Urushihara, H. et al., (2004). "Analyses of cDNAs from growth and slug stages of Dictyostelium discoideum." Nuc. Acids Res. 32(5): 1647-1655.

[8] M. Yasunaga, K. Ushiyama, I. Yoshihara, and J.H. Kim, "Symbolical Kernel-Based Reasoning: Its Application to the Rule Extraction in *Dictyostelium discoideum* DNA", Genome Informatics 2001 pp.413-414, December 2001.