Extracting Transcription Regulatory Elements in *Dictyostelium Discoideum*

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

Abstract

Finding transcription regulatory elements (TREs) is one of the most important tasks in the field of genome informatics. The objective of this research is to extract TREs in DNA sequences of *Dictyostelium discoideum* (*D.discoideum*). Extracting characteristic patterns in DNA sequences is important to find TREs. We propose indices ($Score_M$, $Score_T$) to extract characteristic patterns in *D.discoideum*. $Score_M$ is proposed to extract patterns appeared in common region of DNA sequences. $Score_T$ is proposed to extract repeated patterns in common region of DNA sequences.

Characteristic patterns according to $Score_M$ and $Score_T$ are extracted from non-transcribed region of *Dictyostelium discoideum*. The experimental Result shows that $Score_M$ and $Score_T$ are effective to extract targeted patterns.

Key Words:
Transcriptional regulatory elements, *Dictyostelium discoideum*, Growth stage, Characteristic pattern

1. Introduction

Finding transcriptional regulatory elements (TREs) in DNA sequences is one of the most important tasks in the field of genome informatics. As the first step to find TREs, it is important to extract characteristic patterns in DNA sequences. But we must consider two features of TREs. The first, if known TRE appear in DNA sequences, they are not always exactly same patterns but similar patterns in most cases. The second, a TRE don't appear at common positions of DNA sequences but in common region of DNA sequences. These two features make it more difficult to extract TREs. To resolve these problems, it's necessary to investigate similarities in DNA sequences. Alignment is effective method to investigate it. In this research, indices to extract characteristic patterns by using the results of alignment are proposed and characteristic patterns in DNA sequence are extracted according to proposed indices.

DNA sequences of *Dictyostelium discoideum* (*D.discoideum*) are used as data. In recent years, *D.discoideum* has been researched worldwide, because it has the characteristic life cycle. Life cycle of *D. discoideum* is divided into four growth stages according to periods of gene expression. We assume that specific TREs in each growth stage should exist. So, we extract characteristic patterns in each stage in *D.discoideum*.

2. Gene Expression and *Dictyostelium Discoideum*

2.1. The Mechanism of Gene Expression

2.1.1. Gene Expression

The compounding of protein from DNA is called gene expression. Fig. 1 shows the process of composing protein. When the gene expresses, the gene of DNA is replicated to the Ribonucleic Acid (RNA). This operation is called transcription. The RNA is translated into amino acids and then they are compounded as protein. The starting position of transcription is called transcription start site (TSS) [2].

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 Sciences and Electronics, Univ. of Tsukuba.

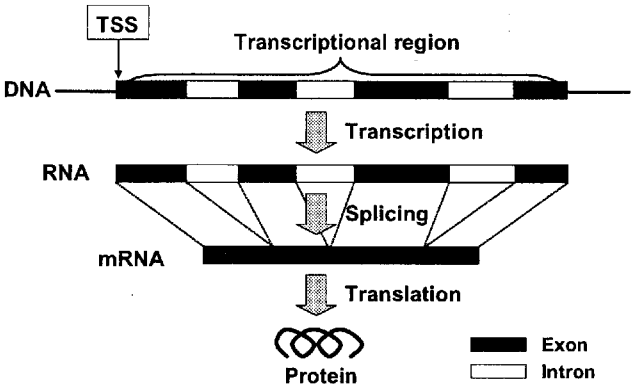

Fig. 1 Process of Compounding Protein

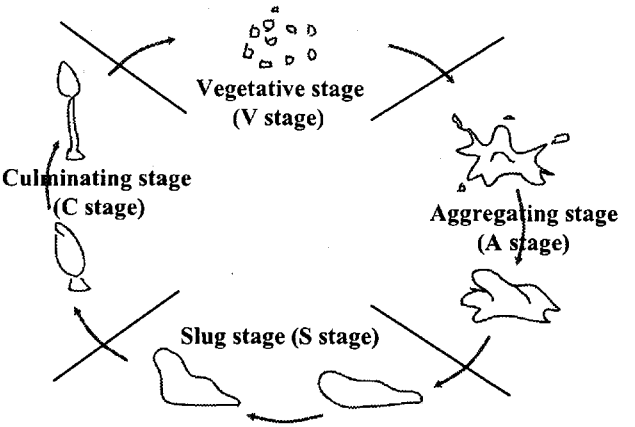

Fig.3 Asexual Life Cycle of *D.discoideum*

CCAAT box (at position between -80 and -60) exist as the famous upstream sequence of promoter. Moreover, the enhancer might exist at position upstream from -100. TATA box, CCAAT box and enhancer, etc is called transcriptional regulatory elements (TREs). Transcription is regulated by combining the transcription factor to the TREs. Many TREs have not been found yet [3].

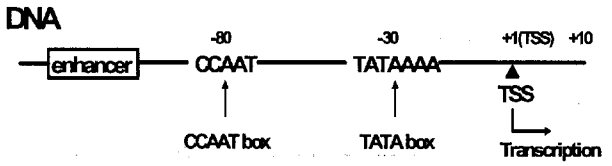

Fig.2 Promoter upstream sequence of eucaryocyte

2.2. *Dictyostelium Discoideum* (*D.discoideum*)

2.2.1. Life Cycle of *D.discoideum*

D. discoideum is a kind of lower eukaryote. Life cycle of D.discoideum consists of asexual life cycle and sexual life cycle. Because DNA sequence data of the asexual life cycle is used in this research, only asexual life cycle is explained in this section.

Asexual life cycle of D.discoideum is divided into four growth stage by timing of gene expression like fig.3.

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 begins to form a fruiting body. Fruiting body has the spore, and amoebas appear by germination of spore.

2.2.2. TREs of *D.discoideum*

D.discoideum has common TREs or specific TREs in four growth stages are not clearly. But, we assume that specific TREs in each growth stage should exist. So, we extract characteristic patterns in each stage. The characteristic patterns are pattern that appear frequently at the specific position in DNA sequences. Extraction of these patterns is first step to search for TREs.

Pattern matching is effective to search characteristic patterns. But, if known TRE appear in DNA sequences, they are not always exactly same patterns but similar patterns in most cases. To resolve this problem, alignment is important. Alignment and pattern matching are explained in next chapter.

3. Pattern Extraction Method from DNA sequences

3.1. Calculation for Alignment Score

3.1.1. Alignment

Search for optimal alignment is necessary for pattern matching. Optimal alignment is optimal correspondence between bases of two sequences. Search for optimal alignment is important to calculate similarities between two sequences [6].

Similarities for corresponded bases between two sequences are defined as table 1. Match or Mismatch shows the matching or mismatching between corresponded bases. Gap is blank symbol to consider insertion or deletion of bases from DNA sequences by process of evolution. Open

gap shows an inserted gap between two bases. Extended gap shows an added gap to the right of gap.

Table.1 Similarity

State	Similarity
Match	+1
Mismatch	-1
Open gap	-2
Extended gap	-1

For example, there is many alignment of "AGTC" with "ATGTC" as follows (- : gap).

1 A- GTC 2 - - A GTC 3 A GT - -C
ATGTC AT - GTC A - TGTC

Each sum of similarities (alignment score) are 2 (for case 1), -2 (for case 2) and -2 (for case 3). Alignment that has maximum alignment score is optimal alignment. This is calculated by DP matching algorithm.

3.1.2. DP matching using Candidate Pattern

For example, we assume "*ACGTATGT*" is known transcription regulatory pattern. We will show how to extract "*ACGTATGT*" from a DNA sequence.

First, "*ACGTATGT*" is matched to a DNA sequence one by one position from the head of sequence, and alignment score is calculated in all position on DNA sequence. Next, the positions with high alignment scores in DNA sequence are extracted. That is, the similar position and their alignment score in DNA sequence for candidate pattern is extracted. Fig.4 shows an example for extraction of alignment score more than 4 and their position.

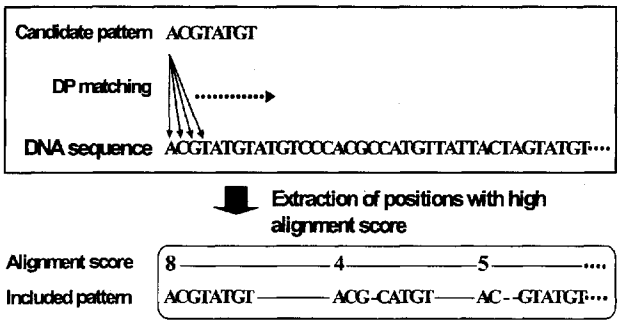

Fig.4 Extraction of positions with high alignment score

3.2. Proposed Relative Score

TREs don't appear at same positions of DNA sequences but in common region of DNA sequences. To cope with this feature, relative score is proposed as index to extract characteristic patterns in DNA sequences. We propose two kind of relative scores ($Score_M$, $Score_T$). These are explained in subsection 3.2.1 and 3.2.2.

3.2.1. Maximum Score ($Score_M$)

A TRE almost appears in common region of DNA sequences. $Score_M$ is index to extract patterns that appear in common region of DNA sequence. Fig.5 shows an example of a distribution of extracted alignment score. The x-axis shows the position on DNA sequence.

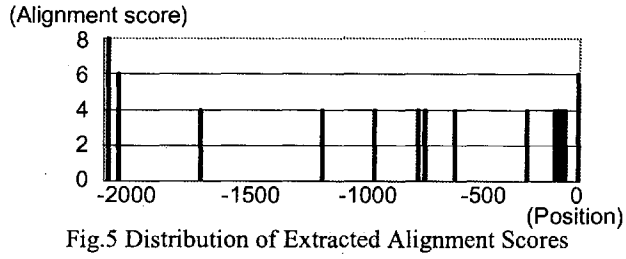

Fig.5 Distribution of Extracted Alignment Scores

We assumed that if a known TRE appears at position k in DNA sequence, it appears at position between k-N and k+N in other DNA sequences. So, $Score_M$ is defined as follows,

$$Score_M(i) = \max_{i-N \le j \le i+N} (Alignmentscore(j)). \quad (1)$$

Fig.6 shows a distribution of $Score_M$ that calculated from alignment scores in fig.5 (N=20).

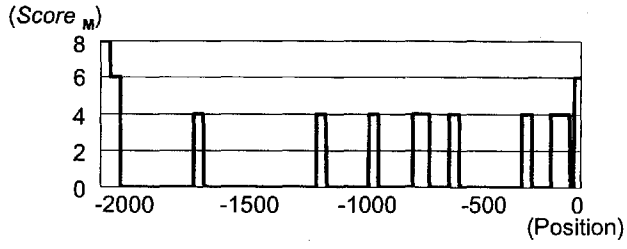

Fig.6 Distribution of $Score_M$

3.2.2. Total Score ($Score_T$)

Repeated pattern like "tttt...", "tatata..." and so on are often included in non-transcribed region of DNA sequence. We think that these repeated patterns are one of the characteristic patterns. Alignment score=4 appears continuously with short intervals at position between -200

and -100 in fig.5. There is possibility of that continued alignment score with short intervals is an influence of repeated pattern. For example, if "accaccacc..." is shifted to the right or the left for 3 bases, it is not different from "accaccacc...". As a result, the alignment score appears continuously in short region. *Score* $_T$ is index to extract repeated patterns in common region of DNA sequences. *Score* $_T$ is defined as follows,

$$Score_T(i) = \sum_{j=i-N}^{i+N} Alignment\ \ score(j) \cdot \qquad (2)$$

If *Score* $_T$ is high at position k, a lot of alignment score is continued at position between k-N and k+N. Fig.7 shows a distribution of *Score* $_T$ that calculated from alignment scores in fig.5 (N=20).

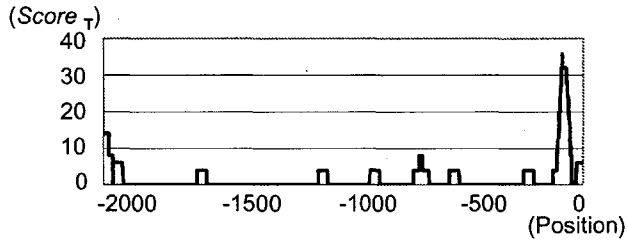

Fig.7 The distribution of *Score* $_T$

Score $_T$ is high at position between -200 and -100 in fig.7. This is equal to region that seems the influence of repeated patterns. In this research, characteristic patterns are extracted according to *Score* $_M$ and *Score* $_T$.

4. Extraction Characteristic Patterns in *D.discoideum*

4.1. Experimental Data and Experimental Conditions

Experimental data is non-transcribed region data of *D. discoideum*. The number of non-transcribed region data in each stage is as follows,

- A stage : 92 data,
- C stage : 58 data,
- S stage : 63 data,
- V stage : 141 data.

These data were created on experiments in Univ. of Tsukuba [7] [8] [9]. These non-transcribed region data consist of 2000 upstream bases from the TSS as shown fig.8.

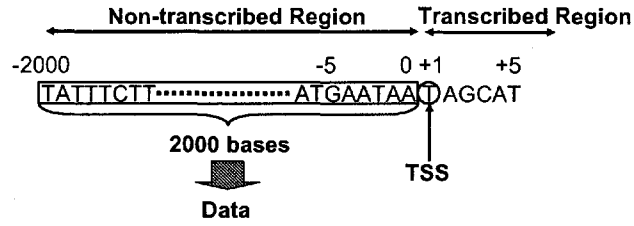

Fig.8 The Contents of Non-transcribed Region Data

Experimental conditions is as follows,

- Candidate patterns are 10-length patterns from "aaaaaaaaaa" to "ttttttttt" (the number of patterns is 4^{10}),
- The positions that alignment score is more than 4 are extracted in DP matching (the total of gap and mismatch is not more than about 3).

4.2. Comparing of four growth stages.

Characteristic patterns are extracted by comparison of *Score* $_M$s or *Score* $_T$s of four growth stages. We will show how to extract characteristic patterns in A stage according to *Score* $_M$.

Average of *Score* $_M$ for a candidate pattern in A stage (*Score* $_M$*_A*) is calculated at first. The *Score* $_M$*_A* is defined as follows,

$$Score_{M\_}A(Cp(j),k) = \frac{\sum_{i=1}^{i\le N(A)} Score(A(i),Cp(j),k)}{N(A)}, \qquad (3)$$

where,

$A(i)$: i^{th} data of A stage,
$Cp(j)$: j^{th} candidate pattern ($0 \le j < 4^{10}$),
k: the position in DNA sequence ($-2000 \le k < 0$),
$N(A)$: the number of data of A stage (92),
$Score_M(A(i),Cp(j),k)$: *Score* $_M$ for $Cp(j)$ at position k in $A(i)$.

The average of *Score* $_M$ in C stage, S stage and V stage (*Score* $_M$*_C*, *Score* $_M$*_S* and *Score* $_M$*_V*) are calculated in the same way.

Next, the percentage of *Score* $_M$*_A* (*P_ Score* $_M$*_A*) is calculated as follows,

$$P\_Score_{M\_}A(Cp(j),k) = \frac{Score_{M\_}A(Cp(j),k)}{Sum\_Score_M(Cp(j),k)} \times 100, \qquad (4)$$

$$Sum\_Score_M(Cp(j),k) = Score_{M\_}A + Score_{M\_}C + Score_{M\_}S + Score_{M\_}V, \quad (5)$$

where $Score_{M\_}A\ (Cp(j),\ k) \geq 4$. This condition to calculate $P\_Score_{M\_}A$ is used to choose only the high $Score_M(A(i),Cp(j),k)$ because the low $Score_M(A(i),Cp(j),k)$ means that few $Cp(j)$ appears in A stage. If $Score_M(A(i),Cp(j),k) < 4$, $P\_Score_{M\_}A(Cp(j),k)=0$. The maximum $P\_Score_{M\_}A$ is calculated as follows,

$$P_{\max}\_Score_{M\_}A(Cp(j)) = \max_{-2000 \leq k < 0} P\_Score_M(Cp(j),k). \quad (6)$$

The candidate patterns with the high $P_{max}\_Score_{M\_}A$ are extracted as characteristic patterns in A stage. Characteristic patterns in other stage are extracted in the same way.

Characteristic patterns according to $Score_T$ are extracted in almost the same way. But, only the condition to calculate P_Score_M is different. In the case of extraction of characteristic patterns according to $Score_T$ in A stage, $P_{max}\_Score_{M\_}A$ is calculated where the following conditions are satisfied.

$$Score_{M\_}A(Cp(j),k) \geq 4 \quad (7)$$

$$Score_{T\_}A(Cp(j),k) \geq \frac{\sum_{k=0}^{2000}(Sum\_Score_T(Cp(j),k))}{4} \times 1.8 \quad (8)$$

Because the maximum $Score_T$ is not 10, the condition only eq.7 is not enough to choose the high P_Score_T. Eq.8 is effective to choose it. "1.8" in eq.8 is decided by preliminary experiments.

4.3. Experimental Results and Discussions

4.3.1. Experimental Results ($Score_M$)

Experimental results according to $Score_M$ are shown in Table.2-Table.5.

Table.2 Characteristic Patterns in A Stage ($Score_M$)

Characteristic Pattern	Position	$P_{max}\_Score_{M\_}A$
(1) acatttataa	-1475	35.9 (%)
(2) acaatattaa	-30	35.3

Table.3 Characteristic Patterns in C Stage ($Score_M$)

Characteristic Pattern	Position	$P_{max}\_Score_{M\_}C$
(1) ttgtgtattt	-109	36.0 (%)
(2) tttgaatttc	-60	35.6
(3) atgtttgatt	-107	35.5

Table.4 Characteristic Patterns in S Stage ($Score_M$)

Characteristic Pattern	Position	$P_{max}\_Score_{M\_}S$
(1) aatatatata	-46	37.8 (%)
(2) acataataat	-46	37.1

Table.5 Characteristic Patterns in V Stage ($Score_M$)

Characteristic Pattern	Position	$P_{max}\_Score_{M\_}V$
(1) tttgcttttt	-43	35.6 (%)
(2) atttggtttt	-43	35.6
(3) ttttgctttt	-44	35.4
(4) ttgctttttt	-45	35.4

These tables show that characteristic patterns according to $Score_M$ appear in region of near the TSS in most cases. This means that the common pattern in each growth stage appear in region of near the TSS in most cases. Characteristic patterns in the same table (Table.6-9) are similar to each other. The positions in the same table (Table6-9) are similar to each other. These mean that $Score_M$ is effective to extract similar patterns in common region of DNA sequence.

4.3.2. Experimental Results ($Score_T$)

Experimental results according to $Score_T$ are shown in Table.6-Table.9.

Table.6 Characteristic Patterns in A Stage ($Score_T$)

Characteristic Pattern	Position	$P_{max}\_Score_{M\_}A$
(1) tgattatttc	-929	42.1 (%)
(2) tcattatttg	-932	41.3
(3) gtgattattt	-935	41.2
(4) tcattatttc	-931	40.5
(5) tgattatttg	-945	40.0
(6) aatactacaa	-87	39.6
(7) aataatatgt	-1950	39.6

Table.7 Characteristic Patterns in C Stage ($Score_T$)

Characteristic Pattern	Position	$P_{max}\_Score_{M}\_C$
(1) attatgatag	-1780	53.0 (%)
(2) attatgatgt	-1781	52.1
(3) atgattatag	-1777	51.5
(4) gattatgata	-1778	50.9
(5) gatgattata	-1778	50.0
(6) attatcatgt	-1779	49.8
(7) atgattatgt	-1776	49.0

Table.8 Characteristic Pattern in S Stage ($Score_T$)

Characteristic Pattern	Position	$P_{max}\_Score_{M}\_S$
(1) aaagacagaa	-1390	47.9 (%)

Table.9 Characteristic Patterns in V Stage ($Score_T$)

Characteristic Pattern	Position	$P_{max}\_Score_{M}\_V$
(1) ttctcttctt	-46	43.0 (%)
(2) ttgttcttct	-47	43.0

These characteristic patterns often include repeating subsequence. For example, "aaagacagaa" in Table.8 includes "aga" twice. Characteristic patterns in the same table (Table.6-9) are similar to each other. The positions in the same table (Table6-9) are similar to each other. These mean that $Score_T$ is effective to extract repeated patterns in common region of DNA sequence.

5. Conclusions

$Score_M$ and $Score_T$ are proposed for extracting characteristic patterns in each growth stage. $Score_M$ is index to extract patterns in common region of DNA sequence. $Score_T$ is index to extract repeated patterns in common region of DNA sequence.

Characteristic patterns according $Score_M$ and $Score_T$ are extracted from non-transcribed region of *D.discoideum*. Experimental results show that $Score_M$ is effective to extract patterns in common region of DNA sequence and $Score_T$ is effective to extract repeated patterns in common region of DNA sequence.

Biological survey is necessary to examine whether extracted characteristic patterns are TRE or not. We will consult Prof. Hideko Urushihara at Univ. of Tsukuba on extracted characteristic patterns.

Acknowledgements

We are most grateful to Prof. 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.

References

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

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

[3] Ikeuchi T, The Science of Protein, *ohmsha* ,Ltd, 1999 (in Japanese).

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

[5] Miller C, McDonald Jand Francis D, "Evolution of Promoter Sequences: Elements of a Canonical Promoter for Prespore Genes of Dictyostelium", J Mol Evol, No.43, pp.185-193, (1996).

[6] Gojohori T, Moriyama E, Naitou K and Kawai M, "Computer analysis of genome information for enormous DNA data", IPS Japan, Vol.31 No.7, pp.878-886, (1990).

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

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

[9] Seo D, Yasunaga M and Kim JH, "A Computational Approach to Detect Transcription Regulatory Elements in Dictyostelium Discoideum", CEC 2004, Paper 1212, (2004).

[10] Yasunaga M, Ushiyama K, Yoshihara I and Kim JH, "Symbolical Kernel-Based Reasoning: Its Application to the Rule Extraction in Dictyostelium discoideum DNA", Genome Informatics 2001, pp.413-414, (2001).