Feature Extraction from Non-transcribed Region of Dictyostelium Discoideum using Moiré Picture

Ikuo YOSHIHARA [1], Toshiro ONITANI [2], Kunihito YAMAMORI [3],
Moritoshi YASUNAGA [4], Hideko URUSHIHARA [5]

Abstract

In recent years, DNA sequences of various living things have been deciphered actively. In this paper, non-transcribed region of *dictyostelium discoideum* used as DNA data, and the method to extract the feature from DNA sequence using Moiré picture is proposed. The proposed method extracts subsequence of high frequency of appearance at all the growth phase and bias of frequency of subsequence appearance at each growth phase.

Key Words:
Moiré, Matching, Dictyostelium Discoideum, Growth phase

1 Introduction

In recent years, DNA sequences of various living things have been deciphered actively. DNA sequence consists of four kinds of bases A (adenine), G (guanine), C (cytosine), and T (thymine). DNA stores information of amino acid of protein,etc.

In genome analysis, development of the technique for extracting the feature of the sequence is one of the important subjects.

We utilize the character of moiré picture that can get coincidence information from a picture, and algorithm based on optical moiré is proposed. Moreover, the method to extracts the feature of sequence is proposed.

2 Matching of Sequence using Moiré Picture

2.1 Dictyostelium Discoideum

There are four growths in the *dictyostelium discoideum* (in Fig.2.1).

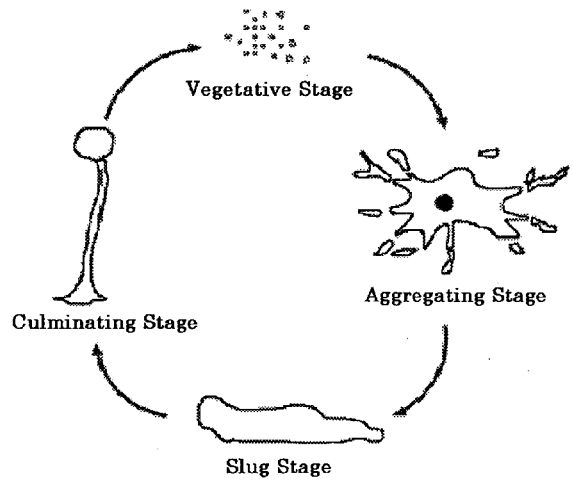

Fig.2.1 Growth of *Dictyostelium Discoideum*

In aggregating stage, the amoeba starts assembly, and pseudoplasmodium is formed. In slug stage, pseudoplasmodium begins to move. In culminating stage, fruiting body is formed, and it has the spore. In vegetative stage, the amoeba starts fission and breeding.

1) Professor, Dept. of Computer Science and Systems Engineering
2) Undergraduate Student, Dept. of Computer Science and Systems Engineering
3) Associate Professor, Dept. of Computer Science and Systems Engineering
4) Inst. of Information and Electronics, University of Tsukuba
5) Inst. of Biological Sciences, University of Tsukuba

2.2 DNA Structure

DNA is a polymer that is called nucleotide. Nucleotide consists of saccharide, a phosphorus acid and a base. There are four kinds of nucleotide A, C, G, and T. Moreover, the amount of adenine and thymine is almost equal and the amount of cytosine and guanine is almost equal in the DNA.

The gene in DNA consists of exon (It is a region where affect coding to the protein) and intron (It is a region between exon), etc. Gene is transcribed into RNA, a part of which is used to translate into the protein. The initiation position of transcript is called "transcription initiation site". Moreover, region that is not transcribed is called "non-transcribed region"(in Fig.2.2).

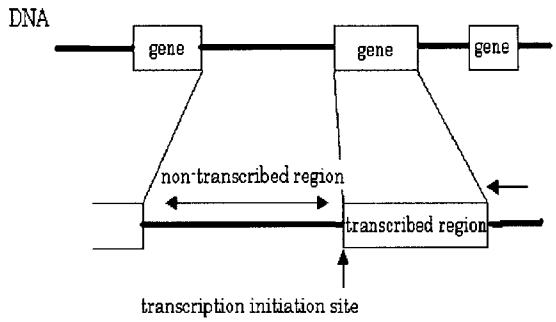

Fig.2.2 DNA Structure

2.3 Moiré Fringes

When the reflection from two objects with almost equal coarseness surface is piled up, stripes appear occasionally. It is called moiré fringes. The moiré fringe often appears when lattices of a curtain overlap each other or when the shadow of a lattice is observed through the lattice itself.

Now, moiré fringes are used at the solid form measurement and the leading means of 3-dimensional measurement.

2.4 Matching of Sequence using Moiré Picture

Moiré picture is created by coding two sequences like the stripes of light corresponding to four-base A, C, G and T and overlapping two coded sequences at a small intersection angle. The coincidence portion included in two sequences is detected from the moiré fringes.

This technique can get coincidence information of wide range visually. So it is effective in the genome analysis that treats many data at once.

At first, each base in sequence is coded to the stripe using four domains. The coding rule of four bases in DNA sequence (adenine: A, guanine: G, cytosine: C, thymine: T) is shown in Fig.2.3.

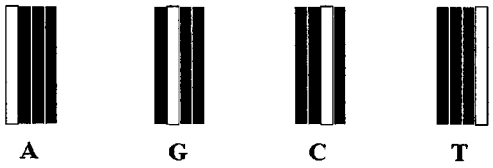

Fig.2.3 Coding Rule

Secondly, "coding picture" which coded all base in sequence is created.

If the coding picture of two sequences (S1, S2) is printed on a transparent sheet respectively and it is overlapped at a small intersection angle, the coincidence portion in sequence will appear as moiré fringe (in Fig.2.4) [ref 3)-7)]

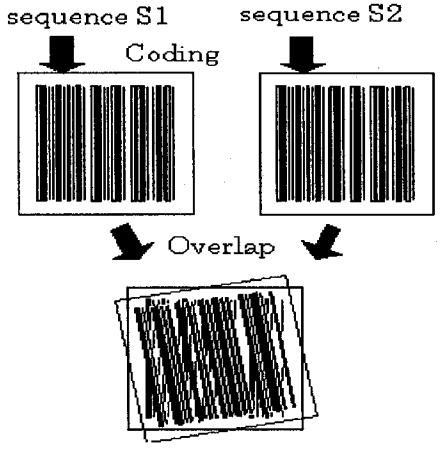

Fig.2.4 Processing Procedure of Creating Moiré Picture

An example of moiré picture obtained from processing procedure of Fig.2.4 is shown in Fig.2.5.

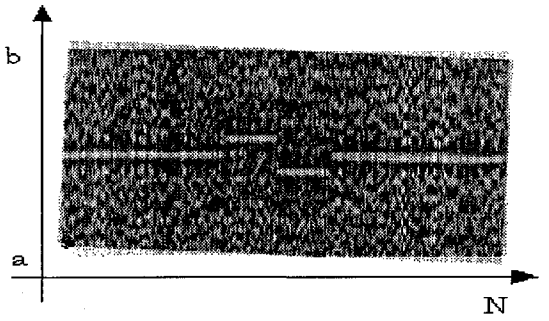

Fig.2.5 Moiré Picture

If there is a coincidence portion in the sequence S1 and sequence S2, the coincidence portion becomes bright stripe and appears in a moiré picture (in Fig.2.5). If it takes point P (i,j) on Fig.2.5, point P shows the matching result of the i'th element of sequence S1 and the (i-j)'th element of sequence S2.

But, N, a, and b in Fig.2.5 are determined by the resolution of a picture, the size of the sheet which prints a picture and the angle that piles up a sheet.

Using of moiré picture has an advantage that cording is easy to process and coincidence information can be got in a picture visually.

But, coding long sequence on the sheet and examining coincidence portion in detail from optical moiré picture takes much time and many efforts. So, making a long moiré picture on the sheet is not always convenient.

In order to compensate this weak point, we propose a "computed moiré (tentative name)" in stead of optical moiré. "Computed moiré" whose outputs are calculated and it is like a moiré picture.

2.5 Computed Moiré

The proposed method by us is called "computed moiré". Computed moiré has the feature that inputs the sequence to computer and outputs the coincidence portion of sequence like a moiré picture by using the line (-). If m'th element of sequence S1 and n'th element of sequence S2 is the same elements, it is assumed "match", and outputs the line (-) to the image. If they are different elements, it is assumed "mismatch", and does not output anything to the image. Fig2.6 shows matching result of all elements in sequence S1 and S2.

Algorithm

$$S1[m]=\{A, C, G, T\},\ S2[n]=\{A, C, G, T\} \quad (1)$$

$$O(S1[m], S2[n])=\delta\ S1[m],S2[n]=\begin{cases}1 \text{ (black)}\\ 0 \text{ (white)}\end{cases} \quad (2)$$

S1[m]: m'th element of sequence S1

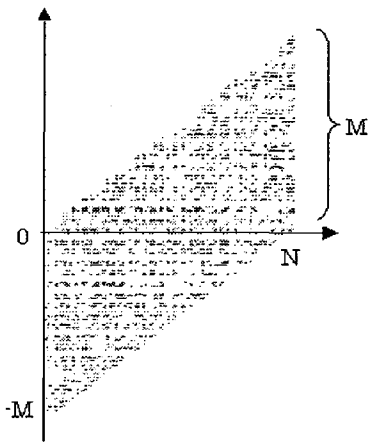

Fig.2.6 Computed Moiré Picture

If it takes point P(i,j) on Fig.2.6, point P shows the matching result of the i'th element of sequence S1 and the (i-j)'th element of sequence S2.

$$P(i,j)=\text{result of } O(S1[i],S2[i-j]) \quad (3)$$

In this image, a shortest line shows the coincidence of an element. That is, a coincidence portion in the sequence becomes a long line by connecting the short line and appears in the image.

However, the coincidence of an element is not suitable for the coincidence of the arrangement.

In order to show only the suitable one as a coincidence of the array, we propose "LPF (Low Path Filter)". "LPF (Low Path Filter)" shows only the long coincidence such as coincidence of the array by putting the short coincidence away. (in Fig.2.7).

Thus, putting the filter on the image can get only necessary data.

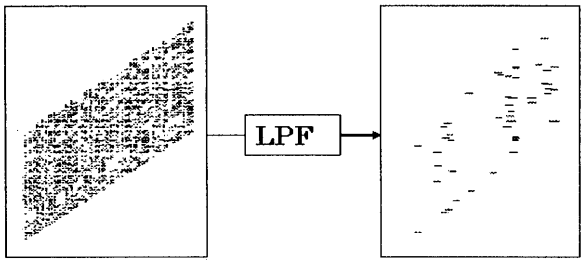

Fig.2.7 LPF

3 Analysis of Dictyostelium Discoideum

3.1 Experimental Data

Non-transcribed region data of *dictyostelium discoideum* is used as experimental data. These data were created based on experiments in Tsukuba University and genome database "cDNA project" [ref 8]. There are four kinds of data of growth phases A (Aggregation stage), C (Culminating stage), S (Slug stage) and V (Vegetative stage). Moreover, amount of data is as follows.

- Growth phase A: 91data
- Growth phase C: 89 data
- Growth phase S: 82 data
- Growth phase V: 102 data

But, V is not used, because it due to imperfection. Non-transcribed region data of *Dictyostelium discoideum* has 2000 bases from the front of transcription initiation site. When there is data that does not come up to 2000 bases, the part is buried by asterisk (in Fig.3.1).

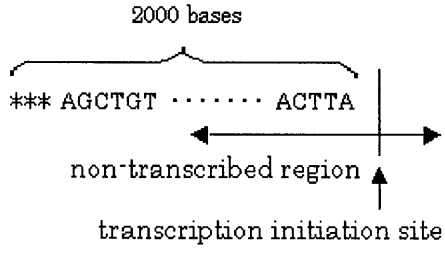

Fig.3.1 The Contents of Data

3.2 Detection of Coincidence Portion

A datum of growth phase A is taken out (it is defined as "data-a1". As an example of detecting the coincidence portion, we experiment by using data-a1. The number of "AAAA" included in the data-a1 and the appearing positions of it are searched.

First of all, computed moiré picture is made from data-a1 and "AAAA" (in Fig.3.2).

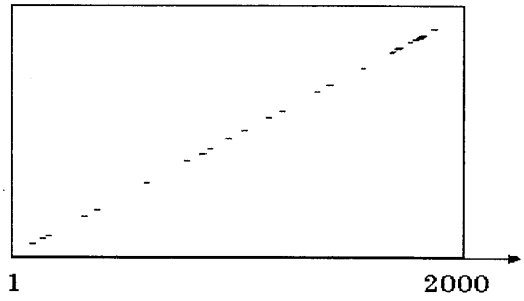

Fig.3.2 Computed Moiré Picture Created From Data-a1 and "AAAA"

The X-axis of Fig.3.2 shows the position of sequence of data-a1. That is, X=i shows the i'th position of data-a1 from transcription initiation site. A line (-) in Fig.3.2 shows that "AAAA" is included in the data-a1.

The number of line (-) means the number of "AAAA" included in data-a1.

By using this operation, the number of "AAAA" included in the sequence is searched.

3.3 Analysis of Non-transcribed Region Data of Dictyostelium Discoideum

Section number is attached from transcription region site of data each 100 bases (in Fig.3.3).

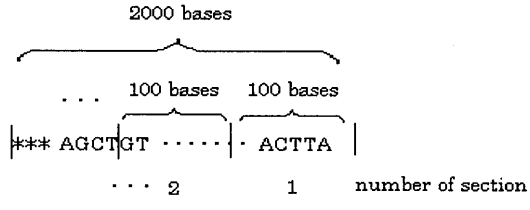

Fig.3.3 Section Number

Short sequence included in the sequence is called "subsequence". The number of four-base subsequence ("AAAA～TTTT") in each section of experimental data is searched, and average of each growth phase is

calculated. The following two experiments are conducted using average of each growth phase.

- Experiment 1

  Objective; Investigating subsequence of high frequency of appearance at all growth phase.

  Method; Section number of the subsequence included twice or more in the section is extracted from average of each growth phase.

- Experiment 2

  Objective; Investigating how the frequencies of subsequence appearances change at each growth phase.

  Method; The subsequence that satisfies the two requirements as follows is extracted.

  - Included once in the section.
  - Bias of frequency of appearance at each growth phase is more than 40 percent.

3.4 Experimental Result and Discussion

Table 1 shows the subsequence and the section number extracted by the experiment 1.Table 2 shows the subsequence and the section number extracted by the experiment 2.

From Table 1, sequences that the number of section appeared more than four times in all the growth phase are extracted. They are assumed to be subsequences with a high frequency of appearance at all the growth phases. Subsequences of high frequency of appearance are as follows.

AAAA, AAAT, AATA, AATT, ATAA, ATTA, ATTT, TAAA, TAAT, TATT, TTTA, TTTT

Table 2

	Growth phase A	Growth phase C	Growth phase S
subsequence	section number		
AAAA	7,9	14	15
AAAC	7		13
AAAG	20	15,19	16,17
AACA	14	4	8,9,11,20
AAGA		19	16,17
AATA	15	17	7,12,19,20
AATG		20	
AATT	12,17		
ACAA	14	4,14,19	8,9,11,20
AGAT			16
AGTA		15	
ATAA	15	17	7,10,12,19,20
ATAT		7,11	2,5,10,15,16,20
ATCA			4,8,9,19
ATGA	18		
ATTA	8,9,11,12,17	14	19
ATTC			18
ATTG	10		
ATTT	9		
CAAA		16,19	5,9,12,20
CAAC			7,8,9
CAAT	14,17		13,20
CATT	13		
CCCC	8		
CTTT			20
GAAA	7		9
GGGG	7		
GTTG	14,17		
GTTT			17
TAAA	9		12
TAAT	14,15		7,9,20
TATA	9	7,18	2,10,15,16,19,20
TATT	8,9,10,11		13,15
TCAA		11	1,3,4,5,9,18
TCAT		13	8
TGAA	18,20		
TGAT	15,18,20	14	
TGGT	20		
TGTT	12,14,17,18,19		
TTAA	12,13		16
TTAT	8,10,11,17		13
TTCA	19		16,18
TTGG	5		
TTGT	12,17,18	6	
TTTA		18	15
TTTC	8,9	16	20
TTTG	5,20		
TTTT	1	18	11,14,16

Table 1

	Growth phase A	Growth phase C	Growth phase S
subsequence	section number		
AAAA	all	all	all
AAAT	all	all	all
AACA			8
AATA	34,5,6,8,9,10,11,12,14,15,16,17,18,19,20	2,3,4,5,6,7,8,9,10,11,15,16,17,18,19,20	2,3,4,5,6,7,8,9,10,12,13,18,19,20
AATT	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,17,18	1,2,3,4,5,6,7,8,9,10,11,13,14,18,19,20	1,2,3,4,5,6,7,8,9,10,11,13,15,16,18,19
ACAA			8,20
ATAA	3,4,6,8,9,10,12,15,16,18	1,2,3,4,5,6,7,8,9,10,11,15,16,17,18,19	2,3,4,5,6,7,8,10,12,16,19,20
ATAT		7	15
ATTA	4,8,9,10,11,12,16,17,18,19	17,18,19,20	8,11,12,13,16,17,18,19,20
ATTT	1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,19,20	1,2,3,4,5,6,7,8,9,10,13,17,18,19,20	8,11,12,13,16,17,18,19,20
TAAA	1,2,3,4,5,6,7,8,9,10,11,13,16	1,2,3,4,5,6,7,8,9,10,11,18,19,20	1,2,3,4,5,6,8,9,10,11,12,13,14,15,18,19,20
TAAT	6,9,10,12,14,15,17,18	2,8,9,15,17,18,19	2,3,4,6,7,8,10,12,18,19,20
TATA			15
TATT	1,2,3,4,5,6,7,8,9,10,11,12,16,17,18,19,20	1,2,4,5,7,9,17,18,19,20	1,2,3,8,9,12,13,15,16,17,18,19,20
TTTA	1,2,3,4,5,6,7,8,9,10,11,12,13,14,17,19	1,2,3,4,5,6,7,8,9,10,13,18,19,20	1,2,3,4,5,6,8,10,11,12,13,15,19
TTTT	all	all	all

According to biological interpretation obtained subsequences, bases of A and T are united easily than the base of G and C. From the result of experiment 1, all the obtained subsequences consist of only A and T. The result of experiment 1 is does not contradict from the view point of biology.

From Table 2, subsequences that the number of sections appeared more than three times in each growth phase are extracted. They are assumed to be subsequences that the frequencies of subsequence appearance change at each growth phase as follows.

- Growth phase A: ATTA, TATT, TGAT, TGTT, TTAT, TTGT
- Growth phase C: ACAA
- Growth phase S: AACA, AATA, ACAA, ATAA, ATAT, ATCA, CAAA, CAAC, TAAT, TATA, TCAA, TTTT

The number of discovered subsequence is as follows.

- Growth phase A: Six subsequences
- Growth phase C: One subsequence
- Growth phase S: Twelve subsequences

According to biological interpretation obtained subsequences, the growth S is active and the growth C is non-active. From the result of experiment 2, growth phase S includes many special subsequences and growth phase C includes a few special subsequences. The number of discovered subsequence does not contradict from view point of biology.

4 Conclusion

Computed moiré that extracts feature of DNA sequence is proposed. The algorithm of computed moiré is based on optical moiré picture.

In order to extract the feature from non-transcribed region of *dictyostelium discoideum*, subsequences of high frequency of appearance and subsequence that the frequencies of appearances change at each growth phase are searched. Their subsequences are discovered. As a result, it is discovered that there is the different feature in sequence of each growth phase.

Future works are to expand program so as to search longer subsequences and so as to search subsequences with alignment.

Reference

1) Miyake N., Kanehisa M.: Project of Genome of Homo and Knowledge Information Processing System, *baifukan* publishing, 1995 (In Japanese)

2) Yamada A., Yokozeki S.: Moiré Fringes and Interference Fringe Application Mensuration, *korona-sha* publishing, 1996 (In Japanese)

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

4) Nitta K., Togo H., Tanida J.: "Matching Information Terminal Based on a Spatially Coded Moiré Technique", Optical Engineering, Vol.40, pp.2386-2391, (2001).

5) Nitta K., Togo H, Yahata A, Tanida J.: "Genome Information Analysis Using Spatial Coded Moiré Technique", Pacific Rim 2001, Technical Digest, pp.494-495, 2001.

6) Nitta K., Yahata A., Tanida J.: "Information Extraction from Amino Acid Sequence Using Spatial Coded Moiré Matching Technique", Optics Japan 2001, Geoinforum, pp.321-322, 2001 (In Japanese)

7) Nitta K., Yahata A., Togo H., Tanida J: "Genome Information Analysis Using Special Coded Moiré Technique", The 48th Applied Physics Relation Union Lecture Geoinforum, pp.102, 2001 (In Japanese)

8) http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html