A Calibration Technique for Single-Dimensional Distribution of Local GC Content

Tomoyuki Fujimoto
tomoyu-f@is.naist.jp
Tadao Sugiura
sugiura@is.naist.jp
Takeshi Nagaya
t-nagaya@is.naist.jp
Tetsuo Sato
t-sato@is.naist.jp
Kotaro Minato
kotaro@is.naist.jp

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0101, Japan

Keywords: GC content, Lambda DNA, single molecular DNA, correlation coefficient

1 Introduction

Recent years have seen growth in research into single-molecule measurement due to progress in optical microscopy techniques[1]. Consequently, information on genome sequences can now be obtained by the measurement of single-molecule DNA using optical microscopes[2], revealing that GC content of genome DNA is not uniform. Using this feature, we previously analyzed single-molecular DNA with single-molecular imaging of DNA stained with GC-specific dye fixed on a glass surface and observed through an evanescent field microscope. From the analysis, it appears that the number of base pairs within the unit length of the DNA on the glass substrates was different from what was estimated, and measured results did not closely match with genome database. Here we report on a novel method of calibrating the number of base pairs within the unit length of DNA.

2 Method and Results

2.1 GC Content Distribution Measurement

We measured GC content distribution using Lambda DNA, with a length of 48,502 bp, using an optical microscope. The steps involved in measuring the GC content distribution were as follows:
1). Prepare DNA stained with YOYO-1 after being stained by 7-AAD. Here, YOYO-1 and 7-AAD are fluorescent molecules and the latter is GC specific.
2). Stretch the DNA by water flow and fix it to a glass surface[3].
3). Obtain DNA images by exciting of each fluorescent molecules by using an evanescent field microscope.
4). Calculate the GC content distribution by pixel value within two images obtained above:

\[
\text{GC content}(i) = \frac{f_{7\text{-AAD}}(i)}{f_{YOYO-1}(i)} \times 100\% \tag{1}
\]

where \(i\) is the pixel position within an image, \(f_{7\text{-AAD}}(i)\) is a pixel value within an image obtained by excitation of a 7-AAD molecule, \(f_{YOYO-1}(i)\) is a pixel value within the other image. Calibration of the base pair number within the unit length is performed in the next step. To compare measured data with theoretical data from the genome database, the theoretical GC content is defined as:

\[
\text{GC content}(i) = \frac{N_G(i) + N_C(i)}{L} \times 100\% \tag{2}
\]

where \(N_G(i)\) is the quantity of Guanine within the given unit length, \(N_C(i)\) is for Cytosine, and \(L\) denotes the number of base pairs within one pixel. There are approximately 600 base pairs depending on the microscope’s resolution.
2.2 Length Calibration

Since the number of base pairs within a given unit length is variable according to the DNA structure, the pixel value distribution of a YOYO-1 image is not uniform. It is thought that a small pixel value indicates a small number of base pairs. Thus, we calculate maximum($f_{\text{max}}$) and minimum($f_{\text{min}}$) values within the YOYO-1 image. First, for length scaling using Eq. (3), we expand the pixel length:

$$
ex = \frac{f_{\text{YOYO-1}(i)} - f_{\text{min}}}{f_{\text{max}} - f_{\text{min}}} \times d, \quad (3)$$

where $d$ represents the number of steps between the maximum and minimum pixel values. In this study we set $d$ to be 10. The calibration step is as follows:
1. Using Eq. (3), expand the pixel length based on $f_{\text{YOYO-1}(i)}$.
2. Shift expanded data using several feature point of theoretical data.
3. Shrink the expanded data linearly searching the maximum correlation coefficient.

2.3 Results

Here we present the results of measuring the GC content of DNA. In Fig. 1 we compare calibrated data of GC content, uncalibrated data and theoretical data from the genome database. Table 1 shows a comparison of correlation coefficients for the theoretical calibrated data and uncalibrated data. The table indicates that the correlation coefficients of all calibrated data have improved.

Table 1: Comparison of correlation coefficient

<table>
<thead>
<tr>
<th>DNA</th>
<th>correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uncalibrated</td>
</tr>
<tr>
<td>1</td>
<td>0.550</td>
</tr>
<tr>
<td>2</td>
<td>0.676</td>
</tr>
<tr>
<td>3</td>
<td>0.112</td>
</tr>
<tr>
<td>4</td>
<td>-0.358</td>
</tr>
<tr>
<td>5</td>
<td>-0.432</td>
</tr>
</tbody>
</table>

Figure 1: Comparison of GC

3 Discussion

In this study, we performed calibration of base-pair numbers within DNA images to determine GC content distribution. From the results we confirmed that calibrated data are improved compared with uncalibrated data. This result can be improved further by stretching DNA using tools such as optical tweezers. There results show the potential for the measurement of GC content distribution to be useful for homology search. In addition, DNA characteristics such as conformation and stiffness can be analyzed.

References

