Hierarchical clustering matrix method (HCM) applied to DNA barcode assembly for bacterial chromosomes

Master Thesis

Wensi Zhu

LUND
UNIVERSITY

Thesis for the degree of Master of Science
Supervisor: Assoc. Prof. Tobias Ambjörnsson

Contents

Popular Science

DNA sequencing, which is a process determining the specific order of nucleotides of a DNA molecule, has a pivotal role in a wide range of scientific fields such as medical diagnosis, forensic biology and biotechnology. DNA barcoding is a complementary tool to traditional DNA sequencing which provides coarse-grained sequence information and it can be used, for instance, for species identification, in the same way a scanner uses the black and white stripes of the barcode for goods in the supermarket.

How do the DNA barcodes work? After staining with dye molecules, DNA molecules with different sequence information have different fluorescent patterns observed with fluorescence microscopy. Then we convert these fluorescent patterns into DNA barcodes, which serve as a sequence-dependent 'ID card' for each DNA molecule. The process takes either a few minutes or hours, depending on the number of molecules. This approach is faster than traditional ones, which can take days or weeks from sample preparation to results. Therefore it has potentials in rapid diagnostics.

However, many unsolved problems still hinder the wide use of DNA barcodes. For example, the fluorescent images are always affected by experimental noise so that the patterns of the same regions of a DNA molecule all have slight differences. Furthermore, DNA molecules are fragmented during extraction from cells, and thus the fluorescent patterns that are collected reflect only the fragmented molecule. Like pieces of a puzzle, the fragmented barcodes need to be joined together to get a barcode for the intact DNA molecule.

This thesis reports on developing an algorithm to piece up DNA fragmented barcodes for chromosomal bacterial DNA while handling the problems mentioned above. This work provides an important opportunity to improve the ways of piecing full chromosomal barcodes together. In the future, chromosomal barcoding might open up new ways to profile the genetic dynamics in various fields like the fast diagnosis of bacterial infections, embryonic diagnostics and cancer diagnosis.

Abstract

DNA barcodes carry coarse-grained genetic information of DNA sequences taken from a genome. Potential applications include bacteriology, medical diagnosis and taxonomy. However, the current state-of-the-art tools for extracting DNA molecules from cells provide only fragmented pieces of chromosomal DNA. As a consequence, also DNA barcodes are fragmented. This calls for the development of complementary computational methods to piece up the fragments which help to restore the intact barcodes. Challenges for such developments are noise effects, an influence of DNA structural variation and experimental errors.

This thesis presents a new method for assembling DNA fragments of large sizes (300 kilobase pairs in mean length). We develop a matrix-based hierarchical clustering algorithm to piece together the DNA fragments by assembling the overlapping DNA regions. Two barcodes are compared by sliding one to another to find the best alignment position. Following this step, we average the overlapping regions and stitch two barcodes together into an assembled barcode. By repeating the above process, we could get a near-intact full barcode of an intact chromosome. We demonstrate that our method works quite well for assembling fragments of theory barcodes with added noise. For the experimental barcodes, we only get several large pieces instead of an intact barcode. In the last section we discuss possible improvements of our method and future applications of DNA barcode assembly of large-sized DNA barcodes.

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Associate Professor Tobias Ambjörnsson, for providing me with a great opportunity to study here with such an interesting project and always being patient and supportive with my endless questions. Special thanks to Albertas Dvirnas for the programming support, mathematical discussion and patient explanations to all my ridiculous problems. I am honored to conduct my thesis in this group where I came to know about so many new things. I would like to thank Björn Linse for the great working atmosphere in our office and Swedish support during the long-lasting winter. Furthermore, I would also like to thank all the amazing people from CBBP for providing Fika and nerdy jokes during my office time.

I would like to pay my heartiest thanks to my friends Jojo, Yuhe, Cesare and others for the company and support.

Nobody has been more important to me in the pursuit of my study than my family. I would express my deep gratitude to my parents whose love and guidance are with me in whatever I pursue. I love you.

Chapter 1

Introduction

The recent determination of numerous bacterial and eukaryotic genome sequences poses new challenges for comparative sequence analysis[1], and thus obtaining a complete genome sequence is one of the most significant tasks in genome biology field([1],[2]). As the global scientific community has gradually paid more and more attention to genome sequencing, the relevant technologies have been rapidly evolved in recent years. Major problems arose because it is still impossible to sequence a whole genome as a single piece. Therefore, a successful recovery of a genome DNA sequence requires about eight copies of each piece to be assembled[3]. Shotgun sequencing helps to break long DNA sequences into smaller fragments which are then sequenced into larger fragments[4]. After this, one applies the computational process of sequence assembly which joins all the reads together to make longer continuous sequences named contigs. And many contigs can be assembled into scaffolds which are ordered collections of contigs and represent a continuous region in DNA or a genome[5]. The whole reconstruction process is known as genome assembly.

Numerical improvements have made genome assembly methods more mature, however, it is widely accepted that several challenges still remain[6]. Here we address three of those challenges. The first challenge is that there are always a large number of fragments: some pieces might be missing altogether, some pieces contain unpredictable errors and some might have multiple copies which increase the complexity of the computations. The second challenge is to find the correct position for the fragments. As the only information for determining the correct position for a fragment is from the neighbors, it requires a high standard for the assembly quality. The third challenge is the ambiguity made by the positioning of similar fragments, which leads to the difficulty of finding the correct location for each of them[7].

A great source of aid for our problem is the optical DNA mapping technique, which was initially developed using restriction enzymes to sequence specifically cleave DNA[8]. The purpose of optical mapping is providing coarse-grained sequence information with a spatial resolution of the order kilobase pairs which are from a large DNA molecule (molecular size typically between 50 kbp and several megabases) at scales larger than typical

contig sizes[7]. The optical mapping technique remedies the defect of the short sequencing fragments and allows the DNA to be imaged in the nanofluidic channels on glass surfaces and later converted into barcodes([8], [9]). Another advantage is that it helps to detect and quantify large-scale structural variations occurring in the genome[10].

Developments in the field of optical DNA mapping have led to an increasing interest in assembling medium-sized (hundreds kbp in length) fragments for unsequenced DNA. In this study, we develop an effective approach for assembling the barcodes from bacterial chromosomal DNA. Our hierarchical clustering matrix method (HCM), is designed to assemble large amounts of DNA barcodes in a quick and accurate way.

Chapter 2

Thesis Overview & Outline

This thesis reports on the development of an algorithm for assembling fragmented DNA barcodes: hierarchical clustering matrix method (HCM). I have implemented this method after a suggestion made by my supervisor Tobias Ambjörnsson. This method aims to piece together a number of experimental DNA barcodes, which are from cells with identical DNA sequence and are randomly fragmented, in order to get relatively intact barcodes with the help of the optical DNA mapping technique. The methodological approach taken in this thesis relies on the assumption that highly similar DNA fragments originate from the same position within a genome[5]. Therefore, the highly similar fragments could be joined through overlapping regions into a longer barcode. Due to the insufficient number of the experimental barcodes, we first develop our method using noisified theory barcodes, which imitate the experimental barcodes, generated from an identical sequence.

The major challenge of this work is the noise effect. The two barcodes are slightly different in the presence of experimental noise even though they are from the same underlying DNA sequence[11]. Beyond this, factors like polymorphisms, repetitive sequences, thermal fluctuations and missing data also limit the accuracy of the assembly[3].

This thesis is structured as follows: background information about different kinds of DNA barcodes and the method to characterize genetic sequencing is given in Chapter 3. In Chapter 4, principles of barcode generation are described. This part includes the methods of theory barcode generation from the known DNA sequence and from the experimental raw kymographs. The method of finding and comparing the most similar barcodes are explained in Chapter 5. Chapter 6 presents the method of generating the assembled barcodes, as well as the HCM method which serves as a tool helping to arrange the data effectively. Chapter 7 provides the test results after using our method, and some commentary on these results. The last chapter presents the conclusion of our work, some possible improvements and an outlook for the future.

Chapter 3

Background

3.1 Plasmids vs. chromosomal DNA

The past several decades have seen increasingly rapid advances in the research of plasmids. Plasmids are small circular DNA molecules in bacteria. Compared to chromosomal DNA a plasmid is fairly short, and its length ranges from a few to hundreds of kilobase pairs. The reasons that plasmids play a critical role in biology are related to its particular characteristic: they can replicate independently. Based on it, many research groups have developed various methods for studying plasmids. The inappropriate use of antibacterial drugs has become a growing threat to human health as the abuse leads to antibiotic resistance[12], which widely spreads in mobile genetic elements such as bacterial plasmids.

A method that can rapidly identify and trace plasmids is needed, as plasmids can autonomously transfer and replicate between bacterial cells. Various methods have been developed for rapid identification and characterization of plasmids. For example, S1-coupled pulsed field gel electrophoresis (S1/PFGE) can survey the plasmid numbers and sizes[13]. Another widely-used method is polymerase chain reaction (PCR) which detects genes and plasmid types[14]. Such approaches, however, have failed to address long-range sequence information from the sample and some require existing information of the plasmids. All these problems require the development of a cost-effective method for rapid identification and characterization of plasmids.

A chromosome is a fairly complex DNA molecule compared to a plasmid. An intact DNA molecule is usually hundreds of kilobase pairs in length. Take the haploid human genome(23 chromosomes) as an example. According to the results of the Human Genome Project, the haploid human genome is estimated to be about 3.2 billion bases long and contain around 20,000-25,000 distinct protein-coding genes[15]. Therefore, the sequence analysis process could be quite challenging.

3.2 Optical DNA Mapping

In recent years, optical DNA mapping combined with nanofluidic channels has been attracting a lot of interest and has proved to be a fairly versatile, cheap and efficient method for coarse-grained visualization of the sequence at the molecule level. The concept of optical mapping was first pioneered by Schwartz[16] and his team in the 1990s as a potentially efficient and versatile method for scaffolding genome sequence. The main idea was to use restriction enzymes to cut individual DNA, which has been fluorescently stained and elongated in agarose gel, and gets images from the fluorescence microscopy. Then they used the images to create DNA barcodes, where the cut positions serve as sequence-specific marks.

Even though the optical mapping provides means for obtaining long-range sequential information of DNA fragments, due to the technical constraints, the resulting images from using this method do not yield full sequence information. Single Molecule Real Time sequencing(SMRT), developed by Pacific Biosciences, could yield gene fragments with much longer lengths. However, this method is less popular as it lacks the ability to provide high-quality data due to its high error rate, as well as being an expensive method[17].

During the 2000s, many improvements to the Optical Mapping method were developed. In 2004 Tegenfeldt, who worked as a postdoc in Bob Austin's group at Princeton University, showed that genomic-length DNA can be extended along nanochannels , which allowed them to conduct further statistic measurements and analysis[18]. This breakthrough then helped the optical mapping method to find applications such as detection of structural gene variations[19] and long genome assembly[20].

The whole DNA barcoding workflow can be divided into three main parts: sample preparation, nanofluidic insertion and data analysis. The optical mapping method mentioned in this thesis is based on the experiments led by Fredrik Westerlund's group from Chalmers University of Technology (Sweden) and the schematic illustration is shown in figure 3.1.

Bacteria are cultivated and then two types of ligands, netropsin and YOYO, are added to the sample solution. The competition of the fluorescent YOYO-1 (bind non-specifically) and the sequence-specific non-fluorescent netropsin creates a sequence-specific pattern along DNA molecules. Next, the sample solution is inserted into nanofluidic channels with pressure. As the local conformational changes and center-of-mass diffusion of DNA also contribute to the raw kymographs, a downstream analysis is required for minimizing the effect of such fluctuations on the DNA barcode[8]. There are two main algorithms to align the features of the kymographs: WPAlign (Weighted Path Align)[21] and SSDAlign (Sum-Squared based Align)[4]. After the kymographs have been aligned, we calculate time-averaged experimental kymographs. Then these time-averaged kymographs are converted into a fluorescent intensity curve, which we refer to as an experimental barcode. In Figure 3.2, the process of obtaining experimental barcodes is illustrated.

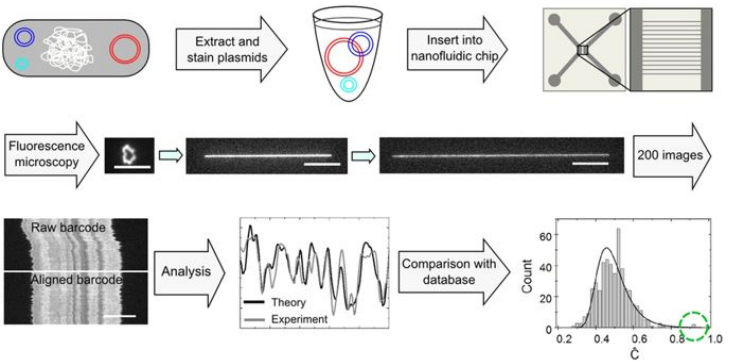


Figure 3.1: Workflow for the optical mapping method for plasmid characterization. Firstly, plasmids are extracted and stained with dye molecules. Then the dyed samples are inserted into a nanofluidic chip using the high pressure of nitrogen gas. Incident light causes the double strands of the DNA structure to break resulting in a linear configuration. Following this step, the fluorescence microscopy will take ~ 200 *images for each plasmid and through a software, the images are aligned and time-averaged. Then the experimental barcode is compared with a set of theory barcodes. The best match to the barcode is identified as shown in the green circle[11].*

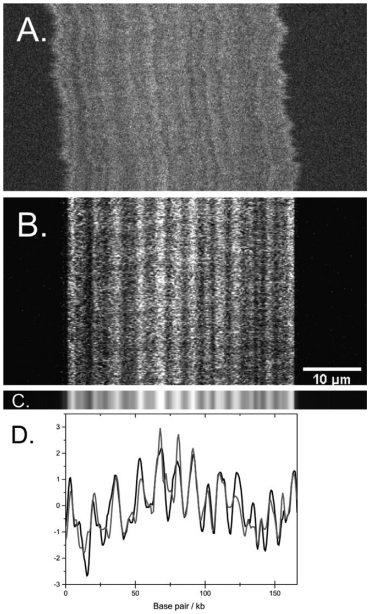


Figure 3.2: (A) Experimental raw kymograph (fluorescent image) for T4 DNA. (B) Aligned kymograph for T4 DNA. (C) DNA barcode consisting of 20 rows generated from the mean of the aligned kymographs. (D) Comparison of the theoretical barcode (gray) and the experimental barcode (black)[22].

Chapter 4

DNA barcode generation

To validate the method of assembling DNA barcode fragments into long barcodes, we generate theory barcodes and noisified theory barcodes to imitate the experimental barcodes. Thus the whole process involves three kinds of barcodes: experimental barcodes, noisified theory barcodes and theory barcodes without noise. Theory barcodes are directly generated from DNA sequences and are used as a reference. The first step of our work is to generate barcodes and this section is explained in detail below.

4.1 Experimental barcodes

The optical DNA mapping experiments were conducted by the Westerlund group at the Division of Chemical Biology of Chalmers University of Technology. The experimental setup and the steps for handling experimental barcodes are introduced below.

4.1.1 Optical DNA mapping experiments

The Westerlund group use *Escherichia coli Strain BL21* as the experimental sample and the DNA molecules are stained with netropsin and YOYO-1, ratios 30:1 (netropsin) and 1:10 (YOYO-1) with respect to DNA. The samples were mixed in 5x TBE (Tris-Borate-EDTA, Medicago AB, diluted with ultrapure water from 10x tablets). Typically the circular DNA molecules are then inserted to nanochannels for linearization (See Figure 3.1). However, our sample molecules are genomic bacterial DNA which is already linear so the linearization step is skipped. 2% (v/v) Beta-mercaptoethanol (BME, Sigma-Aldrich) was added to reduce the photonicking of the samples before the beginning of the experiment. Then the samples were forced into the nanochannels with a dimension of $100 \times 100\ nm^2$ or $100 \times 150\ nm^2$ by pressure-driven flow generated from nitrogen gas.

A combination of an EMCCD camera with a pixel size 130nm and an inverted fluorescence microscope (Zeiss AxioObeserver.Z1) with 100x oil immersion objective (NA

= 1.46) is used to take images of the molecules with an exposure time of 100ms, each molecule would then have approximately 200 images.

4.1.2 Edge identification & bit-weighting

As discussed in Sec. 3, we have to align the features in the kymographs before time-averaging. To that end, we use a sum-square based alignment method, SSDAlign[4].

The time-averaged barcodes contain both the fluorescence signal of the DNA molecule and the lower-intensity background. The main challenge is to define the signal edges. We fit a function $g(x)$ to the data:

$$g(x) = a + (tanh(x-b)d) - tanh((x-c)e)f \tag{4.1}$$

where b and c are the "start" and "end" points; the other parameters are fitting parameters helping to minimize the sum of the squared residuals. b and c are set to the first and last zero-point crossings of the difference between the data signal and the mean data signal. Then we define the effective signal regions L where $L = c - b$. For the downstream analysis, the barcode $B(x)$ can be represented by its effective signal region defining by $0 \leqslant x \leqslant L-1$ with the signal value from b to c. An example fit is shown in Fig. 4.1.

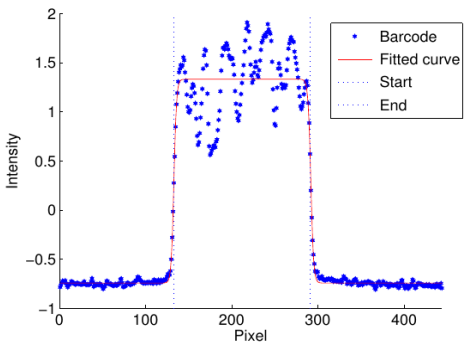


Figure 4.1: The function $g(x)$ (See Equation 4.1) is fitted to a time-averaged barcode of R100 plasmid containing both fluorescence signal and the background signal[8]. The start and end point, b and c, are indicated by the vertical dash lines.

Associated with each barcode $B(x)$, we also have bit-weights[8]. Such bit-weights provide the effective length of the barcodes. The bit weights $W(x)$ have the same range $0 \leqslant x \leqslant L-1$ as the barcode $B(x)$. For a given pixel x, $W(x)$ equals to 0 if x is in an end region and $W(x)$ equals to 1 otherwise. The end regions are the regions where the

distance from b and c points is less or equal to a predefined length $\triangle$.

$$W(x) = \begin{cases} 1, & \text{if } \triangle \leqslant x \leqslant L - 1 - \triangle \, . \\ 0, & \text{if } 0 \leqslant x \leqslant \triangle - 1 \text{ or } L - \triangle \leqslant x \leqslant L - 1. \end{cases} \tag{4.2}$$

We here set $\triangle = 7$ pixels. Due to the system's point spread function, the blurry effects could cause the background signal regions to intermix. The bit-weights effectively mask the ends of the DNA barcodes, thereby minimizing the effect of the intermixing.

4.2 Noisified theory barcodes

4.2.1 Generating randomized barcodes

We here use noisified theory barcode for validating our assembly method.

The noisified theory barcodes we use consists of by two parts: the theory barcode and the randomized barcode. The theory barcode is generated with the competitive binding method mentioned in the next section, and the generation of the randomized barcodes is based on the method called phase randomization ([23],[24]). We here use randomized DNA barcodes instead of random sequence barcodes, which could be generated from a random sequence with equal A/T/C/G characters, due to the fact that DNA genome fragments from organisms have periodic regions of A-T bonds or C-G bonds[25]. This fact explains that the realistic experimental barcodes would have fewer features than the random sequence barcodes do. Figure 4.2 illustrates the comparison of all types of barcodes we mentioned above.

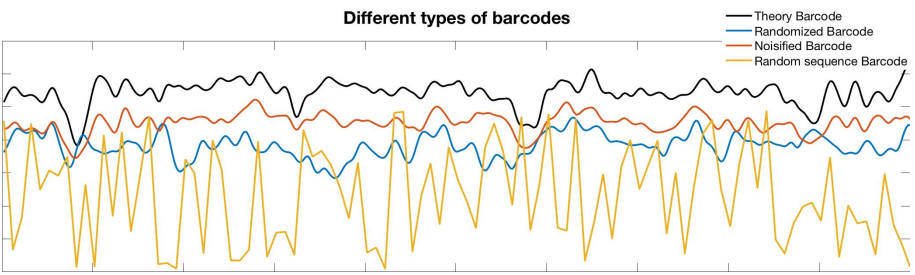


Figure 4.2: Comparison of different types of barcodes mentioned in the thesis. The theory barcode is generated from Escherichia coli Strain BL21 sequence; the randomized barcode is generated based on the periodic nucleotides' bonds of Escherichia coli Strain BL21 sequence; the noisified theory barcode is generated from a combination of a randomized barcode and the theory barcode; the random sequence barcode is generated from the sequence where the probabilities to observe A, T, C, G are independent and equal (= 1/4).

The main idea of the phase randomization is to take real barcodes as input. On completion of calculating their autocorrelation function, output randomized barcodes are produced which have the identical autocorrelation function as the inputs.

Let us consider a database of M real barcodes $p_m(k)$, m=1,2,..., M at base-pair resolution. All these input barcodes are from the RefSeq plasmid database with plasmids longer than 100 base-pairs[8]. Then with the interpolation scheme, we compute database-averaged Fourier amplitudes in Fourier space

$$\hat{p}^{est}(w_l) = \sqrt{\frac{1}{M}\sum_{m=1}^{M} |\hat{p}_m^{interp}(w_l)|^2} \tag{4.3}$$

with spatial frequencies $w_l = \dfrac{2\pi l}{L_{max}}$, l,..., $L_{max}-1$, where L_{max} is the maximum length of the input barcodes.

Then we interpolate the database-averaged Fourier amplitudes to the same length as the theory barcode. Following this process, we draw the uniformly distributed random numbers α, multiply the result from Equation 4.3 by a set of random phases $e^{i\alpha_l}$. Now we are able to get a randomized barcode from the inverse Fourier transformation of

$$\hat{I}^{random}(w_l) = \hat{p}^{est}(w_l)e^{i\alpha_l} \cdot \hat{\phi}(w_l) \tag{4.4}$$

The final step is to convert the resolution from base-pairs to pixels. By repeating the interpolation and the inverse Fourier transform n times, we can generate n randomized barcodes as a null model for the later use.

4.2.2 Noise factor definition

To better mimic the experimental DNA barcodes, we use a noise factor α for adding noise (the randomized barcodes) to theory barcodes (See Sec.4.4):

$$B_{noisified} = \alpha B_{randomized} + (1-\alpha)B_{theory} \tag{4.5}$$

The noise factor α affects the Pearson correlation coefficient value of the barcodes in comparison. If α=1, the barcode is 'all noise'; if α = 0, the barcode is the theory barcode. The two barcodes will be considered the same only if their Pearson correlation coefficient value is above a certain threshold. After doing quantities experiments on plasmids, the Westerlund group have found that the correlation coefficient between the theory barcode and the experimental barcode is typically larger than 0.8. A Pearson correlation correlation value of 0.8 corresponds to an α value of 0.1149 (See Figure 4.3). Even though the α value slightly changes(±5%) with simulations, we do not think the results are altered significantly by the change, and therefore it is ignored in this thesis.

After generating the noise factor, the noisified theory barcodes are produced so that the algorithm can be tested. The noisified theory barcodes are produced from the sequence of *Escherichia coli Strain BL21*. After generating the noisified theory barcodes, they are chopped into short fragments with a mean length of 300 kilobase pairs to imitate the experimental bacteria barcodes.

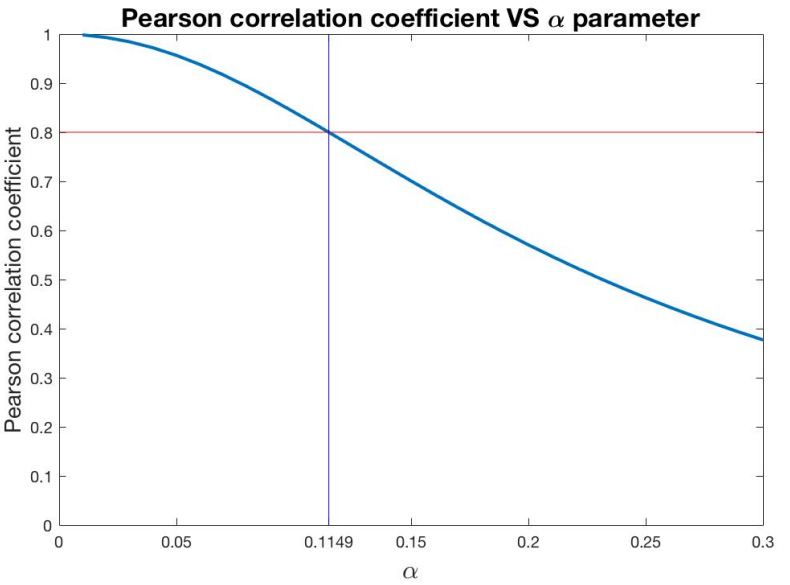


Figure 4.3: Average Pearson correlation coefficient values between 100 noisified theory barcodes and the theory barcode of the BL21 Bacteria as the function of the noise factor α.

4.2.3 Bit-weighting

As described in Sec.4.2, experimental barcodes need the corresponding bit-weighting function to eliminate edge effects. Therefore we also need to generate bit weights for our noisified theory barcodes. Accordingly, we generate the bit weights using Equation 4.2. Here we use $\triangle = 6$ for noisified theory barcodes.

4.3 Theory barcodes

Theoretical barcodes from the molecules whose sequence is already known play a significant role in this thesis, providing a reference for the experimental barcodes. Here we use the competitive binding method for the theory barcode generation. The theory predicts the experimental intensity patterns of the known DNA sequence using a statistical physics framework[4]. The current setup uses two types of ligands: netropsin and YOYO-1. Netropsin has a strong preference for binding A-T rich regions of DNAs, which prevents YOYO-1 to bind to these regions. The two types of ligands both occupy four base pairs when they are bound to DNA. When the DNA sample is added to the mixed solution of netropsin and YOYO-1, the two molecules will compete, leading to emission

intensity variations along the DNA molecule, where A-T rich regions are dark and the C-G rich regions are bright[22]. The competitive binding method provides the probability that a base pair i is occupied by the monomers of netropsin or YOYO-1 respectively.

However, there is a problem when comparing the theory barcodes with experimental barcodes: the resolution is different. The theory barcode has base-pair resolution while the experimental barcode can only be obtained at pixel level due to the limitation of microscope optics. Another problem is that the molecules in the nanochannels are not fully extended.

To deal with this difference, the probability is convolved with a Gaussian kernel, with an experimentally determined standard deviation σ, to imitate the Point Spread Function of the experiments. Finally, the interpolation is applied to achieve pixel resolution and a theory barcode is thus produced[4].

Chapter 5

Barcode comparison

After introducing the method for generating noisified theory barcodes and experimental barcodes, Chapter 5 is concerned with quantitatively comparing barcodes and our method is based on a previous method for merging the intact circular barcodes.

5.1 Length rescaling

We will herein only consider linear barcodes instead of circular barcodes (like plasmids). When comparing two linear barcodes, the method is different from the circular barcodes: the shorter barcode of the two slides from the very left position to the very right position. So during the procedure as we slide one barcode along another, the overlapped region will first increase to a constant number and then decrease. The overlapped region is the only part we need when we compare the similarity of two barcodes and merge them.

5.2 Pearson correlation coefficient

Consider two different barcodes, B_1 and B_2, assume that the length of B_1 is longer or the same length as B_2. The lengths are N_1 and N_2 respectively. We introduce a similarity score C to indicate the similarity level of the two barcodes, which is defined by the Pearson correlation coefficient[26] (PCC) of all "possible alignments" of B_1 to B_2. C depends on two parameters:

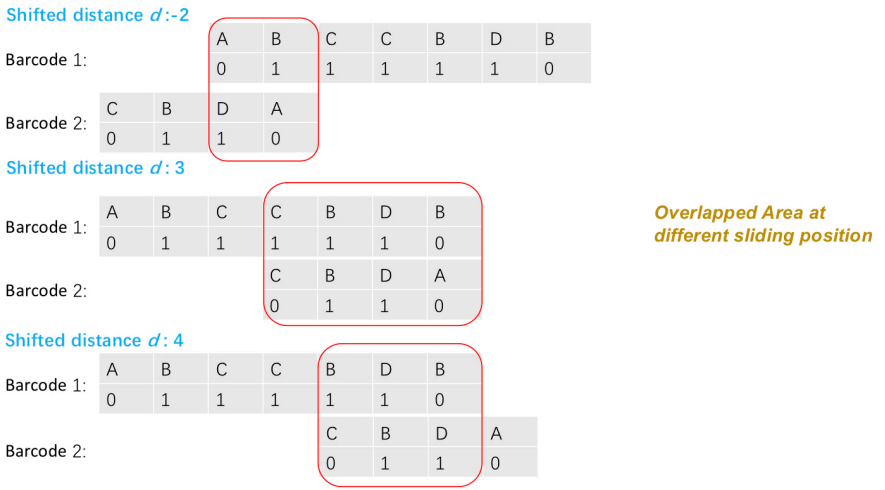


Figure 5.1: Sliding of a pair of barcodes of our hierarchical clustering matrix method. A, B, C represent intensity levels and 0, 1 *represent the values of bit-weights. The longer barcode* B_1 *is fixed and the shorter barcode* B_2 *slides along* B_1*. The design of our method allows two barcodes to be fully overlapped, partially overlapped or not overlapped at all. The two barcodes being compared do not have to be of the same length.*

- Shift parameter d: When comparing two barcodes, to compute all possible alignments of the barcodes, B_2 is slided across B_1. The shift parameter d is defined as an integer in modulo N_1 and it measures the shift distance of the short barcode.

- Flip parameter f: Since the fragments of DNA molecules are inserted with a random orientation into the nanochannels, we do not know their orientations when comparing them. In our model, the flip parameter f works on the shorter barcode B_2 indicating if it is in the 'forward' direction (f = 1) or not (f = 2). Figure 5.1 illustrates how C is calculated.

Compared with an earlier method for merging intact circular barcodes (Appendix A), a featured difference of our barcodes comparison method is the range of shift parameter d. Shift distance in circular barcode comparison is always positive, however, as for our sliding scheme the same parameter ranges from positive numbers to negative numbers. Thus for a circular comparison, there are $2N_1$ valid alignments in total for a given pair of f and d, while for a comparison based on our method there are a total of $2(N_1+N_2-1)$.

Let us now express our similarity score mathematically, also taking bit-weights into account. Barcode 1 $B_1(x)$ and barcode 2 $B_2(x;f)$ are at the pixel positions $x = 1, ...N_i$ (i = 1,2), same for bit-weights $W_1(x)$ and $W_2(x;f)$ for the two barcodes. The bit-weighted

correlation coefficient is defined by:

$$C(d,f) = \frac{1}{n(d;f)-1} \sum_{x=1}^{n(d;f)} A(x, x+d; f) \tag{5.1}$$

where the effective size $n(d;f)$ is the number of overlapping pixels of the barcodes at a flip f and shift d. It is given by:

$$n(d;f) = \sum_{x=1}^{N} W_1(x) W_2(x+d;f) \tag{5.2}$$

Also(f dependencies are left implicit below):

$$A(x, x+d) = \frac{[W_1(x)W_2(x+d)B_1(x) - \bar{I}_1(d)]}{\sigma_1(d)} \times \frac{[W_1(x)W_2(x+d)B_2(x+d) - \bar{I}_2(d)]}{\sigma_2(d)} \tag{5.3}$$

where

$$\bar{I}_i(d) = \frac{1}{n(d)} \sum_{x=1}^{n(d)} W_1(x)W_2(x+d)B_i(x) \tag{5.4}$$

and

$$\sigma_i^2(d) = [\frac{1}{n(d)} - 1] \sum_{x=1}^{n(d)} [W_1(x)W_2(x+d)B_i(x) - \bar{I}_i(d)]^2 \tag{5.5}$$

are the sample estimate of the mean and the (unbiased) sample estimate for the standard deviation for the barcode i (i=1,2).

In Sec.4.2.3 we introduced the bit-weighting scheme. According to the equations 5.1 to 5.5, the bit-weights act as a filter: a bit-weight value represents whether the corresponding pixel is included or not when computing the Pearson correlation coefficients, that is, for a given pair of f and d, any pair of values in B_1 and B_2 which are associated with a weight of 0 in either W_1 or W_2 would be excluded.

Using the two parameters d and f, a Pearson correlation coefficient of two comparing barcodes $C(d,f)$ measures the cross correlation of the barcodes by sliding the shorter barcode B_2 across longer one B_1. Note that expected value ranges from -1 to 1. A value of 0 means the data is uncorrelated and a value of 1 means two barcodes match perfectly.

Now we define the maximum correlation coefficient:

$$\hat{C} = max\{C(f,d)\} \tag{5.6}$$

This equation has an expected value > 0. f and d parameters are denoted by $\hat{f}$ and $\hat{d}$ at the $\hat{C}$ value.

Note that unlike noisified theory barcodes or experimental barcodes, theory barcodes do not have the "blurry ends", which means that their associated bit-weights equal to 1 for all positions in the barcode.

5.3 Threshold of the overlapped length

The comparison of two barcodes requires that the barcodes have a same value of overlapped lengths. However, this causes a major issue, as for how can we define the minimum length of an overlapped region? When the overlapped regions decreases, the results of the pairwise comparison becomes less reliable. Additionally, the lengths of two barcodes also affect this threshold length.

To identify the relationship between the threshold length and the barcode length, we introduce a method for turning Pearson correlation coefficients into a probabilistic framework defined by a p-value. This p-value framework indicates the significance of the matches based on the length of the overlapped region of the comparing pair. The p value is defined as

$$p - value = \int_{\hat{C}}^{\infty} \phi(\hat{C}')d\hat{C}' \tag{5.7}$$

where $\phi(\hat{C})$ is the probability density function for the similarity scores when matching a set of randomized barcodes. From the equation, we get that the p-values are distributed on [0,1]. A p-value smaller than a threshold p_{thre} means that there is a significant resemblance of the comparing barcodes[11].

To begin this process, we use the phase randomization method (Sec.4.3.1) to generate quantities of randomized barcodes and then divide these barcodes into two groups: one group we combine the noise factor (Sec.4.3.2) to get the noisified theory barcodes, while the other group is the originally randomized barcodes which serve as a reference group. Every time we cut the barcodes from two groups into fragments with a certain length, and compute the correlation coefficient values of the fragments and the theory.

For a large number of fragment barcodes, the correlation coefficient values are expected to follow the maximum match scores' distribution defined in the paper [4] from the Ambjörnsson group and the results are shown in Figure 5.2.

The experimental barcodes gathered from the Westerlund group have a mean length of 300kbp. After discussing with my supervisor, we decide to use 60kbp as the threshold-overlapped reagion for the experimental barcodes. The reason is that, comparing to the mean length 300kbp, 60kbp is a good boundary for both discarding the very short barcodes and keeping sufficient amounts of data.

For each length of the overlapped region we test, we set p_{thre} = 0.01 and then use inverse cumulative distribution function to calculate the accordant PCC value. Taking barcodes with 60kbp-overlapped region as an example; we apply p_{thre} = 0.01 to those

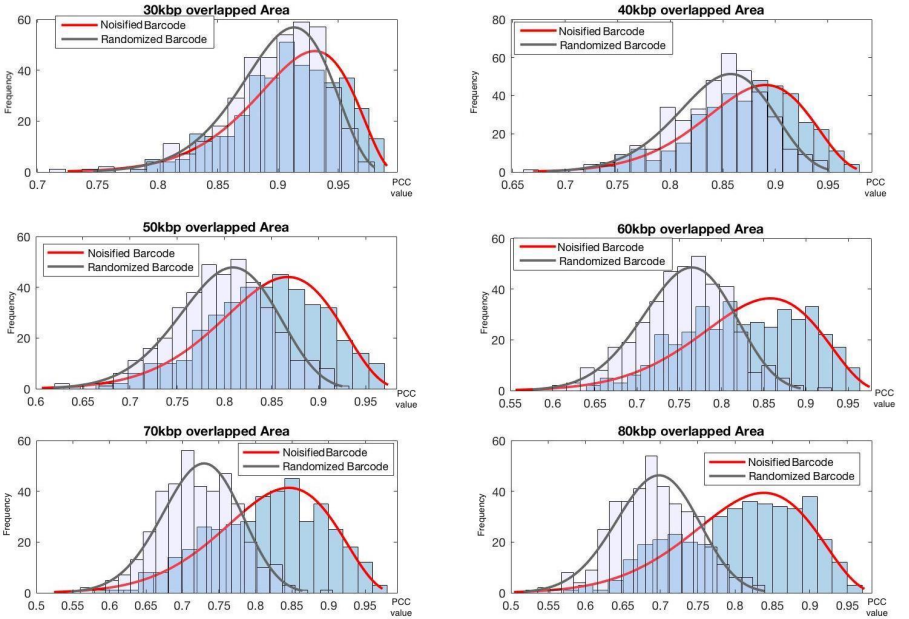


Figure 5.2: Histogram of PCC values obtained by comparing the two types of barcodes (noisified theory barcodes and randomized barcodes) with the theory of different overlapped lengths. The x-axis represents PCC values and the y-axis represents the frequency. These figures help us to decide a threshold for overlapped length when comparing two barcodes (See Sec 5.3). We finally use 60kbp as the threshold for both discarding the very short barcodes and keeping sufficient amounts of data.

randomized barcodes and calculate the accordant PCC value to be 0.8738 (Table 5.1). The result indicates that when the overlapped length is set 60kbp, the probability of a randomized barcode has a similarity score (compared with the theory) that is greater than 0.8738 is approximately 0.01. Under these conditions, when we only keep the barcodes with similarity scores greater or equal to 0.8738, it is less probable that the barcode is a randomized one. This is quite meaningful for the later application on the experimental barcodes. As the experimental barcodes are much more unpredictable, the p-value scheme could help us filter out unrelated barcodes in the comparison process.

The same procedures are carried out to barcodes with different overlapped lengths. Since it is computationally demanding to apply the calculation for every single length, we compute the p-values for every 50 pixels in the range. Following this treatment, we interpolate the data vector to get a function of the two variables: p-value and the overlapped length. We name it as p-value function. This function works as a threshold function and is later used after the calculation of similarity scores of all the possible pairs. When each pair is at its best match position, those with PCC values lower than the threshold

Table 5.1: Inverse of the cumulative probability (60kbp)

$P(X \geq x)$	x
1-p_{thre} = 0.99	0.8738

An example calculation result after using inverse cumulative distribution function. P represents the probability, X represents the variable - the PCC value and x is the x-value. And p_{thre} = 0.01 is applied in our case. This result indicates that when the threshold of overlapped lengths is set to 60kbp, the probability of a randomized barcode has a similarity score (compared with the theory) that is greater than 0.8738 is approximately 0.01.

at the corresponding overlapped length will be discarded. The full p-value function used throughout the rest of this study is provided in Figure 5.3.

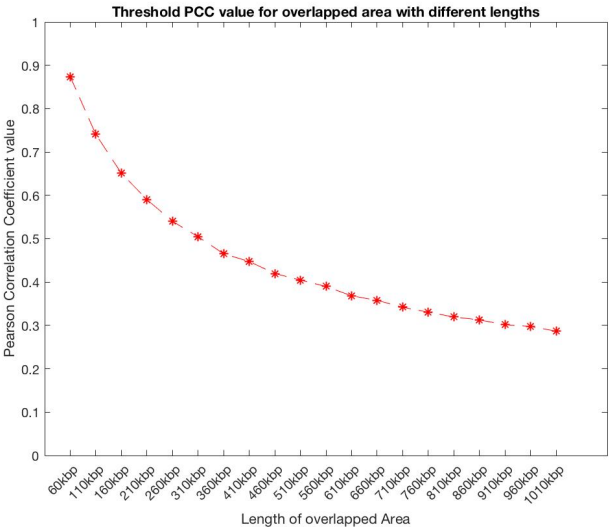


Figure 5.3: Pearson correlation coefficient corresponding to the p-value threshold = 0.01 as a function of the overlapped length of the barcodes. This p-value function serves as a filter so that we only keep the pairs with PCC values higher than their thresholds within the corresponding overlapped length.

Chapter 6

Assembled DNA barcode

There are many factors that might influence individual experiments. One of the main influencing factors on the experimental results is molecular fluctuations; random changes in molecular conformation or concentration. Along with the thermal motion, background noise, experimental error of the DNA molecule staining process, all this stochasticity might result in an unpredictable difference in kymographs and barcodes. Thus, we introduce our method for aligning the multiple DNA barcodes from the same DNA sequence to "assembled DNA barcode". The assembling method used here is an extension of the previous method for averaging intact plasmid barcodes (consensus barcodes). Following on from this, let us introduce our Hierarchical Clustering Matrix (HCM) method.

6.1 Assemble the barcodes

Previous ways for consensus barcode generation, which mostly focuses on plasmid barcodes, require the pair which is being compared to be of the same length[8]. As the object of our study is the genomic bacteria barcode, we made an improvement to our method which meant that the two barcodes could be compared in any linear position. This is to say, they could be totally overlapped, partially overlapped, or even not overlapped at all. The advantage of the design is that the two barcodes could be fully compared in most circumstances.

Now we introduce the overview of our method for generating assembled barcodes through a hierarchical clustering approach. Prior to merging the barcodes, we consider a pool containing M barcodes, $B_a(x)$ and the associated bit weights $W_a(x)$ (a = 1,..., M) that we hope to merge into an assembled barcode. We calculate the Pearson correlation coefficient values for every two barcodes. After replacing the most similar pair with an averaged weight value, there are $M - 1$ barcodes left. After this, we repeat the same process many times, until there is either only one barcode left or no more similarity scores which are higher than the threshold values. The resulting one or several barcodes are the

assembled barcodes of the pool. A detailed explanation of the implementation of this method is below.

Let us now make the illustration more precise in mathematical terms. Assume that we have M barcodes, $B_a(x)$ and the associated bit weights $W_a(x)$ (a = 1,..., M) that we hope to merge into an assembled barcode. Note that in order to reduce the effects of end regions of the barcodes, we use bit-weights to remove the end regions. And the procedure is described as following:

- Compute the similarity score $\hat{C}_{ij}$ by applying equations 5.1-5.6 for every possible pair B_i and B_j.

- Filter all the barcode combinations and only keep those whose similarity scores are higher than the threshold PCC value at their overlapped length. The step is achieved by inputting each $\hat{C}_{ij}$ to the interpolated p-value function (Figure 5.3). If a pair fails to pass the threshold function (Sec 5.3), we would discard this pair for further comparison. Once there is not a pair with a higher similarity score than its threshold, the assembling iteration procedure will stop. Note that here we only consider the pair rather than a specific barcode. Either barcode of the discarded pair can make up another pair with other barcodes.

- Choose the most similar barcodes. In most cases, there are always a few pairs passing the threshold filtering. However, we only merge one pair for each iteration, and thus the next challenge will be choosing the most similar barcodes. Assume that all the pairs we have now have passed the previous threshold test. Consider the couple barcodes B_i, B_j with overlapped length N_i and their Pearson correlation coefficients $\hat{C}_{ij}$. Additionally, the pair has a threshold Pearson correlation coefficient value C'_{ij} obtained from the p-value function. Then we calculate a 'q-value' for the pair according to

$$q = \frac{\hat{C}_{ij} - C'_{ij}}{1 - C'_{ij}} \tag{6.1}$$

  'q-value' in the Equation 6.1 helps to estimate the levels that the correlation coefficient value of a pair is higher than the corresponding threshold. In this way, the pair with highest q-value is the most similar barcodes and they will be merged in the following process.

- Assemble the most similar barcodes at their best alignment position. Here we introduce a counter $\omega_i(x)$ which represents the number of 'non-zero' contributions at a given pixel x of B_i. When merging two most similar barcodes, a counter can be written as

$$\omega_{ij}(x) = \omega'_i(x) + \omega'_j(x) \tag{6.2}$$

where $\omega_i'(x)$, $\omega_j'(x)$ are the bit weights for the shifted (possibly flipped) version of the bit weight of the counter $\omega_i(x)$, $\omega_j(x)$. From the equation above we can easily find that the bit weights for the merged barcode $W_{ij}(x)$ have zeros only at where both original counters have 0s. The merged barcode and associated bit weight are denoted as

$$W_{ij}(x) = \begin{cases} 1, & \text{if } \omega_{ij}(x) > 0. \\ 0, & \text{if } \omega_{ij}(x) = 0. \end{cases} \tag{6.3}$$

$$B_{ij}(x) = W_{ij}(x)\frac{\omega_i'(x)B_i'(x) + \omega_j'(x)B_j'(x)}{\omega_{ij}(x)} \tag{6.4}$$

where $B_i'(x)$, $B_j'(x)$ are the shifted (possibly flipped) version of $B_i(x)$, $B_j(x)$. The factor in Equation 6.4 guarantees that we only merge the barcodes at pixels where there are at least one non-zero contribution(at pixels where $W_{ij} \neq 0$).

For the barcodes from the initial pool, the counters are the same as the initial bit weights $\omega_i(x) = W_i(x)$. While for barcodes produced by the merged barcodes, the counters are no longer limited to 0 and 1, instead are the sums of the aligned weights for the pair of barcodes. With this weighted-merging method, the order of the barcode-merging does not have any impact on the final result as the assembled barcodes are the average values with non-zero bit weights at each pixel position of each aligned barcode.

Here we prepare two examples to explain our method more precisely. Two barcodes B_1 and B_2 and their length N_1 and N_2. Suppose the best alignment of this pair is given by Figure 6.1. Dashed lines divide the pair into three pairs depending on the overlapped regions: left region, overlapped region and right region. Unlike the overlapped region, the other two parts are not involved in the merging procedure. Thus, after we merge the overlapped region, we need to stitch three parts together by the original order, resulting in the final merged barcode. Fig 6.1 presents the case of fully overlapped barcodes of which the lengths of merged barcodes do not change. However, in some cases, the pair might be partially overlapped and the length of the merged barcodes will increase (Figure 6.2). This is quite a significant improvement in our method which could piece up fragments into longer pieces resulting in a nearly intact chromosomal DNA barcode.

The Figure 6.3 shows the assembling process of five fragmented noisified theory barcodes ($M = 5$). First, we chose the most similar barcodes amongst all of the combinations, and we merge B_3 and B_4 into $B_{(3)(4)}$, which replaces the 'parent' barcodes with an averaged barcode value and an averaged associated bit-weight. For the next iteration, there are only four barcodes in the pool. Until the final iteration, the similarity score of $B_{1(2((3)(4)))}$ and B_5 is not equal to or higher than the threshold, so the merging procedure has to be ceased and $B_{1(4((3)(5)))}$ is the final assembled barcode. However, this figure is more idealized than in practice with a large barcode pool, as in large barcode pools there is always

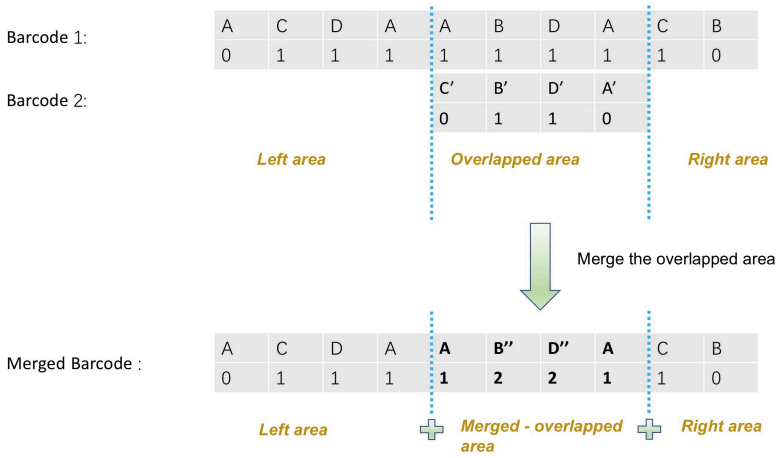


Figure 6.1: Schematic illustration of interior region merging. A, B, C, D represent the intensity levels ($B_i(x)$) and 0, 1, 2, 3 represent the bit-weights ($\omega_i(x)$) in Equations 6.2-6.4. When merging two barcodes of different lengths, at the best alignment position, we divide the overlapping barcode into three parts: left region, overlapped region and right reagion. We only merge the overlapped region and leave the other two parts aside. The newly merged barcode is made up by two original left & right regions and the merged overlapped-region.

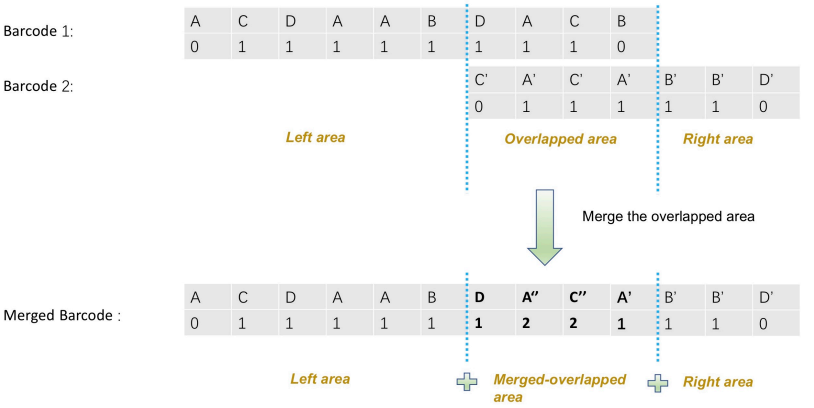


Figure 6.2: Schematic illustration of end region merging. Comparing with Figure 6.1, this figure shows an example of another situation. At the best alignment position, the right region is missing so the final merged barcode is made of the original left region and the merged overlapped-region. The same principle applies to the situation with a missing left region.

more than one assembled barcode left when the merging procedure has stopped. The procedure explained above will be repeated until there is either one barcode left or no more pairs have the positive q-value.

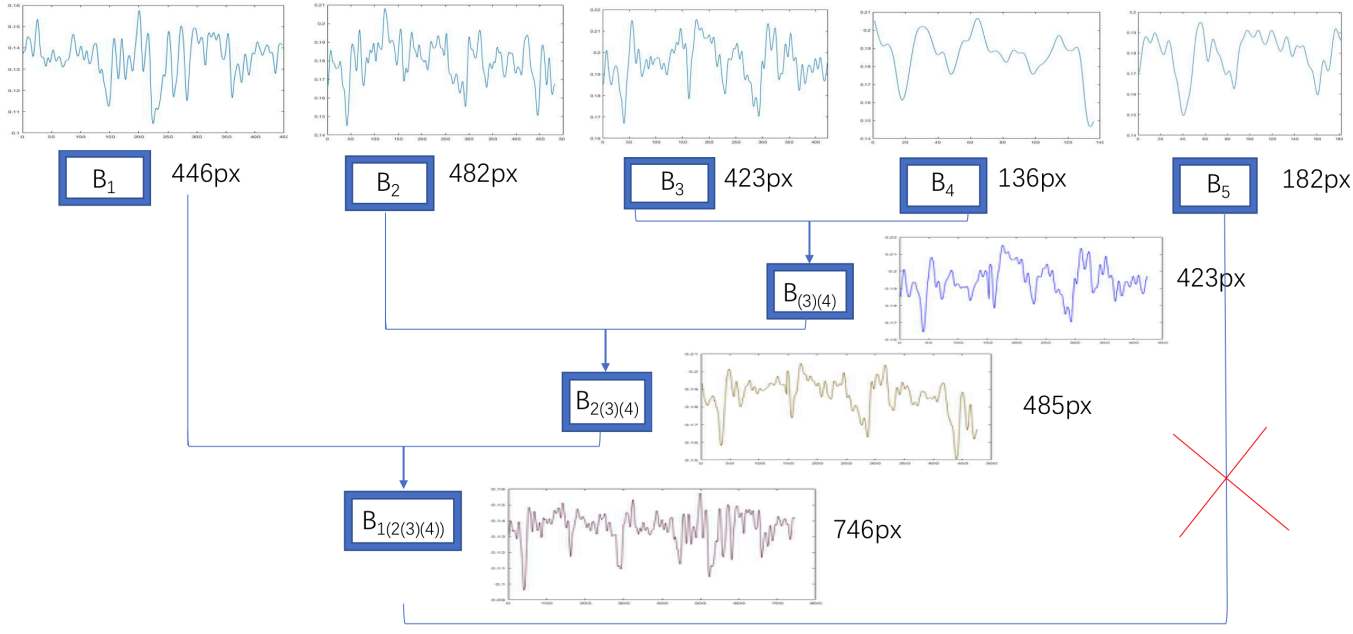


Figure 6.3: Schematic illustration of the method for generating assembled barcodes. B_1-B_5 are five fragmented noisified theory barcodes where four are similar and one is not. For each barcode in the process, the corresponding length is given next to its image. By replacing the similar two barcodes with the weighted average values, four barcodes are merged and B_5 is left aside. Relative orientation and the cutting position of the two barcodes are determined by $\hat{f}$ and $\hat{d}$ in Equation5.6. After analyzing the similarity score (maximum Pearson correlation coefficient $\hat{C}$), B_5 does not belong to this group. $B_{1(2((3)(4)))}$ is the assembled barcode of the other four barcodes. Noted that the final assembled barcode $B_{1(2((3)(4)))}$ is much longer than any original barcodes. With this method, we could finally get a fully assembled barcode by repeating this process.

6.2 Practical implementation of the Hierarchical Clustering Matrix (HCM) method

In this section, we will illustrate how we implement our HCM method in codes. Because our work focuses on large-sized barcodes, which are usually hundreds or thousands of pixels in length, a large amount of calculations is required and hinder the speed of the analysis. A major advantage of the HCM method is that it offers an effective way for parameter manipulation.

The matrices in our method are used to store five types of parameters used in assembly process: Pearson correlation coefficients ($\hat{C}$), shift values (d), orientation values (f), overlapped lengths of the comparing barcodes and p-values. All kinds of the parameter matrices are functionally similar despite that the parameter types stored in the matrices are different. The HCM method is designed to provide a clear and precise way for handling large quantities of information. Figure 6.4 gives an example illustration of the Pearson correlation coefficient matrix.

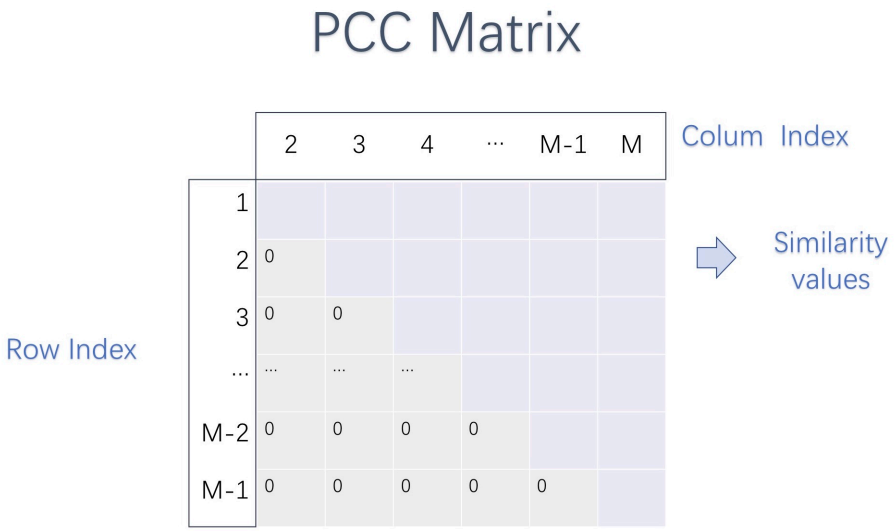


Figure 6.4: Schematic diagram of a Pearson correlation coefficient matrix. Assume that we have M indexed barcodes, the parameter matrix is defined as a $(M-1)*(M-1)$ matrix. The gray region where entries are below the main diagonal is zero-region. Parameters only occupy the upper triangle matrix including the main diagonal. This matrix can store the PCC values of all the possible combinations of M barcodes. The row index represents the barcode with a smaller index of the pair while the column index represents the barcode with the larger index.

The principle of five types of matrices is the same, herein we first give a general ex-

planation of a parameter matrix. Consider M barcodes and associated bit weights, and they are indexed first as B_i and M_i (i=1,2,3,...,M). It is noted that the indexing cannot be changed during the calculation. The parameter matrix is designed as a square matrix with $M-1$ in length side. The row index represents barcodes B_i (i=1,2,3,..., M-1), which have the smaller index of the pair, and the column index represents barcodes B_i (i=2,3,4,..., M) which have the larger index of the pair. So the parameter matrix $P(i,j)$ provides the parameter value of the pair that row i and column j represent.

Note that here the row index or column index of the parameter matrix is different from the barcode index. For instance, see the third matrix in Figure 6.5. Since B_4 has been replaced and deleted, $P(3,3)$ denotes the parameter value of B_3 and B_5. In this thesis $P(i,j)$ denote the parameter value of barcodes that row and column represent not the matrix index. Additionally, it is important to notice that as the matrix is symmetrical, $P(i,j)$ and $P(j,i)(i \neq j)$ have the same meaning. Thus, the entries under the main diagonal are set to be 0s. The upper triangle region and the main diagonal are used for value storage.

Let us now introduce how we implement our algorithm with the matrices.

- Compute the similarity scores for each pair. $\hat{C}_{i,j}$ finds the best alignment of the pair B_i and B_j and then it is recorded at $PCC(i,j)$ of the Pearson correlation coefficient matrix. Accordingly, at the best alignment position, the corresponding $\hat{d}$, $\hat{f}$, the overlapped length are recorded in the same position (i,j) of shift value matrix, orientation Matrix and the overlapped length matrix.

- Apply p-value function for PCC threshold filtering and find the pair with maximum q-value. We apply the p-value function for each pair with their $\hat{C}_{i,j}$ and overlapped lengths and fill the results in the position (i,j) of the p-value matrix. On completion of the p-value Matrix, the process of finding the most similar pair is carried out. According to Equation 6.1, the pair with maximum q-value would be the most similar barcodes in the current iteration. The position of the pair in p-value matrix is denoted as (ki,kj).

- Find the corresponding parameters d, f, barcode indexes and merge the pair. With the matrix index (ki,kj), we could locate and find other parameters needed for the assembling process.

- Replace the barcode and bit-weight with the small index of the parent pair with the new merged barcode and bit-weight. Meanwhile, the barcode and bit-weight with the larger index are removed from the barcode pool. Likewise, we also need to apply the deletion to the rows and columns related in all five parameter matrices.

By now we have illustrated a whole calculation process of an iteration. For the next iteration, we only need to update the all parameter values of pairs including the newly merged barcode from last round and the process is same as described above.

Here we present two examples to better explain the replacement process of the HCM method. Because we have demonstrated that the principle of all five matrices is the same,

we use parameter matrices instead of specific types of the parameter matrix. Assume that we have five barcodes (M = 5) and the parameter matrix in Figure 6.5. x_i represents the parameter value. Suppose that we want to merge B_2 and B_4. Based on the idea introduced above, after assembling the pair into a merged barcode, we need to delete the rows and columns related to B_4. Thus, the fourth row and the third column are removed, and the parameter matrix reduces from 4*4 matrix to 3*3 matrix. For the next round, we just need to recalculate x_1,x_5 and x_7 as only these values are related to the merged barcode B_2 from the last iteration. Nevertheless, if the deleted barcode has the largest index value in the pool, we can only find the corresponding column, not the row. As shown in Figure 6.6, if B_3 and B_5 need to be merged, the fourth column is deleted but there is no corresponding row. However, after the column has been removed, it is obvious to find that the last row is occupied by 0s which make no difference for the following analysis. Hence the last row can be deleted as well so that the resulting matrix is still square-shaped. Until no more p-values are higher than their thresholds or one single barcode is left, the merging procedure stops and the remained barcodes are the assembled barcodes we want.

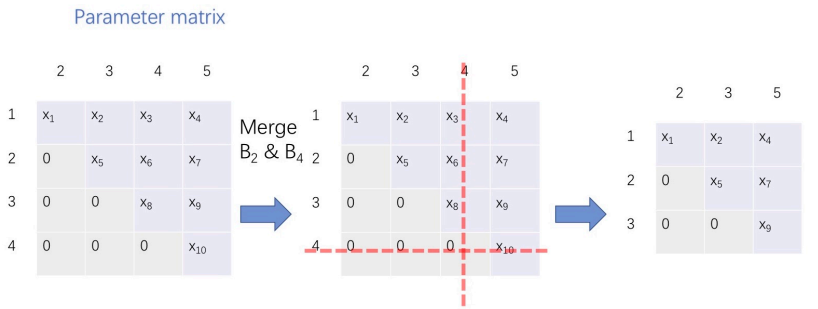


Figure 6.5: Schematic illustration of the replacement process of the HCM method. We use five barcodes (M = 5) as an example. Assume that we want to merge B_2 and B_4, both the row and the column containing B_4's information is deleted after the merging procedure. Then the size of the matrix will decrease by one row and one column. The newly merged barcode will replace the barcode with the smaller index in the pair. In our example, the merged barcode will become the new B_2. For the next iteration, we only need to calculate the parameters related to the new-merged barcode B_2 and then update its row and column.

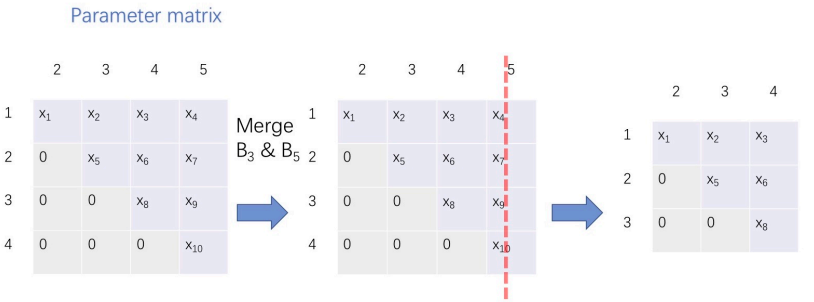


*Figure 6.6: Schematic illustration of one step in the replacement process of the HCM method. We use 5 barcodes (M = 5) as an example. Assume that we want to merge B_3 and B_5, we only need to delete the last column as there are no rows correspond to B_5. Following this step, it is obvious that the last row has all 0 entries so it can be removed as well, resulting in a 3*3 matrix. Likewise, for the next iteration, we only need to update all the parameters related to the new merged barcode B_3.*

Chapter 7

Results and Discussions

7.1 Noisified theory barcodes

We applied our algorithm to the noisified theory barcodes. The procedure involves examining different numbers of copied noisified theory barcodes. As before, we have defined the minimum length of the overlapped region to be 60 kbp, and any compared pairs with an overlapped length shorter than the threshold are discarded. Table 7.1 provides the results obtained from our analysis. We completed many tests on different lengths and the table only shows five representative results for each condition. Note that we only take the largest merged barcodes as the result in Table 7.1.

Notice that with more noisified theory barcodes, we are more to likely get a nearly-complete assembled barcode and this observation is indicated by the coverage rate, which represents the coverage rate of the longest merged barcodes to the theory barcode. In contrast, we find that with fewer noisified theory barcodes, we are more likely to get several shorter assembled barcodes matching the different region of the theory barcode. Besides, if there are sufficient noisified theory barcodes, not only the similarity score PCC values but the merging rate is a little higher. In contrast, the positional accuracy, which represents the accuracy rate of the original fragments ending up at the right position in the final assembled barcodes, is less sensitive to the number of fragments provided at the beginning. Note that before the merging process, we record the best alignment position by comparing original noisified barcodes with the theory barcode. By 'right position', here we mean that after we compare the merged barcode with the theory, these noisified barcodes are merged in the same position as recorded in the beginning. This shows that there are more fragments being involved with the assembled barcode generation process when providing enough noisified theory barcodes. And the coverage rate, which indicates the coverage ratio of the largest merged barcode to the theory barcode at the best alignment position, increases a lot when there are more noisified theory barcodes.

Figure 7.1 presents an example of the comparison of an assembled barcode (from 10

noisified theory barcodes) and the theory barcode. Since the barcodes are over 8000 pixels in length, the figure is divided into several subfigures. Additionally, we append a figure of the comparison of original fragments and the theory barcodes in Appendix B.

No. of noisified barcodes : 10

Exp. No.	Total No. of fragments	No. of merging barcodes	Merging Rate	Length of overlapped Area (pixel)	Coverage Rate	PCC(compared with the theory barcode)	Positional accuracy
1	137	109	79.56%	8307	99.38%	0.9526	100%
2	135	109	80.74%	8355	99.95%	0.9424	91.30%
3	124	97	78.23%	8251	98.71%	0.8455	93.81%
4	130	112	86.15%	4734	56.63%	0.9631	77.68%
5	130	109	83.85%	8248	98.67%	0.9191	98.16%

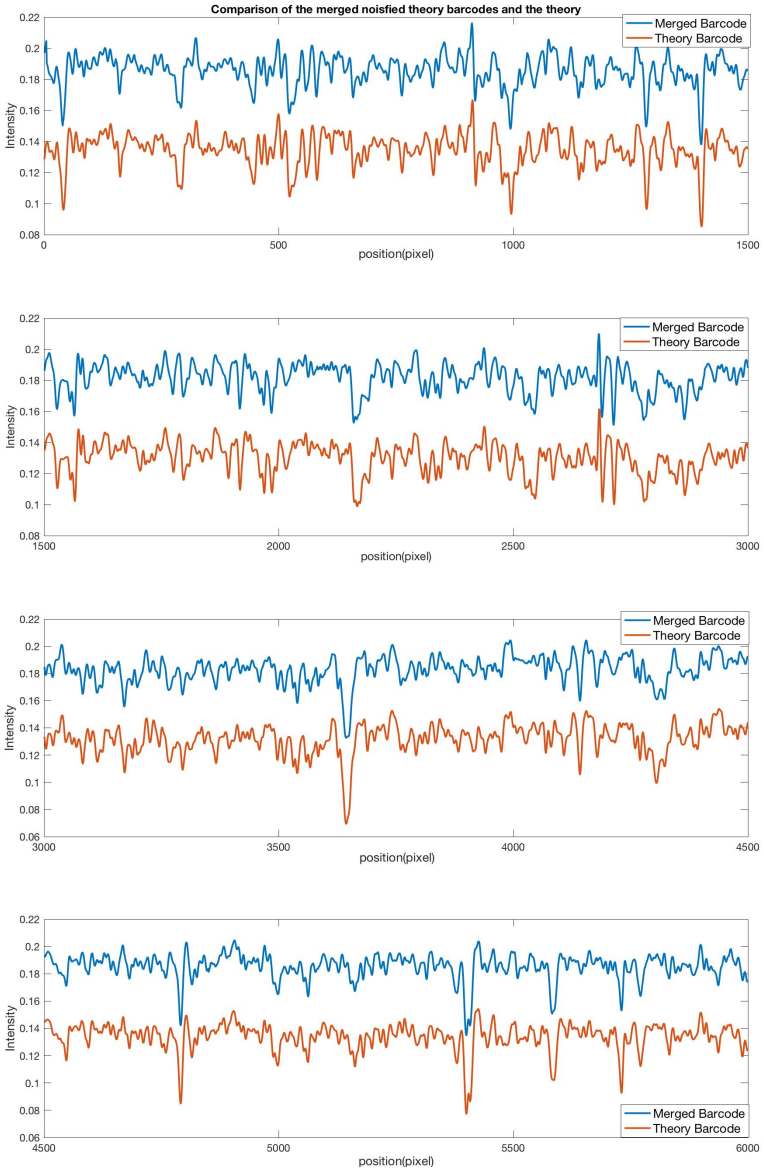
No. of noisified barcodes : 7

Exp. No.	Total No. of fragments	No. of merging barcodes	Merging Rate	Length of overlapped Area (pixel)	Coverage Rate	PCC(compared with the theory barcode)	Positional accuracy
1	91	72	79.12%	6350	75.97%	0.9357	91.67%
2	85	74	87.06%	6877	82.27%	0.9424	100%
3	93	71	76.34%	8088	96.76%	0.8409	100%
4	99	70	70.71%	4011	47.98%	0.9341	94.29%
5	81	66	81.48%	8323	99.57%	0.9389	97.00%

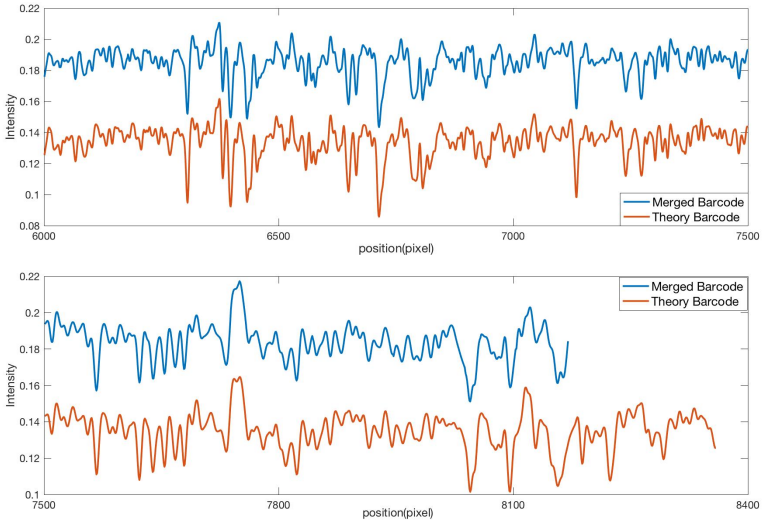
No. of noisified barcodes : 4

Exp. No.	Total No. of fragments	No. of merging barcodes	Merging Rate	Length of overlapped Area (pixel)	Coverage Rate	PCC(compared with the theory barcode)	Positional accuracy
1	54	38	70.37%	4223	50.52%	0.8179	97.37%
2	52	38	73.08%	3726	44.57%	0.8600	84.21%
3	46	31	67.39%	3636	43.50%	0.8670	87.10%
4	60	46	76.67%	3876	46.37%	0.6622	86.96%
5	51	38	74.51%	3347	40.04%	0.8939	100%

Table 7.1: These charts show some of the main parameters of DNA assembly of noisified theory barcodes vary under different situations. A merged rate is the ratio of the number of merged barcodes to the total number of the barcodes. The positional accuracy is the ratio of the number of fragments that are assembled at the right position (See Sec 7.1) and the total number of the fragments. For each experiment, there are always several assembled barcodes after the merging process. The coverage rate and PCC value represent the corresponding parameter values of the largest piece of the resulting assembled barcodes. Coverage rate is the ratio of the overlapped region to the theory barcode when an assembled barcode and the theory barcode were at the best alignment position.



(a) First four panels of a continued Figure 7.1.



(b) Continuation of Figure 7.1.

Figure 7.1: An example of a comparison of a merged noisified theory barcode and the theory barcode. The result is from the first experiment providing 10 noisified theory barcodes in Table 7.1. The merged barcode could match the theory barcode within first 8307 positions (pixel) and the similarity score is 0.9526. The coverage rate of the merged noisified theory barcode is 99.38% and the positional accuracy of the fragments is 100%.

7.2 Experimental barcodes

After testing with our noisified theory barcodes, we applied our method to the data obtained from the Westerlund group. The kymographs are movies captured with 1.6x Optovar (100x total magnification). In short, we find that the merging process is not as successful as that of noisified theory barcodes.

We first use the codes written by the Ambjörnsson group to convert all the kymographs to barcodes and then implemented the method described in Chapter 6. We apply 10 timeframes to the 22 barcodes from the first experiment ('day 1 experiment') to get time-averaged individual barcodes. Following this step, there are 18 barcodes left and 4 are discarded because they do not have enough time frames.

On the completion of our Hierarchical Clustering Matrix method, there are 9 barcodes involved in the merging process and the merging rate is 50%. After the merging process, there are 3 merged barcodes eventually and the PCC values (compared with the theory barcodes) are 0.6151, 0.5588 and 0.7424, respectively. The coverage rate for the largest merged barcode is only 2.45% and the positional accuracy is 33.33%. That is, 3 barcodes (9 barcodes in total) are merged in the right position. All quantities mentioned here have the same definition as in Sec. 7.1. Figure 7.2 shows the comparison of the assembled

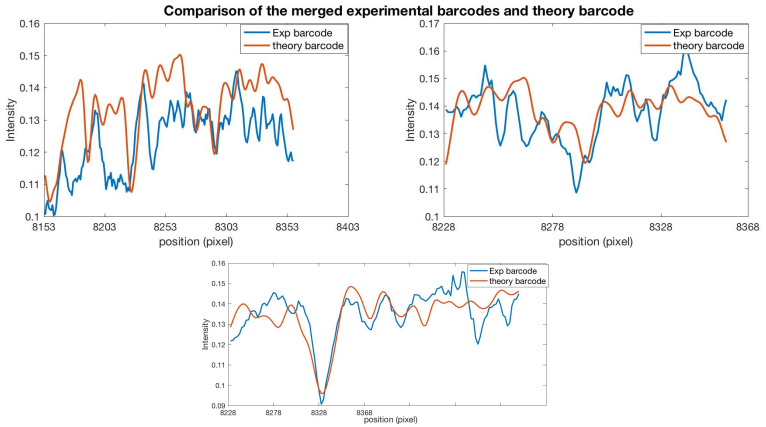


Figure 7.2: Comparison of merged experimental barcodes and the theory barcode generated from the first experimental data ('day 1 experiment'). There are 9 experimental barcodes involved in the merging process (18 experimental barcodes in total) and there are three merged barcodes eventually. The upper left figure shows the merged barcode with 205 pixel-long overlapped region and the similarity score of this matching part is 0.6151; the upper right figure shows the second merged barcode with 130 pixel-long overlapped region and the similarity score of this matching part is 0.5588; the below figure shows the third merged barcodes with 128 pixel-long overlapped region and the similarity score of this matched part is 0.7424.

Similarity Scores of Merged barcodes from the 2rd experiment

Merged Barcode Index	PCC(compared with the theory barcode)	Length of overlapped Area (pixel)
1	0.5087	109
2	0.5413	126
4	0.7893	547
5	0.7931	625
10	0.4996	195
27	0.5753	322

Table 7.2: Similarity scores and the associated overlapped lengths of the merged barcodes from second experiment data.

barcodes and the theory of this experiment.

On the second attempt, we analyzed the data from the same species of the DNA molecules ('day 2 experiment') obtained from the Westerlund group. This time the fits of the experimental barcodes with the theory are slightly improved. Like last time, the kymographs were acquired in the same conditions and the only difference is the experiment date.

Firstly, we use 20 timeframes of the raw kymographs and then get 34 experimental barcodes left (43 barcodes at the beginning). We append a figure of the comparison of the original experimental barcode and the theory in Appendix C. The results of the assembling process are shown in Table 7.2 and Figure 7.3.

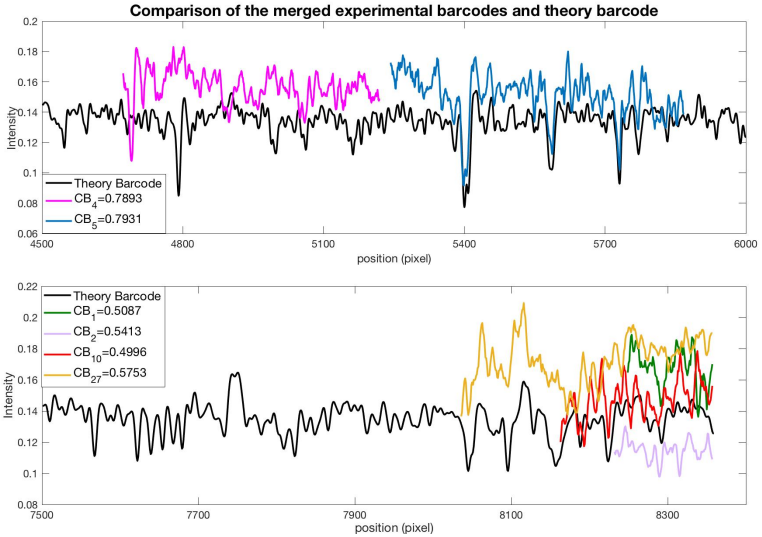


Figure 7.3: Comparison of the merged experimental barcodes and the theory from the second experiment ('day 2 experiment'). There are 23 experimental barcodes involving the assembling process and 6 assembled barcodes left. 10 barcodes are assembled in the right position and the positional accuracy is 43.48%. Note that we compare the original experimental barcodes with the theory barcode, then we record the best alignment positions respectively. The 'right position' means that for the barcodes involved with the merging process, The coverage rate of the largest assembled barcode to the theory is 7.48%.

There are 23 barcodes involved in the assembling process and the merging rate is 67.65%. From the data in 7.2, there are 6 merged barcodes eventually. The similarity scores ranged from 0.49 to 0.79 while the overlapped length are quite short. The coverage of the largest merged barcode (Barcode 5 in Table 7.2) is 7.48%. There are 10 barcodes merged in the right position and the positional accuracy is 43.48%.

When we check the match positions of the assembled barcodes to the theory, there were few clusters instead of a uniform distribution. Figure 7.3 presents the comparison of assembled experimental barcodes and the theory barcode. We can see that the assembled barcodes only gather in the position range [4500, 6000] px and [7500, 8400] px. In order to eliminate possible causes of our method, we calculate the match positions of each experimental barcode before the assembling process and found that the clusters remained, see Figure B.2a - B.2b. This also partially explains that our assembled barcodes cannot be matched with the theory uniformly.

We can see that from position 6000px to 7500px, there are original fragments but no merged barcodes. This fact shows that our method does have limitations in accuracy for the assembled barcode generation. In fact, we traced all the barcodes and compare their matching positions before and after the merging procedure and only 40% of the experimental barcodes ended up at right positions.

Now we are about to make some comments on our method and results. It is obvious that too few barcode fragments will not produce a reasonable result. For the noisified theory barcodes, once we have more original barcodes- which means that we have more fragments- the results of merging will improve accordingly. Especially if providing around 10 noisified theory barcodes, PCC value, coverage rate and the positional accuracy are relatively stable. Even though the results of the experiments are not as good as the noisified theory barcodes, as the second experiment has twice as many as DNA fragments as the first experiment, the parameters also increase correspondingly. According to our results, the total length of the fragment barcodes should be 4-5 times its original theoretical sequencing. In fact, the experimental data we got are few and the total length of the experimental barcodes is approximately twice the theory sequence. This insufficiency could be the part of the reason that there is a big difference in the noisified theory barcode result and the experimental barcode result.

Considering there are many more factors affecting the experimental barcodes, the resulting PCC values are not too low compared with those of noisified theory barcodes. Meanwhile, there are several aspects that we could improve. The p-value function is built based on the comparison results of our noisified theory barcodes and the theory. And our noisified theory barcodes are generated based on our assumption that the threshold of the correlation coefficient between the theory barcode and the experimental barcodes is 0.8 (See Sec.4.3.2). However, the experimental barcodes might be subject to more noise than assumed here and our method does not take account of the molecular fluctuation, so this threshold might be lower than 0.8. Assume that the actual threshold is 0.6, as shown in Figure 5.3, for max_{PCC} = 0.6 we would instead need a length threshold of approximately 200 kbp to make sure that we only get 1% false merging processes. But as we have a length threshold of 60 kbp now, there would be much more false merging processes.

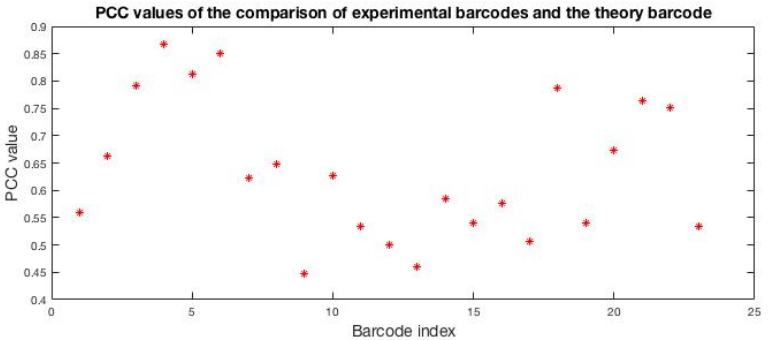


Figure 7.4: PCC values of comparison of the experimental barcodes (from 'day 2 experiments') and the theory barcode. The figure shows the results from 23 experimental barcodes: the mean of these PCC values is 0.6365 and the standard deviation is 0.1286.

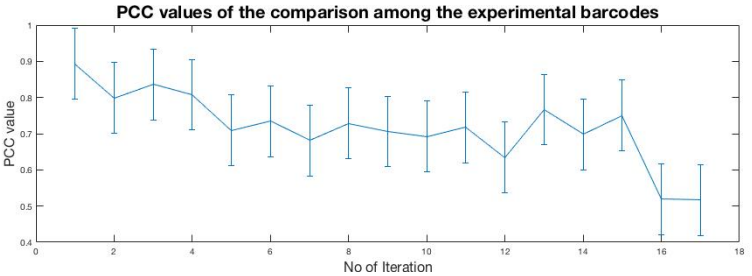


Figure 7.5: PCC values obtained from each iteration during the merging process for 'day 2 experiment'. The x-axis represents the iteration number; the y-axis represents the PCC value of the comparing two experimental barcodes for an iteration. The mean of these PCC values is 0.7170 and the standard deviation is 0.0976.

To prove our conjecture, we calculate the PCC values of the experimental barcodes ('day 2 experiment') and the theory barcode and Figure 7.4 shows the results. The mean of the PCC values is 0.6365 with a standard deviation 0.1286. Now we can conclude that 0.8 is too high for our experimental barcodes.

Additionally, we append a figure (Fig 7.5) reflecting the PCC values of the comparing experimental barcodes for each iteration during the merging process. This figure could be regarded as a direct measure of how much two or more experiments differ from each other.

Another main limitation of DNA assembly is that the conformational fluctuations might affect the barcodes. This motion could cause the molecules to stretch or bend when microscopy was taking images of them. So the raw kymographs acquired for one molecule at different time are not quite the same. Our method does not allow the stretch factors in the calculation which might cause inaccurate results.

7.3 Supplementary discussion

As discussed in Sec. 7.2, a higher threshold of the overlapped length could possibly be an improvement for the HCM method. To verify our speculation, in this section, we increase the threshold of the overlapped length from 60 kbp to 210 kbp, and re-run the merging process for the experimental barcodes from 'day 2 experiment'. The similarity scores of resulting merged experimental barcodes are shown in Table 7.3, the comparison of the merged barcodes and the theory after changing the threshold length is shown in Fig 7.6.

After changing the threshold of overlapped length to 210 kbp, there are still 6 merged barcodes, however, many parameters have slightly changed: the number of barcodes involved in the merging process increases from 23 to 25; the coverage rate of the largest merged barcodes increases from 7.48% to 8.88%; and the positional accuracy increases from 43.48% to 63.16%. Figure 7.7 shows the PCC values of comparing each original experimental barcodes with the theory.

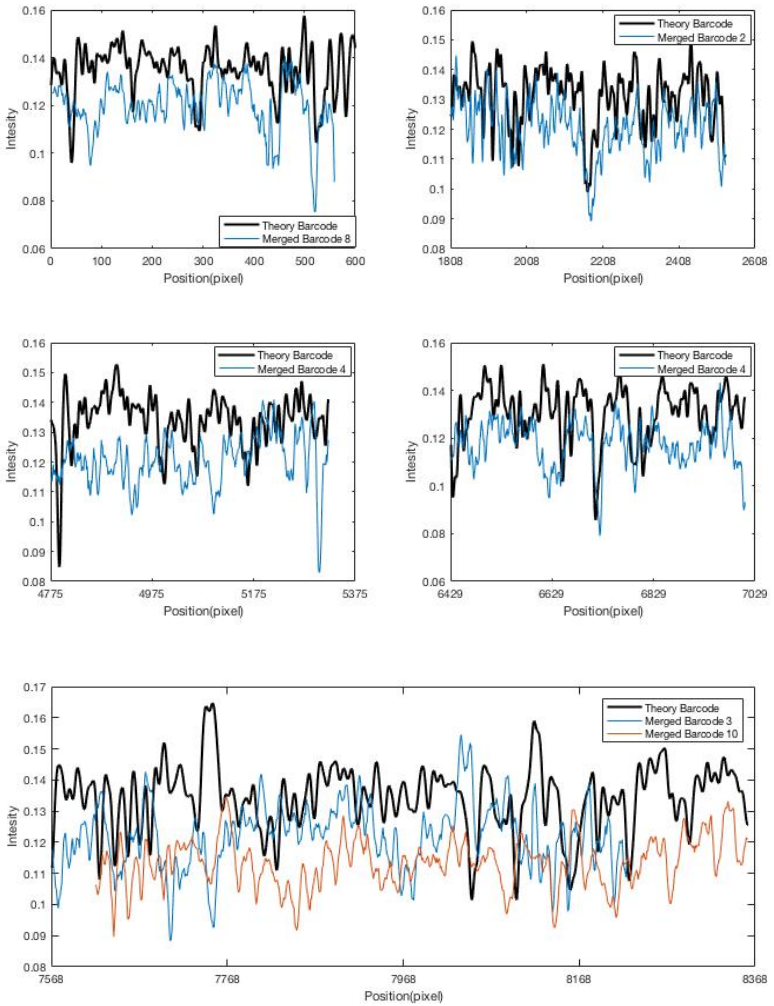


Figure 7.6: After changing the threshold of overlapped length from 60 kbp to 210 kbp, the comparison of the merged experimental barcodes ('day 2 experiment') and the theory barcode.

Similarity Scores of Merged barcodes from the 2rd experiment

Merged Barcode Index	PCC(compared with the theory barcode)	Length of overlapped Area (pixel)
1	0.6219	580
2	0.4658	722
3	0.5082	657
4	0.7893	547
8	0.4296	559
10	0.3655	742

Table 7.3: After changing the threshold of overlapped length to 210 kbp, we implement the merging process for the experimental barcodes from 'day 2 Experiment' and this chart shows the similarity scores and the associated overlapped lengths of the merged experimental barcodes when comparing with the theory barcode.

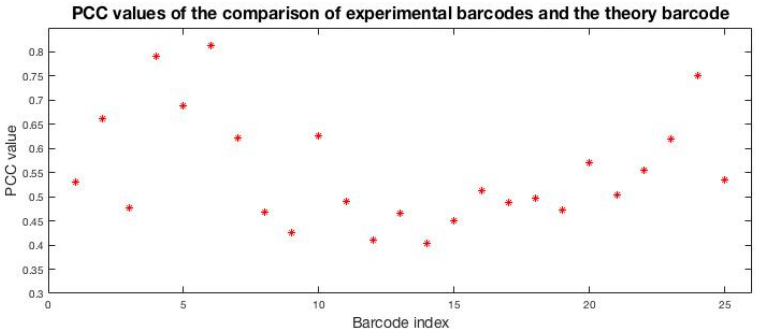


Figure 7.7: PCC values of comparison of the experimental barcodes ('day 2 experiment') and the theory after changing the threshold of overlapped length to 210 kbp. The figure shows the results from 25 experimental barcodes: the mean PCC value is 0.5530 and the standard deviation is 0.1158.

In addition, we also append a figure (Fig 7.8) reflecting the PCC values of the comparing experimental barcodes for each iteration during the merging process.

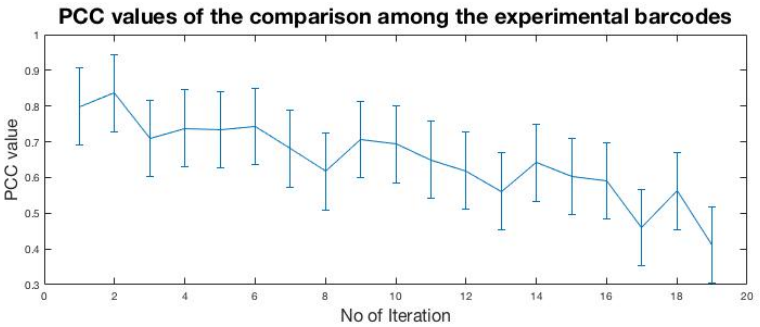


Figure 7.8: PCC values obtained from each iteration during the merging process for 'day 2 experiment' after changing the threshold of overlapped length to 210 kbp. The x-axis represents the iteration number; the y-axis represents the PCC value of the comparing two experimental barcodes for an iteration. The mean value of these PCC values is 0.6501 with a standard deviation 0.1071.

Chapter 8

Conclusion and Outlook

The method presented in this thesis is able to partially align the noisified theory barcodes to each other to recreate the original barcode for the intact chromosome. This can work on large pieces with a mean length of 300 kbp. However, it still remains a challenge to assemble experimental barcodes correctly.

One main reason is that the experimental barcodes have more uncontrolled factors. For example, a single experimental barcode might have a different stretch level at different positions and at a different time. Further research should focus on developing methods how to find the optimal stretch degree of the best alignment.

Another problem is that we cannot guarantee that the preparation work of each experiment is identical. A slight difference in the excitation light of the microscope and the background might cause a change in the kymographs, resulting that assembled barcodes would be aligned in the wrong position. Thus, considerably more work will need to be done to handle the error tolerance while taking care of the accuracy of the algorithm.

In Sec. 4.2.2 we set the threshold of the Pearson correlation coefficient value of an experimental barcode and a theory barcode to be 0.8 and this value is used also for experimental barcodes. Deviations from this value would cause our method to fail.

Besides, our method does not allow barcode stretching when comparing, which will always influence the resulting PCC values. Possibly, 0.8 is too high for our data. From Figure 5.3, we could see that a lower PCC corresponding to a higher threshold of the overlapped length.

Two possible extensions of this implementations are 1) to increase the length threshold and wait for longer DNA barcodes obtained from the lab; 2) to wait for more reproducible experiments where PCC values of experimental barcodes and the theory could be higher; 3) to allow the stretching factors in the algorithm.

This study has left many unsolved problems for future studies. If the development of the method is successful, we could successfully assemble the chromosomal DNA barcodes, and there would be fruitful research that could benefit from it.

HCM method could help to build a chromosomal DNA barcode library. For an

unknown sample, HCM could compare its barcode with the DNA barcode library to find the closest database for taxonomy research[27]. Beyond this, our method combined with the contig scaffolding approach by Dvirnas et al.[4] could work on the full chromosomal DNA sequence assembly.

Bibliography

[1] Aaron CE Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394–1403, 2004.

[2] Zhenyu Li, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1):25–37, 2012.

[3] Monya Baker. De novo genome assembly: what every biologist should know, 2012.

[4] Albertas Dvirnas, Christoffer Pichler, Callum L Stewart, Saair Quaderi, Lena K Nyberg, Vilhelm Müller, Santosh Kumar Bikkarolla, Erik Kristiansson, Linus Sandegren, Fredrik Westerlund, et al. Facilitated sequence assembly using densely labeled optical dna barcodes: A combinatorial auction approach. *PloS one*, 13(3):e0193900, 2018.

[5] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157, 2013.

[6] Sara El-Metwally, Taher Hamza, Magdi Zakaria, and Mohamed Helmy. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS computational biology*, 9(12):e1003345, 2013.

[7] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135, 2008.

[8] Lena K Nyberg, Saair Quaderi, Gustav Emilsson, Nahid Karami, Erik Lagerstedt, Vilhelm Müller, Charleston Noble, Susanna Hammarberg, Adam N Nilsson, Fei Sjöberg, et al. Rapid identification of intact bacterial resistance plasmids via optical mapping of single dna molecules. *Scientific reports*, 6:30410, 2016.

[9] Vilhelm Müller and Fredrik Westerlund. Optical dna mapping in nanofluidic devices: principles and applications. *Lab on a Chip*, 17(4):579–590, 2017.

[10] Robert K Neely, Jochem Deen, and Johan Hofkens. Optical mapping of dna: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95(5):298–311, 2011.

[11] Vilhelm Müller, Nahid Karami, Lena K Nyberg, Christoffer Pichler, Paola C Torche Pedreschi, Saair Quaderi, Joachim Fritzsche, Tobias Ambjörnsson, Christina Åhrén, and Fredrik Westerlund. Rapid tracing of resistance plasmids in a nosocomial outbreak using optical dna mapping. *ACS infectious diseases*, 2(5):322–328, 2016.

[12] WHO. Antimicrobial resistance: global report on surveillance 2014, March 2015.

[13] Bret M Barton, Gordon P Harding, and Anthony J Zuccarelli. A general method for detecting and sizing large plasmids. *Analytical biochemistry*, 226(2):235–240, 1995.

[14] Alessandra Carattoli. Plasmids in gram negatives: molecular typing of resistance plasmids. *International Journal of Medical Microbiology*, 301(8):654–658, 2011.

[15] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931, 2004.

[16] David C Schwartz, Xiaojun Li, Luis I Hernandez, Satyadarshan P Ramnarain, Edward J Huff, and Yu-Ker Wang. Ordered restriction maps of saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262(5130):110–114, 1993.

[17] Michael J Levene, Jonas Korlach, Stephen W Turner, Mathieu Foquet, Harold G Craighead, and Watt W Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *science*, 299(5607):682–686, 2003.

[18] Jonas O Tegenfeldt, Christelle Prinz, Han Cao, Steven Chou, Walter W Reisner, Robert Riehn, Yan Mei Wang, Edward C Cox, James C Sturm, Pascal Silberzan, et al. The dynamics of genomic-length dna molecules in 100-nm channels. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30):10979–10983, 2004.

[19] M Neto, G Skorski, D Thevenot, and E Loukiadis. Optical maps: methodology and applications in microbiology. *EuroReference*, 5:38–46, 2011.

[20] Phil Latreille, Stacie Norton, Barry S Goldman, John Henkhaus, Nancy Miller, Brad Barbazuk, Helge B Bode, Creg Darby, Zijin Du, Steve Forst, et al. Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC genomics*, 8(1):321, 2007.

[21] Charleston Noble, Adam N Nilsson, Camilla Freitag, Jason P Beech, Jonas O Tegenfeldt, and Tobias Ambjörnsson. A fast and scalable kymograph alignment algorithm for nanochannel-based optical dna mappings. *PloS one*, 10(4):e0121905, 2015.

[22] Adam N Nilsson, Gustav Emilsson, Lena K Nyberg, Charleston Noble, Liselott Svensson Stadler, Joachim Fritzsche, Edward RB Moore, Jonas O Tegenfeldt, Tobias Ambjörnsson, and Fredrik Westerlund. Competitive binding-based optical dna mapping for fast identification of bacteria-multi-ligand transfer matrix theory and experimental applications on escherichia coli. *Nucleic acids research*, 42(15):e118–e118, 2014.

[23] Thomas Schreiber. Constrained randomization of time series data. *Physical Review Letters*, 80(10):2105, 1998.

[24] Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3-4):346–382, 2000.

[25] Douglas Poland. The phylogeny of persistence in dna. *Biophysical chemistry*, 112(2-3):233–244, 2004.

[26] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[27] W John Kress and David L Erickson. Dna barcodes: methods and protocols. In *DNA Barcodes*, pages 3–8. Springer, 2012.

Appendices

Appendix A

Plasmid barcode comparison

Plasmids are circular DNA molecules, so their DNA barcodes are also circular. When computing the similarity of the two barcodes, the algorithm from Pearson correlation coefficient calculation requires pairwise comparisons of barcodes are of the same length. Previous researchers first determined a target length and then interpolate the barcodes so to achieve the target length. However, we use another method, which uses "zero" to cover the positions that shorter barcodes lack (FigureA.1). We use zero elements to fill in the gaps of the short barcodes and corresponding bit-weights so the two barcodes can meet the conditions of the pairwise comparison.

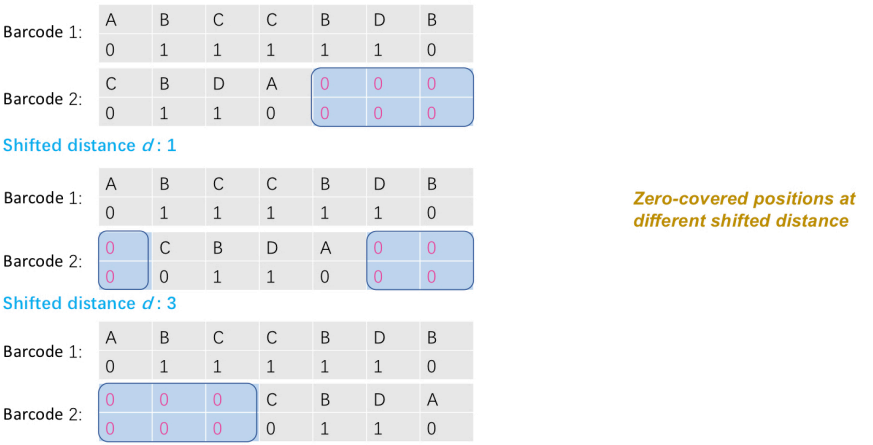
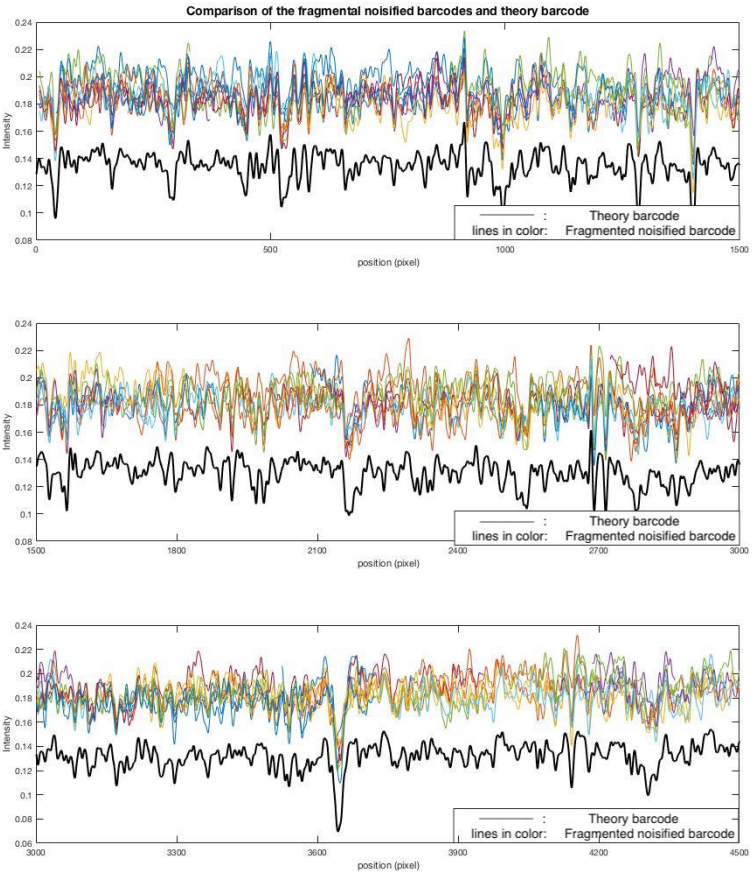


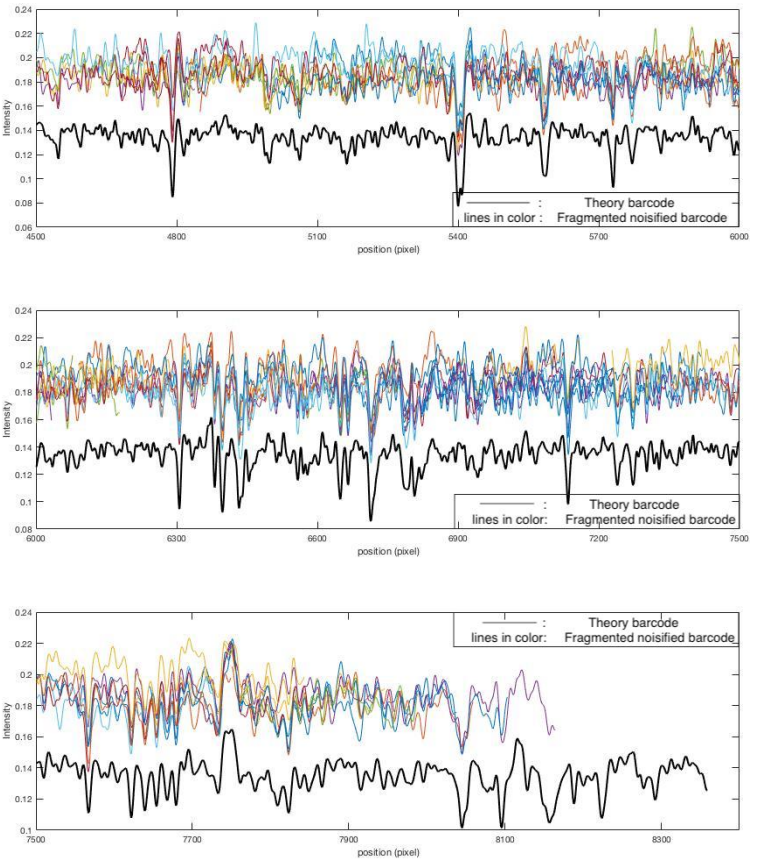
Figure A.1: Rescaling the comparing circular barcodes by inserting the "zero-covering positions".

Appendix B

Supplementary Figures

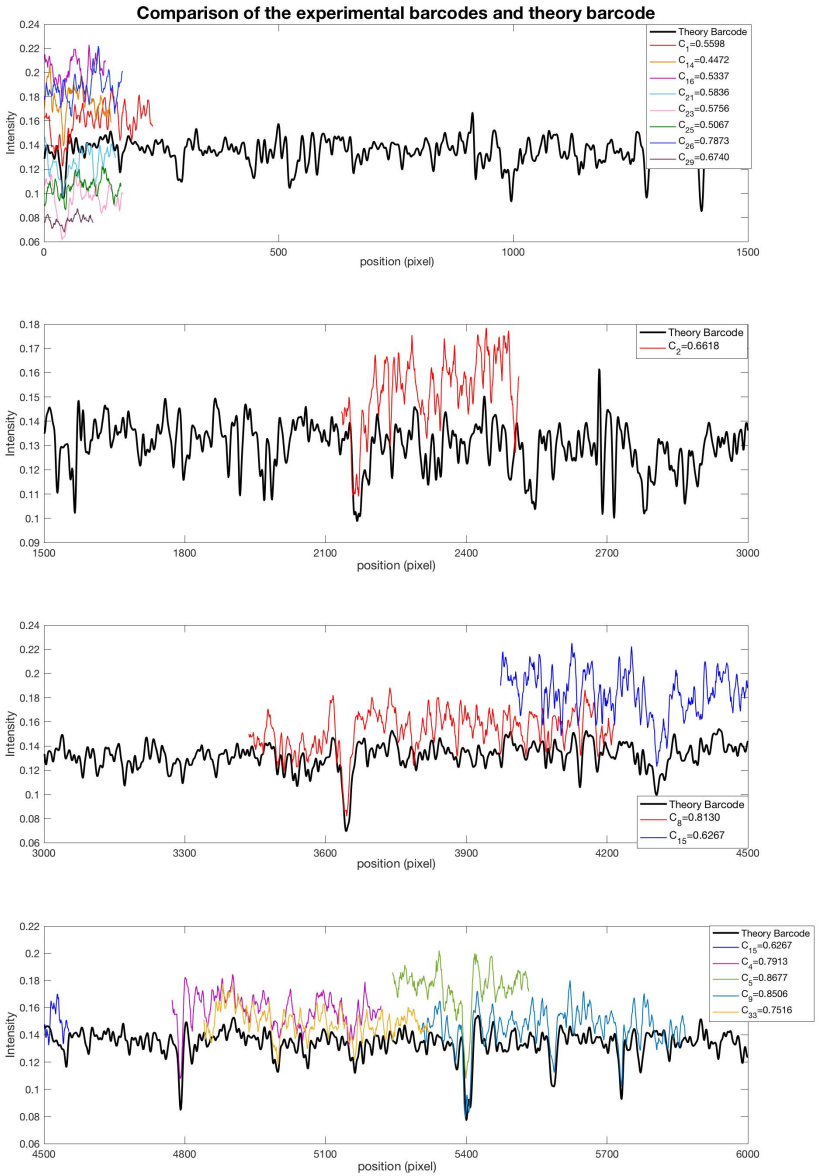


(a) Here are the first three panels of a continued Figure B.1.

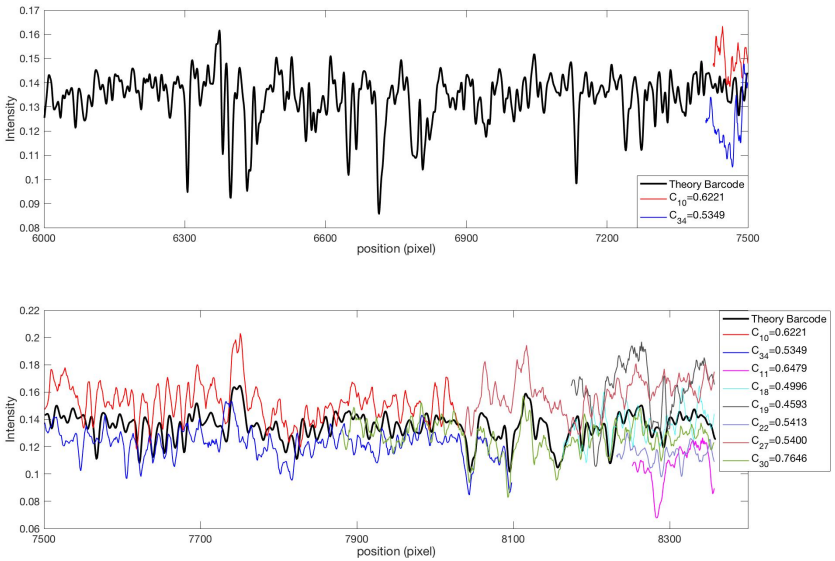


(b) Continuation of Figure B.1a

Figure B.1: The figure shows the comparison of the theory barcode and the fragmented noisified barcodes from 10 noisified barcodes. The black fit is the theory barcode and the others are fragmented noisified barcodes. After the assembling process, the assembled barcode is given in Figure 7.1a and 7.1b. The fragmented noisified barcodes are approximately uniformly distributed along the theory, and thus it is more likely to get a nearly intact assembled barcode at the end.



(a) These are the first three panels of a continued Figure B.2.



(b) Continuation of Figure B.2a.

Figure B.2: The figure shows the comparison of the theory barcode and the original experimental barcodes from the second experiment ('day 2 experiment'). The black fit is the theory barcode and the others are experimental barcodes. The assembled barcode is given in Figure 7.3. It is obvious that the original experimental barcodes have already clustered. And compared with fragmented noisified barcode (Figure B.1a and Figure B.1b), there are too few fragments. These factors partially explain that the resulting assembled barcodes are several pieces instead of a nearly intact barcode.