LU TP 17-06
March 2017

GENE-ID USING SIMULTANEOUS DNA BARCODING AND ENZYMATIC LABELING

FYTM03

Henrik Nordanger

Department of Astronomy and Theoretical Physics, Lund University

Master thesis supervised by Tobias Ambjörnsson

Abstract

Antibiotic resistance is an ever growing problem, and is considered one of the major challenges of modern medicine. In order to avoid further development and spread of multiresistance among bacteria, the use of antibiotics should be more restrictive, and more tailored to individual cases of infection. To facilitate more directed treatments, it is crucial to be able to quickly and reliably identify the strain of bacteria. A logical next step would be to directly be able to identify any genes coding for antibiotic resistance.

To this end, I investigate the possibility of combining two methods of optical mapping of plasmids - DNA barcoding, and enzymatic labeling of specific genes. The end goal is to develop a single experiment in which both the plasmid type, and any genes coding for antibiotic resistance, can be identified.

I focus mainly on three tasks - all related to the processing of experimental data. (1) Improving an earlier method for kymograph alignment, necessary for processing data from experiments using DNA in nanochannels. (2) Creating reproducible time averages of sparsely labeled barcodes, by attempting to exclude non emitting fluorophores. (3) Determining the number of fluorophores in a sparsely labeled barcode, by fitting a number of Gaussians to its reproducible time average.

The results of all three tasks were promising, but as no data was available of plasmids labeled using both techniques of optical mapping, no definitive conclusions could be made about the approach.

Populärvetenskaplig Sammanfattning

Plasmider är korta cirkulära DNA-molekyler, som kan hittas i de flesta bakterier. Detta projekt handlar om att skapa och testa ett nytt sätt att identifiera plasmider, samt specifika gener på dessa. Principen bygger på att kombinera en tidigare metod kallad "Competitive binding-based optical DNA mapping", och infärgning av en längre DNA-sekvens med fluoroscerande (självlysande) molekyler, för att kunna avgöra huruvida specifika gener är närvarande eller inte. Detta tillvägagångssätt vore potentiellt betydligt snabbare än traditionell DNA-sekvensering, vilket är av stor relevans bland annat då det ska avgöras huruvida antibiotikaresistens är närvarande i en population av bakterier. I längden kan detta leda till effektivare behandlingar av infektioner, vilket blir allt mer relevant allteftersom antibiotikaresistens blir vanligare. Mina uppgifter i projektet är att utveckla metoder och programvara för analys av filmer av färgade DNA-molekyler, som tagits med hjälp av fluorescens-mikroskop.

Competitive binding-based optical DNA mapping går ut på att låta de två ämnena YOYO-1 och netropsin binda till en DNA-molekyl. YOYO är en fluorescerande molekyl, som kan binda till vilken DNA-sekvens som helst. Netropsin däremot är icke-fluorescerande,

men binder endast till specifika fyra baspar långa sekvenser. Om en längre bit utsträckt DNA, efter markering med YOYO och netropsin, observeras under mikroskop så kommer den att likna en ”streckkod” med mörka band där netropsin bundit, och lysande band där endast YOYO är närvarande. DNA från olika organismer kommer att uppvisa olika band, eller olika streckkoder, och kan därmed identifieras.

Inom detta projekt undersöks möjligheten att sedan även använda ett enzymbaserat molekylkomplex för att färga in längre DNA-sekvenser, framförallt de som kodar för antibiotikaresistens. Genom att observera om och var dessa komplex bundit till DNA:t, och hur den uppkomna streckkoden ser ut, kan det snabbt identifieras vilken bakterieart det handlar om, och huruvida den är resistent mot en viss typ av antibiotika. Detta arbete syftar till att undersöka hur pålitlig och användbar en sådan metod skulle vara. På längre sikt skulle tekniken kunna bli tillräckligt snabb och praktisk för att användas för att undersöka de bakterier enskilda patienter infekterats av. Med hjälp av detta skulle skräddarsydda antibiotikakurer kunna väljas ut, vilket skulle kunna spara tid, lidande och resurser, och dessutom minska riskerna för ytterligare utvecklande och spridning av antibiotikaresistens.

De experimentella data som används har erhållits från Fredrik Westerlunds grupp vid Chalmers tekniska högskola och Robert Neelys grupp vid Universitetet i Birmingham, i form av korta filmer av utsträckta DNA-plasmider under mikroskop, då de färgats med ämnena beskrivna ovan. Dessa filmer har bearbetas med tidigare utvecklad programvara, så att en tvådimensionell bild, så kallad ”kymograf”, av varje DNA-sträng erhålles. X-ledet i dessa bilder representerar positionen längs med DNA-strängen, medan y-ledet representerar tid, med varje rad (varje pixel i y-led) innehållande data från en bildruta (”frame”) ur filmen.

Då DNA-molekylerna inte är helt stilla under experimentet (Westerlund-gruppens experiment), utan sträcker ut och drar ihop sig både lokalt och globalt, och flyttar sig i sidled, är de uppkomna streckkoderna inte särskilt tydliga i kymograferna innan vidare behandling. Den första delen av mitt arbete bestod därför av att förbättra en metod för att kompensera för dessa rörelser, eller att ”räta ut” streckkoderna. Därefter utvecklades en metod för att även ta i beaktning att fluoroforer slumpmässigt kan ”slå på eller av” under ett experiments gång (experiment från Robert Neelys grupp). För att finna intensiteten av emitterat ljus, måste medelvärden tas endast från den del av experimentet då molekylen ”är på”.

Contents

List of Figures

List of Tables

1 Purpose & Aims

Optical mapping is a group of techniques for evaluating the genetic content of DNA molecules, by staining them with fluorescent dyes. Such methods have begun to emerge as a complementary or alternative approach to traditional DNA sequencing. While offering significantly lower resolution, typically in the order of kilo base pairs (kbp), optical mapping shows promise as it has the potential to be significantly faster and practically simpler to perform than its alternative. Several applications exist in which the lower resolution would, at least potentially, suffice. These include for example the identification of large-scale structural variations, or determination of bacterial species or strain, both of which are relevant for combating the growing problem of antibiotic resistance.

Another use of optical mapping is to determine whether a specific gene is present in the DNA under examination. This is what is being considered in this thesis, where we examine the possible advantages of combining two types of optical mapping - competitive binding barcoding, and enzymatic labeling. It is explored whether this approach could be applied to bacterial plasmids - circular DNA molecules found in bacteria, to create an experiment in which both the plasmid type itself as well as genes coding for antibiotic resistance can be identified. The envisioned platform for accomplishing this is displayed in figure 1, and is similar to previous essays on related topics[1][2]. More details on all the steps are presented in following sections (2.2-2.4).

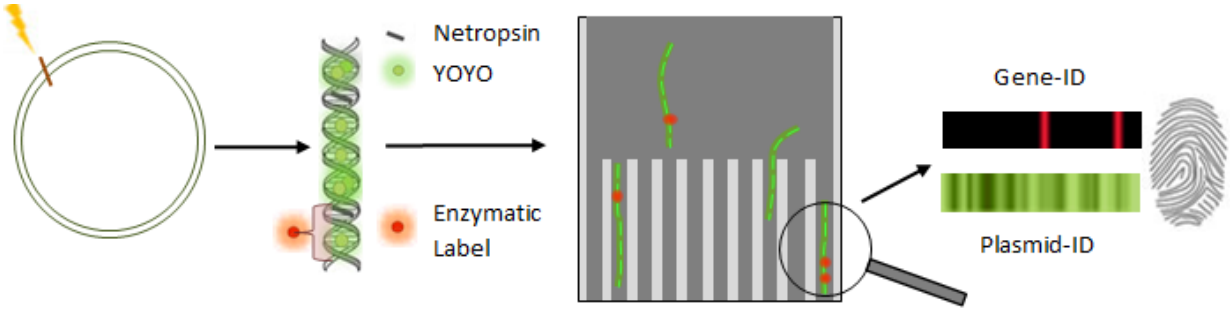

Figure 1: **Envisioned process for single-experiment identification of both plasmid type and individual genes.** An illustration of the steps involved in the envisioned process. A circular plasmid is first cut open at a random point along its contour, by irradiating it with light. The linearized string of DNA is then labeled using both a competitive binding scheme including netropsin and YOYO-1 (green), and enzymatic labels (red). These two techniques are described in sections 2.3 and 2.4 respectively. In the third step the plasmid is inserted into a nanochannel, stretching it out so that it can be properly observed using a fluorescence microscope. The last step is the processing of images taken of the plasmid, which is required for producing a DNA-barcode (green) that can be used as a "fingerprint" to identify the plasmid type. New to this thesis is the inclusion of the enzymatic labels, which can bind to specific genes which therefore become visible under the microscope. If several different enzymatic labels using different colors are used, several different genes can be identified. (Parts of image taken from [2].)

The focus of this thesis is on the processing of the experimental data, meaning the last step in figure 1. Unfortunately, no data was available of plasmids labeled using both competitive binding barcoding and enzymatic labeling. Therefore two separate sets of

data had to be used - one for each type of labeling scheme. This posed limitations when developing and testing new methods.

Experiments with plasmids isolated from K. pneumoniae, dyed using netropsin and YOYO-1 creating competitive binding barcodes, were performed by Fredrik Westerlund's group at Chalmers University of Technology. Data of enzymatically labeled T7 phage DNA, using "nick-labeling", was provided by Robert Neely, University of Birmingham. Note that Neely's lab used stretching due to surface adsorption rather than due to confinement in nanochannels as displayed in figure 1. We mainly address three issues, with main focus on the first one:

Task 1 - Kymograph alignment of densely labeled DNA
When performing optical mapping of DNA in nanochannels, thermal motion leads to blurring of any fluorescent patterns along the DNA, such as the green barcode seen in figure 1 (see section 2.2 for details). This loss of information constitutes a major problem, for example when attempting to identify a plasmid. Image processing used for compensating for this effect is referred to as kymograph alignment. I develop and test an improved version of the alignment algorithm described by Noble et.al. (2015)[3].

Task 2 - Generating reproducible time averages of sparsely labeled barcodes
DNA stained with a comparatively small number of dye molecules are referred to as sparsely labeled. An example of this are the enzymatic labels (red dots) in figure 1, and is discussed further in section 2.4. Images of such a DNA molecule, taken at different times, may differ significantly due to individual fluorophores 'blinking', i.e. turning off and on in a seemingly random manner. It is therefore a challenge to obtain an image that is representative of the whole molecule. Simply taking the average of a large number of images is not feasible, as fluorophores have a limited lifetime due to bleaching. I develop a method for creating *reproducible time averages* of sequences of images (or rather, of kymographs, see section 2.2), by identifying and excluding regions where fluorophores are off. The resulting intensity profile along the DNA molecule should then ideally be independent of blinking.

Task 3 - Fitting Gaussians to barcode time averages, with application to gene identification
Using the reproducible time average of a sparsely labeled barcode (task 2), the positions of individual fluorophores along the DNA under examination is determined by fitting a number of Gaussians to the intensity profile along the molecule, using weighted least squares minimization. As the fluorophores in question are bound to a specific gene, the number and positions of possible copies of that gene can be determined.

The problem addressed in task 1 is common to most experiments with DNA in nanochannels, and so an improved kymograph alignment algorithm will have more widespread use than just as a prerequisite for the following tasks. The time average of an aligned kymograph (the *barcode*) can reliably be used as a "fingerprint" for identifying the plasmid[4][5].

Also completing the following two tasks ideally means that it can simultaneously be determined how many copies there are of a specific gene. It can thus be found whether for example a gene coding for antibiotic resistance is present, and if that gene has undergone any amplification (see section 2.1).

2 Introduction

In this introductory section we introduce the readers to bacterial plasmids and their role in antibiotic resistance, optical DNA mapping, and the two types of fluorescent labeling techniques used in this study - competitive binding based labeling and enzymatic labeling.

2.1 Bacterial Plasmids

In bacteria, all genetic material is partitioned into chromosomal DNA and plasmid DNA. All essential genes are gathered in a single chromosome, somewhat similar to those that can be found in animals and other eukaryotes. Genes that are only useful in some situations however, are usually located on plasmids, which are circular strings of DNA that range from a few kbp (kilo base pairs) to hundreds of kbp in length. Each individual bacterium can have several different plasmids, and several copies of each type. Examples of genes commonly found on plasmids include those coding for transporter proteins, which have the function of removing hazardous molecules from the cell. Genes coding for antibiotic resistance are also usually located on plasmids, and are of high interest in many modern medical applications.

Bacterial plasmids are highly dynamic with respect to their genetic content, and can vary greatly even within a single population. Genes that are not of any use at a given time will, to a degree, be selected against as they only takes up space without serving any purpose. On the other hand, when a specific gene becomes useful, it can become duplicated, leading to its expression becoming faster and more effective. This is called gene amplification.

Plasmids can also be transferred horizontally between bacterial cells. This means that the plasmid can be moved from one individual bacterium to another, which can even be of a different bacterial species, without any cell division occurring. A gene that has suddenly become useful can thus spread in an environment very quickly. For example, this can happen in the body of a patient being treated with antibiotics. The bacteria will effectively be bred to survive in the suddenly dangerous environment, and any genes coding for antibiotic resistance will be spread, rendering the treatment more and more useless. For this reason it is essential to select an antibiotic for which there is no resistance among the bacteria. Consequently, it would be an enormous advantage to be able to identify any resistance genes, and then tailor the treatment accordingly.

Current techniques for identifying genes, such as DNA sequencing, require cultivation of the bacteria under examination. In addition to being time consuming, having the bacteria reproduce outside the environment under interest (i.e. the patient's body) may drastically

decrease the frequency of resistance genes. Optical mapping, which utilizes single DNA molecules, circumvent this problem, potentially making it a more reliable technique for this purpose[6].

2.2 Nanochannel-based Optical Mapping & Kymographs

In optical mapping, DNA molecules are stained with one or more types of fluorescent dye. Each molecule is then observed under a microscope while the fluorophores are being excited, creating a visible fluorescence intensity pattern that can ideally be used to tell something about the genetic content of the molecule. For any pattern to be visible, the DNA has to be stretched, for which there are a number of methods. An electric field or hydrodynamic flow can be used[7][8], or optical tweezers in cases where more precise handling is required[9]. Here we look mainly on DNA stretched in nanochannels (experiments by Westerlund's group)[10] or using molecular combing and surface adsorption (data from Neely's group)[11].

Nanochannel-based optical mapping (the main focus herein) involves inserting the DNA molecule into a nanochannel (typically of width ∼100 nm), after which they stretch out due to the confinement[12]. Figure 2 illustrates the principle, but also shows that the stretching is far from perfect. Indeed, the molecule is elastic and has some room to contract, extend, and bend slightly due to thermal motion. As it is not fixed at any end, it can also undergo global motion along the channel. This motion can be problematic during experiments, as the configuration of the molecule at any given time is unknown. If a single image is taken, any local stretching or contraction can make patterns difficult to identify. A discussion on the physics of nanoconfined DNA can be found for example in [10].

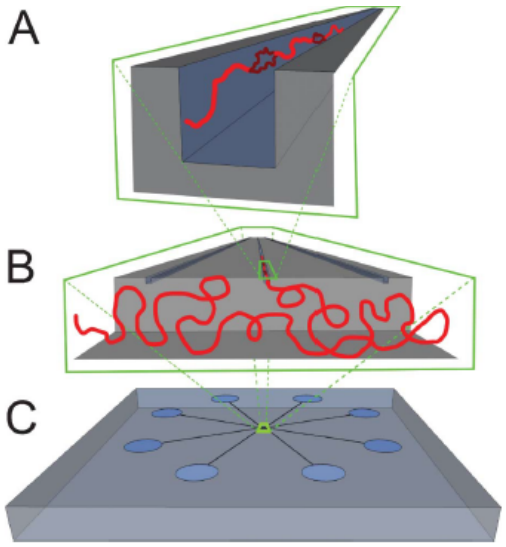

Figure 2: **DNA in a nanochannel device.** (A) A (linear) DNA molecule inside a nanochannel. (B) The same molecule as it enters the nanochannel. (C) A *lab-on-a-chip* device with several channels, which is commonly used in experiments. Before entering the channel the molecule is coiled, but mostly stretches inside the channel. (From [12].)

When working with nanoconfined DNA, it is possible to see if the molecules in the channels are circular as opposed to linear, as circular ones display twice the intensity per length. This can be a significant advantage when working with plasmids, as it is possible to ensure that the whole molecule is intact when performing an experiment. This opportunity does not arise for DNA molecules that are naturally linear, and the experimenter may run the risk of unknowingly continuing with just a fraction of the DNA that was intended.

After confirming that a plasmid inside a nanochannel is intact, it can be cut open and brought to a linearized form suitable for mapping. A simple way to accomplish this is to expose the molecule to light and wait for a photo-cut to occur, which will cause it to get "nicked" at random points, meaning that one of the two strands will break. When both strands are cut sufficiently close, the plasmid will break open, and the molecule will unfold inside the channel. This approach was taken by Fredrik Westerlund's group at Chalmers University of Technology, who generated most of the data used in this thesis. A significant problem with this method however, is that the dye molecules get bleached over time, as they are exposed to the light. It is also possible that more than one cut will appear, meaning the plasmid will be split into several fragments. These problems are incentives to find a new way of cutting plasmids. One such method is to introduce a cutting enzyme to the molecule, which will target a specific sequence, and which can be removed after the plasmid opens. This will reduce the effect of bleaching, and the risk of a second cut.

The common approach in optical mapping experiments is to take several images in short succession, typically around 200, with around 0.1 s intervals. The images are cropped so that only the region of the images containing the molecule remains, typically only being a few pixels wide (as the nanochannel containing the DNA is very narrow), and orders of magnitude larger in the other dimension. The averages of the intensity values across the width of the channel is taken, forming a one dimensional array of the same length as the cropped image. These arrays (one from every image) are arranged on top of one another to form a new image. This is what is called a *kymograph*, where each row contains an intensity profile from one frame of the original movie (with the first frame at the top). See figure 3A for an example. The general look of the kymograph is reminiscent of a barcode, which is the origin of the term DNA barcode. If there was no movement of the molecule during the capturing of the images, the barcode would be entirely straight, meaning each row would be identical. Thermal motion however leads to the barcode being "fuzzy", and while features are detectable, simply taking an average of each row would not yield very useful results. To interpret the pattern, and compare them to one another, kymographs have to be aligned. Methods for doing this is discussed in section 3.1.

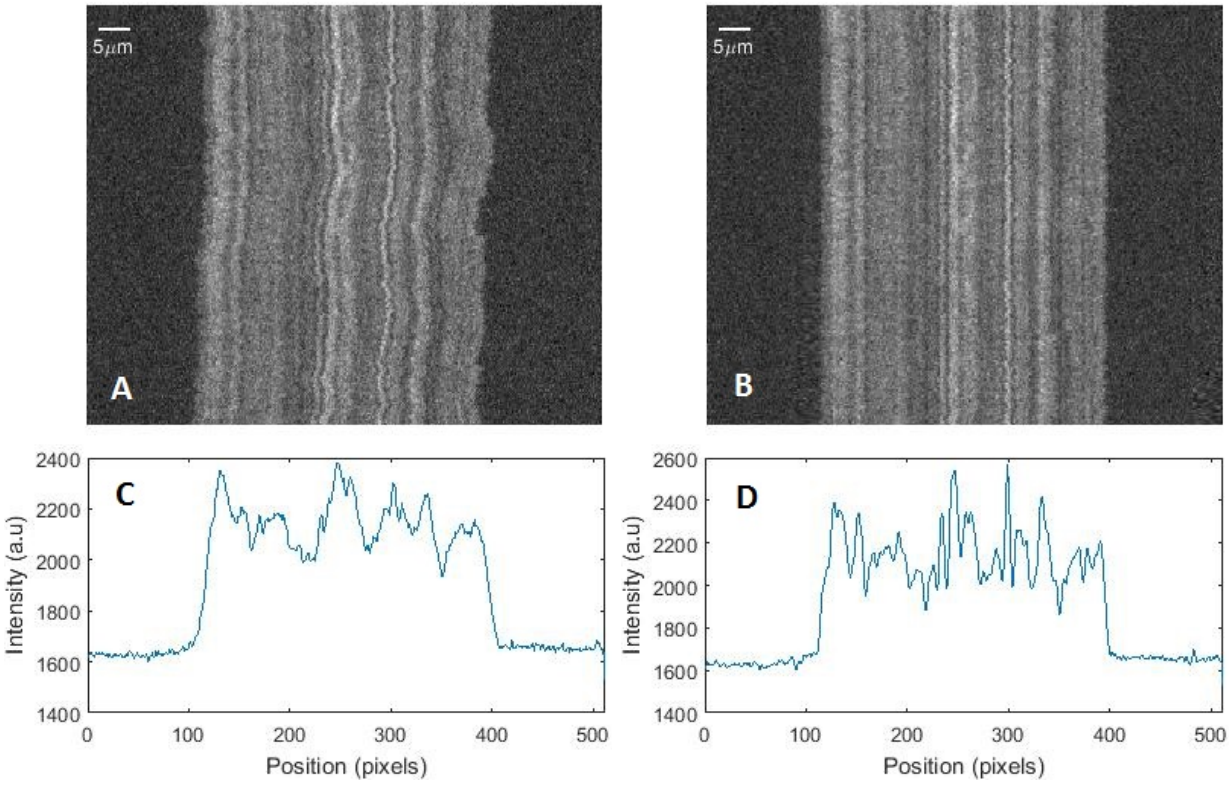

Figure 3: **Task 1 - Kymograph alignment.** (A) An example of an unaligned kymograph of a *pUUH239.2* type plasmid (from K. pneumoniae), dyed with netropsin and YOYO-1 to create a competitive binding barcode (see section 2.3). (B) The aligned kymograph, after using the alignment algorithm "WPAlign" as described in [3]. (C) Time average of the unaligned and (D) aligned kymographs. As can be seen, more features are discernible in the time average of the aligned kymograph, than in the unaligned one. Data provided by Fredrik Westerlund's group, Chalmers University of Technology.

2.3 Competitive Binding Barcodes

There are several different methods of fluorescently labeling a DNA molecule, that are used in optical mapping. The one being considered here is so-called competitive binding[13], illustrated in figure 4. Its name arises from the fact that a second chemical is used in addition to the fluorescent dye, which "competes" with the dye in the sense that it blocks certain binding sites at the DNA molecule. The chemicals used in the experiments that this thesis is based on, are called netropsin and YOYO-1. Netropsin, which is the non-fluorescent blocking chemical, binds primarily to any four-basepair long sequence containing only A and T. The dye YOYO-1 can bind to any four-basepair sequence, and is thus able fill up many remaining sites. The parts of the molecule that will light up under microscope is then those that contain few four-basepair sequences of A and T[13]. The resulting pattern (the DNA *barcode*) can be used as a "fingerprint" for the DNA molecule. This competitve binding scheme can therefore be used to identify for example a plasmid, without having to fully sequence it.

Other dying methods used in optical mapping of DNA include DNA melting, which involves heating the molecule to a temperature such that the weaker AT-bonds break up, while the stronger GC-bonds remain intact[14]. A fluorescent dye is then introduced to the partly melted DNA-strand, where it will bind only to the still closed regions. Enzymes can also be used to create barcodes by cutting[15] or "nicking" the DNA (meaning one of the two strands gets cut)[16]. Another type of method is enzymatic labeling, to be discussed in section 2.4.

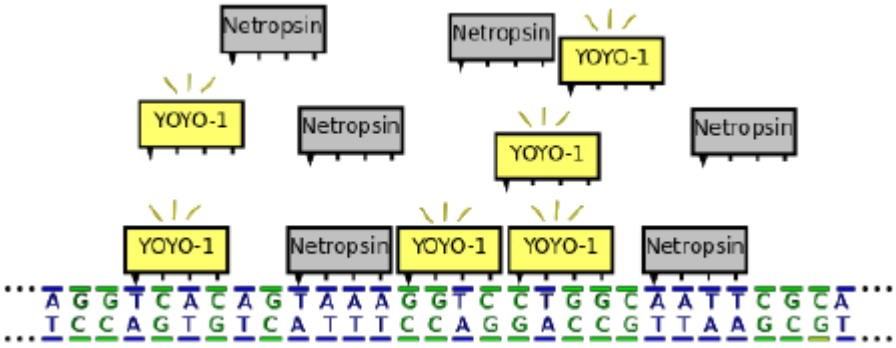

Figure 4: **Competitive binding.** Illustration of competitive binding involving YOYO-1 and netropsin. Netropsin binds only to four-basepair sequences containing only A and T, while the fluorescent YOYO-1 can bind to any 4-basepair sequence. When viewed under a fluorescence microscope, regions with less A and T will generally appear bright, creating a DNA barcode. (From[17].)

2.4 Optical Mapping of Sparsely Labeled DNA

Optical mapping schemes in which the number of dye molecules binding to the DNA molecule is small, are said to use sparse labeling. More specifically, if there is little or no overlap between the intensity profiles of the individual fluorophores, the DNA can be considered sparsely labeled. This typically occurs when the number of fluorophores is less than of the order of 10/kbp, but no well defined differentiator exists. Sparse labeling can be achieved using for example enzymatic labeling, in which an enzyme is used for binding a fluorophore to a specific sequence of DNA[18]. If this sequence is long enough, it can even be used to label a specific gene. In such cases, as few as one fluorophore can be bound to the DNA (if the gene only occurs once), creating a very simple pattern with only a single "dot" along the molecule.

When fluorophores are as far apart as in sparse labeling, the phenomenon called fluorescence intermittency, or *blinking*, plays a significant role. When a fluorophore is under continuous excitation, it may randomly switch between emitting and non emitting states, meaning that it will intermittently go dark. It is a universal property of nanoscale emitters, and is related to the competition between radiative and non-radiative relaxation pathways of the molecule[19]. In the case of densely labeled DNA, individual fluorophores have less of an impact on the overall pattern, and blinking generally does not have to be compensated for. A kymograph showing the effects on a sparsely labeled barcode, can however be seen in figure 5. Most of the fluorophores turn off at some point, seen as a "stripe" going

dark. Simple time averages of such kymographs (averages of all rows) will not be useful measures, as they can look significantly different from experiment to experiment, due to the random nature of the intermittency. A method for compensating for this problem is described in section 3.2, where reproducible time averages are created by excluding parts of kymographs, where fluorophores seem to be non-emitting.

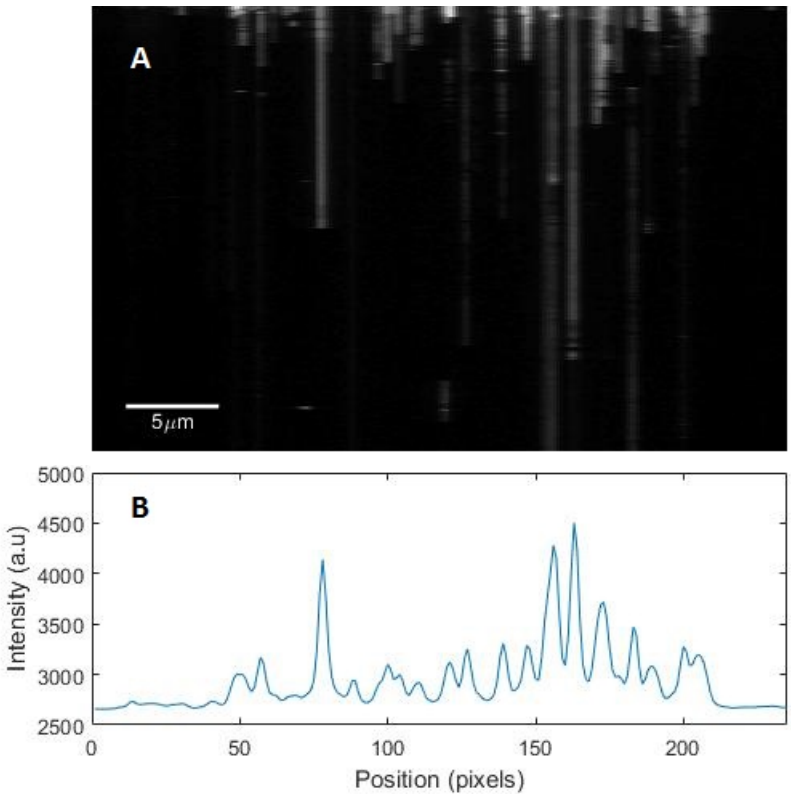

Figure 5: **Task 2 & 3 - Sparsely labeled kymograph and its time average.** (A) Kymograph of T7 phage DNA, nick-labeled using an enzyme binding to the sequence TCGA. Based on the full genetic sequence of the DNA, there should be 107 binding sites along the molecule[20]. It can be seen that the dye molecules only fluoresce intermittently over the course of the video capture. The plasmid is here immobilized on a surface as opposed to confined in a nanochannel, therefore showing little movement during the video capture. However, a kymograph of nanoconfined, sparsely labeled DNA is expected to look very similar after alignment (task 1). (B) Raw time average of the kymograph. Note that due to intermittency in fluorescence, the peak heights vary substantially. Data provided by Robert Neely, University of Birmingham.

In figure 5 it can be seen that the overall intensity is greater towards the top of the kymograph, meaning the level of fluorescence is higher earlier on in the experiment. The reason for this is *photobleaching*, which is a photochemical alteration of a fluorophore molecule after a period of continuous excitation, such that it permanently loses its ability to fluoresce[21]. This means that as the experiment progresses, more and more fluorophores permanently go dark.

3 Methods

We here consider optical mapping of plasmids, simultaneously using both a competitive binding scheme (section 2.3) and enzymatic labeling (section 2.4). As no such data was currently available, the methods described here were developed and tested using two sets of data - each one consisting of plasmids labeled using either of the two methods. The goal in mind was to make a procedure, including the three steps outlined in section 2 and further described here, to determine the presence and positions of specific genes along plasmids, as well as the plasmid type itself.

The first step of our approach relies on the fact that the two methods of optical mapping can utilize different fluorophores, which can be excited by light of different wavelengths. Therefore, an experimenter can select to view either the competitive binding barcode, or the enzymatic labels, at any given time by changing an excitation filter[22]. When taking several images in short succession to create a kymograph, every other image should be of the competitive binding barcode, and the others of the enzymatic labels. It is then a simple procedure to separate this kymograph into two, extracting every other row to form one competitive binding kymograph, and one enzymatic labeling kymograph. If the images were taken with small enough interval between them, the two kymographs should be aligned in the same way, meaning that it is sufficient to find the alignment of the densely labeled one, and then apply the same alignment to the sparsely labeled one. This is very practical as it is generally much more difficult to find the alignment of a kymograph of sparsely labeled DNA.

The aligned, sparsely labeled kymograph can then be used to create a time average, a problem described in section 2.4 and addressed in section 3.2. Lastly, to determine the position of the enzymatic labels along a plasmid, based on the time average, a number of Gaussians are fitted to the intensity curve, by performing weighted least squares minimization.

3.1 Task 1 - Kymograph Alignment of Densely Labeled DNA

The method for aligning kymographs (see figure 3) described here is largely based on the algorithm developed in [3], called WPAlign. As such it should be considered an improvement on that method, rather than a separately developed one.

The approach involves treating the kymograph alignment as a shortest path problem in a weighted directed graph. Kymographs are considered as energy landscapes (see Figure 3), where the best paths correspond to features such as ridges or valleys reaching from the top to the bottom of the image. The process can be divided into four steps, described in the subsections below. In WPAlign (**W**eighted **P**ath **Align**) only the last three steps were present, and those were gone through recursively, finding and aligning one feature at a time, and then repeating the whole process. In our approach all detectable features are found in one go, meaning the steps are only performed once, following an initial center-of-mass alignment step. As the recursion present in WPAlign is removed, the newer algorithm is called NRAlign - **N**on **R**ecursive **Align**. As opposed to WPAlign, it avoids the repeated

interpolation of intensity values in the kymographs (see section 3.1.4). Naturally one of the steps, namely feature detection and alignment, has to be performed slightly differently in the two algorithms. Apart from this, the description in section 3.1.2-3.1.3 follow that of [3] closely.

3.1.1 Center-of-mass Alignment

In nanochannel experiments a plasmid can sometimes undergo significant global motion during the video capture, either due to thermal motion or because of a net flow through the channel. This can for example be seen in figure 3A, where a kymograph shows the plasmid moving slightly towards the left. To compensate for this, each row in the kymograph can be shifted so that the center of mass of the molecule remains still. A simple way to do this is to move each row horizontally, without any stretching, so that the Pearson correlation coefficient with the first row is maximized. The global motion can then be considered compensated for, leaving only the local fluctuations of the molecule. This is the first step of NRAlign, but was not present in WPAlign.

3.1.2 Image Pre-processing

For the main part of the alignment, the kymograph is first smoothed using a 2D Gaussian lowpass filter in order to remove noise due to random intensity fluctuations. A square filter of side length 10 pixels was used, as in [3]. Then a Laplacian of Gaussian filter is applied to the smoothed kymograph, resulting in a *Laplacian response image*, which will be referred to as K. It is a matrix of the same size as the kymograph, but has large positive values in dark bands and large negative values in light bands. K is linearly rescaled by dividing all elements by $\max(\text{abs}(K))$.

A feature in the original kymograph, such as a valley or ridge, should be represented by continuous regions in K, of high and low values respectively. Now, to avoid the detection of features composed partly of ridges and partly of valleys, we generate two images, K_+ and K_-. K_+, which emphasizes dark regions, is calculated using

$$K_+(x,y) = \begin{cases} 1 - K(x,y) & \text{for } K(x,y) > 0 \\ 0 & \text{for } K(x,y) \leq 0 \end{cases} \tag{3.1}$$

Similarly K_-, emphasizing bright regions, is calculated using

$$K_-(x,y) = \begin{cases} 1 + K(x,y) & \text{for } K(x,y) < 0 \\ 0 & \text{for } K(x,y) \geq 0 \end{cases} \tag{3.2}$$

Examples of K_+ and K_- can be seen in figures 6A and 6B. Elements set to zero in K_+ or K_- will later be ignored, when paths are to be found. All other values are positive, with the smallest ones representing the most pronounced features. The separation of K into two parts means that a particular feature cannot be composed of both a valley and a ridge, as one of those types of elements are zero in either K_+ or K_-. However, this does

not fully eliminate the risk of detecting "false" features, as a path through K_- may cross a region in K_+, if that region is narrow enough that the crossing can occur in a single step. (See section 3.1.3.)

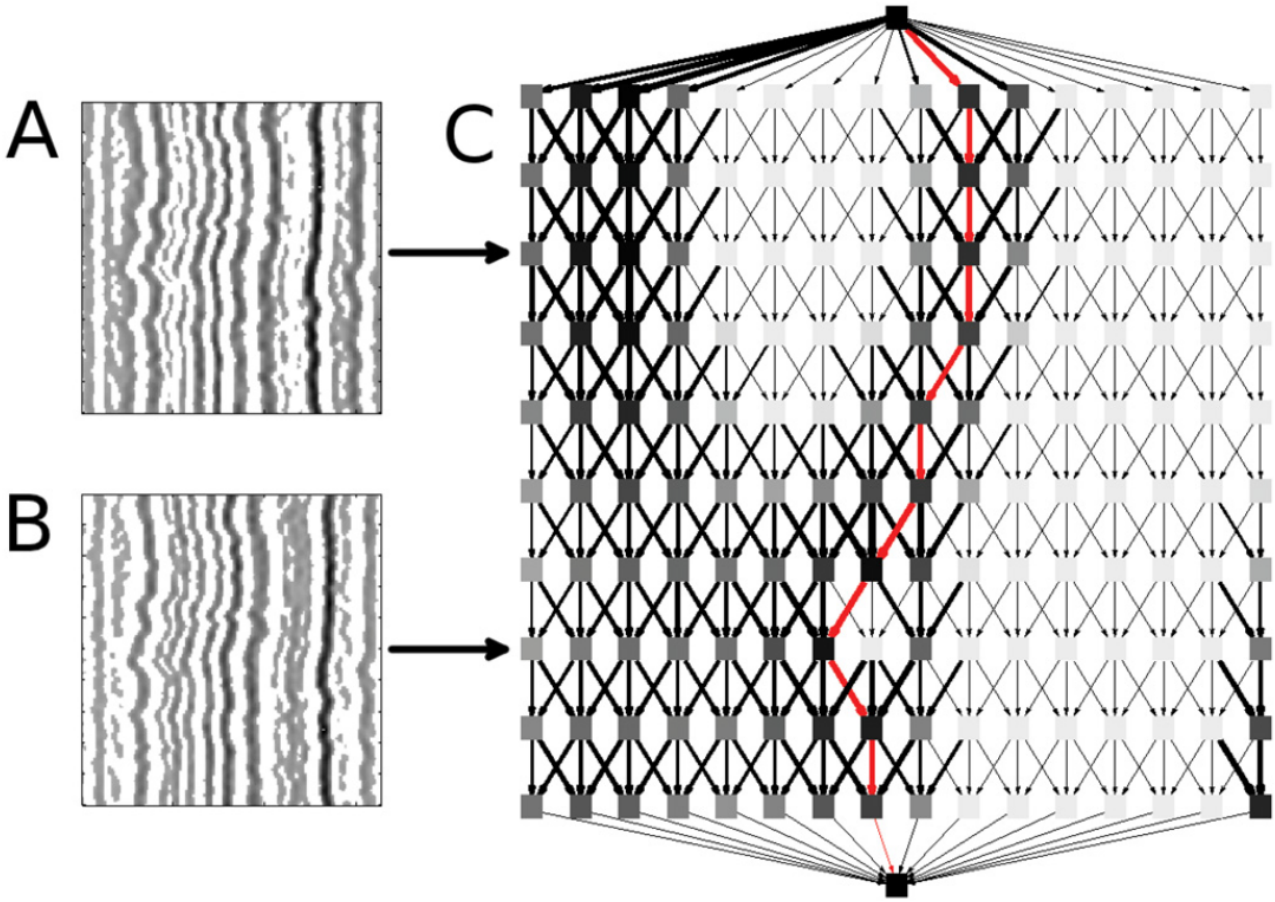

Figure 6: **Network assembly and feature detection.** (A) & (B) Examples of K_+ and K_-, rescaled from a Laplacian response image K using eqs. 3.1 and 3.2. These images emphasize dark and bright regions in the original K, respectively. White pixels represent barriers that potential features cannot cross, while continuous dark regions stretching vertically through the images indicate likely features. (C) An example of a network assembled from one realization of K_+ or K_-. Note that two separate networks are assembled for any K, one for each of K_+ and K_-, although only one is shown here. Each node (square) within the rectangular region of the network represents a pixel in K_+ or K_-. The top and bottom nodes lying outside the bulk of the network are the peripheral nodes, constituting the starting and ending points of the network. The width of the edges (arrows) corresponds to the inverse of the edge weight, and the darkness of the nodes represents the average inverse weight of incoming edges. The red line illustrates the shortest path through the network. This illustrative example network was created using a small subsection of an actual Laplacian response image. Potential features were here restricted to move no more than 1 pixel left or right, in any given step from one row to the next. This corresponds to having a parameter value of $k = 1$, see section 3.1.3. (From [3].)

3.1.3 Network Assembly

K_+ and K_- can now to be considered energy landscapes, where detecting the best feature (ridge or valley in the original kymograph) is the same as finding the lowest energy path from the top to the bottom. To set this up as a weighted shortest path problem, we assemble two acyclic, directed networks G_+ and G_- from K_+ and K_- respectively. The procedure for this is described for K_+, but is applied identically to K_-. See figure 6C for an illustration.

1. One node is created for each pixel in K_+, plus an additional two nodes, referred to as peripheral nodes. The total number of nodes in G_+ is thus $(M \times N) + 2$, where M and N are the number of rows and columns respectively.

2. One of the peripheral nodes, representing the starting point of the network, is connected to each node in the first row of G_+. Similarly, each node in the last row of G_+ is connected to the second peripheral node, which represents the end point. All edges so far are given equal weights.

3. All other nodes are connected in the following manner: The node at row i and column j is connected to nodes in the row $(i+1)$ directly below, belonging to columns $(j-k)$ to $(j+k)$. Each node is thus connected to $2k+1$ nodes in the next row. The value of k used was 1, see section 5.1. Exceptions are nodes close to the left or right border of the kymograph, for which connections to non-existing nodes, outside the kymograph, are ignored.

4. Finally, the weight of each edge is assigned a value equal to the intensity of the pixel in K_+, corresponding to the node the edge is directed to.

3.1.4 Feature Detection and Alignment

In our assembled network, every path between the peripheral nodes represents a potential feature. The most prominent valley should correspond to the shortest path through K_+, and the most prominent ridge to the shortest path though K_-.

Finding the shortest path through a directed, acyclic network is achievable using algorithms that are readily available as part of software packages. Here, the function `graphshortestpath` was used, which is part of the Bioinformatics toolbox in MATLAB(2016b). The function uses Dijkstra's algorithm to find the shortest path, and the computational time required scales linearly with the sum of the number of edges and nodes in the network. This sum grows bilinearly with the number of time frames (rows in the kymograph) and the width of the kymograph. As the width is proportional to the length of the DNA molecule, the time required for kymograph alignment should be proportional to the length of the genome that is to be mapped.

The two shortest paths are found, one through each of K_+ and K_-, and the shortest of these two is accepted. Thereafter WPAlign and NRAlign differ.

WPAlign After a single feature has been found, the kymograph is aligned by setting $F(i) = \langle F(i) \rangle$. $\langle . \rangle$ denotes the mean horizontal position of the feature (rounding to the nearest integer) over all rows i, and the parts of the row to the left and right of F(i) are stretched (or compressed) linearly to accommodate this. Due to the discrete nature of the kymograph, it is not elementary to assign new intensity values to the pixels in the stretched or compressed regions. To calculate these new values, cubic spline interpolation is used.

After having aligned a single feature, we wish to return and find the next one. To do this, the resulting partly aligned kymograph is split vertically along the newly aligned feature, and w columns are removed from each of the new smaller kymographs, on the side adjacent to the split. w is half the width of a typical feature, and is set to 5 in our applications.

These previous steps are then performed on each of the smaller kymographs, so that one feature can be found and aligned in either one. The images are split again, and the process is repeated recursively, until no parts remain that are wider than $2w$. Then the kymograph is considered fully aligned.

NRAlign As in WPAlign, the process starts with simply finding one single feature. This feature is however not aligned, but its position is saved for later. Instead, all edges leading to any node within w columns to the left or right of the feature, are removed. Thereby the kymograph itself is not changed, but future features are prevented from coming within w columns of an already detected feature. Another feature is then found, without having to go through pre-processing or network assembly again.

When no additional paths can be found through the network, the alignment begins. Each feature l is straightened by setting $F_l(i) = \langle F_l(i) \rangle$, where again $\langle . \rangle$ denotes the mean horizontal position of the feature (rounding to the nearest integer) over all rows i. As in WPAlign, cubic spline interpolation is used.

A clear advantage of the latter method is that interpolation of intensity values is only performed once - after all features have been found. In the previous method however, interpolation would have been performed repeatedly and recursively on many parts of the aligned kymograph.

3.2 Task 2 - Generating Reproducible Time Averages of Sparsely Labeled Barcodes

To create a reliable time average of a sparsely labeled (aligned) kymograph, such as the one in figure 5, two phenomena have to be compensated for. One is photobleaching - a photochemical alteration of a fluorophore molecule after a period of continuous excitation, such that it permanently loses its ability to fluoresce[21]. The other is blinking, previously described in section 2.4. Both are compensated for by only taking regions of kymographs into account, where no fluorophores are in non-emitting states. The method requires the

width of the point spread function, σ_{PSF} to be known for the experiment. We here use 472nm.

To identify the *signal regions* of a kymograph where one or more fluorophore is emitting, a multilevel thresholding method based on the Otsu method was used[23][24]. A threshold intensity value was found, separating the signal from the background. Regions were only considered if they had a minimum vertical length, determined by a parameter we refer to simply as the *minimum segment length*, and a width equal or greater than σ_{PSF}. Also, small gaps in the signal regions were closed, if they had a smaller vertical length than the parameter we call the *maximum gap length*. See appendix B for further details.

The *reproducible* time average of the kymograph is then calculated by taking the mean of each column, only taking into account the elements considered part of the signal regions. For any column of the kymograph where no element belongs to the signal region, the corresponding time average element is calculated as the average of the background in that column.

3.3 Task 3 - Determining Positions and Number of Enzymatic Labels

After having obtained a reproducible time average of a sparsely labeled kymograph (task 2), the resulting intensity curve should include a number of peaks, situated at the locations of any bound emitters (such as enzymatic labels) along the DNA. The approach here is to fit a number of Gaussian peaks to this data, using weighted least squares minimization.

We denote the horizontal coordinate along the curve x. The intensity of the time average at pixel i (at $x = x_i$) is then denoted y_i. Each fluorophore j, situated at a position μ_j, is observed with an integrated single emitter intensity $a_j = a_0 \pm \sigma_{a_0}$. The mean intensity for a single emitter (a_0) as well as its standard deviation σ_{a_0} should be known beforehand, and its point spread function (PSF) is assumed to be a Gaussian of known width σ_{PSF} (not to be confused with σ_{a_0}). The total integrated intensity over the whole time average is the sum of the contributions of all the fluorophores, and an additive background noise assumed to also be normally distributed around a (known) mean value b. If there are M emitters along the DNA (each of which observed with integrated intensity a_j) the total intensity at pixel x_i can be approximated by the expression

$$I_i = I(x_i|\boldsymbol{a}, \boldsymbol{\mu}) = \sum_{j=1}^{\mathrm{M}} a_j \frac{e^{-(x_i-\mu_j)^2/2\sigma_{PSF}^2}}{\sqrt{2\pi\sigma_{PSF}^2}} + \mathrm{b} \tag{3.3}$$

The intensities a_j are constrained by

$$a_0 - p\sigma_{a_0} \leq a_j \leq a_0 + p\sigma_{a_0} \tag{3.4}$$

where p is the number of standard deviations chosen for the desired interval of confidence. Also, naturally the fluorophores are restricted to lie inside the observed region,

$$x_1 \leq \mu_j \leq x_{\mathrm{N}} \tag{3.5}$$

Here, N is the number of pixels, meaning the length, of the time average. The parameters to be fitted are then the positions and intensities of all emitters, $\{\boldsymbol{a}, \boldsymbol{\mu}\} = \{a_1, ...a_M, \mu_1, ..., \mu_M\}$, and possibly also the number of emitters M. If M is unknown, it is constrained to lie between the values $M_{min/max}$, so that the sum over a_j can match the integrated intensity of the time average.

$$M_{max/min} = \frac{\sum_{i=1}^{\mathrm{N}} y_i}{a_0 \mp p\sigma_{a_0}} \tag{3.6}$$

When attempting to fit the parameters, the algorithm is run for each $M_{min} < M < M_{max}$ to chose the one that gives the smallest value for the *reduced chi-square statistic*, χ^2/ν, see 3.7. $\nu = N - 2M$ is the degrees of freedom of the model.

The weighted least squares minimization itself is performed by minimizing the function

$$\chi^2 = \sum_{i=1}^{\mathrm{N}} \frac{(y_i - I_i)^2}{\sigma_i^2} \tag{3.7}$$

with respect to the parameters $\{\boldsymbol{a}, \boldsymbol{\mu}\}$. For this, the non-linear optimization algorithm *Trust-Region-Reflective* in MATLAB(2016b) was used. The inverse weights σ_i of the data points i were calculated as the standard deviation in column i of the (aligned) kymograph. After having found the optimal parameter values $\{\boldsymbol{a}^*, \boldsymbol{\mu}^*\}$, their estimated errors were computed using an approach presented in appendix C.

4 Results

In this section we present results for the three tasks (see Section 1 and 3) addressed in this thesis.

4.1 Task 1 - Kymograph Alignment

The newly developed alignment algorithm NRAlign, as described in 3.1, yields kymographs that look very similar to the results of the previous WPAlign. In cases of either method however, it can be seen that alignment errors sometimes occur. Several instances were observed where WPAlign produced such errors, and in some of these NRAlign did as well. No clear example could however be found where NRAlign yielded a visibly worse result than WPAlign. An example of a kymograph aligned using either method can be seen in figure 7, in which it can clearly be seen that NRAlign produces a better alignment than the previous method. Please note however that in most cases, very little difference could be seen.

A more quantitative approach to testing and comparing the two methods was to generate a so called *consensus barcode*, and compare this to the individual barcodes. Consensus

barcodes are described in A.2, and can in simple terms be seen as the average of the barcodes from several molecules (meaning the average of the time averages of several aligned kymographs). They were generated using the approach presented in [17], using 10 individual barcodes for each consensus. The individual barcodes were then compared to their consensuses, for which some of the results can be seen in figure 8. The measure used for comparing two barcodes was the *similarity score*, described in appendix A.1. This score is a number between -1 and 1, where 1 corresponds to the two barcodes being identical, and 0 to them being uncorrelated.

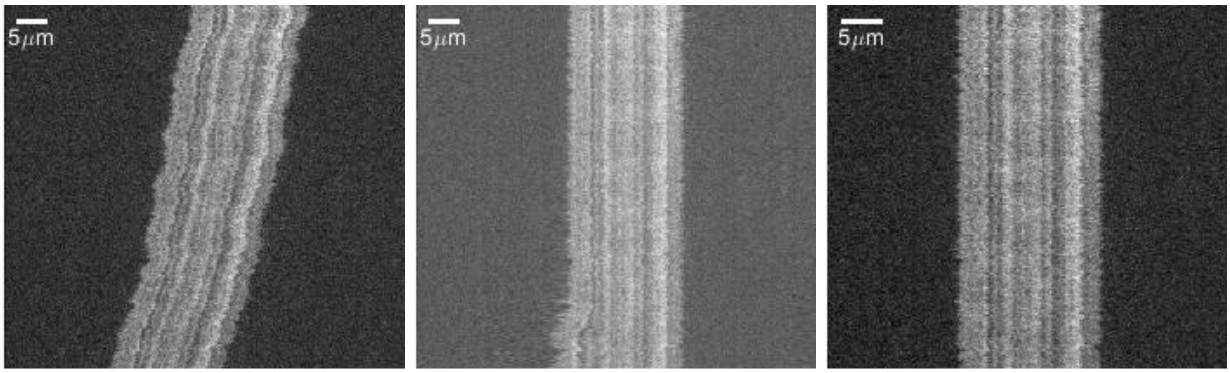

Figure 7: **Results of kymograph alignment, using both WPAlign and NRAlign.** (A) A raw, unaligned kymograph of a 500415-type plasmid (from K. pneumoniae). (B) Result after alignment using WPAlign. (C) Result after aligning using NRAlign. The difference in apparent brightness is an artifact that sometimes arises in WPAlign. Note the apparent alignment error towards the end of (B).

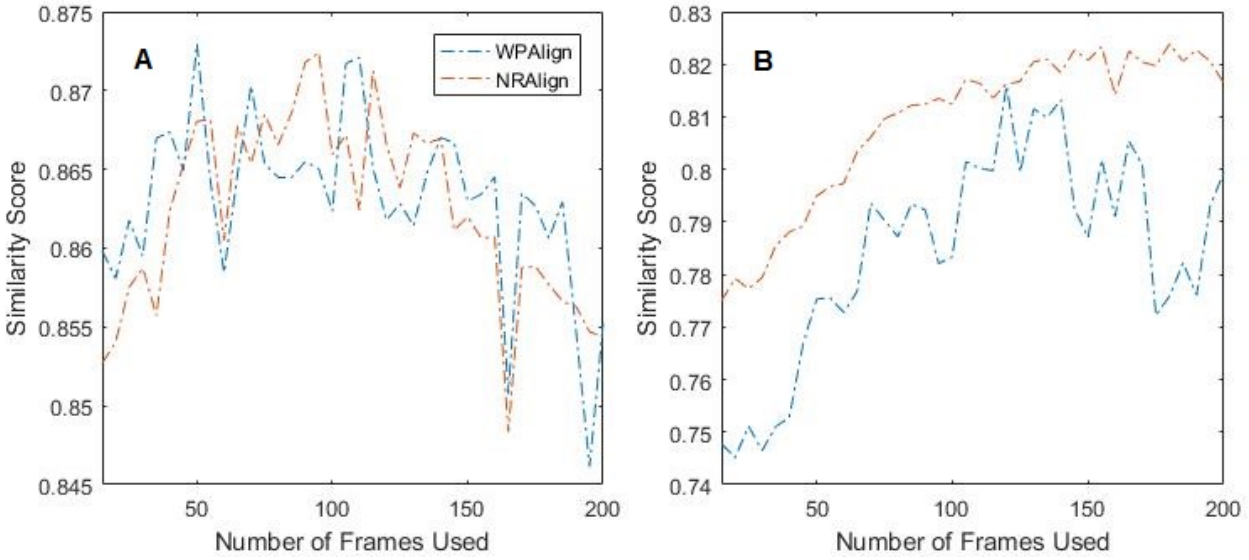

Figure 8: **Similarity scores of aligned kymographs, using both WPAlign and NRAlign.** The similarity score is shown as a function of the number of frames that was used in the kymograph. The displayed scores are averages of the results of comparing 10 aligned kymographs with a consensus barcode. (A) The results of kymographs of 49173-type plasmids. (B) The results of kymographs of 500415-type plasmids. Both plasmid types were isolated from K. pneumoniae.

In figure 8A, there is very little difference between the two methods (note the scale on the vertical axis). In 8B however, NRAlign clearly gives superior results, regardless of the number of frames used. Also, the similarity score stays somewhat more constant, indicating smaller sensitivity to details in the kymograph (such as pixels of especially low or high intensity). Steps where the similarity score drops significantly in figure 8A, such as between 160 and 165 frames, could be seen to correlate with alignment errors occurring. Also, it was clear that alignment errors occurred much more often when the unaligned kymograph showed greater global motion of the DNA molecule (see for example figure 7).

4.2 Task 2 - Reproducible Time Averages

Reproducible time averages of kymographs displaying T7 phage DNA were created using the method presented in 3.2. The similarity scores between these and a theoretical barcode were calculated. See appendix A.1 for details. The results were compared to the similarity score obtained when comparing the theoretical barcode to raw time averages of the kymographs. The difference in similarity score, as a function of the minimum segment length and maximum gap length can be seen in figure 9. In figure 10 two single time frame barcodes, as well as one raw and one reproducible time average, can be seen compared to the theoretical intensity profile of the plasmid (for the optimal choice of parameters).

The theoretical barcode used for comparisons was generated by starting with a flat intensity profile, with a length such that a resolution of 191bp/pixel was achieved. Then, Gaussian functions were added at the points along the curve corresponding to where the

relevant binding sites for the enzymatic labels are (in this case the sequence TCGA). These were determined from the known sequence of the T7 phage DNA under consideration[20]. The intensities a_j and widths σ_{PSF} were set constant for all added Gaussians.

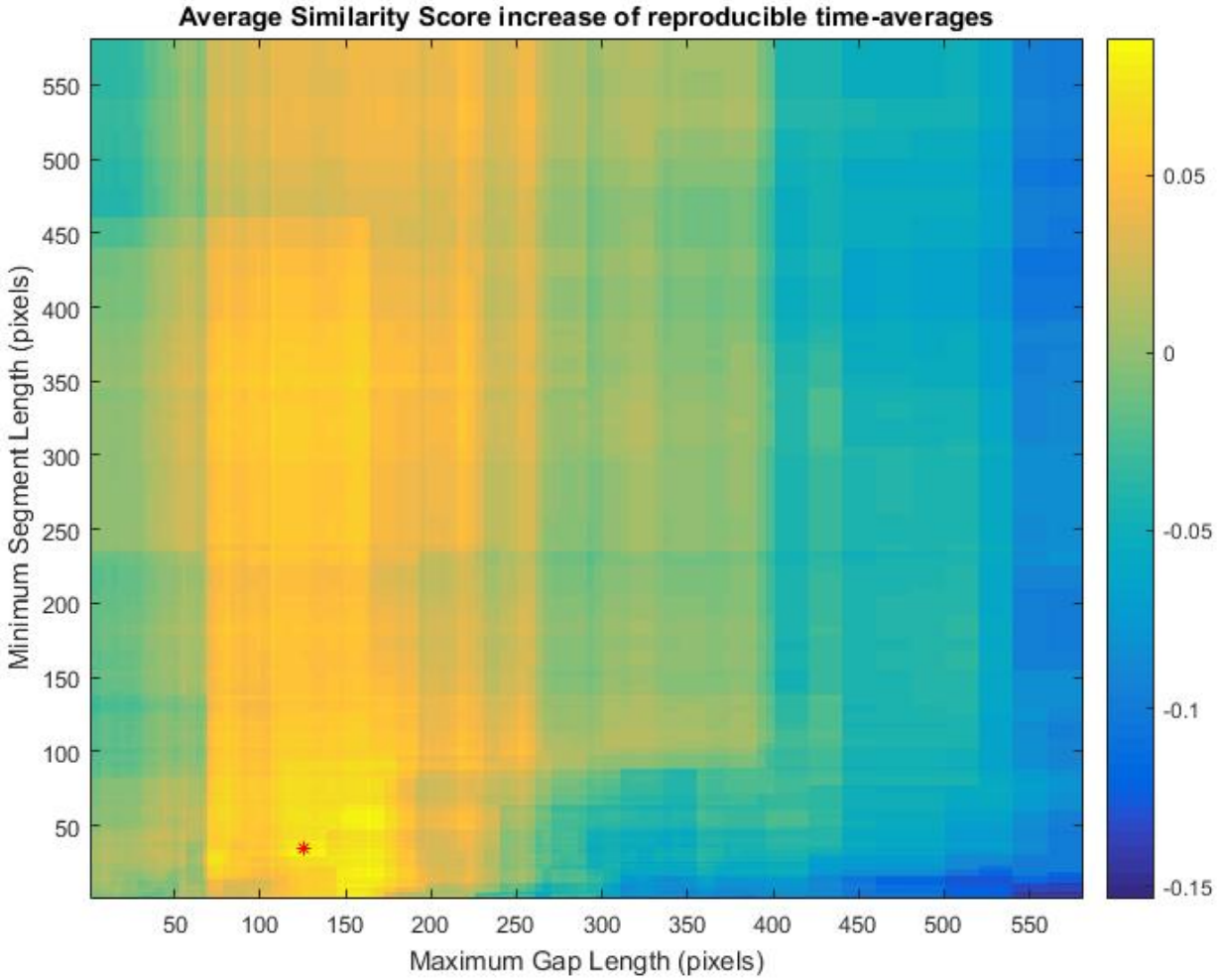

Figure 9: **Heatmap showing similarity score increase for different parameter values.** The similarity score increase is the difference between the score when using reproducible time averages (section 3.2), compared to raw averages. Brighter colors therefore correspond to improvement. Reproducible time averages were created for 4 kymographs of enzymatically labeled T7 phage DNA, and compared to a theoretical intensity curve. The highest mean increase in similarity score was 0.88, and occurred at a maximum gap length of 125, and a minimum segment length of 35 (red "*"). As can be seen, the maximum gap length affects the similarity score to a significantly larger degree than the other variable.

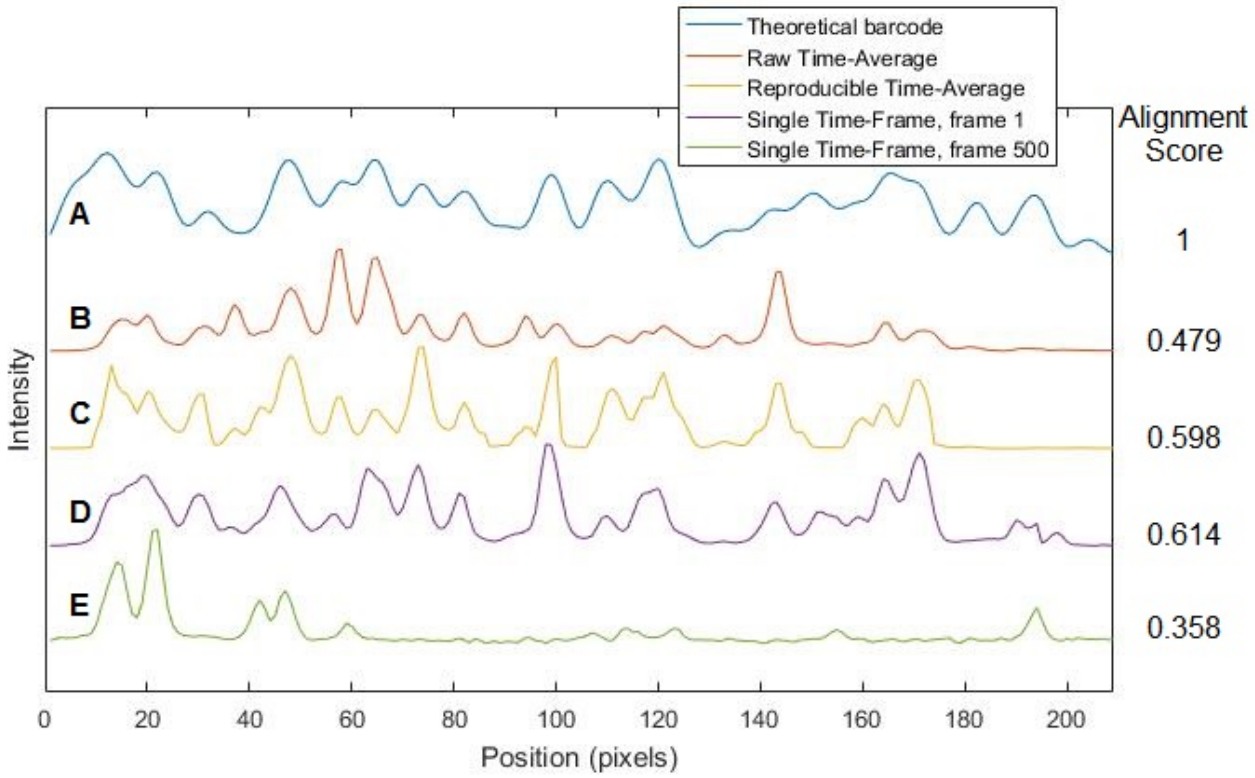

Figure 10: **Comparisons between theoretical and experimental barcodes** (A) A theoretical barcode for T7 phage DNA, with labels placed at any copy of the sequence TCGA. (B) Raw time average of a kymograph of T7 phage DNA, nick-labeled using an enzyme binding to the sequence TCGA. (C) Reproducible time average of the same kymograph. (D) Single time frame barcode, using the first frame in the kymograph. (E) Single time frame barcode, using the 500:th frame in the kymograph. The full kymograph can be seen in figure 5, and had a total of 1000 frames. Note that as T7 phage DNA is naturally linear, the barcodes have not been horizontally shifted (only flipped) when fitted to one another.

In figure 10 we can see that the reproducible time average and the first single time frame barcode look reasonably similar to the theoretical barcode. The raw time average and particularly the second single time frame barcode (frame 500) are however not very similar to the theoretical barcode at all.

Figure 11 shows the mean similarity scores of different single-frame barcodes, with increasing frame number towards the end of the kymograph. It can be seen that the first frames produce fairly good results, that are superior to the reproducible time average. After around 100 frames however, individual frames are not very reliable, and quickly become worse than the raw time average.

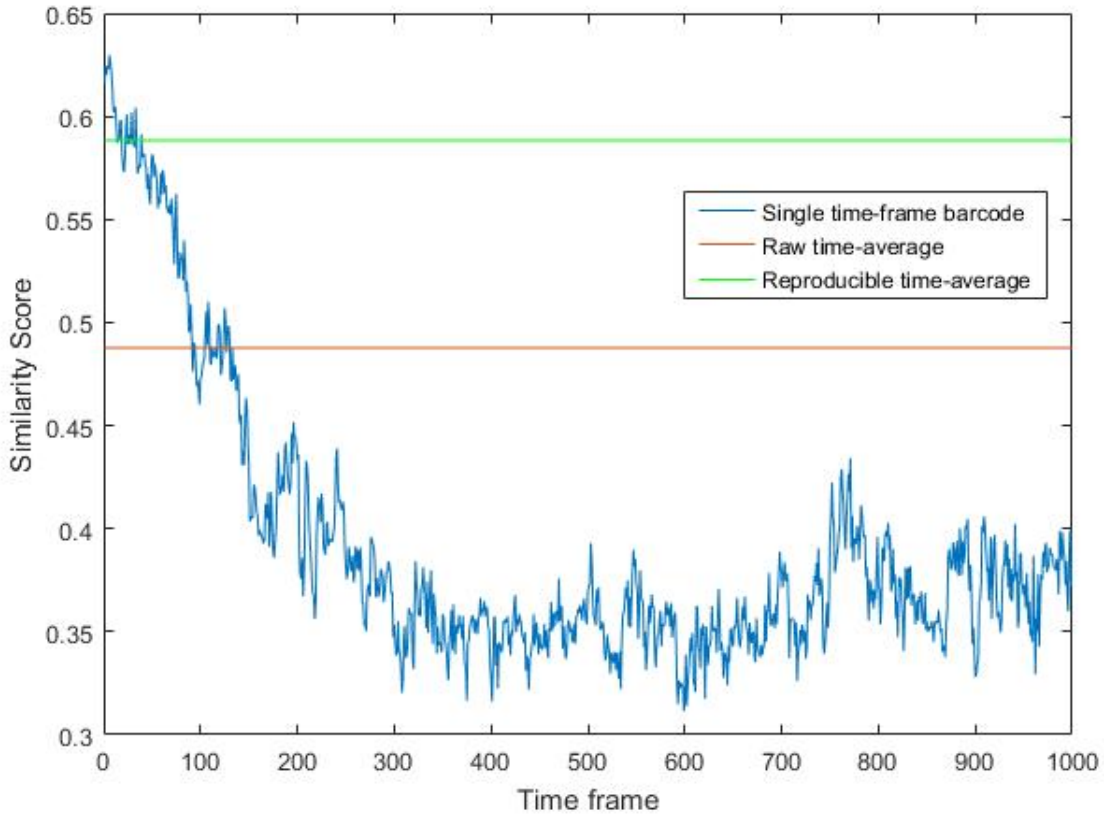

Figure 11: **Similarity scores of single-frame barcodes, raw time averages and reproducible time averages when compared to a theory barcode.** The plot shows the average similarity score from 4 kymographs of T7 phage DNA. For the single time frame barcodes, one row (or frame) from the kymograph was simply extracted, and compared to the theoretical sequence. The reproducible time averages were generated using a maximum gap size of 125 pixels, and a minimum segment length of 35 pixels. Note that perfectly matching barcodes would have similarity score 1, while uncorrelated ones would have score 0.

4.3 Task 3 - Weighted Least Squares Fitting

The weighted least squares approach to determining the positions of enzymatic labels on sparsely labeled DNA could unfortunately not be tested on experimental data. The barcodes provided by Robert Neely, University of Birmingham, were created from T7 phage DNA, which has 107 TCGI-loci to which the enzymatic labels can bind. Fitting 107 Gaussians to such a barcode (one for each loci) would be much too computationally demanding. Therefore a theoretical barcode was generated, based on the sequence of the last 7527 bp of T7 phage DNA[20]. This piece contains 13 binding loci, which was much more manageable.

The theoretical barcode was generated in the same manner as in section 4.2. To simulate an experiment, we here also added Gaussian noise with mean zero, and standard deviation equal to the standard deviation of the background in one of the kymographs obtained from Robert Neely (107.9 a.u.).

The fitting to the theoretical curve can be seen in figure 12, and the positions of the fitted peaks can be seen in table 1.

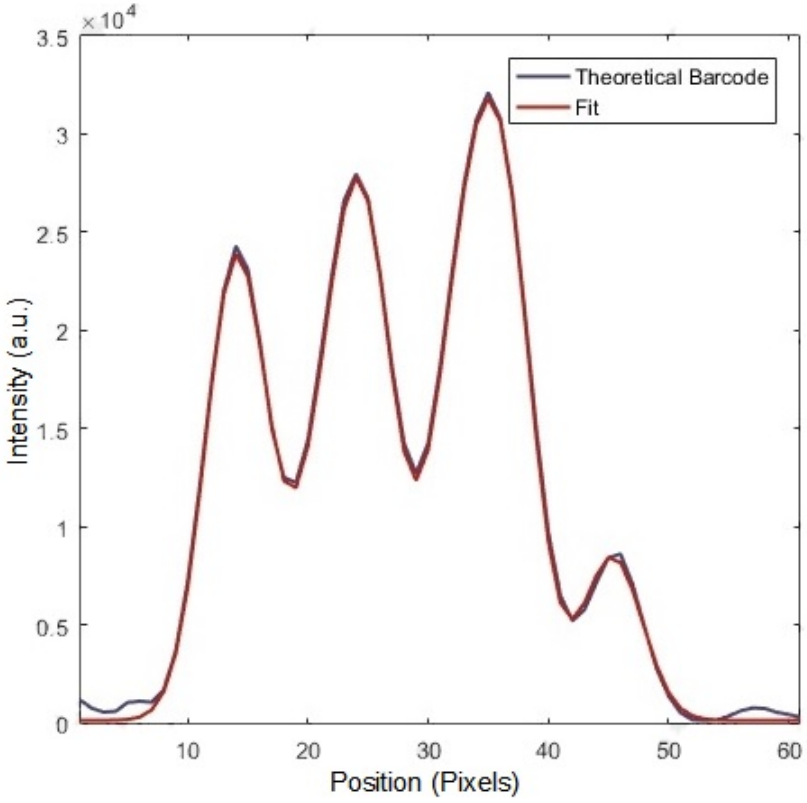

Figure 12: **Fit to theoretical barcode of partial T7 phage DNA.** The theoretical curve was generated based on a part of the sequence of T7 phage DNA[20], containing 13 TCGA loci, to which enzymatic labels could bind. The area per fitted peak was 51530±5000 a.u.×pixels, the mean background intensity was 126.2 a.u., and the width of the point spread function was 2.473 pixels. The noise added to the theoretical barcode before fitting had standard deviation 107.9 a.u. The fit was performed according to section 3.3, with resulting fluorophore positions as seen in table 1. The range of allowed values for the area per peak allowed for 12-14 peaks, but the best fit was found for 13 peaks (see section 3.3).

Table 1: **Results of fitting.** Fitted and true positions of fluorophores on a segment of T7 phage DNA. Note that while the fitted values lie fairly close to the true values, only two (peak #1 and #7) are so close that they lie within one standard deviation of one another.

Peak index	Fitted Position	True Position
1	13.546 ± 0.037	13.544
2	13.549 ± 0.116	14.340
3	15.270 ± 0.095	14.870
4	21.242 ± 0.051	21.315
5	23.704 ± 0.004	24.577
6	24.722 ± 0.051	24.629
7	25.500 ± 0.167	25.534
8	32.528 ± 0.004	32.795
9	32.528 ± 0.028	33.243
10	35.547 ± 0.028	36.223
11	36.605 ± 0.102	36.972
12	36.605 ± 0.030	37.040
13	45.350 ± 0.066	46.117

Only two of the peaks lie within one standard deviation of the true positions, while typically 67% should do so given that the model is correct. Four peaks lie within two standard deviations, while typically 94% should do so. While the fit in figure 12 looks very successful, the deviations from the correct positions indicate that the procedure is somehow flawed. It is however noteworthy that the procedure yielded the correct number of peaks, even though the range of allowed peak intensities was fairly large. 12 or 14 peaks would also have satisfied condition 3.6, but 13 yielded the best fit. This should be considered a significant success as the number of bound fluorophores, rather than their exact positions, is the most relevant result.

5 Discussion

In this thesis, three different tasks have been addressed. Here follows three sections in which the results are discussed, and possible improvements are suggested.

5.1 Kymograph Alignment

Based both on the better matching between consensus- and individual barcodes when using NRAlign compared to WPAlign, and the more rare occurrence of visible alignment errors, the newly improved algorithm seems to yield superior results to the previous one. Still, it was clear that alignment errors still occur in NRAlign. Tweaking the image pre-processing might improve the results somewhat, but for more significant progress a more sophisticated

way of determining the "reasonability" of detected features should be used. For example, features could be discarded if alignment would lead to highly non-uniform stretching.

The reason why center-of-mass alignment was necessary in NRAlign, but was not previously used in WPAlign, seems to be due to a "built-in" feature of WPAlign, taking its place. In kymographs where significant global motion occurs, straightening a single feature detected in WPAlign leads to the rest of the kymograph also being partly straightened.

The parameter k, describing by how many columns a feature can shift sideways from row to row, was set to 1 in our testing. According to [3] however, the value 2 would be recommended for the type of data used here. When attempting alignment with $k = 2$ in NRAlign, it quickly became clear that the main change was that alignment errors became more common. Without using center-of-mass alignment, the net effect of a larger k might have been more clearly positive.

5.2 Reproducible Time Averages

Our algorithm for producing reproducible time averages of sparsely labeled barcodes shows clear signs of being better than simply taking a raw time average of the barcode. This can clearly be seen in figure 9, where the similarity score with a theoretical barcode increases by almost 0.1 for some parameters. Figure 10 also show that the reproducible time average looks much more like the theoretical barcode.

A possible way of significantly improving the process could be to try and identify how many fluorophores are emitting at different points in the kymograph. In our approach, all parts of the signal region are treated equally as long as the intensity is above the threshold value. For example, one particular column can be considered, where two fluorophores are emitting during the first half of the kymograph, but only one during the second half. The reproducible time average would include the entire column, yielding an intensity in between that of one and two fluorophores.

5.3 Weighted Least Squares Fitting

The weighted least squares fitting of Gaussians to a theoretical barcode, as seen in figure 12, is visually appealing. However, as the positions of the fitted peaks deviate more than predicted from the true positions (table 1), it is reasonable to conclude that some of the assumptions made were not strictly correct. Notably, the parameter error estimation presented in appendix C was done under the assumption that there was no correlation between pixels. In reality there most certainly is, meaning that the number of effective data points is smaller than the number of pixels. Therefore, a better method for computing error estimations for the parameters should be developed, taking correlations between data points into account. In addition, a measure of the uncertainty in the number of peaks would be highly useful, as it is the number of peaks rather than their exact positions that are of the greatest relevance.

6 Conclusions

The results of all three main tasks addressed in this thesis showed some promise. Firstly, the new alignment algorithm, NRAlign, tends to produce better results than the previously developed WPAlign, without displaying any clear drawbacks. In cases where kymographs show little global motion of the molecule, the two algorithms produce very similar results, but when more significant global motion is present NRAlign seems significantly superior.

The method for creating reproducible time averages would seem to generate better and more reliable barcodes of sparsely labeled kymographs, than simply taking raw time averages.

The fitting of Gaussians also seems to work as intended, but this was much harder to test as no data of kymographs having few enough fluorophores was available. Applying the method to a theoretical barcode yielded the correct number of fluorophores, but their fitted positions generally deviated by more than one standard deviation from the true positions. As the most sought after result was the number of fluorophores (or rather the sequence they bind to), rather than their exact positions, this result can however also be considered satisfactory.

A natural next step would be to apply the procedure, including all three tasks, to plasmids that are simultaneously labeled with both YOYO-1/netropsin and enzymatic labels. This would definitively show if it is possible to identify both plasmid type and the presence of specific genes in a single experiment, using our approach. Our result show promise for this, but improvements may have to be developed, particularly for generating reproducible time averages of sparsely labeled barcodes.

A Similarity Score & Consensus Barcodes

A.1 Similarity Score

The measure used for quantifying the similarity between two barcodes is the *Similarity score*, closely related to the *Pearson correlation coefficient*. The latter is defined as

$$C(i_{shift}) = \frac{1}{\mathrm{N}-1} \sum_{i=1}^{\mathrm{N}} Y_1^*(i) \cdot Y_2(i + i_{shift} - 1) \tag{A.1}$$

where Y_1 and Y_2 are the two barcodes, renormalized so that they both have mean zero and standard deviation 1. They are stretched to the same length (N pixels), meaning it is assumed that no part of either molecule is missing. The factor before the sum in eq. A.1 ensures normalization so that $-1 \leq C \leq 1$. Due to the cyclic nature of plasmids and the fact that they may have been cut at any point along their contours, one is circularly shifted along the other by changing i_{shift} to find the optimal fit.

This is achieved by first repeating Y_1 twice, forming Y_1^*. The length of Y_2 is tripled by adding zeros to the ends. Circular shifting can now be simulated by linearly shifting Y_2 along Y_1^*. All values of i_{shift} between 1 and N are then tried, and the best fit is denoted $\widehat{C} = \max(C(i_{shift}))$. As the directions of the plasmids are unknown, $\widehat{C}$ is also calculated with one of the curves horizontally inverted. The orientation generating the highest value is considered correct, and this $\widehat{C}$ is what we refer to as the similarity score between the two barcodes.

A.2 Consensus Barcodes

When generating experimental barcodes, results may vary due to differing experimental conditions or variations in the genetic content of the DNA under consideration. In order to avoid working with an uncharacteristic barcode, a *Consensus barcode* can be generated from a set of several individual barcodes. This was for example done to verify the results of task 1, where 10 individual barcodes were compared to a consensus, using either the newly developed NRAlign to the previously developed WPAlign[3]. See section 4.1.

We used an automatic procedure for creating consensus barcodes, developed by [17]. It consists of pairwise comparison of the individual barcodes, to find the best similarity scores. After repeated reorientation, all barcodes will be optimally shifted and inverted relative to one another. The consensus is then calculated by taking the average of each barcode. See [17] for details. The consensus barcode will then be more reliable than the individual ones, for use as a "fingerprint" identifying for example a plasmid.

B Creating Masks for Reproducible Time Averages

To create a mask covering the signal regions of a kymograph, an image thresholding scheme using a multilevel Otsu method was used[23][24]. Three intensity threshold values were found, separating the kymograph into 4 levels. The lowest of those should correspond to the background level, and the others to 1, 2, and 3 fluorophores overlapping, respectively. In that sense, the thresholding should work if there is no more than 3 overlapping fluorophores at any point in the kymograph.

All except the lowest level is initially considered to constitute the signal regions of the kymographs. However, scattered about there will usually be several small fragments surrounded by background, and several small "holes" in the signal regions, where the intensity is locally lower than in the surroundings. Two parameters, the *minimum segment length* and the *maximum gap length*, determine how many rows a part of the signal region has to cover in order not to be discarded, and how short a gap has to be in order to included in the signal region, respectively.

A number of morphological operations are now performed on the mask, illustrated in figure 13. Morphological operations are procedures performed on binary images, usually to reduce noise[25]. To begin, gaps in the signal regions are filled by morphologically closing the mask, using a rectangular structuring element to remove any gaps that are one pixel wide, and have a length smaller than the maximum gap length. It is then morphologically opened, removing any segments that are narrower than the width of the point spread function, σ_{PSF}, or shorter than the minimum segment length. The resulting mask is then morphologically dilated, horizontally widening any feature by $\sigma_{PSF}/2$. Lastly, any remaining gaps being one pixel wide or long, are filled through morphological closing. As σ_{PSF} is usually an non-integer, terms including this factor is rounded to the nearest integer, before use in the morphological operations just described. For details on opening, closing, or dilating, see [25].

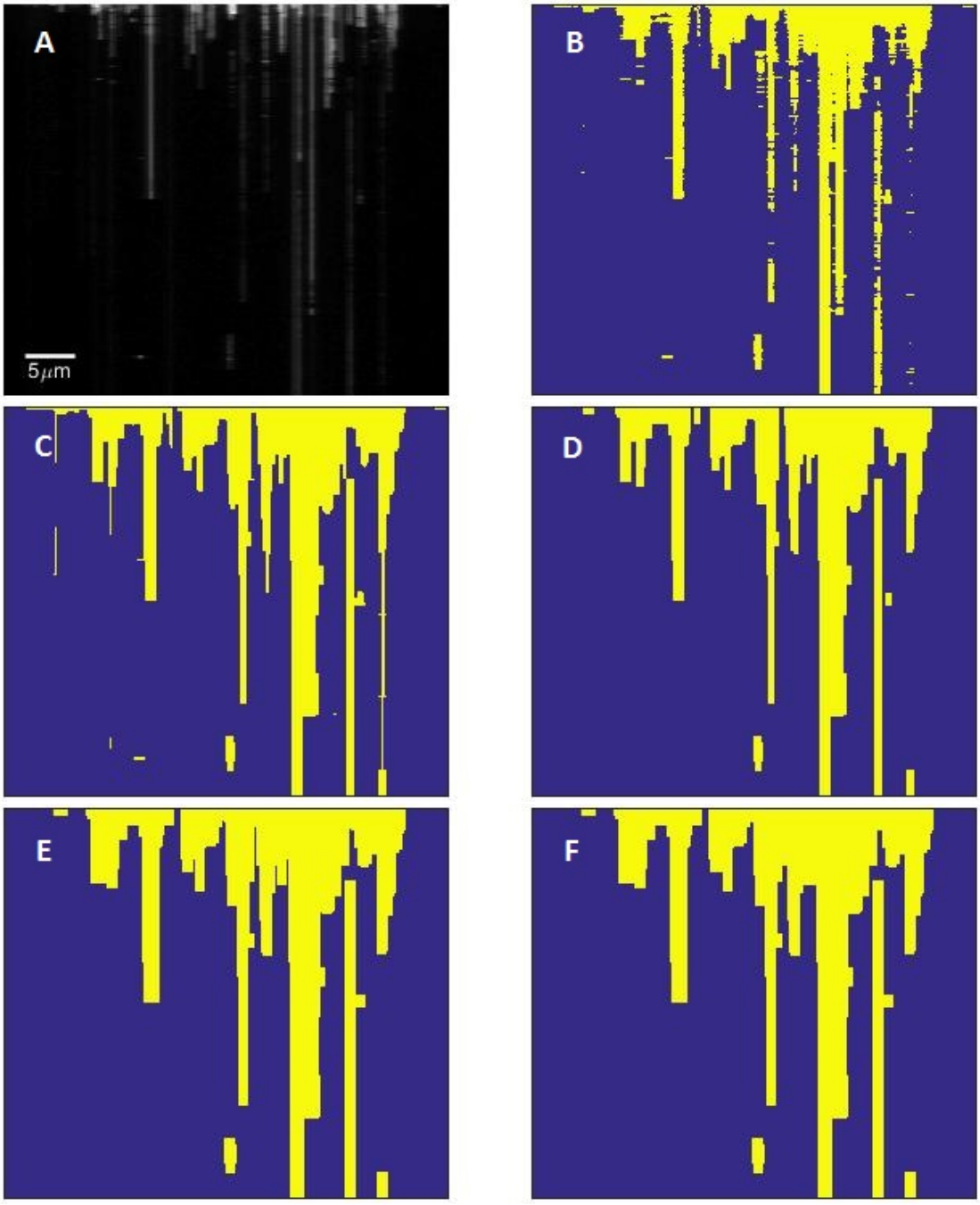

Figure 13: Illustration of the morphological operations performed to generate a reliable mask, to cover what should be considered the signal regions of a kymograph. (A) The original kymograph. (B) The mask, after having included all pixels except the background level. (C) The mask after morphological closing, using a maximum gap length of 125. (D) The mask after morphological opening, using a minimum segment length of 35. (E) The mask after morphological dilation. (F) The final mask, after having morphologically closed remaining 1-pixel gaps.

C Parameter Errors in Weighted Least Squares Estimation

In weighted least squares estimations, previously described in section 3.3, optimal parameter values $\boldsymbol{\beta}^*$ are found by minimizing the chi-square function

$$\chi^2 = \sum_{i=1}^{\mathrm{N}} \frac{(y_i - I(x_i|\boldsymbol{\beta}))^2}{\sigma_i^2} \tag{C.1}$$

with respect to the parameters $\boldsymbol{\beta}$ of the model I. $\boldsymbol{y} = \{y_1, ..., y_{\mathrm{N}}\}$ are the experimental data points at positions $\boldsymbol{x} = \{x_1, ..., x_{\mathrm{N}}\}$. The least squares estimation corresponds to maximizing the so called *likelihood function*, L. The likelihood of a set of parameters $\boldsymbol{\beta}$, given the data $\boldsymbol{y}$, equals the probability of obtaining those data points given those parameter values,

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{\beta}) \tag{C.2}$$

In our case, the probability of obtaining the set of data $\boldsymbol{y}$ is the product of the probabilities of obtaining each individual point y_i,

$$P(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\mu}) = \prod_{i=1}^{\mathrm{N}} P(y_i|\boldsymbol{a}, \boldsymbol{\mu}) = \prod_{i=1}^{\mathrm{N}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{\left(y_i - I(x_i|\boldsymbol{a},\boldsymbol{\mu})\right)^2}{2\sigma_i^2}} \tag{C.3}$$

$I(x_i|\boldsymbol{a}, \boldsymbol{\mu}))$ is the predicted intensity at point x_i, given by eq. 3.3, and y_i is the experimental intensity at the same point. The set of parameters $\{\boldsymbol{\beta}\} = \{\boldsymbol{a}, \boldsymbol{\mu}\}$ are the intensities and positions of the emitters respectively. In the last step of eq. C.3 we have assumed that the experimental noise is normally distributed with mean zero and known standard deviations σ_i at each point (standard deviation in column i of the kymograph).We also assumed that fluctuations around mean values are uncorrelated between pixels, and that intensity values at different time frames are independent quantities. Now we can further rewrite C.3;

$$\prod_{i=1}^{\mathrm{N}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - I(x_i|\boldsymbol{a},\boldsymbol{\mu}))^2}{2\sigma_i^2}} = \left[\prod_{i=1}^{\mathrm{N}} \frac{1}{\sqrt{2\pi}\sigma_i}\right] \exp\left(-\frac{1}{2}\sum_{i=1}^{\mathrm{N}} \frac{(y_i - I(x_i|\boldsymbol{a}, \boldsymbol{\mu}))^2}{\sigma_i^2}\right) = \mathrm{W} \exp\left(-\frac{1}{2}\chi^2\right) \tag{C.4}$$

In the last step we have used eq.C.1, and $\mathrm{W} = \prod_i^{\mathrm{N}} \left(1/\sqrt{2\pi}\sigma_i\right)$. It is now evident that minimizing χ^2 is equivalent to maximizing the likelihood function L, as the exponential function is monotonic.

We now Taylor expand χ^2 around the optimal parameter values $\{\boldsymbol{\beta}^*\}$. We consider only the terms up to second order:

$$\chi^2 = \chi^2\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}*} + \sum_{a=1}^{2\mathrm{M}}(\beta_a - \beta_a^*)\frac{\partial\chi^2}{\partial\beta_a}\Big|_{\beta_a=\beta_a^*} + \frac{1}{2}\sum_{a,b=1}^{2\mathrm{M}}(\beta_a - \beta_a^*)(\beta_b - \beta_b^*)\frac{\partial^2\chi^2}{\partial\beta_a\partial\beta_b}\Big|_{\substack{\beta_a=\beta_a^*\\ \beta_b=\beta_b^*}} \tag{C.5}$$

This can be inserted into C.4, yielding

$$P(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{\mu}) = \mathrm{W}\exp\left(-\frac{1}{2}\chi^2\right) = \mathrm{W}^*\exp\left(-\frac{1}{4}\sum_{a,b=1}^{2\mathrm{M}}\boldsymbol{H}_{ab}(\beta_a - \beta_a^*)(\beta_b - \beta_b^*)\right) \tag{C.6}$$

where

$$\boldsymbol{H}_{ab} = \frac{\partial^2\chi^2}{\partial\beta_a\partial\beta_b}\Big|_{\substack{\beta_a=\beta_a^*\\ \beta_b=\beta_b^*}} \tag{C.7}$$

is the Hessian matrix, and we have used that the first order derivatives of χ^2 are zero at $\boldsymbol{\beta} = \boldsymbol{\beta^*}$. The constant term in the Taylor expansion has been absorbed into W^*. Using eq. C.6, it can be shown that

$$\langle(\beta_a^* - \widehat{\beta}_a)(\beta_b^* - \widehat{\beta}_b)\rangle = 2(\boldsymbol{H}^{-1})_{ab} \tag{C.8}$$

Here "$\widehat{\ }$" denotes the true value of a parameter. Eq. C.8 means that the inverse of the Hessian matrix yields the covariance of the estimated parameters. The standard deviations of the parameters j is then given by the square root of the diagonal elements of the inverted matrix,

$$\sigma_{\beta_j} = \sqrt{2(\boldsymbol{H}^{-1})_{jj}} \tag{C.9}$$

To calculate the Hessian, we need the second order derivatives of χ^2 with respect to the parameters. We have

$$\frac{\partial\chi^2}{\partial\beta_a} = -2\sum_{i=1}^{\mathrm{N}}\frac{y_i - I_i}{\sigma_i^2}\frac{\partial I_i}{\partial\beta_a} \tag{C.10}$$

$$\frac{\partial^2\chi^2}{\partial\beta_a\partial\beta_b} = -2\sum_{i=1}^{\mathrm{N}}\left(\frac{y_i - I_i}{\sigma_i^2}\frac{\partial^2 I_i}{\partial\beta_a\partial\beta_b} - \frac{1}{\sigma_i^2}\frac{\partial I_i}{\partial\beta_a}\frac{\partial I_i}{\partial\beta_b}\right) \tag{C.11}$$

The derivatives of the model function

$$I_i = I(x_i|\boldsymbol{a},\boldsymbol{\mu}) = \sum_{j=1}^{\mathrm{M}} a_j\frac{e^{-(x_i-\mu_j)^2/2\sigma_{PSF}^2}}{\sqrt{2\pi\sigma_{PSF}^2}} + \mathrm{b} \tag{C.12}$$

are given by

$$\frac{\partial I_i}{\partial a_j} = \frac{\partial I}{\partial a_j}\bigg|_{x_i} = \frac{e^{-(x_i-\mu_j)^2/2\sigma^2_{PSF}}}{\sqrt{2\pi\sigma^2_{PSF}}} \tag{C.13}$$

$$\frac{\partial I_i}{\partial \mu_j} = \frac{\partial I}{\partial \mu_j}\bigg|_{x_i} = a_j \frac{x_i-\mu_j}{\sigma^2_{PSF}} \frac{e^{-(x_i-\mu_j)^2/2\sigma^2_{PSF}}}{\sqrt{2\pi\sigma^2_{PSF}}} \tag{C.14}$$

$$\frac{\partial^2 I_i}{\partial a_j \partial \mu_j} = \frac{\partial^2 I}{\partial a_j \partial \mu_j}\bigg|_{x_i} = \frac{x_i-\mu_j}{\sigma^2_{PSF}} \frac{e^{-(x_i-\mu_j)^2/2\sigma^2_{PSF}}}{\sqrt{2\pi\sigma^2_{PSF}}} \tag{C.15}$$

$$\begin{aligned}\frac{\partial^2 I_i}{\partial \mu_j^2} = \frac{\partial^2 I}{\partial \mu_j^2}\bigg|_{x_i} &= -a_j \frac{1}{\sigma^2_{PSF}} \frac{e^{-(x_i-\mu_j)^2/2\sigma^2_{PSF}}}{\sqrt{2\pi\sigma^2_{PSF}}} + a_j \frac{(x_i-\mu_j)^2}{\sigma^4_{PSF}} \frac{e^{-(x_i-\mu_j)^2/2\sigma^2_{PSF}}}{\sqrt{2\pi\sigma^2_{PSF}}} = \\ &= \frac{a_j}{\sigma^2_{PSF}} \left(\frac{(x_i-\mu_j)^2}{\sigma^2_{PSF}} - 1 \right) \frac{e^{-(x_i-\mu_j)^2/2\sigma^2_{PSF}}}{\sqrt{2\pi\sigma^2_{PSF}}}\end{aligned} \tag{C.16}$$

All other second order derivatives are zero. To calculate the elements of the Hessian (eq. C.7), the expressions C.11-C.16 all have to be used.

References

[1] Y. Michaeli and Y. Ebenstein. (2012). Channeling DNA for optical mapping. Nat. Biotechnol. 30: 762–763.

[2] V. Müller, F. Rajer, K. Frykholm, L. K. Nyberg, S. Quaderi, J. Fritzsche, E. Kristiansson, T. Ambjornsson, L. Sandegren, and F. Westerlund. (2016). Direct identification of antibiotic resistance genes on single plasmid molecules using CRISPR/Cas9 in combination with optical DNA mapping. Scientific Reports. 6:37938.

[3] C. Noble, A.N. Nilsson, C. Freitag, J.P. Beech, J.O. Tegenfeldt, T. Ambjörnsson. (2015) A Fast and Scalable Kymograph Alignment Algorithm for Nanochannel-Based Optical DNA Mappings. PLoS ONE 10(4):e0121905.

[4] L. K. Nyberg, S. Quaderi, G. Emilsson, N. Karami, E. Lagerstedt, V. Müller, C. Noble, S. Hammarberg, A. N. Nilsson, F. Sjöberg, J. Fritzsche, E. Kristiansson, L. Sandegren, T. Ambjörnsson, and F. Westerlund. (2016). Rapid identification of intact bacterial resistance plasmids via optical mapping of single DNA molecules. Scientific Reports 6:30410.

[5] V. Müller, N. Karami, L. K. Nyberg, C. Pichler, P. C. Torche Pedreschi, S. Quaderi, J. Fritzsche, T. Ambjörnsson, C. Åhrén, and F. Westerlund. (2016), Rapid Tracing of Resistance Plasmids in a Nosocomial Outbreak Using Optical DNA Mapping. ACS Infectious Diseases 2, 322–328.

[6] A. Samad, E.F. Huff, W. Cai, and D.C. Schwartz. (1995) Optical mapping: a novel, single-molecule approach to genomic analysis. Genome research, 5(1): 1–4.

[7] C.-h Lee, and C.-c Hsieh. (2013) Stretching DNA by electric field and flow field in microfluidic devices: An experimental validation to the devices designed with computer simulations. Biomicrofluidics. 7:14109.

[8] J. F. Marko and E. D. Siggia. (1995) Stretcing DNA. Macromolecules. 28, 8759-8770.

[9] M. D. Wang, H. Yin, R. Landick, J. Gelles, and S. M. Block. (1997) Stretching DNA with optical tweezers. Biophys J. 72(3): 1335–1346.

[10] W. Reisner, J. N. Pedersen, and R. H. Austin. (2012) DNA confinement in nanochannels: physics and biological applications. Rep. Prog. Phys. 75, 106601

[11] R. K. Neely, J. Deen, and J. Hofkens. (2011) Optical mapping of DNA: Single-molecule-based methods for mapping genomes. Biopolymers 95.5:298-311.

[12] R. L. Welch, R. Sladek, K. Dewar and W. W. Reisner. (2012), Denaturation mapping of Saccharomyces cerevisiae. Lab Chip, 12, 3314–3321.

[13] L. K. Nyberg, F. Persson, J. Berg, J. Bergström, E. Fransson, L. Olsson, M. Persson, A. Stålnacke, J. Wigenius, J.O. Tegenfeldt, F. Westerlund. (2012). A single-step competitive binding assay for mapping of single DNA molecules. Biochemical and Biophysical Research Communications 417, 404–408.

[14] W. Reisner, N. B. Larsen, A. Silahtaroglu, A. Kristensen, N. Tommerup, J. O. Tegenfeldt, and H. Flyvbjerg. (2010) Single-molecule denaturation mapping of DNA in nanofluidic channels. Proceedings of the National Academy of Sciences, 107(30):13294–13299.

[15] D. C. Schwartz, X. Li, L. I. Hernandez, S. P. Ramnarain, E. J. Huff, and Y. Wang. (1993) Ordered restriction maps of saccharomyces cerevisiae chromosomes constructed by optical mapping. Science, 262(5130):110–114.

[16] S. K. Das, M. D. Austin, M. C. Akana, P. Deshpande, H. Cao, and M. Xiao. (2010) Single molecule linear analysis of dna in nano-channel labeled with sequence specific fluorescent probes. Nucleic acids research, 38(18):e177–e177.

[17] E. Lagerstedt. (2014) Nanochannel-based DNA Barcoding for Plasmid Characterisation Theoretical Transfer Matrix Calculations, Bio-informatics Tools and Joint Theory-Experiments. LUP Student Papers `http://lup.lub.lu.se/student-papers/record/4647948`

[18] J. Temsamani, and S. Agrawal. (1996) Enzymatic labeling of nucleic acids. Mol. Biotech. 5:223–232.

[19] P. Frantsuzov, M. Kuno, B. Jankó, and R. A. Marcus. (2008). Universal emission intermittency in quantum dots, nanorods and nanowires. Nature Physics. 4: 519–522.

[20] NCBI Reference Sequence: NC 001604.1

[21] A. J. Berglund. (2004) Nonexponential statistics of fluorescence photobleaching. Journal of Chemical Physics 121:2899-2903.

[22] J. R. Lakowicz, Principles of Fluorescence Spectroscopy (3rd ed.). New York: Springer Science+Business Media. p. 41.

[23] N. Otsu. (1979). A threshold selection method from gray-level histograms. IEEE Trans. Sys., Man., Cyber. 9 (1): 62–66.

[24] P.-s Liao, T.-s Chen and P.-c Chung. (2001) A Fast Algorithm for Multilevel Thresholding. J. Inf. Sci. Eng. 17 (5):713–727.

[25] J. Serra. (1982) Image Analysis and Mathematical Morphology, Volume 1, Academic Press.