UNIVERSIDAD NACIONAL DE COLOMBIA

A Computational Methodology for the Generation of Genomic Maps from Fluoroscanning Images

Alberto Mario Ceballos-Arroyo

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión.
Medellín, Colombia
2022

A Computational Methodology for the Generation of Genomic Maps from Fluoroscanning Images

Alberto Mario Ceballos-Arroyo

Tesis presentada como requisito parcial para optar al título de:
Magíster en Ingeniería de Sistemas

Director:
Ph.D., Juan Pablo Hernandez Ortiz
Línea de Investigación:
Visión Artificial, Biología Computacional
Grupo de Investigación:
Center for Research and Surveillance of Tropical and Infectious Diseases (CRS-TID)

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2022

Acknowledgements

Much of the work done as part of this thesis was supervised by Professors Juan Pablo Hernandez Ortiz and David C. Schwartz. Their profound knowledge and careful guidance led to many exciting discussions that helped me translate my Computer Science expertise into solutions that now help carry out experiments in the Laboratory for Molecular and Computational Genomics (LMCG) at the University of Wisconsin - Madison. Without them, this work would not have been possible.

An equally important part of my work was the result of close collaboration with LMCG Ph.D. students and interns such as Yumin, Alejandro, Wenkai, Susana, and Eamon. In particular, this work relied heavily on the laboratory experiments carried out by Yumin, the theoretical insight and computational implementations provided by Alejandro, and Wenkai's first version of WScan that was the basis for the first chapter of this thesis.

I must also thank my dear wife Geraldine, who was always present during my time as a M.Sc. student and accompanied me through many moments of happiness and sadness. My father Gustavo, my step-mother Yolanda, my aunt Andrea, as well as my siblings Alejandro, Angelica, Juan Carlos, Gustavo, Fabian, and Juan David were also part of this achievement by supporting me as I moved from my small town to Medellín to undertake this journey.

Finally, I want to thank the Universidad Nacional de Colombia for receiving me as a M.Sc. student and funding my internship at the University of Wisconsin-Madison, through which I was able to collaborate much more closely with the folks at LMCG.

Una metodología computacional para la generación de mapas genómicos a partir de imágenes de Fluoroscanning

Abstract

Fluoroscanning is a novel system for quickly generating genomic maps. Unlike preceding systems like optical mapping and nanocoding, Fluoroscanning relies only on the intensity signals produced by dye fluorochromes when bound to DNA nucleotides, which we deem Fluoroscans. As part of this work, we wanted to develop and evaluated a fast digital image processing pipeline for extracting Fluoroscan signals from fluorescence microscopy images, to devise and implement a parallel and highly optimized algorithm for simulating the physical principles behind Fluoroscanning, and to guide laboratory experiments using such a tool in order to enable the generation of genomic maps through alignment algorithms. As a result of our work, we were able to set up a workflow in which real Fluoroscans extracted from digital images were used to adjust the parameters of a Monte Carlo simulation of Fluoroscanning which was then leveraged to guide further laboratory experiments and to generate a synthetic human-genome-scale dataset which will enable the development of signal alignment algorithms for genomic map generation.

Keywords: image processing, DNA, genomics, simulations, signal processing.

Resumen

El Fluoroscanning es un sistema novedoso para la generación rápida de mapas genómicos. A diferencia de sistemas anteriores como el optical mapping y el nanocoding, el Fluoroscanning solo se basa en la intensidad de las señales (que llamamos Fluoroscans) producidas por fluorocromos de tinte cuando se adhieren a nucleótidos de ADN. Como parte de este trabajo, se desarrolla y se evalúa una serie de pasos que incluyen procesamiento de imágenes para extraer señales Fluoroscan de manera rápida a partir de imágenes de microscopía de fluorescencia, un algoritmo paralelo y altamente optimizado para simular los principios físicos detrás del Fluoroscanning y una metodología para guiar experimentos de laboratorio a partir de dicho algoritmo. Como resultado de nuestro trabajo, pudimos establecer un flujo de trabajo en el que Fluoroscans reales extraídos de imágenes digitales se utilizaron para ajustar los parámetros de las simulaciones, que a su vez fueron utilizadas para guiar experimentos de laboratorio y para generar un conjunto de datos sintético a escala genómica que permitirá ayudar al desarrollo de algoritmos de alineamiento de señales para la generación de mapas genómicos.

Palabras clave: procesamiento de imágenes, ADN, genómica, simulaciones, procesamiento de señales.

Contents

Acknowledgements v

Abstract vii

1 Introduction 2
- 1.1 Background and motivation 2
- 1.2 Previous work 4
- 1.3 Problem statement 5
- 1.4 Contributions 6
- 1.5 Organization 6

2 Fluoroscanning 8
- 2.1 Introduction 8
- 2.2 Conceptual and theoretical framework 9
  - 2.2.1 Key terminology 9
  - 2.2.2 Previous work 11
- 2.3 Fluoroscanning and its long term objectives 12

3 Extraction of Fluoroscans from digital images 14
- 3.1 Introduction 14
- 3.2 Previous work 15
- 3.3 Methods 16
  - 3.3.1 Experimental data 16
  - 3.3.2 Data acquisition 17
  - 3.3.3 Exploratory data analysis 20
  - 3.3.4 Pre-processing 22
  - 3.3.5 Image stitching 23
  - 3.3.6 Detection of regions of interest 24
  - 3.3.7 Segmentation 24
  - 3.3.8 Characterization 26
  - 3.3.9 Evaluation metrics 26
- 3.4 Results 27
  - 3.4.1 Pre-processing 27
  - 3.4.2 Image stitching 28

3.4.3 Segmentation 28
3.5 Discussion 29

4 Monte Carlo simulations of Fluoroscanning 31
4.1 Introduction 31
4.2 Methods 31
4.2.1 Underlying principles 31
4.2.2 Implementation 36
4.2.3 Random number generation 37
4.2.4 Input data and Monte Carlo parameters 37
4.2.5 Data pre-processing 38
4.2.6 Monte Carlo steps 39
4.2.7 Optical signal (Monte Carlo Fluoroscan) generation 41
4.2.8 Consensus MC Fluoroscan aggregation 44
4.2.9 Comparing Fluoroscans and GC profiles with MC Fluoroscans 44
4.3 Results 46
4.4 Discussion 46

5 Studying MC Fluoroscans and their usefulness for genomic map generation 48
5.1 Introduction 48
5.2 Comparing Monte Carlo Fluoroscans with Fluoroscans derived from data-driven approach 48
5.2.1 Measuring feature importance in Monte Carlo Fluoroscans 50
5.2.2 Results of the importance analysis 51
5.3 Fitting MC parameters to experimental data 55
5.3.1 Experimental data and setup 55
5.3.2 Fscan preprocessing workflow 56
5.3.3 Parameter fitting workflow 57
5.3.4 MC parameter fitting results 60
5.3.5 Understanding noise factors 60
5.4 Discussion 61

6 Conclusions and future work 63
6.1 Conclusions 63
6.2 Future work 64

Bibliography 65

1 Introduction

1.1. Background and motivation

DNA encodes genetic information and is comprised of two anti-parallel, complementary strands of nucleotides — e.g., cytosine (C) and guanine (G), adenine (A) and thymine (T). Sets of three nucleotides, called codons, are transcribed by RNA polymerase into messenger RNA. The latter is in turn decoded and translated by the ribosome into amino-acids, resulting in the proteins that shape living beings as we know them. This process is known as the central dogma of molecular biology (Alberts et al., 2008, p. 331), as depicted in Figure **1-1**. Although most of DNA consists of repeated sections, the frequencies of individual nucleotides fluctuate significantly across genes (Louie et al., 2003; Majewski et al., 2002). Furthermore, variations can occur within the whole set of DNA — i.e., the genome — of a living being. These alterations include single nucleotide variants (SNVs) and larger structural variants associated with standard genetic polymorphism as well as diseases like cancer (Li et al., 2016; Valouev, Schwartz, et al., 2006).

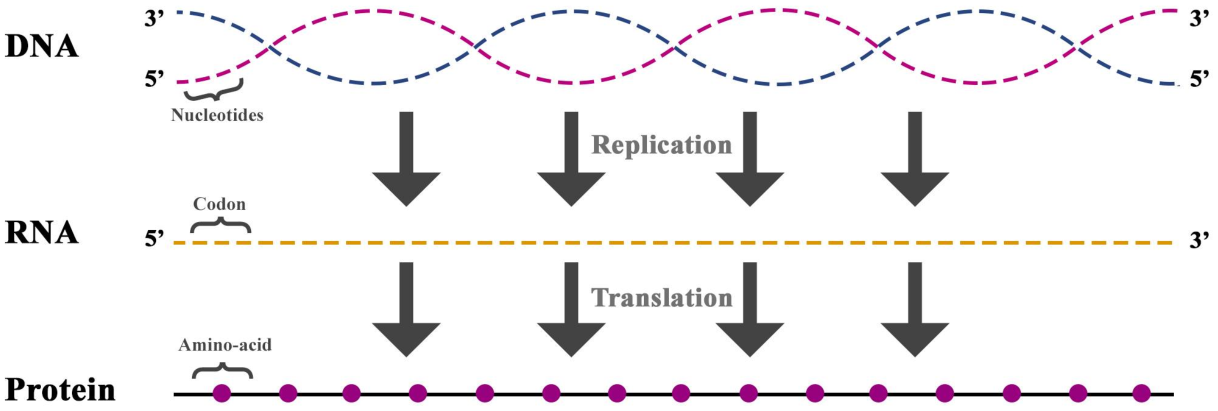

Figure 1-1: Abstracted representation of the central dogma of molecular biology. Source: the authors, inspired by Alberts et al. (2008, p. 331).

Genomic background analysis is key for gaining a comprehensive understanding of an individual's genome structure and tailoring treatments for subject-specific conditions (Nandi, 2017; Valouev, Li, et al., 2006). Modern genomic map assembly and sequencing algorithms, in conjunction with readily available computational resources, have allowed biomedicine experts to handle vast genomics datasets. In fact, the analysis of human genomes has revealed that sequence variation is highly prevalent among individual genomes (Gupta et al., 2015). However, these advances have not been carried out comprehensively nor are they informative enough across the entire human genome (Kounovsky-Shafer et al., 2017; Lesho et al., 2016). On the other hand, research efforts such as the Precision Medicine Initiative (PMI) are intended to analyze genomic data from around a million volunteers (Precision Medicine Initiative (PMI) Working Group, 2015), resulting in a pressing need for developing novel methods to improve the acquisition and processing of genomic data (Kounovsky-Shafer et al., 2017).

Among the wide array of approaches for the analysis of genomic data, single DNA molecule analysis systems such as optical mapping and nanocoding have proven useful for their capability to quickly generate genome maps. In such systems, DNA molecules are analyzed directly, which enables faster generation of large-scale datasets while having no noise from DNA amplification steps, and allows researchers to study longer sequences. More specifically, in optical mapping, DNA sequences are identified by their detecting cuts caused by restriction enzymes (Valouev, Schwartz, et al., 2006; Valouev, Li, et al., 2006; Zhou & Schwartz, 2004); whereas, in Nanocoding, nicking endonucleases are used to incorporate fluorescent labels in specific locations of the DNA molecules (Gupta et al., 2016; Jo et al., 2009). In both approaches, distance-based barcodes covering long molecular spans are produced based on the distances between restriction/nicking sites. Such barcodes can be employed for refining existing genomes, as a basis for de-novo genome assembly, and for detecting mutations in the genomes of specific individuals.

Fluoroscanning, developed at the Laboratory for Molecular and Computational Genomics (LMCG), is a novel single DNA molecule analysis system (Nandi, 2017). This system builds on the single-molecule-based nanocoding system for whole-genome analysis, also pioneered at LMCG (Gupta et al., 2016). Unlike other genomic assembly approaches such as optical mapping and nanocoding, fluoroscanning requires little manipulation of genomic DNA and supports extensive DNA molecule analysis. Furthermore, no cloning is involved, which means that variations across the genome are fully accounted and that artifacts related to mapping small fragments from repeat regions can be avoided (Nandi, 2017). Fluoroscanning is intended to allow for the generation of genomic maps based on the features of luminosity signals (Fscans) extracted from the backbone of dyed DNA molecules using image processing algorithms. The rationale behind this is to exploit fast approaches from the signal processing domain in order to produce such maps faster than the distance-based alignment approaches in which other systems rely. If this is achieved, it will be possible to detect structural mutations in an individual's genome and to pinpoint areas that should receive further attention in order to provide personalized healthcare.

1.2. Previous work

Single molecule genome analysis systems, such as optical mapping and nanocoding, are capable of detecting and describing structural variants and polymorphisms, which can result in a myriad of diseases and related complications (Jo et al., 2009; Shiguo et al., 2007). These techniques also serve as scaffolds to guide and validate DNA sequencing-based genome assemblies. Notably, sequencing approaches for genomic data acquisition have proven to be precise and provided the input for many state-of-the-art sequence alignment algorithms, but they are not scalable enough for fulfilling neither the large genomic assembly needs of many laboratories nor the objectives of the Precision Medicine Initiative, which entail analyzing data from over a million volunteers in order to develop patient-specific treatments based on genomic background (Kounovsky-Shafer et al., 2017; Leung et al., 2017; Precision Medicine Initiative (PMI) Working Group, 2015).

In order to develop the optical mapping and nanocoding systems, algorithms employed for sequencing-based genomic analysis were modified to process distance-based barcodes, resulting in faster genomic analysis (Gupta et al., 2016; Jo et al., 2009; Nagarajan et al., 2008; Valouev, Li, et al., 2006). However, in the aforementioned systems, genomic composition is calculated based on restriction gaps and nicking sites which must be detected with specialized algorithms after a molecule's backbone is segmented (Ravindran & Gupta, 2015). On top of being more computationally expensive to analyze, restriction site densities constrain the resolution of the generated genomic maps to around 1kb on a human cell line (Chan et al., 2018). In contrast, Fluoroscanning only requires detecting the DNA molecules and extracting intensity signals from them prior to inferring the genomic composition of a DNA-sequence. Furthermore, the fact that the system is based on continuous signals means that its resolution is higher (around 550bp), and has the potential of improving with the use of better quality optical sensors. However, Fluoroscanning currently depends on the nanocoding system for carrying out alignment, meaning that its full potential is yet to be realized (Nandi, 2017).

Due to the reasons described before, it is necessary to develop new algorithms or to adapt existing ones (Nandi, 2017) to better analyze Fluoroscanning data. We expect that a computational pipeline comprising Fluoroscan extraction, reference Fluoroscan generation based on Monte Carlo simulations, and consensus Fluoroscan assembly (i.e., barcoding) based on alignment will contribute to the state-of-the-art in computational approaches for solving biological problems (Ching et al., 2018; Min et al., 2017). More specifically, a computational approach for the generation of DNA barcodes based on fluoroscanning will open the door for future work on the detection of genetic anomalies in specific patients and will constitute a step towards the goal of building a fully functional fluoroscanning system which outperforms previous DNA barcoding methods. With such a system, candidate mutation areas in the genome (that is, areas of the reference genome for which no signal fragments could be aligned) could be pointed out in order to carry out an in-depth analysis which in turn would help determine the exact causes of such mutations.

1.3. Problem statement

Single molecule DNA analysis systems like optical mapping and nanocoding enable the generation of genomic maps and the detection of structural variations based on DNA barcodes (Gupta et al., 2016; Jo et al., 2009). However, running times for both systems are lengthy: the alignment of a (healthy) human DNA molecule to a reference genome takes up to two months with optical mapping and six weeks with nanocoding even when relying on high-performance clusters for compute. Ideally, Fluoroscanning-based genomic analysis should be faster than the former systems, since extracting intensity signals is a less computationally intensive task than detecting restriction gaps or nicking sites and measuring the distances between them (Nandi, 2017).

Nandi (2017) showed that DNA fragment intensity signals extracted from digital images correlate with GC/AT content and the presence of certain motifs, meaning that Fluoroscanning data could provide important sequence composition information and allow for the generation of DNA barcodes. However, no method exists yet that allows Fluoroscanning signals to be aligned to each other or to a reference genome (Nandi, 2017); the existing pipeline involves running a complete nanocoding pipeline and extracting noisy luminosity signals from DNA molecules that have not been prepared with Fluoroscanning in mind. Since nanocoding does provide an accurate alignment, Nandi (2017) was able to prove that signals extracted from DNA molecules belonging to the same region had statistically significant similarities. Thus, further work must be directed toward developing tools that enable extracting Fscans from DNA molecules prepared specifically for the Fluoroscanning system. More specifically, this problem can be divided into three sub-problems: extracting Fscans from DNA images, generating a “reference signal” from the reference genome that serves to guide laboratory experiments, and aligning Fluoroscans extracted from images to the reference signal in order to produce consensus Fluoroscans for the full genome. As part of this thesis, we focus on the former two and leave the task of designing alignment algorithms for future research.

Fluoroscan intensity signal extraction requires the design of a computational pipeline of algorithms designed specifically for treating Fluoroscanning images. This way, we will avoid the computational cost of adapting techniques derived from nanocoding such as those employed in the approximation proposed by Nandi (2017). As per Ravindran and Gupta (2015), the first part of the proposed computational pipeline will be comprised by several phases: image pre-processing, stitching of adjacent images, identification of regions of interest, and extraction of intensity signals. Each of these steps will be done based on digital image processing techniques, since machine learning approaches based on convolutional neural networks and transformer-based architectures would require a significant human annotation effort.

In order to provide enough data for guiding the laboratory experiments carried out at the LMCG, we propose generating *in-silico* Fluoroscans. This can be done by means of chemically and theoretically informed Monte Carlo simulations. Knowledge about physical, chemical (Kounovsky-Shafer et al., 2017), and statistical (Nandi, 2017) properties of DNA-fluorochrome interactions — e.g., stretch, dye fingerprint, binding and release probabilities, etc. — will be used to generate accurate intensity simulations from any nucleotide sequence; we named the resulting signals straightforwardly: Monte Carlo Fluoroscans. The advantage of this approach is that any arbitrary sequence that can be used as part of laboratory experiments can be also modeled to obtain a Monte Carlo Fluoroscan. However, a prerequisite for guiding laboratory experiments with Monte Carlo Fluoroscans is to fit the parameters of our Monte Carlo to the conditions of the experiments. Naturally, the experiments are subject to much more noise than the Monte Carlo under ideal conditions. Therefore, we must develop a pipeline for introducing noise into the simulated Fluoroscans. By modeling a sufficient amount of sources of noise, it should be possible for the distribution of Monte Carlo Fluoroscans to be fitted to that of Fluoroscans obtained from experiments. One added benefit to this is that it will enable the generation of massive datasets of simulated Fluoroscans which can be used to test future Fluoroscan alignment algorithms without carrying out a massive amount of experiments and to have a well-defined ground truth for evaluation.

1.4. Contributions

As part of this work, we developed a solid framework for the extraction of Fluoroscans from digital images, which can be employed as part of the Fluoroscanning whole genome analysis system. We proposed Monte Carlo simulations of the dye-DNA interactions which allowed us to simulate the dynamics of Fluoroscanning and better adjust actual experiments. Finally, we proposed a set of signal processing algorithms and workflows that allows us to validate the usefulness of our simulations for guiding experiments in the LMCG, which in turn constitute a first step toward the alignment of Fluoroscans for genomic map generation.

1.5. Organization

The document is organized as follows: in Chapter 2, we briefly summarize the proposed Fluoroscanning whole genome analysis system. In Chapter 3, we outline the image processing pipeline employed for extracting fluoroscans from digital images. Next, in Chapter 4, we describe the framework for *in-silico* simulations of fluoroscans. In Chapter 5, we report the methodology for validating the usefulness of Monte Carlo simulations for guiding Fluoroscanning experiments. Finally, in Chapter 6, we provide conclusions and insight for future work.

2 Fluoroscanning

2.1. Introduction

Fluoroscanning is a single DNA molecule analysis system in which the composition of DNA molecules is estimated based on fluorescence profiles extracted from the molecules' backbones after dying them with fluorochromes. Unlike optical mapping and nanocoding, Fluoroscanning does not rely on enzymatic reactions and its resolution is not limited by the frequency of specific restriction/nicking sites. Even when compared to other approaches for DNA sequence composition analysis, Fluoroscanning reflects the GC content of DNA molecules with higher resolution and accuracy, with its estimated resolution being of around 550 bp as estimated from the point spread function (PSF) of the microscope.

The data acquisition workflow is fairly simple: first, dye binding is carried out in a controlled manner using fluorochromes; next, the molecules are presented on positively-charged PDMS microchannels and surfaces that cause them to stretch; and, finally, digital images can be acquired either manually or automatically with a microscope, which enables the signals to be extracted from the backbone of the molecules using image processing algorithms similar to those employed for optical mapping (Ravindran & Gupta, 2015). A more detailed description of this workflow can be visualized on Figures **3-2** and **3-3**, while the complete image processing suite is described in Chapter 3.

For DNA sequence profiling, Fluorosanning relies on the increased quantum yields and binding probabilities of bis-intercalating fluorochromes (such as YOYO-1) when interacting with GC-rich regions, as opposed to AT-rich ones (Netzel et al., 1995; Rye et al., 1992), resulting in favorable signal-to-noise ratios (Günther et al., 2010). In essence, this means that, upon extracting fluorescence profiles from dyed DNA molecules, GC-rich regions tend to be brighter (and show up as peaks) whereas AT-rich regions tend to be darker (and show up as valleys). Moreover, a simple GC% profile can be derived from a reference DNA sequence and used to validate the accuracy of Fluoroscans.

However, the above does not paint the full picture: numerous photophysical factors impact noise and part of the experimental work at LMCG involved carefully tuning the amount of dye, binding conditions, and ionic strengths through a two-step binding scheme. This helped them reduce the inhomogeneous binding and signal phase variation caused by an excessive amount of dyes being present while the DNA molecules are stretched. Parallel to this, our our efforts in Chapters 4 and 5 centered on fitting the Monte Carlo simulations to the experimental conditions at the LMCG by introducing noise sources like variable dye loading, dye luminosity variation, degraded dark fluorochromes, phase shifts, and imaging noise.

2.2. Conceptual and theoretical framework

2.2.1. Key terminology

- DNA: Deoxyribonucleic acid is a molecule composed of two nucleotide chains which form a double helix carrying the genetic instructions used in the growth, development, functioning, and reproduction of all living organisms and a number of viruses. DNA contains four complementary deoxynucleotide bases: cytosine and guanine, adenine and thymine. The strands which makes up the double helix are not only complementary, but also anti-parallel. The RNA polymerase transcribes these into messenger RNA, replacing thymine with uracil. Messenger RNA is translated into the amino-acids which make up proteins necessary for living beings (Alberts et al., 2008, p. 331).

- Codon: A sequence of three DNA or RNA bases which encode a specific amino-acid. Many codons can translate into the same aminoacid, however, in most cases altering a single base can result in a different aminoacid. There also exist start and stop codons which delimit coding regions (Alberts et al., 2008, p. 367).

- Genome: The complete set of genes of a living being. A genome contains all of the information necessary for building and maintaining an organism. In the case of human beings, the genome is made up by more than 3 billion pair bases, and there is a copy of it in every cell with a nucleus (Roy et al., 2017).

- Single Nucleotide Variation (SNV): A variation which occurs when a single base in a genome is altered with respect to the reference genome. Notably, some SNVs have been associated to diseases (Katsonis et al., 2014).

- Single Nucleotide Polymorphism (SNP): SNVs which occur with a known frequency (greater than 1 per cent) within a given population. Despite the importance of both SNPs and SNVs, single nucleotides are too small to be detected by fluorescence microscopy, and thus will not be considered in our work.

- Structural Variation (SV): Large-scale (larger than 1 kb) variations in a genome with respect to the reference genome. Some examples of SVs are insertions, duplications, deletions, translocations, complex genomic rearrangements, and aneuploidy (Gupta et al., 2015).

- Structural Polymorphism (SP): Similar to SNPs, SPs are structural variations which occur with a known frequency within a given population. They have been recently discovered to be rather frequent within the human genome (Gupta et al., 2015). Unlike SNVs and SNPs, SVs and SPs are big enough to be analyzed at the 200 nucleotide per pixel resolution provided by fluoroscanning (Nandi, 2017).

- Fluorescence microscopy: Fluorescent molecules absorb light at one wavelength and emit it at a longer one. When fluorescent dyes interact with nucleic acids they have a probability of intercalating and fluorescing. This means cells that lack color can be stained with dyes for visualization. This is done by means of fluorescence microscopes which possess two filters: one that only allows wavelengths that excite the dyes through, and another one that only lets through the wavelengths emitted by the dye when it fluoresces. This way, fluorescence microscopy can be used to visualize specimens that otherwise would be impossible to see (Alberts et al., 2008, p. 586).

- Dye-DNA interactions: Dyes used for fluoroscanning, such as YOYO-1 (named from Oxazole Yellow, abbreviated as YO), exhibit very large degrees of fluorescence enhancement when they bind to nucleic acids. As per some studies, fluorescence intensity of YOYO depends on the base sequence and GC-rich DNA sequences have twice the quantum yield of those rich in AT (Larsson et al., 1994; Netzel et al., 1995). This indicates that the probability of dyes intercalating between DNA bases and emitting fluorescence is non-uniform.

- Digital image processing: A branch of signal processing focused on the processing of digital (that is, discrete) images by means of a digital computer. Tasks typically classified as digital image processing receive an image as input and output another, modified digital image. Examples include noise removal and image enhancement. Algorithms used for feature measurement and extraction are categorized as image analysis methods, but for the sake of simplicity we will group them together with lower-level image processing techniques (Gonzalez & Woods, 2008).

- Monte Carlo (MC) simulations: A family of non-deterministic or numerical statistical methods employed for approximating complex mathematical expressions whose exact evaluation is complex to carry out. MC methods rely on repeated random number sampling to generate numerical results. These kinds of simulations are useful because they allow modeling physical systems with many parameters, such as the Dye-DNA interactions which allow us to obtain Fluoroscans (DeGroot et al., 2011, p. 787).

- Machine Learning: A field built on computer science, mathematics, and probability. It is centered around the design of (sometimes biologically inspired) models for automated large-scale data analysis. Unlike algorithms created for obtaining exact solutions, machine learning models have a strong focus on approximate predictions based on large sets of data. There exist main two sub-fields within machine learning: supervised learning, where the focus is on accurate prediction based on labeled datasets, and unsupervised learning, where the aim is to find accurate descriptions of data Barber (2012). For the purposes of this thesis, we focus on the random forest algorithms used by Nandi (2017) as part of his work.

2.2.2. Previous work

- Optical mapping: this system for single molecule genome analysis was pioneered at the LMCG (Aston et al., 1999; Dimalanta et al., 2004; Schwartz et al., 1993; Teague et al., 2010), and it has enabled scientists to carry out comprehensive genome analyses that complement DNA sequencing by detecting structural variations at much larger scales while sacrificing basepair-level resolution. Optical mapping allows for the construction of ordered restriction map of enzyme cut sites spanning whole genomes, without any need for cloning. Microfluidic devices and automated imaging and computational workflows are combined to generate large-scale datasets from millions of DNA molecules that are first stretched on a charged surface and then ”cleaved” by a restriction enzyme targeting specific DNA sub-sequences (from 4 to 8 bp in length). The gaps left by the enzymes are imaged by means of fluorescence microscopy, and the distances between the gaps are used as a descriptor for every imaged molecule. These ”ordered restriction maps” (also called Rmaps) are, to some extend, similar to sequencing data, and many methods have been developed for aligning them, either among themselves for *de-novo* assembly, or against a reference for genome refinement or structural variant detection (Valouev, Schwartz, et al., 2006; Valouev, Li, et al., 2006; Valouev, Zhang, et al., 2006). Importantly, one can easily ”simulate” Rmaps by cutting reference genomes at restriction sites and use them as a noiseless reference for optical mapping.

- Nanocoding (Jo et al., 2009): also created in the LMCG, this is a more advanced single molecule genome analysis which uses fluorochrome labeling instead of restriction cuts for characterizing DNA molecules. The process is as follows: first, a long DNA molecule (up to several Mb) is acquired from a test genome. The molecule is stained with a fluorochrome that embeds into DNA and produces fluorescence light. Next, the DNA molecule is nicked with restriction enzymes that only cut one strand of the cleavage site and marks it with a fluorochrome that can be visualized as a red punctate in contrast with the green DNA backbone (Kounovsky-Shafer et al., 2017). The molecules are presented on nanoslits which confine and stretch DNA strands, thus facilitating visualization (Kounovsky-Shafer et al., 2017). Images of lengthy DNA sequences with a resolution of 200 nucleotides per pixel are then acquired by fluorescence microscopy. Once the molecule backbones and the punctates are identified, alignment methods designed for optical mapping can be used without extensive modifications.

- Nanocoding-based Fluoroscanning (Nandi, 2017): prior to this and several concurrent works at LMCG, Fluoroscanning was based on the existing nanocoding system. Once images are captured and processed, areas marked by nanocoding punctates — the nanocoding equivalent of restriction gaps — are exploited for aligning the whole set of DNA strands either among themselves or to a reference genome. Under this framework, intensity signals (Fluoroscans) are extracted for every nanocoding map and those belonging to the same genome region are averaged. The set of aligned, averaged Fluoroscans spanning the whole genome is called a consensus Fluoroscan and is expected to provide genomic information about organisms similarly to consensus maps obtained from optical mapping and nanocoding.

- Other approaches for enzyme-free DNA sequence profiling: Fluoroscanning is not the sole system to attempt to reveal DNA sequence composition without any enzymatic reactions. For instance, Reisner et al. (2010), Marie et al. (2013), and Marie et al. (2018) introduced and developed denaturing mapping, a method which exploits the relative differences in melting temperatures of DNA regions based on their GC content. This enabled them to obtain fluorescence intensity profiles that reflected composition at a resolution of around 1 kb, but the temperature manipulation and the subsequent kymograph acquisition are lengthy processes. Concurrent work by the Jo laboratory (Lee & Jo, 2016; Lee et al., 2018; Park et al., 2019) has shown that fluorescent proteins that bind to AT-rich areas can serve as a constrast to non-specific fluorescent DNA binder to produce sequence-specific DNA composition profiles. On the other hand, researchers at the Westerlund laboratory (Dvirnas et al., 2018; Müller et al., 2019) have used netropsin (a sequence-selective ligand) to block YOYO-1 dyes from binding with AT-rich regions, making them darker. Notably, the resolution of approaches based on sequence selective binding such as those of the Jo and Westerlund groups are lower (around 10 kb), likely due to false-negative and false-positive bindings.

2.3. Fluoroscanning and its long term objectives

Currently, at LMCG, Fluoroscans can be only aligned to generate a genome map by relying on the nicking sites provided by nanocoding system. Ideally, Fluoroscanning should work independently from nanocoding, that is: individual Fluoroscans should be aligned to a reference genome based only on fluorescence signal information. Indeed, Fluoroscanning is intended to allow for the generation of genomic maps based on the features of the Fluoroscan signals. The idea behind this is to exploit fast algorithms from the signal processing domain in order to produce such maps faster than the distance-based alignment approaches in which other systems rely. If this is achieved, we should be able to detect large scale mutations in an individual's genome and to pinpoint areas of it that should receive further attention in order to provide personalized healthcare.

3 Extraction of Fluoroscans from digital images

3.1. Introduction

The first step in the Fluoroscanning pipeline is to extract signals from digital images of elongated, dyed DNA molecules presented on a charged surface (see Figure **3-1**). We acquire such images following several steps, which comprise the experimental procedure in which the sample is prepared and placed on a glass surface, and image acquisition using a MD scanner. This way, it is possible to acquire about 400 images 2560x2160 16-bit gray-scale images per experiment. However, prior to signal extraction, it is necessary to address several issues, including the presence of noise in images as a result of noise particles being present on the DNA fragment backbone, DNA fragments being crossing over each other, DNA fragments overlapping, lighting from noise particles interfering with the intensity profile of DNA fragments, lighting from DNA fragments interfering with the intensity profile of other nearby fragments, and images being out of focus, among others. In addition to noise, we also need to account for the need to stitch together several images in order to be able to analyze DNA fragments which extend across two or more images, which also requires us to define the amount of overlap to be included in every pair of consecutive pictures belonging to the same channel.

In order to address the above issues, we build upon **Wscan**, a computer vision program to automatically render Fscan profiles from image datasets of DNA molecules bound with fluorochrome dyes. The microscope images were flat field corrected then input to Wscan. The workflow comprises several stages that are detailed below: image overlapping, region of interest detection, molecule segmentation, DNA molecule backbone identification, molecule background correction, and lastly, Fscan output (Ravindran & Gupta, 2015). This improved version of **Wscan** is implemented with Python 3 and several popular libraries, such as **NumPy** (Harris et al., 2020) and **scikit-image** (Van der Walt et al., 2014).

This chapter is organized as follows: in Section 3.2 we describe previous work on the segmentation of elongated DNA molecules as well as thin, elongated items in general in digital images. Next, in Section 3.3 we describe the methodology followed in this work. Then, we show our results in Section 3.4, and finally we analyze them and provide some ideas for future work in Section 3.5.

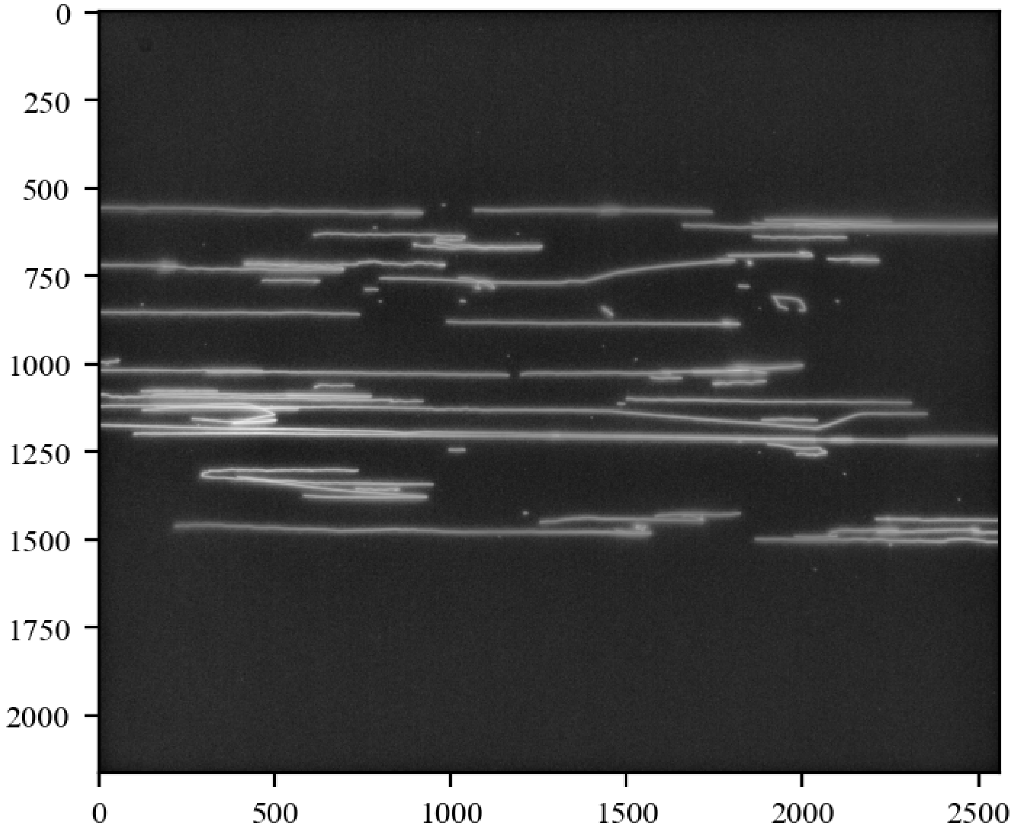

Figure 3-1: Elongated DNA molecules imaged using the MD system; as in most Fluoroscanning experiments, they have an horizontal orientation which we can leverage for Fluoroscan extraction. The image was processed for visualization using Contrast Limited Adaptive Histogram Equalization (CLAHE) with clip limit = 0.001. Source: the authors.

3.2. Previous work

Previous work on segmentation of dyed DNA fragments in digital images comprises proposals for the segmentation of optical mapping and nanocoding images, which is centered mainly on the use of traditional digital image processing algorithms. Other relevant work includes both classic digital image processing and convolutional neural networks for the segmentation of thin elongated objects in fields such as medical and concrete structure image analysis.

In their review, Ravindran and Gupta (2015) described the framework for the processing of Optical Mapping images in research and commercial-grade systems. The described method consists of first segmenting and skeletonizing the backbone elongated DNA molecule fragments. This is done by taking as part of the backbone the pixels with the highest intensity value along the axis the molecules are deposited in a 5-pixels neighborhood perpendicular to such axis. In addition, every backbone pixel has to fulfill the condition of having an intensity value higher than the intensity of the extremes of its neighborhood by manually defined falloff value δ. Then, the endpoints of the the molecule fragments are used to stitch adjacent images. Next, the fragments are grouped as belonging to the same molecule based on their proximity and orientation, and the distance maps are calculated using integrated fluorescence across a 5-pixel neighborhood surrounding the fragments.

Bahadar et al. (2016) proposed a method for the segmentation of retinal blood vessels in digital images using a Hessian based approach and Otsu-based region thresholding. First, they used CLAHE to enhance contrast. Then, they used the Hessian matrix and eigenvalues transform at two different scales in order to extract wide and thin vessels, which is coupled with a modification of global and local Otsu thresholding in order to classify vessel and non-vessel pixels. Finally, they employed an area-based threshold method to clean up noisy pixels and unconnected regions up to 30 pixels in size, resulting in the final segmented image.

Shit et al. (2020) developed *clDice* a novel loss function for the segmentation of thin tubular structures in digital images. In essence, the authors defined an improved loss function based on the classic morphological skeletonization algorithm and the well-known DICE loss function, and implemented a differentiable version of it. This way, they were able to train convolutional neural networks in such a way that the ”backbone” of thin elongated objects could be prioritized over other features such as the absolute number of correctly classified pixels, resulting in better performance for such architectures.

3.3. Methods

3.3.1. Experimental data

In this section, we analyzed images belonging to two kinds of samples in order to account for two key needs in the Fluoroscanning pipelines:

- BAC 876A24: We study this bacterial artificial chromosome since it contains several AT-rich areas which turn into noticeable features in Fluoroscanning experiments. Furthermore, the size of the sequence is such (around 224 kb) that most fragments fit in a single image, thus allowing us to carry out Fluoroscanning signal quality experiments prior to the implementation of more elaborate digital image processing algorithms.
- HEK: We use DNA material from Human Embryonic Kidney cells since it allows us to easily image very long DNA fragments which span several images. Such fragments can belong to a much more varied set of DNA sequences, meaning the features of the extracted Fluoroscans cannot be readily associated without robust signal alignment.

However, this data does allows us to verify the effectiveness of digital image processing algorithms in extracting large signals.

3.3.2. Data acquisition

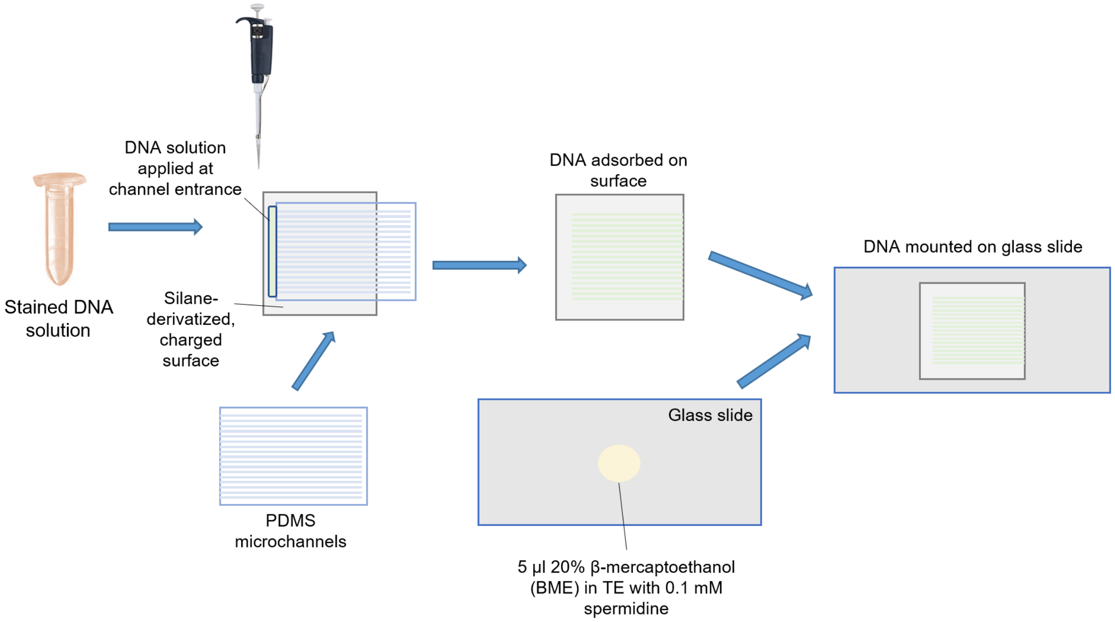
DNA solution applied at channel entrance
Stained DNA solution
Silane-derivatized, charged surface
PDMS microchannels
DNA adsorbed on surface
DNA mounted on glass slide
Glass slide
5 μl 20% β-mercaptoethanol (BME) in TE with 0.1 mM spermidine

Figure 3-2: The workflow developed at LMCG for preparing DNA samples. First, the PDMS microchannels are adhered to a charged surface. Next, the DNA solution is spread at the entrance of the microchannels with a wide bore pipet tip and the capillary action pulls the solution, stretching the DNA molecules on the channels. The microchannels are removed and the surface is mounted on a glass slide with a drop of BME/TE for visualization. Source: Yumin Lian at the LMCG.

Sample preparation

In Fluoroscanning experiments, the sample is prepared per the following protocol (see Figure 3-2 for a visual description):

- PDMS micro-channels are adhered to a derivatized, positively charged glass surface. The micro-channel is cut to create the opening.
- The DNA solution is placed at the channels' entrances with a wide bore pipette tip.
- Capillary action sucks the solution into the channels and deposits the stretched DNA molecules.

- Once the solution reach the end of the channels, the PDMS is peeled off and any leftover solution left on the glass surface is cleaned.
- A drop of BME/TE is added at the center of a glass slide.
- The side of the surface with the DNA is placed on the BME/TE drop and centered on the glass slide.

Image acquisition

Prepared samples are placed in a MD fluorescence microscopy which takes between 20 and 100 2560x2160 16-bit gray-scale images of each channel in the charged glass surfaces. In some cases, fragments can not be imaged in a single picture, so a degree of overlap (around 700 microns) between images belonging to the same channel is included as part of the acquisition process. See Figure **3-3** for a visual overview of this process.

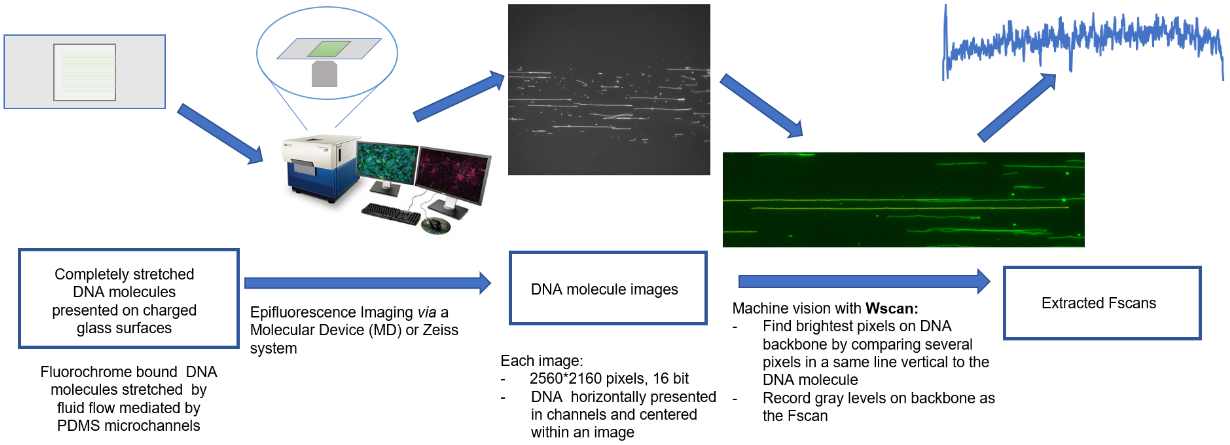
Completely stretched DNA molecules presented on charged glass surfaces
Fluorochrome bound DNA molecules stretched by fluid flow mediated by PDMS microchannels
Epifluorescence Imaging *via* a Molecular Device (MD) or Zeiss system
DNA molecule images
Each image:
- 2560*2160 pixels, 16 bit
- DNA horizontally presented in channels and centered within an image
Machine vision with **Wscan:**
- Find brightest pixels on DNA backbone by comparing several pixels in a same line vertical to the DNA molecule
- Record gray levels on backbone as the Fscan
Extracted Fscans

Figure 3-3: The workflow developed at LMCG for acquiring images from a prepared sample. The sample is first put in the microscope and the images are captured either manually or automatically depending on the system. Once the images are acquired, the Fluoroscans are extracted using the methods described in the rest of this chapter. Source: Yumin Lian at the LMCG.

Data labeling: stitching

In order to be able to quantify the automated stitching procedure, 50 pairs images belonging to several DNA samples were selected and aligned manually by two human annotators in order to determine the number of pixels that need to be considered in both the x and y-axes for lining up and stitching overlapping images (see Figure **3-4**).

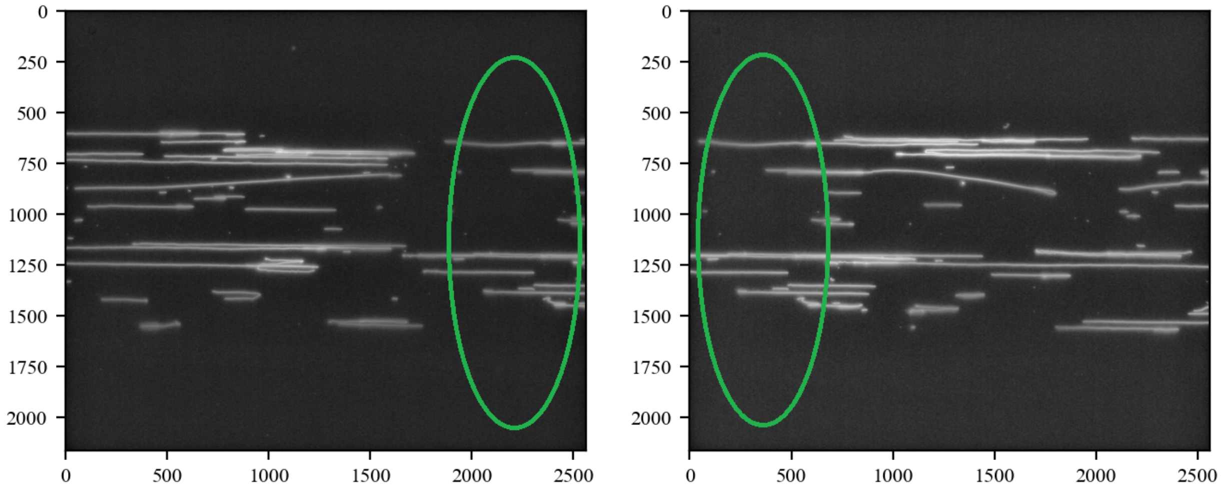

Figure 3-4: Two overlapping images acquired from the same channel in a sample. The overlapping areas (circled in green) for each image are of around 730 pixels and are useful for putting two consecutive images together. The images were processed for visualization using CLAHE with clip limit = 0.001. Source: the authors.

Data labeling: pixel-wise segmentation

20 images corresponding to several DNA samples (experiments with images deemed to be good, bad, and very noisy) were selected and segmented manually by two human annotators after carefully evaluating a set of conditions for discerning usable DNA molecules from unusable ones as well as noise elements. Such conditions were deemed to be as follows:

- The backbone of an usable DNA molecule is defined as the pixel with the highest intensity level in its 5-pixel neighborhood perpendicular to the orientation of the molecule (pixel class 1, color green).
- Very small particles or DNA molecules (less than 30 pixels) are classified as noise (pixel class 2, color yellow).
- Any increase of more than 1000 in intensity values is determined to be an overlap between two molecules, a knot within a single molecule, or a particle of noise embedded on/near the molecule (pixel class 3, color red).

Situations where two molecules were too close to each other were not considered to be overlaps unless the two molecules actually touched each other, due to issues with subjective calculation of proximity between objects in images. Such situations are addressed more appropriately in further sections.

The resulting images (see Figure 3-5) contain information from each of the three aforementioned classes and were used to evaluate the performance of the segmentation algorithms employed in this work.

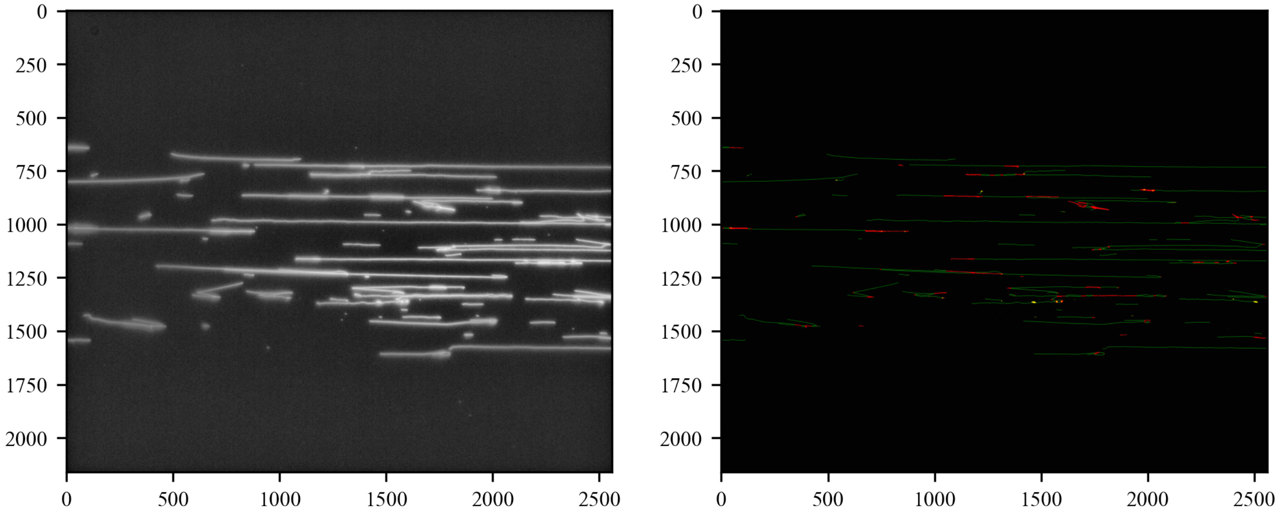

Figure 3-5: An image and its corresponding, manually-labeled segmentation mask. Green, yellow, and red pixels represent usable molecules, noise fragments, and unusable molecules, respectively. Contrast enhanced for visualization of the original image using CLAHE with clip limit = 0.001. Source: the authors.

3.3.3. Exploratory data analysis

Prior to carrying any kind of processing on the images, we carried out a basic statistic analysis on the distribution of pixel intensities for a set of 100 images selected from several experiments. We were able to evidence that the pixel intensity values ranged from 500 to around 40,000, with a mean intensity value of 905.7. The histogram of one of the images is displayed in Figure 3-6.

Analyzing the histograms of several images allowed us to notice that the distribution of pixel intensity values are skewed to the right, that is, there are numerous outlier, high intensity pixel values which often correspond with noise and overlapping areas as shown in Figure 3-7. However, there is no hard-threshold that can be defined from visual inspection of the histogram, so more elaborated approaches were deemed necessary for DNA molecule segmentation.

In addition, the study of several consecutive images belonging to the same channel in Subsubsection 3.3.2 allowed us to determine that, due to particularities with the imaging system

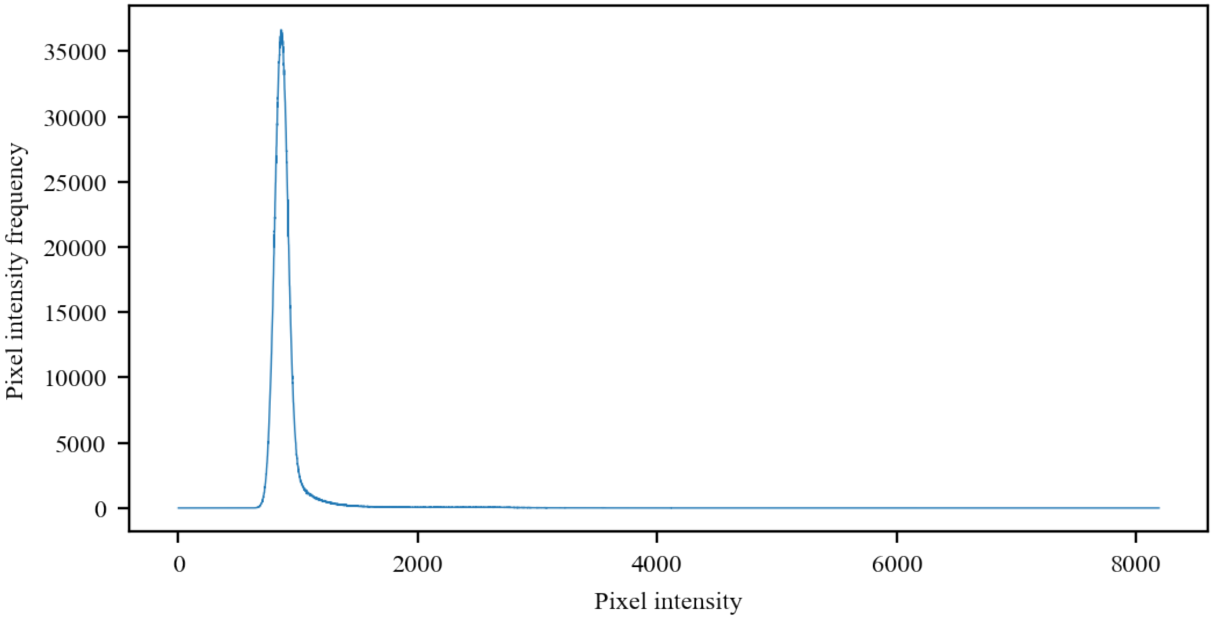

Figure 3-6: Histogram of a single dyed DNA molecules image; while it would be fairly easy to set a threshold for separating background from areas where dye is present, a similar approach can not identify noise particles. The bins are trimmed to around 8000 for visualization. Source: the authors.

and the sample itself, the y-axis offset increases or decreases constantly with each consecutive image, resulting in an effect similar to that occurring with some images stitched together and displayed (see Figure **3-8**) in the work of Ravindran and Gupta (Ravindran & Gupta, 2015).

One last discovery made during the exploratory data analysis was that the imaging system sometimes produced images comprised entirely of white noise. In order to prevent such images from having an impact on the data processing pipeline, they are excluded based on the following criteria:

$$usable(I) = \begin{cases} True, & \text{if } max(I) - min(I) \geq 1500 \\ False, & \text{otherwise} \end{cases} \tag{3-1}$$

where I is the array containing all of the intensity values which represent a given image. In other words, images with a low dynamic range are assumed to be unusable.

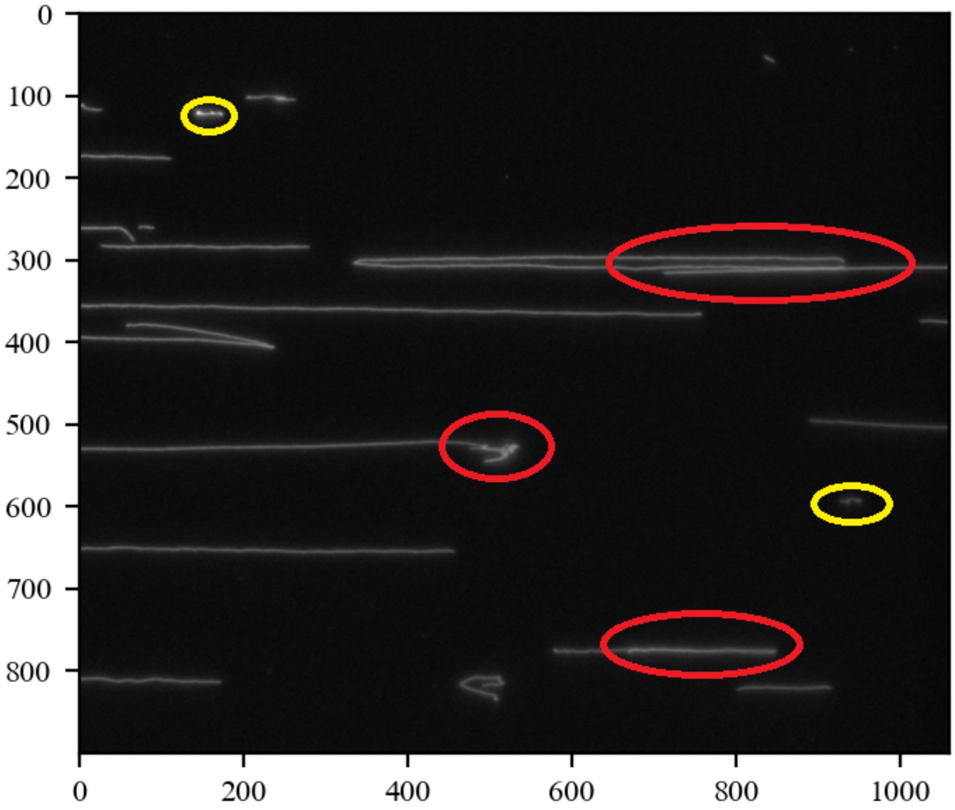

Figure 3-7: Close-up of part of an image in which noise fragments (circled in yellow) and overlapping molecules (circled in red) are present. Ideally, our image processing pipeline should not extract Fluoroscans from such areas. The image was processed for visualization using CLAHE with clip limit = 0.001. Source: the authors.

3.3.4. Pre-processing

Phenomena such as photo-bleaching and variations in the baseline intensity levels across each image result in inaccurate molecule extraction in areas close to the edges of the images. This is problematic since the precision of the Fluoroscanning system depends strongly on the intensity levels associated with each extracted molecule. For this reason, we implemented a two-step strategy for removing the background in our images:

- The background was corrected using simple flat field correction, employing a bright image F in order to get a corrected version C of image I, as in Equation 3.3.4.

$$C = \frac{I}{F} \tag{3-2}$$

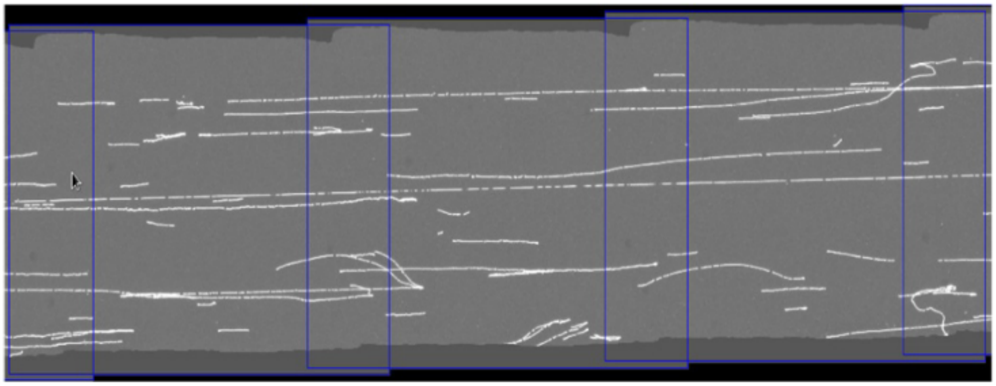

Figure 3-8: Example of three stitched images in the Optical Mapping system. Note that if the stitching is done poorly, information extracted from molecules spanning more than one image will be inaccurate. Source: Ravindran and Gupta (2015).

- Any remaining background irregularities are removed by estimating the background using a median filter with a column-vector shaped kernel of size (100×1). The median filter is very robust to outliers and thus fits mostly to the background, preserving the relative differences between molecule intensity levels.

Other approaches such as fitting second and third order surfaces to the shape of the background (Voigtländer, 2015), while useful for single images, were not as effective near the edges of the images, making the subsequent image stitching step more difficult.

3.3.5. Image stitching

After the images have been pre-processed, the next step is to obtain a single, contiguous image for every channel in the sample. The number of images per channel can range from 10 to 100 depending on the experiment, so this must be done automatically. One way to achieve this is by carrying out image registration. A naive approach is to minimize the Mean Squared Error (MSE) of the differences between pixel intensity values of the overlapped areas of both images, testing for a range of offset values o_x and o_y. More elaborate approaches involve the use of the Fourier transform to detect the offset values which maximize the cross-correlation in the frequency domain (Guizar-Sicairos et al., 2008).

We tested both approaches in addition to a simple but effective improvement of the Fourier transform approach: we determined the vertical and horizontal shift between two contiguous images by using a fast-Fourier-transform-(FFT) phase cross-correlation 2D image registration algorithm (Guizar-Sicairos et al., 2008) on image slices of 800 pixels width in order to reduce computational costs and to reduce the chances of obtaining spurious offsets. Once the offsets for every pair of images I_i, I_{i+1} have been calculated, the images are merged into a single, contiguous array representing the whole channel. In the cases where unusable images are present, such images are discarded and a fixed horizontal offset of $o_x = 10$ pixels with no vertical offset is used instead to signal that the remaining images are not contiguous.

3.3.6. Detection of regions of interest

As seen in Figures **3-1** and **3-4**, each channel (that is, the region of the image where DNA molecules are) comprises about 50% of the images captured with the MD system, with the rest of the image consisting mostly of noise and background areas. For this reason, we employ a simple approach for detecting the region of interest in each channel:

- The image is reduced to 1/5 of its original size to reduce computational costs.
- The reduced image is transformed to a 8-bit representation and thresholding is used with a fixed threshold $t = 10$ to segment it into background and foreground areas.
- A fixed window fitted to the vertical size of the channel is used at every possible position in the y-axis.
- The window which maximizes the amount of pixels segmented as foreground is selected as the window containing the region of interest.
- The window's coordinates are extrapolated to the original image, which is then reduced to the channel only.

Since the maximum vertical size of the channel can vary when using stitched images, the size of the fixed window corresponding to the channel is increased based on the accumulated o_y offsets.

3.3.7. Segmentation

Otsu-based thresholding

The first tested approach for segmentation was to carry out binary thresholding based on Otsu's method (Otsu, 1979), which minimizes intra-class variance based on Equation 3.3.7:

$$\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \tag{3-3}$$

where w_0 and w_1 are the probabilities associated to each class (0: background, 1: foreground) and the σ^2 values are the corresponding intra-class variances of the underlying intensity levels, for a given threshold t. The minimization of the intra-class variances is achieved by iterating through all possible thresholds and choosing the one which minimizes the variances.

Standard-deviation-based thresholding

The second approach for segmenting elongated DNA molecules was to employ standard-deviation-based thresholding. The idea behind using standard-deviations is to detect pixels which are within the range of expected Fluoroscanning signal values, pixels which fall below the usual range of intensities (background), and very intense pixels which are above the usual range of intensities (noise or overlaps).

Morphology and intensity-based segmentation

A third approach for segmentation is to employ more complex digital image processing techniques: after ROI detection, we determine the minimal bounding box for each DNA molecule. Since most molecules were horizontally orientated as they were deposited in microchannels, we identified the brightest pixels along the horizontal axis ("DNA backbone") using a maximum filter with a 1 × 100 pixels kernel. This approach helped obtain (noisy) image masks of candidate molecules. Next, we removed all noise particles and DNA molecules less than 50 pixels in size with 8-connectivity (i.e., pixel a is connected to pixel b if a is in one of the eight positions that surround b) using the Python **scikit-image** package (Van der Walt et al., 2014). The identified bounding boxes in the ROIs helped reducing the computational cost in the subsequent steps.

For each rectangular bounding box containing a DNA molecule, our algorithm identifies the brightest pixel in each column of pixels vertical to the DNA backbone. Next, it checks the connectivity of the brightest pixels using a queue. If the brightest pixels are eight-connected and form a series of interconnected segments, the entire object is defined as the DNA molecule backbone. Otherwise, a quality filter is used to determine how to connect the fragments. Compared with the skeletonization functions usually used for computer vision, this identification algorithm is more effective in detecting the backbone of the molecules, as it includes several inductive biases about the way molecules are visualized in the microfluidic channels.

Despite the accuracy of this approach, we were able to identify artifacts consisting of truncated backbones in the form of dim spots on some molecules. The truncated points were discriminated from non-molecule objects through morphological filtering; more specifically, we used the closing operation on the identified backbones and pinpointed all pairs of molecules that such an operation was able to reconnect. Then we use a modified version of Dijkstra's algorithm for path-finding (Cormen et al., 2001), which helps us link two backbone fragments belonging to the same molecule.

The grey levels on each molecule's backbone pixels were output as a continuous fluorescence intensity signal. These signals, called Fscans, were exported in Excel files with additional information about the DNA molecules, including each pixel's (x, y) position on the overlapped image, assigned molecule numbers, and image numbers. Composite images labeled with the backbone pixels, ROI boxes, and molecule numbers were also exported, along with screen views of Fscan profile plots for quick data inspection.

3.3.8. Characterization

Proximity-based criteria

A neighborhood criteria is used to determine whether the pixels in the detected molecule backbone are usable, this is done by taking a circular area with a radius of 5 pixels around the molecule. The quality of a molecule is penalized based on two criteria: first, the entropy of the neighborhood, and second, the presence or absence of other molecules or noise fragments in the defined neighborhood.

Elongation-based criteria

Ideally, molecules should be well-stretched in order to be able to extract good Fluoroscans from them. For this reason, we calculate the tortuosity of each molecule as per Espinosa et al (2013). We deem those molecules with a low tortuosity score to be well stretched whereas molecules with high tortuosity scores are classified as poorly stretched.

3.3.9. Evaluation metrics

The MSE metric was employed to measure the accuracy of the image stitching approaches, whereas the Jaccard Index metric was employed to quantify the quality of the segmentation algorithms. A more qualitative approach was used in order to facilitate the analysis of the characterization criterions: the Fluoroscans with the best and worst scores extracted from the BAC images were visually compared with their ground-truth in order to verify the results.

Mean Squared Error

The MSE metric measures the squared error for the stitching coordinates of a pair of images is defined as the squared difference between the automatically obtained offsets o_x, o_y and the ground truth, human-annotated offsets $\hat{o_x}$, $\hat{o_y}$ (Eq. 3-4). This metric penalizes large errors with powers of 2, which is preferable to us since large errors would result in being unable to recover Fscans belonging to molecules spanning two or more images.

$$MSE = \frac{(o_x - \hat{o_x})^2 + (o_y - \hat{o_y})^2}{2} \tag{3-4}$$

Jaccard Index

Given a ground-truth binary mask U and an automatically generated segmentation mask V, the Jaccard Index is defined as the intersection over the union of all pixels in both masks (Eq. 3-5); higher Jaccard Index values are better.

$$Jaccard(U,V) = \frac{|U \cap V|}{|U \cup V|} \tag{3-5}$$

3.4. Results

3.4.1. Pre-processing

It is possible to confirm the positive impact of the proposed background correction approach in Figure **3-9**, where the intensity levels of a molecule which crosses from one image to another are displayed prior to and after background correction. In addition, Table **3-1** confirms a noticeable improvement of image stitching accuracy when using pre-processed images.

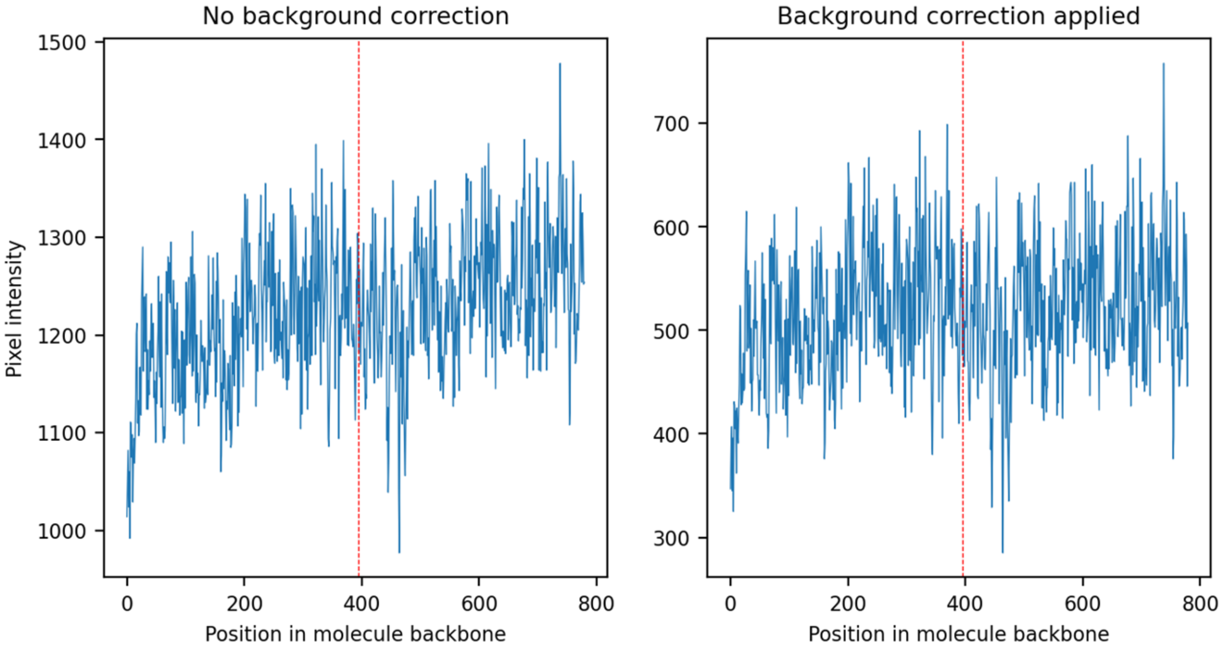

Figure 3-9: The intensity profile of a molecule before (left) and after (right) background correction. The red line indicates a point where two images were combined using the stitching algorithm. We can see that background correction partly removes the baseline effect in the signal. Source: the authors.

3.4.2. Image stitching

As shown in Table **3-1**, restricting the portion of the images based on the number of microns of overlap established in the MD system set-up allows to obtain a more precise estimation of the offset values o_x and o_y while using less computational resources. Figure **??** depicts the results of applying the best stitching approach on a series of 20 pictures taken consecutively.

Table 3-1: MSE and average execution times of the tested image registration approaches; the restricted Fourier-based registration is both faster and much more accurate than all other approaches. Source: the authors.

Name	**MSE**	**MSE (pre-pr.)**	**Execution time**
Naive registration	21.71	14.34	5.10 seconds
Fourier-based registration	7.41	5.22	0.74 seconds
Restricted Fourier-based registration	**0.14**	**0.08**	**0.22 seconds**

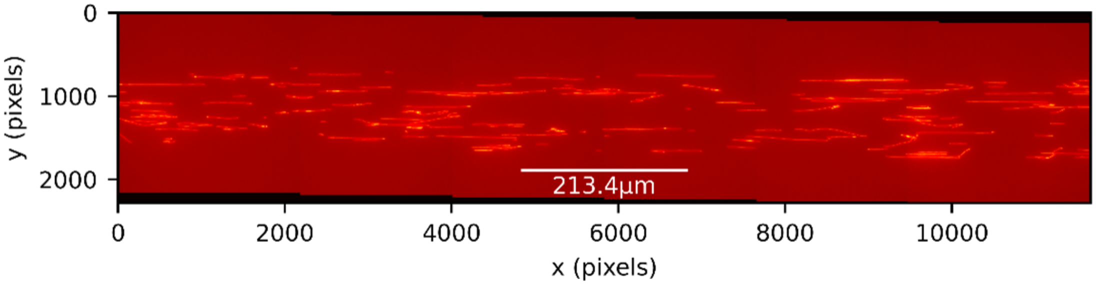

Figure 3-10: A mosaic of 20 stitched contiguous Fluoroscanning images into a super image and a zoomed-in view on a portion of the super image. Since there is usually some degree of vertical shift, there is a step effect in the mosaic. Source: the authors.

3.4.3. Segmentation

The segmentation results can be seen in Table **3-2**, we can see that, despite a slightly higher computational cost, our morphology-based approach significantly improves over the two other proposed methods.

Table 3-2: Performance of molecule detection and segmentation approaches; despite being an order of magnitude slower, the morphology-based approach has a much better Jaccard index. Source: the authors.

Name	**Jaccard Index**	**Execution time**
Otsu's thresholding	0.675	0.12 seconds
Standard deviation-based approach	0.761	**0.11 seconds**
Morphology-based approach	**0.893**	1.77 seconds

3.5. Discussion

The proposed framework for the segmentation and characterization of dyed, elongated DNA molecules using morphology-based techniques allowed us to extract Fluoroscan signals which strongly correlate with the corresponding GC-composition profiles, and to measure their quality based on two criteria: vicinity to noise and other molecules and molecule stretch. Due to the properties of the sensor's point spread function, the resolution achieved by this method can reach up to 550 bp, which, to the best of our knowledge, surpasses that of other approaches centered on enzyme-free DNA content profiling (Marie et al., 2018). Importantly, WScan is not sensitive to changes in the general luminosity of an image and it has been already tested with two microscopes and under a myriad of experimental conditions at LMCG, which means that it has the potential to output even better results as the experimental methods behind Fluoroscanning improve.

In terms of accuracy, the morphology-based approach results in precise segmentation masks although there is a higher computational cost associated to the more complex computational pipeline. The thresholding (Otsu, 1979) and standard deviation-based segmentation approaches are an order of magnitude faster, but often fail due to the fact that the histograms of the Fluoroscanning images do not lend themselves to simple thresholding. On the other hand, when compared to the image processing methods used for optical mapping and nanocoding (Cao et al., 2014; Ravindran & Gupta, 2015), our approach uses several additional filtering steps to ensure that the extracted molecule backbones are not rendered useless by nearby molecules or noise particles.

Future work should comprise the usage of more elaborate, orientation-independent techniques such as matched-Gaussian filters for the morphological segmentation approach, which have been proved to be very effective for supervised segmentation tasks. In addition, modern semantic segmentation neural network architectures could have reasonable performance despite the limited amount of available data. However, we have to note that DNA molecules are usually stretched in a predictable way, so deep learning-based approaches using specialized loss functions (Shit et al., 2020) might not be necessary given the accuracy of our current approach. Moreover, our methodology allows us to easily enforce the fact that we only want to consider a pixel to be part of the backbone if it is the brightest along a perpendicular cut (even if illumination conditions change), which is not necessarily the case with deep learning methods.

In addition to the above, we argue that it is necessary to propose a formal database architecture for the Fluoroscanning pipeline which allows end-users to discard poor Fluoroscans by examining the original images and the extracted signals. Such a structure should also account for noise and overlaps by partially flagging molecules which contain them so that their usable portions can be exploited in further steps. This will probably be a key step as part of a larger research and software engineering effort to turn Fluoroscanning into a fully automatic system for whole genome analysis.

4 Monte Carlo simulations of Fluoroscanning

4.1. Introduction

In order to study the behavior of DNA molecules during the processes on which the Fluoroscanning system is based, it is necessary to carry out computational modeling of the way DNA and dye molecules interact (Nandi, 2017). This is motivated by the fact the experimental procedures on which Fluoroscanning is based have not been fully standardized yet and it is thus necessary to provide tools that allow us to create genome-sized datasets for testing further steps of the pipeline.

Due to the above, we propose the generation of Monte Carlo Fluoroscans based on physical and chemical properties of DNA and dye. Some of the properties we considered aspects were the quantum yield of the dyes, the dye binding and release probabilities, the dye to basepairs ratio, and the variation in length caused by the intercalation of dye fragments between basepairs, among other parameters related to the phenomena on which the Fluoroscanning system is based.

4.2. Methods

4.2.1. Underlying principles

We rely on a Monte Carlo (MC) scheme that incorporates fundamental aspects from the chemistry of the dyes as well as approximated physical aspects of dye intercalation, molecular diffusion and fluctuations. We focus on bichromophore dyes of the cationic cyanine family, namely the pyridinium or oxazole yellow YOYO-1 and the quinolinium or thiazole orange TOTO-1 (see Fig. 4-2 A-B). The general framework of our MC is to generate a random walk of dye binding (intercalating) and releasing events over independent DNA molecules with a specific sequence. Simulated fluorescence signals are averaged over thousands of MC steps and a consensus Fscan is then generated for the DNA molecule of interest. We selected some characteristics from the dye's chemistry and physics to inform the MC random walk. Similarly, we introduce fluctuations that attempt to mimic experimental conditions as close as possible. Consequently, the MC is parameterized with a set of "macroscopic" variables

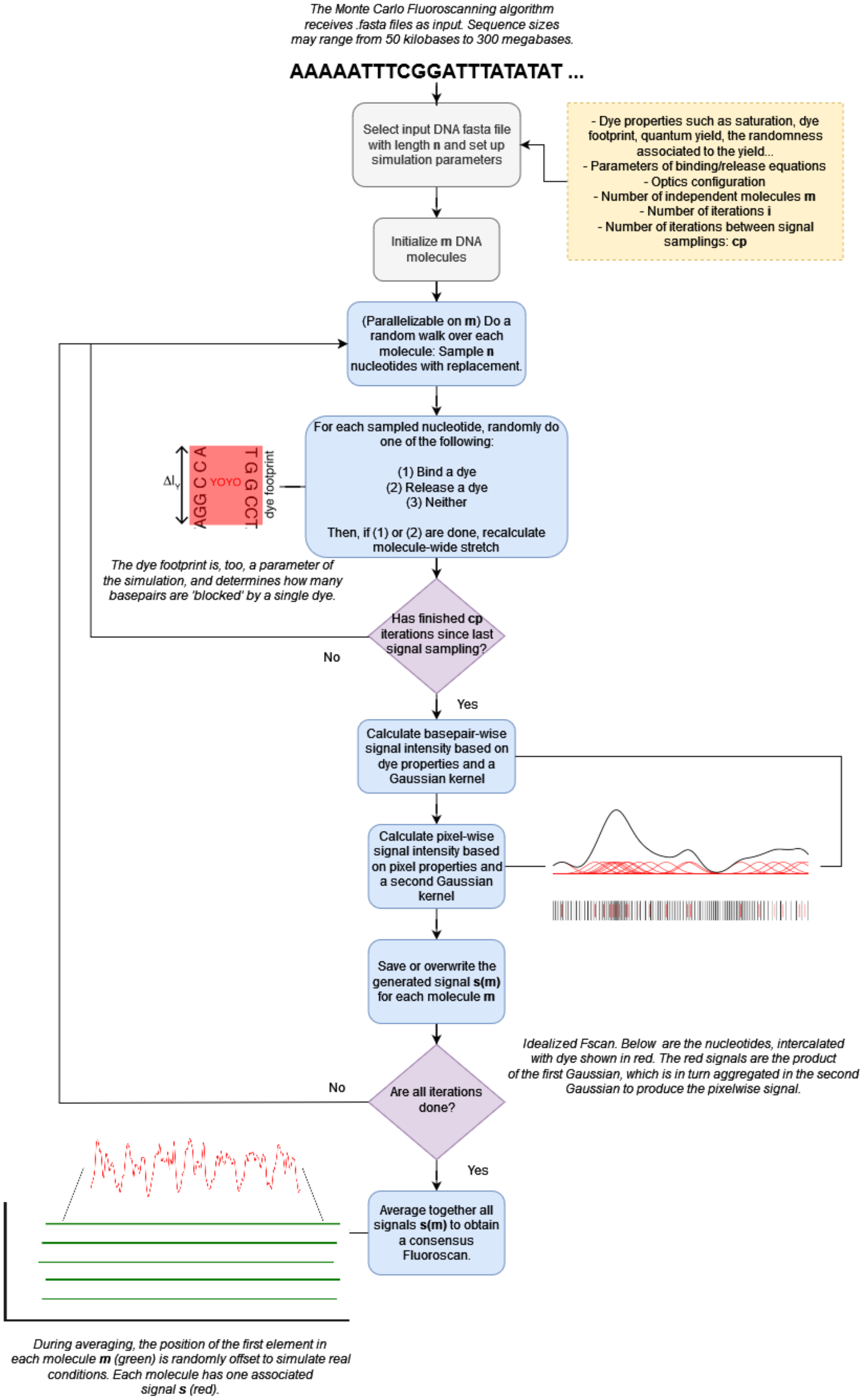
The Monte Carlo Fluoroscanning algorithm receives .fasta files as input. Sequence sizes may range from 50 kilobases to 300 megabases.
AAAAATTTCGGATTTATATAT ...
Select input DNA fasta file with length n and set up simulation parameters
- Dye properties such as saturation, dye footprint, quantum yield, the randomness associated to the yield...
- Parameters of binding/release equations
- Optics configuration
- Number of independent molecules m
- Number of iterations i
- Number of iterations between signal samplings: cp
Initialize m DNA molecules
(Parallelizable on m) Do a random walk over each molecule: Sample n nucleotides with replacement.
For each sampled nucleotide, randomly do one of the following:
(1) Bind a dye
(2) Release a dye
(3) Neither
Then, if (1) or (2) are done, recalculate molecule-wide stretch
ΔI_Y
AGG C C A
YOYO
T G G CCT
dye footprint
The dye footprint is, too, a parameter of the simulation, and determines how many basepairs are 'blocked' by a single dye.
Has finished cp iterations since last signal sampling?
No
Yes
Calculate basepair-wise signal intensity based on dye properties and a Gaussian kernel
Calculate pixel-wise signal intensity based on pixel properties and a second Gaussian kernel
Save or overwrite the generated signal s(m) for each molecule m
Idealized Fscan. Below are the nucleotides, intercalated with dye shown in red. The red signals are the product of the first Gaussian, which is in turn aggregated in the second Gaussian to produce the pixelwise signal.
Are all iterations done?
No
Yes
Average together all signals s(m) to obtain a consensus Fluoroscan.
During averaging, the position of the first element in each molecule m (green) is randomly offset to simulate real conditions. Each molecule has one associated signal s (red).

Figure 4-1: The Monte Carlo Fluoroscanning simulation pipeline. Source: the authors.

that cover the relevant physical-chemical aspects.

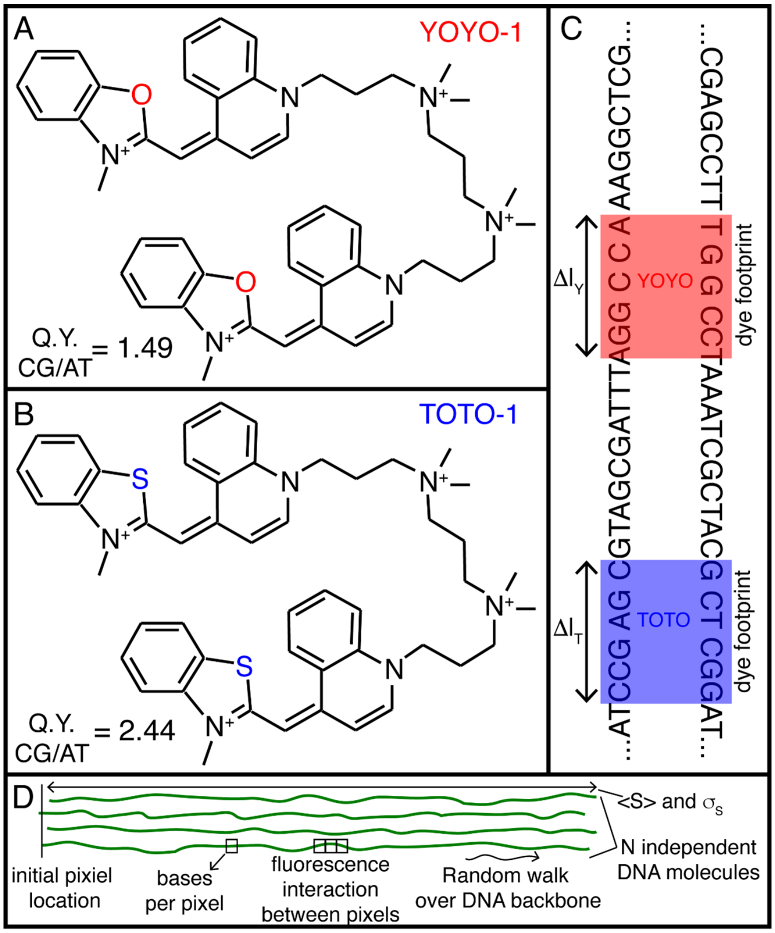

Figure 4-2: **A.** Chemical structure of a bichromophore pyridinium or oxazole yellow YOYO-1 dye. **B.** Chemical structure of a bichromophore quinolinium or thiazole orange TOTO-1. **C.** Schematic of a YOYO and TOTO – DNA intercalation where the DNA lengthens a distance Δl and the dye imposes a footprint based on neighbor exclusion. **D.** Schematics of some of the MC's parameters: location of the molecule end in the initial pixel, size of the pixel, pixel-to-pixel interaction, molecular stretch fluctuations and backbone stretch variations. Source: the authors.

The success behind cationic cyanine dyes is originated on the large degree of fluorescence enhancement when intercalating (binding) on nucleic acids; it has been established that a change in the relative orientation of the benzothiazole and quinolinium rings from skew to planar is responsible of this enhancement and that the intercalation lengthens the DNA molecule (Deligeorgiev et al., 2010; Netzel et al., 1995). In principle, dAdT sequences produce different type of binding than dGdC sequences. Concomitantly, a site that provides greater torsion immobilization with respect to the central methine produces a greater emission enhancement (Netzel et al., 1995). In addition, every dye molecule that binds within two complementary bases imposes an exclusion restriction on the neighboring bases, which defines the dye's footprint (Fig. **4-2** C). Previous experimental studies helped to identify the physical characteristics of the dye intercalation, including how the dye may alter the physical and mechanical properties of the DNA molecules and the process of intercalation (Nyberg et al., 2013). These efforts identified the DNA stretch to be a control variable for the binding/release rate (Bennink et al., 1999). Overall, the probability of a dye molecule to intercalate increases by increasing the stretch due to an increase of the binding rate and a decrease of the releasing one.

From all the properties of fluorescence dyes, there are two that are worthwhile mentioning: (i) the emission lifetime (ELT) and (ii) the fluorescence quantum yield (QY). The excited state of a fluophore exists for a finite time. During this time, it undergoes conformational changes and is subjected to several interactions with its molecular environment. These processes have two important consequences: the energy of the exited electronic singlet state is partially dissipated, yielding a relaxed singlet excited state from which fluorescence emission originates. Second, not all the molecules initially excited by absorption return to the ground state by fluorescence emission. Other processes such as collisional quenching, fluorescence resonance energy transfer and intersystem crossing may also depopulate the relaxed singlet state (Johnson, 2010). The fluorescence quantum yield, which is the ratio of the number of fluorescence photons emitted to the number of photons absorbed, is a measure of the relative extent to which these processes occur. YOYO-1 and TOTO-1 have longer emission lifetimes that other cyanine dyes, with an ELT in the range of 3 to 5 nanoseconds (Netzel et al., 1995; Shapiro, 2004). More importantly, the average emission lifetimes on dAdT and dGdC do not vary (Netzel et al., 1995). Consequently, the only timeline restriction on an experiment is the time needed for the destruction of the excited fluorophore due to photosensitized generation of reactive oxygen species, namely photobleaching. On the other hand, there are differences in the emission QY between the pyridinium and quinolinium dyes when bound to dAdT and dGdC. For instance, QYs for YOYO-1 are reported to be 0.43 and 0.64 when bounded to dAdT and dGdC, respectively, while for TOTO-1 are 0.16 and 0.39 (Netzel et al., 1995). It results in a GC/AT QY ratio of 2.44 for TOTO-1 and of 1.49 for YOYO-1. Therefore, QY becomes a major element of the in-silico Fscan calculation through the MC.

In laboratory experiments, the molecular presentation of the DNA molecules starts with preparation and addition of the dye to the polymer solution. The conditions of ionic strength, temperature, pH and concentration are engineered and controlled with high laboratory standards to ensure consistency and to minimize statistical errors. For the fluoroscanning measurements, the molecules are elongated and presented on a surface, typically following optical mapping mounting techniques (Tang et al., 2015). Here, the molecular stretch is controlled with microscale precision, thereby resulting on an average stretch with a standard deviation between 3 to 8% for an ensemble of mounted molecules. Using the image-processing methods described in Chapter 3, we identify the end of the molecules (initial pixel) and the Fscan signal is measured. The Fscans are then used to organize and orient identical molecules. The optics' resolution dictates the size of the pixel, which is translated to the average number of bases per pixel. Common values of optical resolution determine that the number of bases per pixel is between 90 and 150. The fluorescence intensity of a particular pixel is a collective and coherent optical construction of all the "active" dyes within the pixel and from interactions between intensities from the neighboring pixels.

The MC constructs Fscans through a random walk over the sequence of the molecule of interest, comprising dye fluorochrome intercalation (binding) and release events. These MC Fscans are predicated by the binding and photophysical properties of dye-DNA complexes reported in the literature and measured in the experiments carried out at the LMCG, which include: i) A non-overlapping dye "footprint" [1 dye molecule /4 bp] (Johansen & Jacobsen, 1998; Spielmann et al., 1995), with dye binding ("loading") reflecting experiment d/b ratios (available dye amount limited at any d/b ratios or unlimited, resulting in various dye loading up to a full loading of 1 dye/ 4 bp); ii) sequence dependent quantum yields of YOYO-1 and TOTO-1, as described above (Netzel et al., 1995); iii) MC Fscan length (assuming B-DNA) adjusted for intercalation-mediated elongation effects, 0.51 nm/dye molecule as reported by Günther et al. (2010) by fitting the force extension curves of YOYO-1 bound DNA; and iv) imaging system parameters of pixel size (106.7 nm, 64.3 nm; Molecular Devices, Zeiss, respectively) and measured point spread function (PSF). The PSF was measured by averaged Gaussian fitting on images of 100 nm diameter fluorescent beads (FluoSpheres Size Kit 2, Carboxylate-modified Microspheres, yellow-green fluorescent (505/515); Invitrogen; CAT No. F8888) images: Molecular Devices (σ) = 1.11 +- 0.03 pixels), and the Zeiss system (1.67 +- 0.06 pixels).

During a step, a base is selected randomly along the DNA molecule and a probability tree is used to decide whether the dye is bind or released. The MC is guided by chemical and physical information. The schematics of some of the MC variables are listed in Fig. **4-2** D. The MC simulation starts by defining the sequence and the number of independent and identical molecules to be considered. The MC performs the chemically informed random walk on each molecule independently. The end base of each molecule is placed randomly within the initial *in-silico* pixel, introducing the first level of fluctuation to the MC. The average stretch and its standard deviation are input parameters that determine the second and third level of experimental fluctuations. Each molecule is constructed base-to-base following a random walk through the DNA backbone using 0.34 nm as the maximum distance between bases and satisfying the average stretch and its deviation. The probabilities for a dye to bind or to be released are calculated from a hyperbolic tangent function of the average stretch. The parametrization of these functions follow previous experimental conclusions and experimental Fscans. The quantum yield, the footprint and the intercalating length are central input variables that are collected from the chemistry of each dye.

For our simulations, a MC step is defined as the number of random events to bind or release a dye to each base pair of the DNA molecule of interest. In other words, a MC step for a 21 kbp long DNA covers 21 thousand random events. We found that a consistent cFscan is obtained after 20 to 40 thousand MC steps. Block averaging is performed every 1,000 MC steps, where the dyes that are intercalated are used to generate the in-silico fluorescence signals from Gaussian functions. The optical intensity in a pixel is constructed from all the individual Gaussians within the pixel and from the intensity interaction with neighboring pixels. We used a hyperbolic tangent function for the pixel-to-pixel interaction with a width of 3 pixels. The average molecular stretch defines the binding (PB) and the release (PR) probabilities. During a MC event, the probability of a binding or releasing event is selected with a MC probability of PMC = 50%. Subsequently, a base pair is selected randomly and its status is used to delineate the success of the MC event. The state of a base pair can be either “free”, intercalated with a dye – dyed- or “within the footprint of a neighboring dye”. During a binding event, the dye intercalates a free base pair according to an acceptance probability AB = min(1, PB). Similarly, a release event will free the intercalating dye of a dyed base pair with according to AR = min(1, PR). During the block averaging, we track the average number of dyed base pairs, the average stretch, the MC acceptance/rejection ratio, the base intercalation fluctuations, among other variables. Recall that during a dye intercalation the DNA lengthens, therefore the number of bases per pixel and the base-pixel identity changes during every accepted MC event.

The MC simulation also accounts for experimental noise factors stemming from degraded “dark” fluorochromes, variable luminosities of individual dyes, imaging noise, and DNA stretch. The amount of noise required to closely simulate the experiment is empirically tested in Chapter 5 by comparing MC Fscans to Fscans derived from monomers (monFscans).

4.2.2. Implementation

Monte Carlo experiments are a class of computational algorithms in which problems whose exact solutions are extremely hard to calculate in a deterministic way are solved using random sampling to obtain an approximate solution. Many physical and mathematical problems have been addressed employing this kind of solution, however, this usually implies high computational costs, so highly optimized and/or parallel implementations are often preferred. The MC Fluoroscan simulation algorithm was thus developed in **C++** 11 as it is one of the fastest programming languages used in scientific computing (International Organization for Standardization, 2012; Pereira et al., 2021). We used the **config4cpp** library for parsing configuration files, the **Eigen** numerical computing library for highly optimized array operations (Guennebaud & Jacob, 2010), and the **Open MP** library for parallel processing (Chandra et al., 2001). We compiled the code with the G++ open source compiler with several optimization flags and run most simulations on the CONDOR HTC cluster provided by the University of Wisconsin, Madison as well as in the UNICA cluster of Universidad Nacional de Colombia. The full process is described in Figure **4-1**.

4.2.3. Random number generation

All random numbers were generated using the C++ 11 implementation of the Mersenne Twister pseudo random number generator (Matsumoto & Nishimura, 1998), which has a period of $2^{19937} - 1$. We employ several instances of the random number generator (initialized with different seeds) when carrying out parallel simulations on the same DNA molecule.

4.2.4. Input data and Monte Carlo parameters

For the purpose of carrying out MC Fluoroscan simulations, we first prepare the data and the simulation parameters. The parameters in the Fluoroscan simulation include the following:

- Sequence: the DNA sequence to be used as input for the simulation.
- Region of interest: the first and last basepair of the DNA region to be simulated.
- Number of molecules: the number of molecules to be simulated.
- Distance between basepairs (d_{bp}): the distance between two basepairs, fixed to 0.34 nm as per the literature.
- Pixel size (s_{px}): The size of each pixel in the simulated sensor, in nanometers, fixed to either 106.7 or 64.3 (depending on the optics) as per experimental measurements at LMCG.
- Simulated sensor's Gaussian width (w_{pix}): The width of the Gaussian for the simulated sensor's point spread function, set to 0.75 experimentally.
- Molecular elongation and standard deviation (el, σ_{el}): How stretched the molecule is, a fully stretched molecule will have a higher dye binding probability and a lower release probability.
- Transition of the binding probability function: defines how fast the probability of binding transitions from 0 to 1, set to 0.42 by default.
- Stiffness of the binding probability function: defines how close to a straight line is the slope of the binding probability function, set to 2.0 by default.

- Transition of the release probability function: defines how fast the probability of release transitions from 0 to 1, set to 0.64 by default.
- Stiffness of the release probability function: defines how close to a straight line is the slope of the release probability function, set to 3.0 by default.
- Increase in length for dye intercalation (Δl): The increase in length for each dye particle bound to the DNA molecule, set to 0.51 as per Günther et al. (2010).
- Limited dye simulation flag: Whether to run a simulation with limited dye.
- Basepair to dye ratio for limited dye simulations ($\frac{bp}{dye}$): The basepair to dye ratio for a limited dye simulation.
- GC quantum yield: The intensity levels associated to dyes bound to GC nucleotides, set to 0.64 for YOYO-1 dye and 0.39 for TOTO dye, as per Netzel et al. (1995).
- AT quantum yield: The intensity levels associated to dyes bound to AT nucleotides, set to 0.43 for YOYO-1 dye and 0.16 for TOTO dye, as per Netzel et al. (1995).
- Dye footprint: The number of basepairs which are blocked by a bound dye fragment, set to 4 (including the main basepair) by default as per experimental measurements at LMCG.
- Methylation sites: The specific n-mers which result in methylation. Each methylation site has its own associated binding and release probability parameters and quantum yields.
- Dark fluorochrome rate: the frequency at which a given fluorochrome might ”stay dark” during the simulation, leading to a diminished contribution of dye to the captured intensity values.
- Dye luminosity perturbation: whether to add a random perturbation to the luminosity yield of bound fluorochromes, and the range of possible perturbations.

4.2.5. Data pre-processing

Prior to carrying out the MC steps proper, several operations are carried out. For simplicity, these are going to be described for a single molecule. First, the data from a FASTA file is loaded into memory and the number of basepairs n in the region of interest are counted. The string is parsed into a vector seq of length n, on which G and C nucleotides are labeled as 1, while A and T nucleotides are labeled as 2.

Several values of interest such as the initial length of the DNA sequence, the maximum number of dye intercalations, and the maximum length after dye intercalations are then calculated. The number of pixels is defined as the maximum intercalation length divided by the pixel size s_{px}. Several masks are also generated at this first step, including a mask for methylated areas, which serves as a look-up vector for calculations related to this feature during further steps. If the relevant flag is turned on, the number of available dye fragments is also established during this stage by dividing the number of basepairs by the $\frac{bp}{dye}$ ratio.

Prior to beginning the simulation, an array containing the distances between every couple of adjacent basepairs is created and a random walk is carried out on it as per the following equation:

$$d_{i,i+1} = (r - 0.5) \cdot 0.05 \cdot d_{bp} \cdot el, \tag{4-1}$$

where r is a random number sampled uniformly between 0 and 1, d_{bp} is the standard distance between basepairs, and el is the molecular elongation. This introduces a degree of randomness in the distances between basepairs which simulates the physical phenomenon of some parts of the molecule not being fully elongated on the positively charged surface. In order to simplify further calculations, we carry out the cumulative sum operation on array d, resulting in an array x which contains the cumulative length of the molecule at every basepair i.

4.2.6. Monte Carlo steps

Each step of the Monte Carlo is carried out by sampling n basepairs in the molecule with replacement (which acts as a sort of bootstrapping). Once a basepair is selected, a Bernoulli random variable is sampled with $p = 0.5$ to decide whether to attempt to bind dye to or release dye from the selected nucleotide. Once an action is selected, a uniform random variable will be sampled taking into account the probability of binding/release, as follows:

$$p(bind) = \frac{1 + tanh(el - tS_{binding})SS_{binding}}{2}, \tag{4-2}$$

$$p(release) = \frac{1 - tanh(el - tS_{release})SS_{release}}{2}, \tag{4-3}$$

where $tS_{binding}$ is the transition speed of the binding probability function, $SS_{binding}$ is the stiffness of the binding probability function, $tS_{release}$ is the transition speed of the release probability function, and $SS_{release}$ is the stiffness of the release probability function. Thus, the state of each basepair is modeled as a simple Markov Chain (see Fig. 4-3) where it can either have a dye molecule bound to it (A) or not (B), based on the random probabilities defined before.

There are, however, two conditions which can prevent dye from binding on a basepair during a given iteration of the Monte Carlo:

- The simulation is carried out with limited dye and there is no more dye available (until some dye is released from the rest of the molecule).

- There is already a dye fragment bound to the basepair or its neighbors.

Likewise, there are is one condition that can result in a release event not happening: the basepair not having a dye molecule bound to it in the first place.

In all such cases, the attempt to bind or release will be aborted and the next basepair will be sampled. Essentially, this means that for many of the n basepairs sampled during a given Monte Carlo iteration no action will be taken at all.

It is important to note that the cumulative length of the molecule stored in x is modified whenever a dye fragment intercalates with it, so a molecule which is fully loaded with dye will typically be longer than molecules that are not, which raises the need to normalize molecular length in some manner during further steps. We also note that the probability of binding is higher than the probability of release in most experimental configuration, so in order to prevent the Monte Carlo from getting stuck on a given state for too many iterations, between 10 and 20% of the molecule is stripped of all dye fragments with a probability of 1% at the beginning of every iteration.

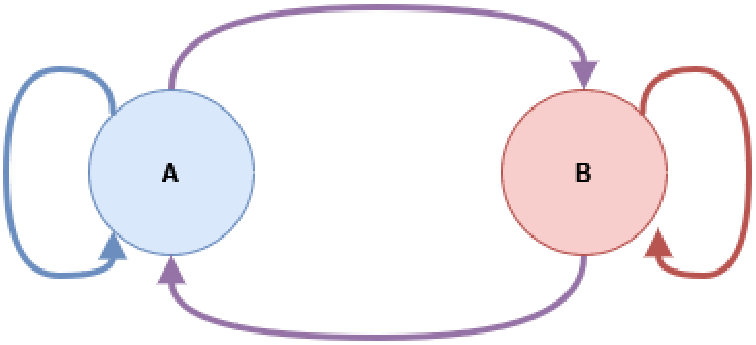

Figure 4-3: Markov Chain of dye intercalation states for a given basepair. Source: the authors.

4.2.7. Optical signal (Monte Carlo Fluoroscan) generation

After a certain number of Monte Carlo steps, two sampling operations are carried out in order to simulate the aggregation of the lighting emitted by each DNA fragment which has some dye intercalated with it and the optics of the MD imaging system, which are described visually in Figure 4-4. The first step consists of aggregating the fluorescence yields of each basepair depending on the local point spread function and the intensity values of their neighborhoods, while the second step entails the aggregation of the intensity levels captured by the simulated camera sensor depending on its point spread function and the established pixel size.

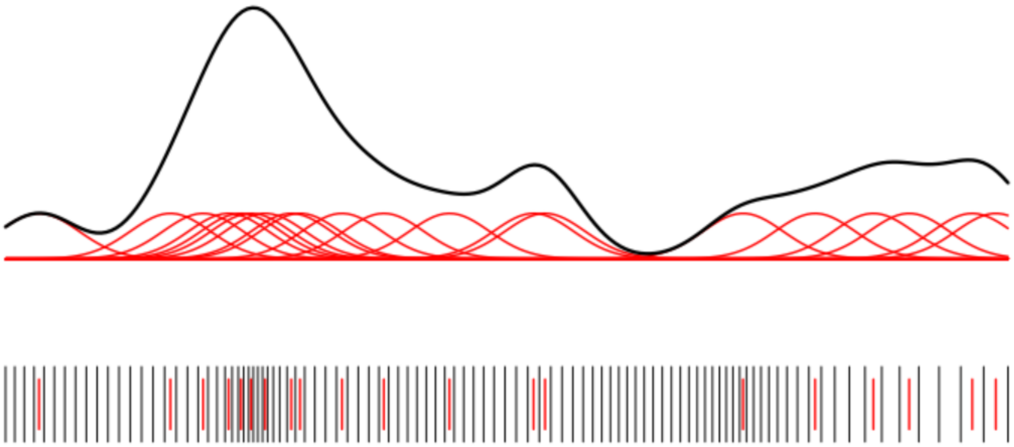

Figure 4-4: An idealized view of fluoroscanning data aligned to its underlying molecule. Bases are marked on the bottom as 100 short vertical black lines. Notably, local distortions of the DNA backbone result in bases being non-uniformly distributed along the interval. Upon staining of the DNA, dye molecules intercalate between neighboring DNA bases; they do so in this example at only 20 spaces (short red vertical lines). Probabilistically, the dye binding sites are assumed to depend in some way on the specific bases that are nearby the binding site. Individual dye molecules fluoresce upon excitation with light; the detectable fluorescence of each dye molecule here is shown as a Gaussian curve (red) positionally centered on the dye location and with a fixed scale. The total fluorescence is a superposition (addition) of the individual curves, and results in the thick black curve in this example. Source: Nandi (2017).

Point spread function at the basepair scale

The perceived yield corresponding to a basepair which has a dye molecule bound to it is impacted by its neighborhood. The point spread function of this phenomenon can be simulated as a Gaussian with the following variance:

$$\sigma_{bp}^2 = (d_{bp}\Delta l) \cdot 0.5, \tag{4-4}$$

where d_{bp} is the standard distance between adjacent basepairs and Δl is the increase in length when a dye binds to a basepair. Likewise, the yield ϕ emitted by a bounded basepair is defined as:

$$\phi(j) = QY(j) \cdot (1 + \zeta(j)) \cdot \big(1 - dark(j)\big), \tag{4-5}$$

where QY is a function defining the standard quantum yield associated to basepair j depending on whether seq_j corresponds to an AT or GC nucleotide and whether or not it underwent methylation, $\zeta(j)$ is a value sampled from a normal distribution $\mathcal{N}(0, \epsilon^2)$[1] which perturbs the yield of basepair j, and $dark(j)$ is a random Bernoulli variable sampled with p defined by the dark fluorochrome rate parameter, and it defines whether the dye bound to basepair j will produce any luminosity; if the rate is set to 0, the simulation will not have any dark fluorochromes. Given the above, the yield Φ emitted by a given basepair is calculated as follows:

$$\Phi(i) = \sum_{j=i-20}^{i+20} \phi(j) \cdot e^{\frac{-d(i,j)^2}{\sigma_{bp}^2}}, \tag{4-6}$$

where j is a basepair in the neighborhood comprising the 40 positions surrounding basepair i, and d is the one-dimensional distance between central basepair i and neighboring basepairs j, which is calculated from the cumulative length vector x. This process results in an aggregated intensity array Φ with one intensity value per basepair, which is used as the input for the next step.

Point spread function at the camera sensor scale

Once the per basepair intensity values are calculated, we can simulate the alignment of the camera to the basepairs and calculate the intensity of each pixel as an iterative process with a pixel-level variance defined based on the following equation:

$$\sigma_{pix}^2 = 2(w_{pix} \cdot s_{px})^2. \tag{4-7}$$

The Gaussian bell which represents this process, with pixel width $w_{pix} = 0.75$, can be seen in Figure 4-5. The result of the described process is an array which contains the intensity

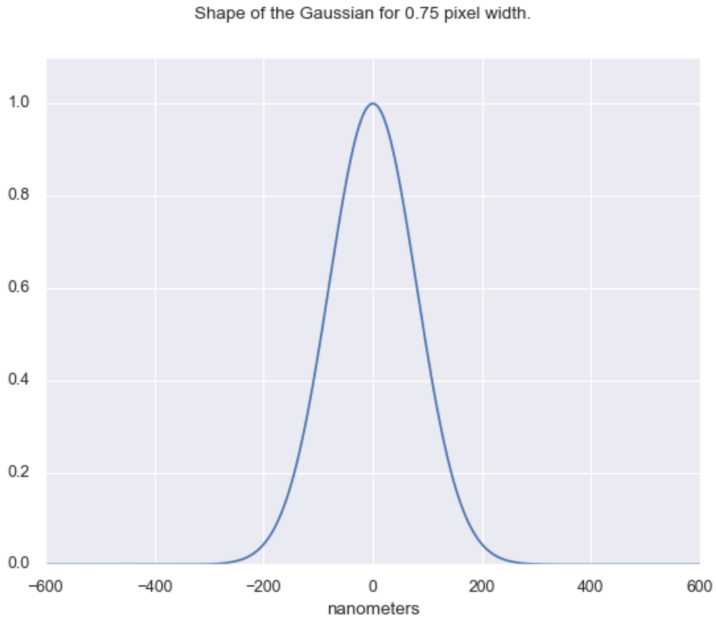

Figure 4-5: Gaussian bell representing the contribution of basepairs to the aggregated intensity of a given pixel. The closer a basepair is to the center of the area covered by the pixel, the more of its intensity will be captured by the simulated camera sensor. Accordingly, basepairs close to the area covered by a given pixel can also impact the total intensity due to the properties of the point spread function.

values equivalent to each pixel that the imaging system would have captured. That is, a Monte Carlo Fluoroscan of the molecule after being dyed with fluorescent dye.

The equation for obtaining the intensity level I associated with camera pixel c is defined as follows:

$$I(c) = \sum_{j=i-d_{pix}}^{i+d_{pix}} \Phi(j) \cdot e^{\frac{-d(x_c,j)^2}{\sigma_{pix}^2}}, \tag{4-8}$$

where d_{pix} is four times the number of basepairs expected to be contained within the span of a pixel $(4\frac{s_{px}}{\Delta l})$,$x_c$ is the estimated position of the center of the camera pixel perpendicular to the molecule, and i is defined as the position of the basepair that is closest to x_c. When applied over all camera pixels, this results in a full signal I: the MC Fluoroscan. One final post-processing step can be carried out in order to simulate imaging noise: we optionally add noise on top of each pixel in signal I, sampled from a Gaussian distribution $\mathcal{N}(0,\, \rho^2)$,

[1] This means that 95% of the time, the perturbation will be in the interval $[-\epsilon \cdot QY(i), \epsilon \cdot QY(i)]$. If ϵ is set to 0, no perturbation will occur.

where ρ is a percentage of the signal's values adjusted as part of the work done in Chapter 4.

4.2.8. Consensus MC Fluoroscan aggregation

When the aforementioned process is carried out for several molecules, the signals can be aggregated onto a single, consensus Monte Carlo Fluoroscan of the molecule, which due to the noise introduced by the randomness in length and dye binding is closer to the experimental results. This is done by first using a simple cross correlation to re-align the molecules and then adding the intensities of each Fluoroscan and dividing it by the standard deviation after subtracting the mean (see Figure **4-6**). Any pixel which is not covered by all of the molecules is discarded in order to avoid sudden dips in intensity at the edges of the consensus signal.

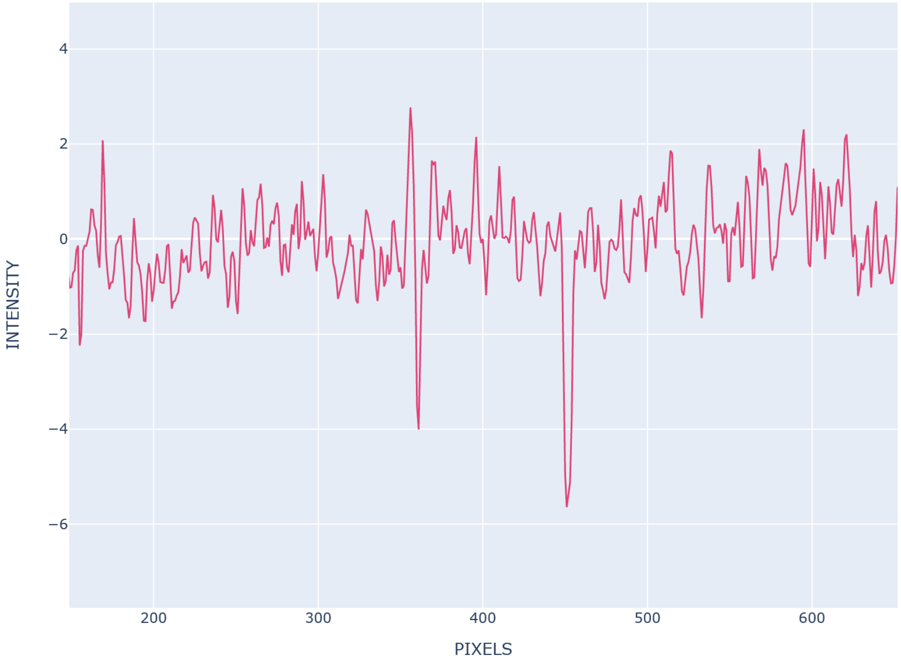

Figure 4-6: Consensus Monte Carlo Fluoroscan aggregated from the Monte Carlo Fluoroscans of 5 simulated BAC 876A24 molecules. The two dips around pixels 360 and 450 correspond to AT-rich areas. Source: the authors.

4.2.9. Comparing Fluoroscans and GC profiles with MC Fluoroscans

In order to compare actual Fluoroscans with MC Fluoroscans we standardize both to have mean 0 and unit standard deviation and we employ cross correlation to align them in order to provide a better comparison. We were able to measure aspects such as the frequency of peaks and the average elongation in order to adjust the best conditions for MC Fluoroscan generation, and we established a Gaussian width of 0.75 for the pixel point spread function resulting in the best fit to the experimental data as shown in Figure 4-7.

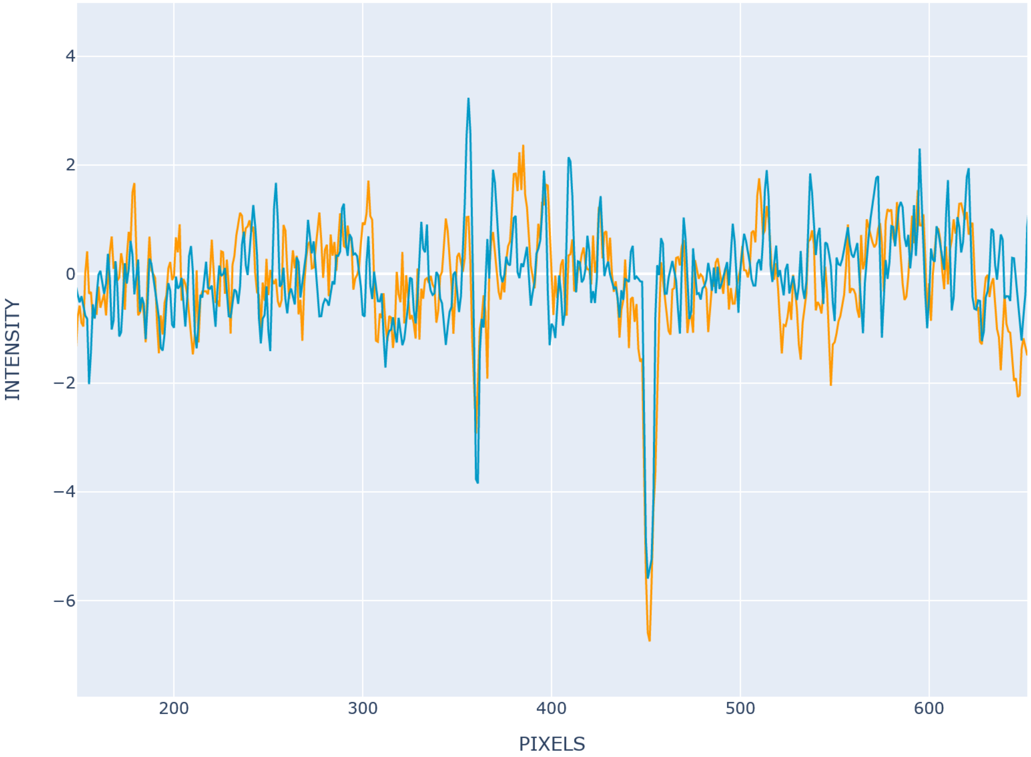

Figure 4-7: Comparison of a Consensus MC Fluoroscan (cyan) and an actual Consensus Fluoroscan (yellow) obtained from laboratory experiments on BAC 876A24. The two big dips around pixels 360 and 450 are easily identifiable in the experimental data. Source: the authors.

The comparison of a GC profile with the MC Fluoroscan is a similar procedure. We read the DNA sequence and establish a window size based on the configuration employed for the computational experiment. This way, we are able to compute a GC profile where high % GC areas are represented as peaks and high % AT areas are represented as valleys. In order to guarantee a comparison independent of the intensity and GC % units, we repeat the standardization procedure (see Figure 4-8).

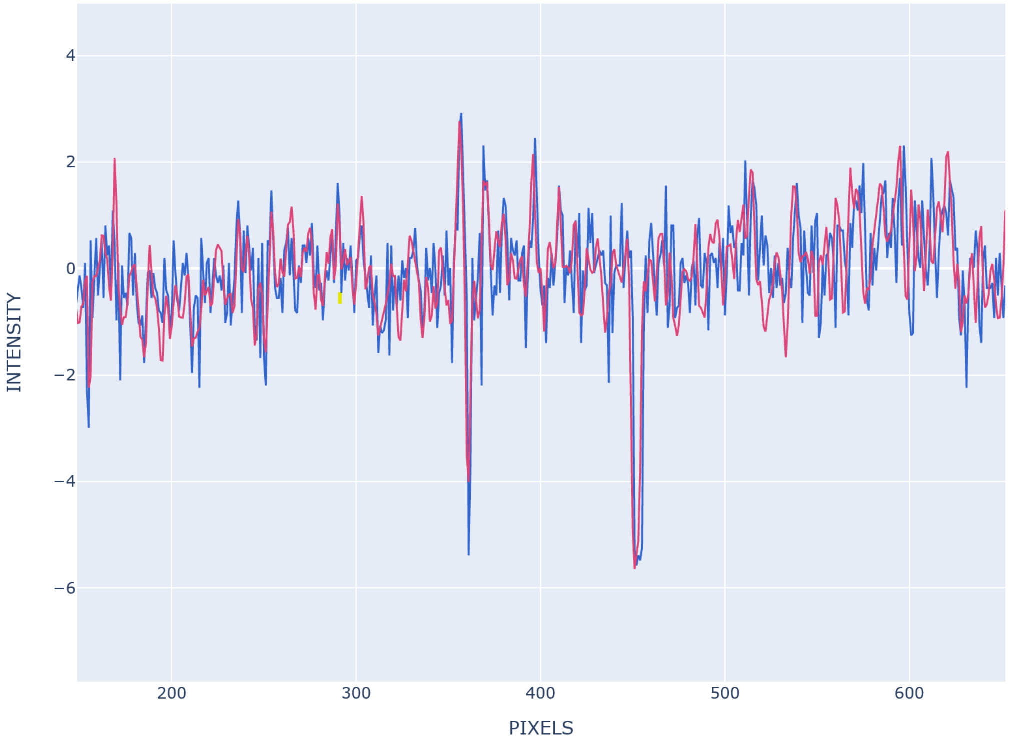

Figure 4-8: Comparison of a Consensus MC Fluoroscan (red) and a GC profile (blue) obtained from the DNA sequence of BAC 876A24. Under ideal conditions, the MC Fluoroscan is very similar to the GC profile. Source: the authors.

4.3. Results

After comparing the Monte Carlo Fluoroscans with Fluoroscans extracted from actual dyed DNA molecules as well as the ”ground-truth” GC profile of the BAC itself (Figures **4-7** and **4-8**), we see that the simulation approach is close to the physical and chemical phenomena behind the acquisition of Fluoroscans. By adjusting the parameters of the simulation it was possible to closely approximate the resolution present in real Fluoroscans. The extent at which our simulated data agrees with experimental data is further explored in Chapter 4. Using our final parameters, adjusted through collaboration with the LMCG, we generated Fluoroscans for most of the human genome.

4.4. Discussion

Monte Carlo Fluoroscans properly reproduce the most relevant landmarks present in both the GC% profiles of the studied sequences and Fluoroscans resulting from the experimental procedures described in Chapter 3. The flexibility of the simulation setup allows us to manipulate physicochemical conditions and noise factors such as varying levels of elongation, different point spread functions for the intensity profiles, differential binding probabilities for specific sequences, variations in dye loading, among others. With the development of this tool, we enable the generation of massive datasets of Fluoroscanning data based on any given DNA sequence instead of being limited to what can be achieved with experimental data. Most importantly, the generation of such datasets is the basis for the next objective: studying the agreement between simulated and experimental data, and determining whether it is possible to use the Monte Carlo to test experimental conditions ideas at a fraction of the cost it would take to test them in the laboratory.

To the best of our knowledge, this is one of the first attempts to develop simulations that can guide experimental procedures for a whole genome analysis system based on sequence fluorescence intensity profiling (that is, using no restriction enzymes). Our approach considers both the physicochemical conditions of the experimental setup and a significant number of noise sources, ranging from phase noise to perturbations in the luminosity output of the dye. In contrast, Müller et al. (2019) used simulations to determine whether a whole human genome could be mapped based on their sequence profiling approach, in a fashion similar to simulating restriction maps in optical mapping work. On the other hand, Lee et al. (2018) focused on the physical properties of DNA molecules confined in nanochannels. Compared to their proposals, ours centers on modeling a wide array of parameters so as to test different experimental conditions in a fast manner.

Finally, we must note that the current version of the Monte Carlo simulation is implemented in C++ and optimized for multi-threaded execution. However, due to the complexity of software development in C++, it is relatively difficult to introduce new features. Due to this, we decided to translate the Monte Carlo to Python, using the **Numba** and **Numpy** libraries for performance-critical sections and the standard library for non-critical tasks. One major advantage is that new features can now be added with a fraction of the time that required for doing so in the original C++ version, and that the code-base is now much more accessible to researchers that are not familiar with C++. As of the writing of this document, the Python version is faster than the C++ version (mainly due to clever use of vectorization for parts of the code that previously were for loops), although it lacks checkpointing capabilities and should be used only for shorter runs. In the future, we expect this missing feature to the Python version so that it can be used for large-scale simulations of complete genomes.

5 Studying MC Fluoroscans and their usefulness for genomic map generation

5.1. Introduction

In previous chapters, we developed a pipeline for extracting Fluoroscans from digital images derived from experiments at the LMCG. Likewise, we developed highly customizable algorithm capable of simulating the Fluoroscan generation process under various conditions. As part of this chapter, we hope to use the tools we built before to address two main questions:

- Whether or not MC Fluoroscans agree with findings derived from large scale experimental data.
- Whether or not MC Fluoroscans parameters can be fitted to achieve maximum similarity to experimental data.

By addressing both questions, we hope to bring the simulation and experimental aspect of our research into a common space where simulated data can be used to guide laboratory experiments, thus saving precious time and resources and driving us closer to the goal of generating genomic maps from fluorescence microscopy images.

5.2. Comparing Monte Carlo Fluoroscans with Fluoroscans derived from data-driven approach

The MC approach was not the first attempt at studying the properties of Fluoroscanning. Previously, Nandi (2017) carried out a data-based approach to the generation of synthetic Fluoroscanning data. He used a huge human chromosome dataset obtained using the nanocoding system was used to train a tree-based Gradient Boosting regression algorithm. As part of this process, Nandi considered a set of features based on DNA n-mers with lengths 1 to 5 (base mers) as well as the same mers present in the neighboring pixels (labelled as + and ++, as in Figure 1). This simulates a Gaussian kernel and allows to consider the importance of mers at varying distances from the pixels that are generated through this approach. The feature importance was calculated using nodal impurity (explained in the first document) and resulted in AT-majority mers with lengths two to three having the most importance.

Monte Carlo Fluoroscan generation, on the other hand, is based on physical intuitions of the behavior of DNA fragments interacting with dye molecules. This approach has been useful to generate big Fluoroscan datasets without requiring us to carry out the huge number of experiments that were necessary for training the Gradient Boosting algorithm. However, this does not mean that the importance analysis results obtained with the Gradient Boosting are inadequate, as recent experimental data from tests done at the Schwartz Lab have shown that AT-rich areas are more informative in terms of characterizing the underlying features of a given DNA molecule. Thus, we expect both approaches to output similar results.

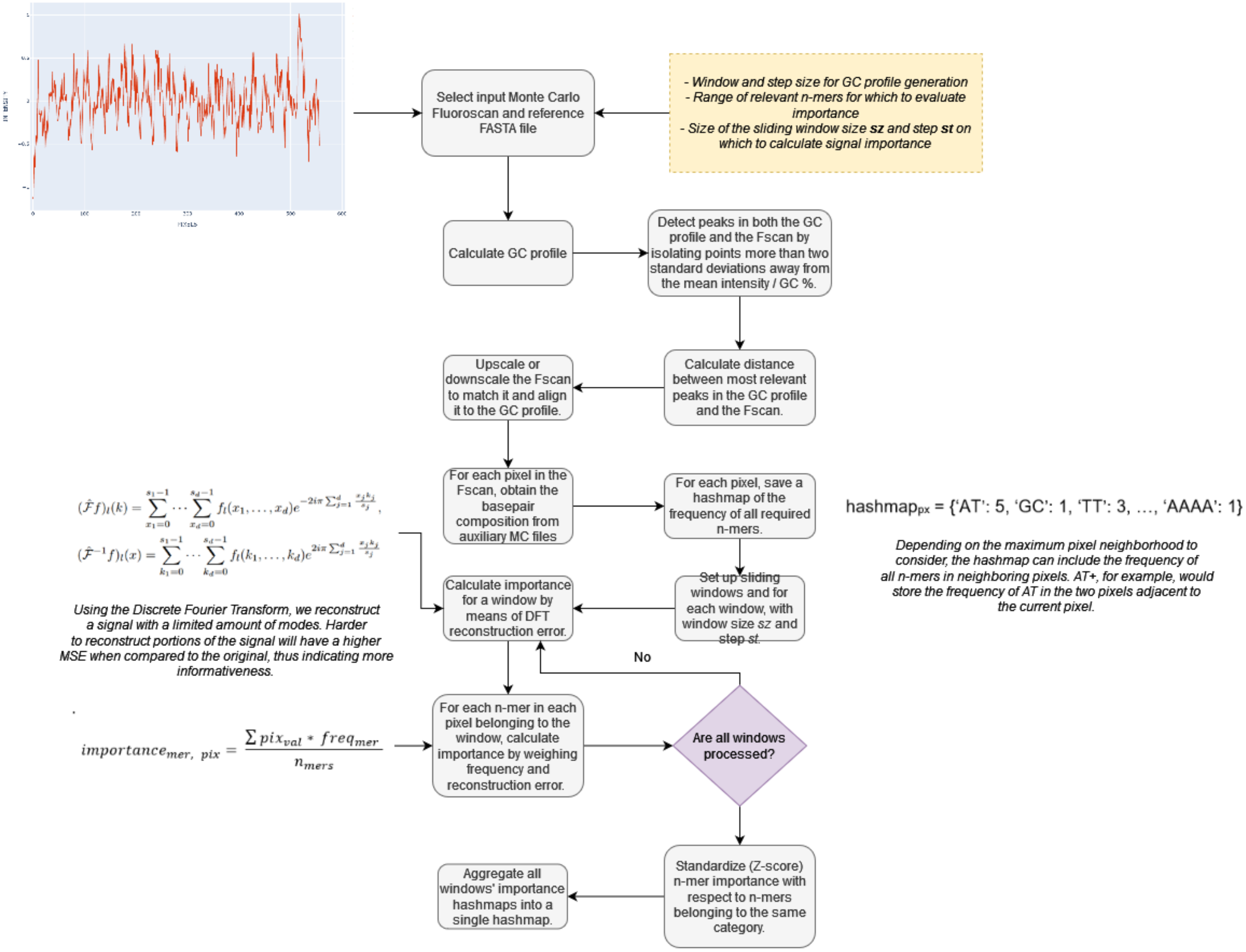
Select input Monte Carlo Fluoroscan and reference FASTA file
- Window and step size for GC profile generation
- Range of relevant n-mers for which to evaluate importance
- Size of the sliding window size sz and step st on which to calculate signal importance
Calculate GC profile
Detect peaks in both the GC profile and the Fscan by isolating points more than two standard deviations away from the mean intensity / GC %.
Calculate distance between most relevant peaks in the GC profile and the Fscan.
Upscale or downscale the Fscan to match it and align it to the GC profile.
For each pixel in the Fscan, obtain the basepair composition from auxiliary MC files
For each pixel, save a hashmap of the frequency of all required n-mers.
hashmap$_{px}$ = {'AT': 5, 'GC': 1, 'TT': 3, …, 'AAAA': 1}
Depending on the maximum pixel neighborhood to consider, the hashmap can include the frequency of all n-mers in neighboring pixels. AT+, for example, would store the frequency of AT in the two pixels adjacent to the current pixel.
$(\hat{\mathcal{F}}f)_l(k) = \sum_{x_1=0}^{s_1-1} \cdots \sum_{x_d=0}^{s_d-1} f_l(x_1,\ldots,x_d) e^{-2i\pi \sum_{j=1}^{d} \frac{x_j k_j}{s_j}},$
$(\hat{\mathcal{F}}^{-1}f)_l(x) = \sum_{k_1=0}^{s_1-1} \cdots \sum_{k_d=0}^{s_d-1} f_l(k_1,\ldots,k_d) e^{2i\pi \sum_{j=1}^{d} \frac{x_j k_j}{s_j}}$
Using the Discrete Fourier Transform, we reconstruct a signal with a limited amount of modes. Harder to reconstruct portions of the signal will have a higher MSE when compared to the original, thus indicating more informativeness.
Calculate importance for a window by means of DFT reconstruction error.
Set up sliding windows and for each window, with window size sz and step st.
No
$importance_{mer,\ pix} = \frac{\sum pix_{val} * freq_{mer}}{n_{mers}}$
For each n-mer in each pixel belonging to the window, calculate importance by weighing frequency and reconstruction error.
Are all windows processed?
Standardize (Z-score) n-mer importance with respect to n-mers belonging to the same category.
Aggregate all windows' importance hashmaps into a single hashmap.

Figure 5-1: Pipeline for determining n-mer importance, we align the MC Fluoroscan to a reference GC profile and, based on the underlying composition of the MC Fluoroscan, we produce an importance score. Source: the authors.

5.2.1. Measuring feature importance in Monte Carlo Fluoroscans

Having stated the above, it is necessary to provide additional detail on the approach used to measure importance in Monte Carlo Fluoroscans. As part of the Monte Carlo, the quantum yield of every basepair in a sequence is first aggregated with its neighbors to create a full resolution signal using a Gaussian, which is further aggregated to pixel-size resolution employing a second Gaussian. The yield data for all basepairs is readily available as an output of the C++ Monte Carlo routines, for example, for a sequence A A A T there exists a vector of yields which would look like [0.78, 0.91, 0.91, 0.78]. For each basepair there is additionally data on its position in the DNA chain measured in nanometers. In addition, there is information on the yield and the position of each pixel's center along the chain.

Taking this into account, we read all of this data and generate a hash-map structure for every pixel where the occurrence of every possible mer from lengths 1 to 5 is registered. The base mer features are counted by considering all basepairs in the following interval:

$$[px_{i,center} - px\_size, px_{i,center} + px\_size] \tag{5-1}$$

Where $px_{i,center}$ represents the center of pixel i and px_size is the pixel size as defined during the simulation. An example of a hash-map for an arbitrary pixel would be as follows:

$$hashmap_{px} = \{AT : 5, GC : 1, TT : 3, \ldots, AAAA : 1\} \tag{5-2}$$

This hash-map is furthermore expanded by including + and ++ features, in which the n-mer counts of neighbouring pixels are added as independent features (e.g., AT++: 10). This way, we simulate Nandi's approach by considering the same amount of possible mers in a 5-pixel window. Once the n-mer data for each pixel is obtained, we can now measure the importance of each feature for a given dataset. This is done by considering several aspects:

- The intensity level of the pixel in relation to the full-length signal
- The frequency of the n-mers
- The degree to which a given pixel which contains the n-mers remains after applying a Fourier filter
- The size of the pseudo-pixel which directly impacts the Gaussian function used to weigh feature importance (this step is relevant during the generation of the hash-map)

The process begin by calculating a Fourier reconstruction of a 50-pixel size window f with the Discrete Fourier Transform F:

$$(\hat{\mathcal{F}}f)_l(k) = \sum_{x_1=0}^{s_1-1} \ldots \sum_{x_d=0}^{s_d-1} f_l(x_1, \ldots, x_d) e^{-2i\pi} \sum_{j=1}^{d} \frac{x_j k_j}{s_j}. \tag{5-3}$$

In order to filter noise from the signal, the inverse $\hat{\mathcal{F}}^{-1}$ is calculated with modes truncated to a set $Z_{k_{max}}$ as per a k_{max} value as follows:

$$(\hat{\mathcal{F}}^{-1}f)_l(x) = \sum_{k_1=0}^{s_1-1} ... \sum_{k_d=0}^{s_d-1} f_l(k_1, ..., k_d) e^{-2i\pi} \sum_{j=1}^{d} \frac{x_j k_j}{s_j}, \tag{5-4}$$

$$Z_{k_{max}} = \{(k_1, ..., k_d) \in Z_{s_1} \times ... \times Z_{s_d} | k_j \leq k_{max,j} \text{ or } s_j - k_j \leq k_{max,j}, \text{ for } j = 1, ..., d\}. \tag{5-5}$$

Where k_{max} is empirically defined as 80% of the number of modes for the window. By doing this, we keep only the most important features of the signal Brady (1992). Once the window f is filtered, the importance for all mers in pixels within the window is calculated as follows:

$$importance_{mer,pix} = \frac{\sum pix_{val} \cdot freq_{mer}}{n_{mers}} \tag{5-6}$$

In addition, to avoid bias toward smaller n-mers, the importance of every with a given length and at a given distance from the center pixel (base, + or ++) is standardized (mean 0, standard deviation 1) with respect to mers belonging to the same category. After this, we obtain a hash-map of importance levels for every pixel which measures quantitative importance instead of the frequencies:

$$imp\_hashmap_{px_i} = \{AT : 1.46, GC : 1.2, TT : 2.33, .., AAAA : 0.1, ...\} \tag{5-7}$$

The final importance hashmap is thus the sum of importances for all pixels in the studied signal (see Figure **5-1** for a summary of the complete procedure, and Figure **5-2** for some examples of the resulting importance scores).

5.2.2. Results of the importance analysis

Ideally, there should be consistency in terms of the mers that appear most frequently as important, but this aspect tends to fluctuate due to the particularities of each DNA fragment or chromosome. For this reason, it is possible to aggregate data from several chromosomes; we thus compare the importance of three datasets:

- Chromosome 1.
- Chromosome 19.
- Chromosomes 1 and 19.

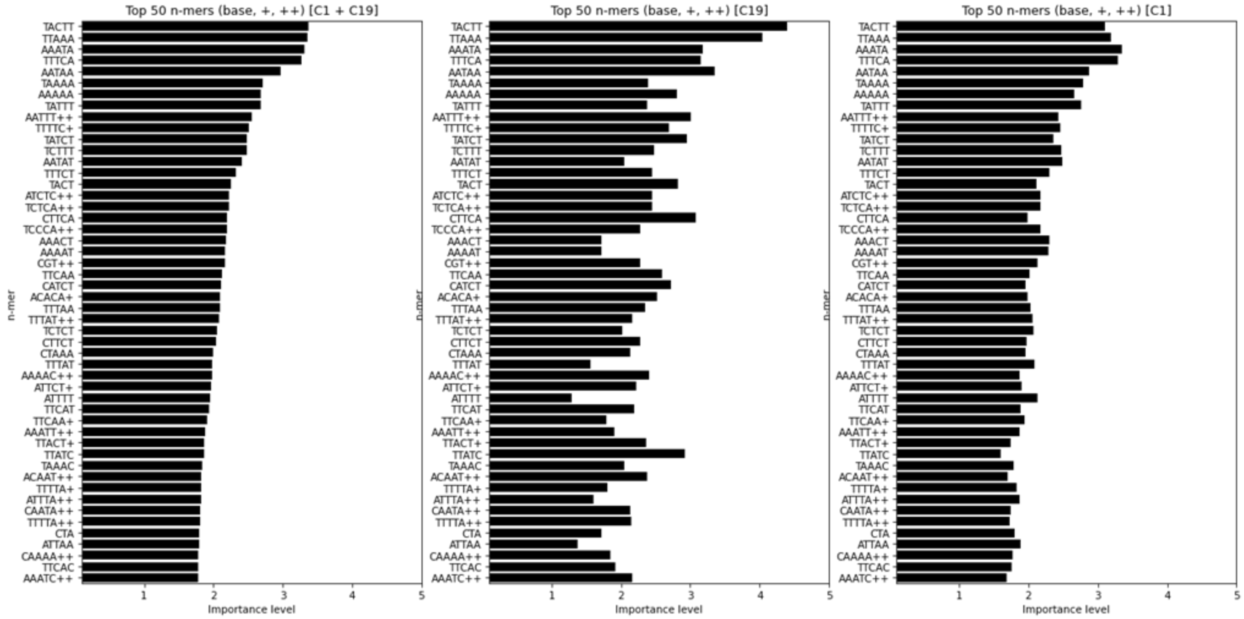

Figure 5-2: N-mer importance results on chromosomes 1 and 19, with 106.7 nm pixel size. We can see that mostly 5-mers are considered important, with little representation of smaller n-mers. Source: the authors.

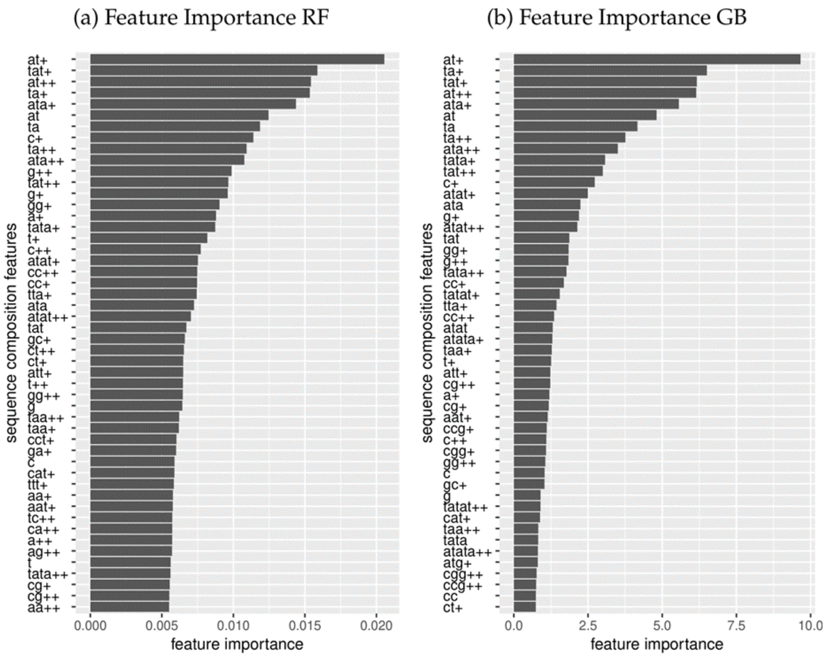

Figure 5-3: Nandi's feature importance results, using Random Forest (left) and Gradient Boosting Trees (right) algorithms fitted to Nanocoding data. Source: Nandi (2017).

The experiments were carried out with the base conditions used for experiments with the Monte Carlo, the most important of these being the pixel size of 106.7 nm that was employed in the simulations (see Figure **5-2**).

The results show that there is agreement between the results on both chromosomes, aside from the expected fluctuations. The most frequent mers are AT-based 5-mers and 4-mers in the central pixel and the ++ pixels to each side of the center, which is also in agreement with Nandi's results. However, this differs from the prevalence of AT-based 1-mers and 2-mers in Nandi's work, as seen in Figure **5-3**.

Considering the above, we hypothesize that the results might be heavily influenced by the pixel size employed in the simulations, which motivates us to carry out simulations with several pixel sizes and re-evaluate our results.

Pseudo-pixel testing

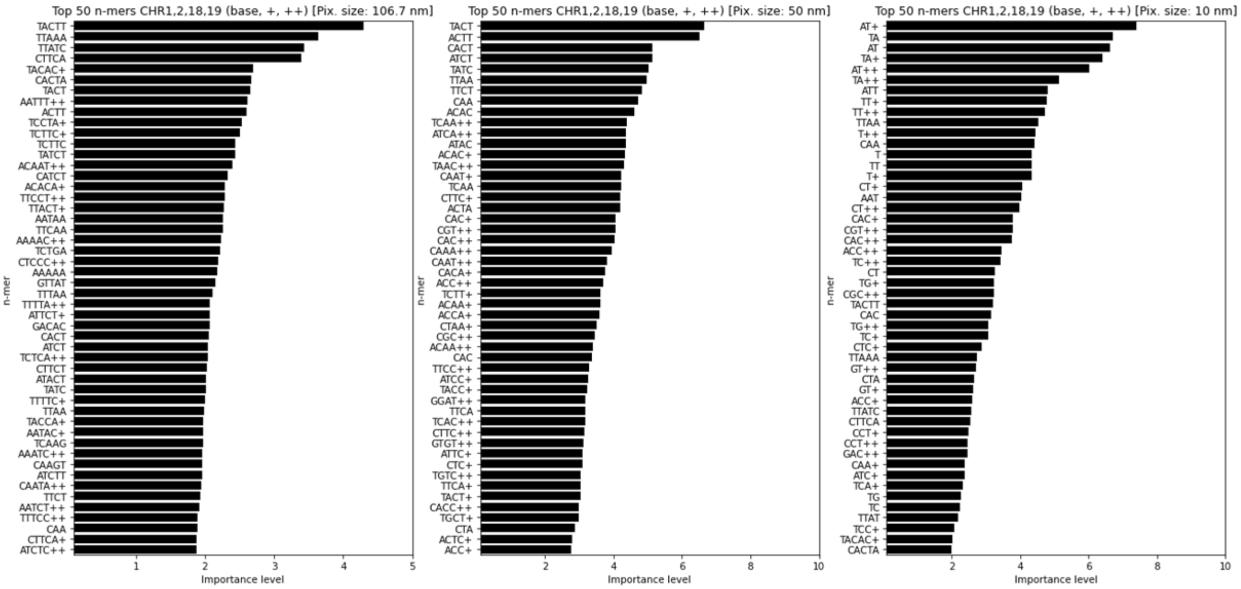

Figure 5-4: Pseudo-pixel size test results on an aggregate dataset comprised by chromosomes 1, 2, 18, and 19 with 106.7, 50, and 10 nm pixel size. Smaller pixel sizes result in a more varied representation of n-mers of different lengths, which is in turn closer to what was obtained by Nandi (2017). Based on this data, we were able to confirm that there was an agreement between both methodologies for computing feature importance. Source: the authors.

To determine if the pixel size employed for the gaussian pseudo pixel in the simulation had an impact on the relevance of specific n-mer sizes, we carried out tests with several values

Table 5-1: Comparison of n-mer frequency in the top 20 n-mers for each pixel size, vs Nandi's Gradient Boosting. Source: the authors.

Configuration	**1-mers**	**2-mers**	**3-mers**	**4-mers**	**5-mers**
Gradient Boosting	10%	40%	30%	20%	0
106.7 nm Monte Carlo	0	0	0	10 %	90 %
50 nm Monte Carlo	0	0	20 %	80 %	0
10 nm Monte Carlo	15 %	60 %	20 %	5%	0

Table 5-2: Comparison of AT frequency in the top 20 n-mers and top 50 n-mers for the 10 nm Monte Carlo vs Nandi's Gradient Boosting. We can see that both approaches have a similar % of AT n-mers in both the top 20 and top 50, which suggests agreement. Source: the authors.

Configuration	**% of AT in top 20**	**% of AT in top 50**
Gradient Boosting	90.19 %	72.18 %
10 nm Monte Carlo	84.09 %	64.49 %

for the pixel size, in this case we the results with 106.7 nm, 50 nm, and 10 nm. In order to account for a wider region of the human genome, we carried out this analysis on chromosomes 1, 2, 18, and 19.

As it can be seen in Figure 5-4, we get a distribution closer to that obtained by Nandi, albeit there is not the same distance between the 1st and 2nd top mers. As a way to better quantitate the effect of both configurations, we calculate the relative frequency of n-mers with a given size in the top 20 positions for each simulation, as per Table 5-1. In addition to, we calculate the relative frequency of AT and GC mers in the Gradient Boosting analysis and our Monte Carlo-based analysis (see Table 5-2).

From these results, we can see that short, AT-rich n-mers are prevalent in our Monte Carlo simulation (chemistry and physics-informed approach) to a similar degree as they are in the Gradient Boosting (data-informed approach). Importantly, this suggests that Monte Carlo simulations of Fluoroscanning can be used to guide wet lab experiments under a controlled environment.

5.3. Fitting MC parameters to experimental data

At this stage of the Fluoroscanning project, our work was focused on small-scale alignment to enable fitting the parameters of the MC Fluoroscan generation algorithm to the experimental data. For this reason, researchers at the LMCG devised a concatemer sequence comprised of 30+ repeats of the 5.5 mb pUC19-H23 plasmid monomer. catFscans resulting from the concatemer sequence can be easily divided into several smaller segments (monFscans) that can be aligned together with relatively simple signal processing algorithms. On the other hand, our MC algorithm is readily available to generate synthetic Fscans from the same concatemer FASTA sequence (catMCFscans) that can also be divided into smaller segments (monMCFscans). These synthetic Fscans can be used as a theoretical reference for processing experimental Fscans under various conditions, resulting in a golden opportunity for understanding Fluoroscan generation parameters (and therefore noise) through quantitative analyses.

As stated in Chapter 3, our Monte Carlo algorithm incorporates noise factors including dye luminosity variation, degraded dark fluorochromes, phase shifts, and imaging noise. By comparing the synthetic data with experimental data, we validated and quantitated these noise factors from experiments. These noise factors are also reduced correspondingly through the data processing workflow including normalization, filtering by length, filtering by similarity to the synthetic reference and averaging for a consensus. The noise-reducing workflow is especially empowered by large datasets collected from DNA concatemers. After the processing steps, we evaluated signal similarity by Pearson correlation coefficient and signal to noise contrast by information theory (Borst & Theunissen, 1999).

5.3.1. Experimental data and setup

In this chapter, we focus on the concatemer sequence resulting from chaining together 30 repeats of the pUC19-H23 plasmid monomer. Each monomer consists of 5627 bp, for a total 168810 bp in the full sequence. The parameters to fit include the stretch, the percentage of dark fluorochromes, the Gaussian dye intensity perturbation, and the imaging noise. The amount of noise to closely simulate the experiment are empirically tested by comparing the resulting monomer MCFscan (monMCFscan) dataset to the monFscan dataset. DNA stretch variation is assessed by the length distributions of monMCFscans and monFscans. The three noise factors affecting fluorescence intensity, including “dark” fluorochromes, variable luminosities of dyes and imaging noise, are quantitated by cumulative information rate using Borst and Theunissen’s informational theory method (Borst & Theunissen, 1999).

5.3.2. Fscan preprocessing workflow

The preprocessing workflow (see Fig. **5-5**) features a series of filters and analyses designed to minimize outliers in Fscans stemming from variances in intramolecular stretch (S = apparent molecule length / theoretical molecule length of DNA with intercalation-mediated elongation), producing “signal phase shifts,” and overlapping DNA fragments that contribute spurious luminosities to analyte molecules. The workflow is largely implemented using tools available in the **SciPy** (Virtanen et al., 2020) Python library.

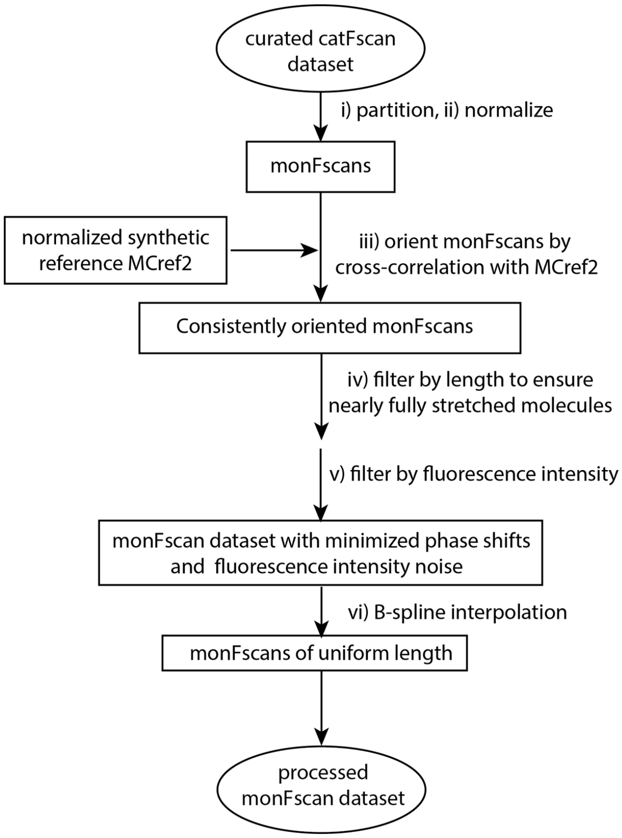
curated catFscan dataset
i) partition, ii) normalize
monFscans
normalized synthetic reference MCref2
iii) orient monFscans by cross-correlation with MCref2
Consistently oriented monFscans
iv) filter by length to ensure nearly fully stretched molecules
v) filter by fluorescence intensity
monFscan dataset with minimized phase shifts and fluorescence intensity noise
vi) B-spline interpolation
monFscans of uniform length
processed monFscan dataset

Figure 5-5: catFscan preprocessing workflow for reducing outliers and partitioning catFscans into consistently oriented, uniformly sized monFscans. Filters attenuate experiment noise in datasets stemming from spurious fluorescence signals, and local variations of DNA stretch (causing phase shifts). Source: LMCG.

Wscan (described in Chapter 2) creates concatemer Fscan datasets (catFscan) from image files which are then manually curated. The resulting catFscans are processed by these steps:

- Partition each catFscan into constituent monomer “monFscans.” The Python package **detecta** (Duarte, 2021) identifies periodic valleys (fluorescence signal minima) harboring AT-rich sequence as the cut-off for bounding each monFscan within a concatemer (catFscan). The maximum valley threshold and the minimum monomer size are adjusted to refine partitioning for each catFscan dataset.

- Normalize each monFscan by dividing signal values by its minimum and then subtracting the mean. The minimum signal values serve as internal standards for correcting fluorescence intensity variances across large datasets.

- Consistently orient monFscan (5' – 3') by cross correlation against a synthetic reference. The reference signal, "MCref2," is created by two steps of a Monte Carlo simulation (see Chapter 3) and normalized in the same way as with monFscan datasets, ensuring catFscans comprise consistently oriented monFscans.

- Ensure nearly full stretch to minimize phase shifts by filtering monFscans (mode length 2 pixels); monFscans typically span 40 pixels, or 2.57 μm.

- Filter monFscans by fluorescence intensity: (+/- 1 SD of the range of pixel grey levels).

- Ensure monFscan datasets are the same length as MCref2 by B-spline interpolation.

Using this workflow, it is possible to generate datasets of monFscans from laboratory experiments, which are one of the inputs for the parameter fitting workflow described in the following section.

5.3.3. Parameter fitting workflow

The pUC19-H23 plasmid concatemer sequence comprising 30 repeats of the plasmid is used for the MC simulation. The noise parameters are fitted through the pipeline described in Fig. **5-6**.

- 40 concatemer molecules (40 29 full monomer Fscans = 1,160) are generated in each simulation. During the first run, the noise parameters are selected naively. In further runs, they are adjusted to better fit the experimental data.

- Partition the concatemer MCFscan (catMCFscan) / catFscan into monMCFscans /monFscans as per the workflow described in Section 5.3.2.

- Compare monMCFscans/ monFscans by length distributions. Outliers with lengths outside the range of 25 to 60 pixels are left out.

- Preprocess the monMCFscans /monFscans datasets by the rest of the workflow described in Section 5.3.2 after partitioning. Both datasets use MCref2 as reference. An exception in the workflow is using a different normalization method. Normalizing using

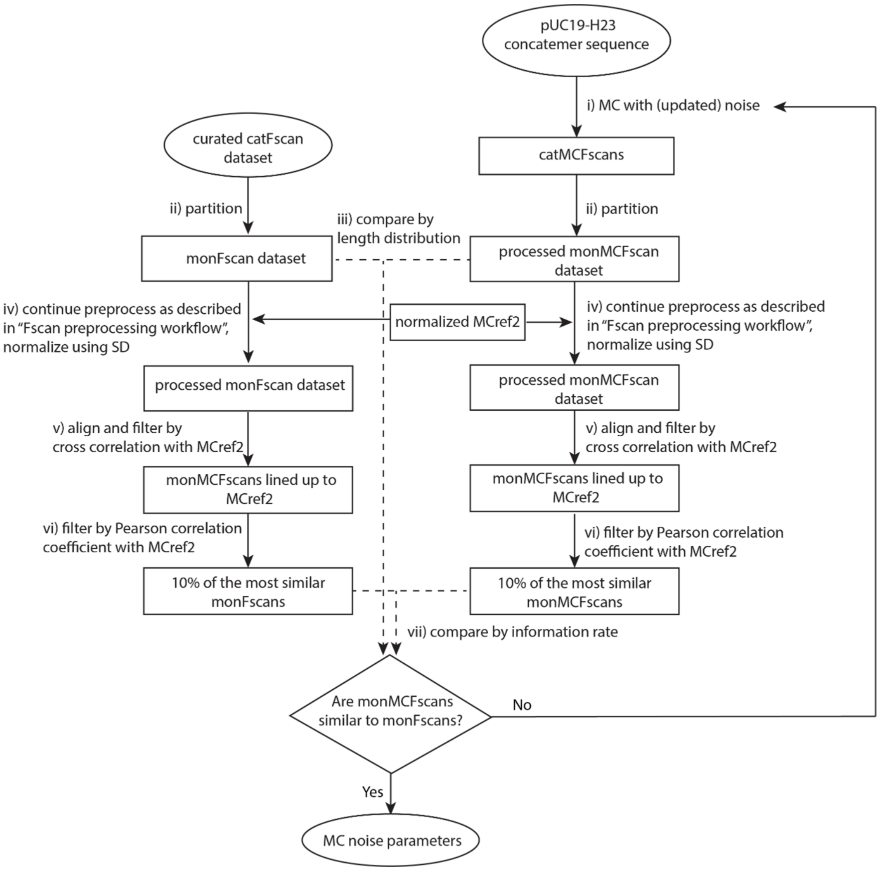
pUC19-H23 concatemer sequence
i) MC with (updated) noise
curated catFscan dataset
catMCFscans
ii) partition
ii) partition
iii) compare by length distribution
monFscan dataset
processed monMCFscan dataset
iv) continue preprocess as described in "Fscan preprocessing workflow", normalize using SD
normalized MCref2
iv) continue preprocess as described in "Fscan preprocessing workflow", normalize using SD
processed monFscan dataset
processed monMCFscan dataset
v) align and filter by cross correlation with MCref2
v) align and filter by cross correlation with MCref2
monMCFscans lined up to MCref2
monMCFscans lined up to MCref2
vi) filter by Pearson correlation coefficient with MCref2
vi) filter by Pearson correlation coefficient with MCref2
10% of the most similar monFscans
10% of the most similar monMCFscans
vii) compare by information rate
Are monMCFscans similar to monFscans?
No
Yes
MC noise parameters

Figure 5-6: Workflow for empirically determining noise for MC by comparing length distribution and information rate in resulted monMCFscans to the monFscans, both datasets preprocessed using the workflow described in Section 5.3.2. Source: LMCG.

the minimum as internal standard can preserve potential fluorescence intensity variances on modified DNA alleles. However, dividing by the minimum also exaggerates variations in the normalized signals. Therefore, we normalize differently to minimize signal variances for the noise analysis by $x_i = \frac{x_i - \bar{x}}{SD(x)}$.

- Align monMCFscans/monFscans with MCref2 by cross correlation. Most of the monMCFscans/monFscans do not need to be shifted. Some signals that are shifted by more than 2 or 3 pixels are filtered out.

- Filter by Pearson correlation coefficient. 10% of the monMCFscans/monFscans most similar to MCref2 are carried forward.

- Compare noise in filtered monMCFscans/ monFscans by cumulative information rate.

The noise parameters are then adjusted to repeat the workflow until the noise in monMCFscans is the same as in monFscans.

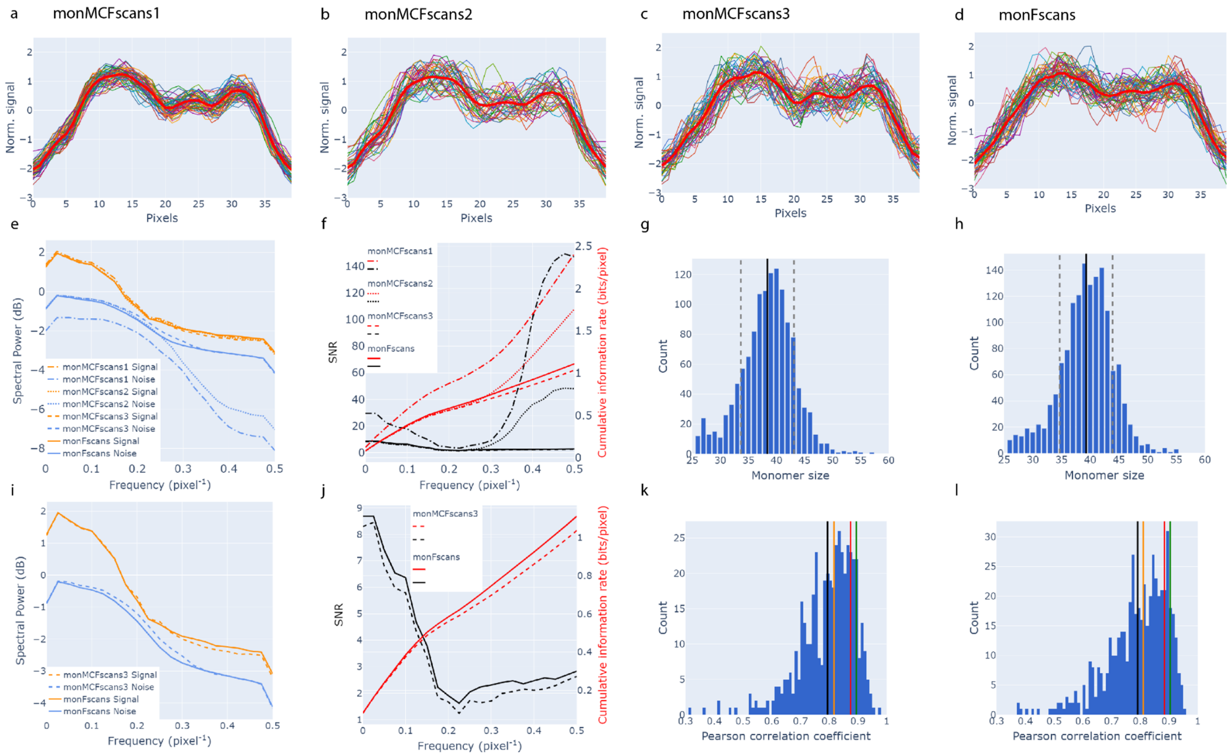

Figure 5-7: monMCFscans are compared to monFscans from one-step binding with a d/b mixing ratio in 1x TE. (a-d). Aligned, preprocessed and filtered individual signals (monFscans) from each dataset and their average signal (thick red) used for the noise analysis (Borst & Theunissen, 1999). (a) monMCFscan1 without noise. (b) monMCFscan2 with noise: 80% -120% DNA stretch, 20% of dark fluorochromes, Gaussian dye luminosity variation with $\sigma = 60\%$. (c) monMCFscans3 with noise same as (b) and pixel level noise($\sigma = 1.3\%$). (d) monFscans (d/b mixing ratio in 1x TE). (e) Overlayed signal and noise power spectra and (f) signal to noise ratio (SNR) and cumulative information rate curves of the four datasets. The overlays of only results from monMCFscans3 and monFscan are plotted for clarity: (i) signal and noise spectra and (j) signal to noise ratio (SNR) and cumulative information rate. (g, h) Length distributions of (g) monMCFscans3 (38.4 4.7 pixels) and (h) monFscans (39.3 4.6 pixels) with the mean showed by black line and the range of SD by gray dashed lines. (k, l) Pearson correlation coefficient distributions of (k) monMCFscans3 (mean: 0.791, median: 0.810, 80th percentile: 0.877, 90th percentile: 0.895) and (l) monFscans (mean: 0.790, median: 0.806, 80th percentile: 0.879, 90th percentile: 0.902). The mean (black), median (orange), 80th percentile (red) and 90th percentile (green) are marked by colored lines. Source: LMCG.

5.3.4. MC parameter fitting results

As per our empirical results, 20% of randomly selected dyes are dark (no fluorescence) and dye luminosity variances can be modeled as Gaussian distribution (σ = 60% of signal values). Furthermore, imaging noise is determined to be a Gaussian distribution (σ = 1.3% of signal values) to be added to final MCFscans on a per pixel basis. The appropriate amount of stretch variation in MC, imitating the effect of DNA elongation in microfluidic channels, is determined based on the length distribution of monMCFscans. Random stretch from 80% to 120% in MC results in similar length distributions between the monMCFscans and monFscans (Fig. 5-7g, h). The normal distributions of monMCFscan lengths are caused by two factors. First, the stretch-based dye-binding probability and release probability modelled in MC affect the intercalation-induced elongation. Second, noise in signals causes errors in automatic partitioning of catFscans into monFscans. Besides stretch variation, the addition of dye luminosity fluctuation and dark fluorochromes in MC both increase the noise spectrum densities (Fig. 5-7e, results of testing the two noise factors separately are not shown). The high frequency noise is manifested at the pixel level by variation of grey levels (Fig. 5-7e,f). This noise simulates background noise from ambient light and image readout on the camera. Fig. 5-7i, j show the closely overlapped signal and noise power spectra, SNR and cumulative information rate curves of monFscans and monMCFscans added noise. As an overall assessment of similarity, Pearson correlation coefficient distributions are compared between monMCFscan and monFscan datasets (Fig. 5-7k, l). The distributions of Pearson correlation coefficients of the two datasets measured with the same reference show very similar shapes and statistics. The close match of monMCFscans and monFscans in length distributions, signal information rates and Pearson correlation coefficient distributions support that our in-silico Monte Carlo simulation accurately captures the signal and noise features of experiment data.

5.3.5. Understanding noise factors

By closely simulating experiment noise factors in MC, we understood the types of noise in Fscans and minimized the noise using the designed catFscan preprocessing workflow. The monFscans from the concatemers provide large datasets for the noise analysis. The knowledge we obtained from monFscans about the dye binding and photophysical properties of the Fluoroscanning system can be applied to general DNA Fscans.

Stretch variation, as the source of phase shifts, is minimized by the length filtering steps which include nearly fully stretched molecules, with the resampling step also helping to bring all monFscans to the same length. Our modeling of stretch variation simulates a homogeneous stretch along each DNA molecule. Although stretch-facilitated elongation introduces some local phase shifts due to random dye binding patterns in MCFscans, the main component of local phase shift in Fscans, probably caused by intramolecular inhomogeneous stretch, is missing in the MCFscans. Local stretch variation is not incorporated because of the complexity in quantifying its extent and frequency in Fscans for a meaningful simulation. Lacking local phase shift explains the more pronounced features in the cMCFscan than in the cFscan like the deeper valleys (Fig. 5-7c, d). Besides the differences in cFscan features, monMCFscans are very similar to monFscans in the evaluations of similarity with respect to the reference and in terms of signal to noise contrast.

We attribute noise in fluorescence intensity to dye luminosity fluctuation, degraded dark fluorochromes, and imaging noise. The first two factors simulated in MC mainly contribute to the medium frequency noise we see in the experimental Fscans. As for the missing high frequency noise component, we introduce it by adding pixel level imaging noise. By combining the three sources of noise, we enable a close match between the noise spectra of monMCFscans and monFscans. Furthermore, since we understand the fluorescence intensity noise originates from both dye luminosities and imaging noise, we added filters in the Fscan preprocessing workflows to minimize these types of noise. A filter of fluorescence intensity range of monFscan is applied first for removing Fscans obviously out of range, mostly due to DNA overlapping. After that, we use the synthetic reference from MC for similarity filters for the Fscans using the Pearson correlation coefficient and cross-correlation. The synthetic reference is an average of MCFscans with incorporated noise to represent the common features of Fscans.

5.4. Discussion

In this chapter, we discussed two questions: first, whether or not there was an agreement on the importance of specific n-mers for our physicochemically informed approach and for Nandi's machine learning-based approach; and second, whether or not we were able to adjust the parameters of our Monte Carlo simulations to closely match the distribution of real Fluoroscans.

Regarding feature importance, our first attempt at comparing the Random Forest (RF)-based approach with our simulations only resulted in partial agreement: while it was clear that AT-rich regions were the most important according to both methods, the RF prioritized smaller n-mers whereas the Monte Carlo focused on 4 and 5-mers. One hypothesis was that due to the relatively big size of our simulated pixel (106.7 nm, which is equivalent to between 125 and 325 bp depending on dye load), we were giving more importance to larger n-mers, whereas the RF has direct access to all n-mer scales as part of its regression mechanism. Once we reduced our pixel size to 10nm (equivalent to between 10 and 30 bp) we saw a much bigger similarity both in the frequency of all n-mer sizes for both methods. Another significant takeaway is that, the importance of AT-rich regions across all of our tests agrees with recent work on enzyme-free whole genome analysis, where AT-rich regions are prioritized as targets for differential binding (Dvirnas et al., 2018; Park et al., 2019).

In regards to noise parameter adjustment, using concatemers to extract catFscans allowed us to circumvent the lack of a proper alignment system by using simple alignment approaches to align hundreds of smaller DNA fragments expected to have the same features to a reference signal generated with our Monte Carlo. By doing this and filtering based on several criteria as described in Figure **5-5**, we were able to develop a comparison framework for tweaking the Monte Carlo noise parameters until finding the correct conditions so that the distribution of monomer Monte Carlo Fscans matched that of experimental monomer Fscans. Since we only had to generate experimental Fscans once and every Monte Carlo run took around an hour, we were able to iterate quickly. One noise component that we are not yet able to manipulate is the intramolecular inhomogeneity that causes local phase shifts. Future work should consider ways to address this in order to more accurately fit the experimental distribution.

6 Conclusions and future work

6.1. Conclusions

In this work, we proposed a series of steps toward a comprehensive methodology for generating genomic maps from Fluoroscanning images. Compared to other whole genome analysis systems, Fluoroscanning requires a different set of techniques to guarantee the generation of reliable maps for big genomes. Therefore, we explored the segmentation and characterization of molecules from images in order to obtain Fluoroscans, the generation of synthetic Fluoroscans using a chemically and theoretically informed Monte Carlo algorithm, and the development of approaches to verify the agreement between experimental and simulated data to better guide the experiments carried out at the Laboratory for Molecular and Computational Genomics.

The main highlights of our work include the fact that, by using morphology-based digital image processing algorithms, DNA molecule backbones can be extracted reliably. Moreover, our Monte Carlo algorithm can model the majority of the sources of noise that impact experimental Fluoroscan data, including imaging noise, dark fluorochromes, methylation factors, and stretch variation, among others. Thanks to parallelization, the Monte Carlo can also be scaled up by using high-performance computing clusters to generate massive datasets of synthetic Fluoroscans. This aspect was particularly useful when fitting the noise parameters of the MC to those of real data, as it allowed us to quickly iterate through sets of parameters.

Although our work did not encompass the implementation of signal alignment algorithms required for aligning thousands of smaller Fscans into a single genome map, the Monte Carlo Fluoroscan simulation algorithm with fitted noise parameters can be a powerful tool for validating any future proposals for Fluoroscan alignment; any evaluation protocol will be simplified by the fact that we can easily retrieve the underlying composition and location of any Monte Carlo Fluoroscan, no matter the noise conditions that we employ. Motivated by this, we are planning to run our algorithm with the discovered noise parameters on the majority of the human genome. Essentially, this means that our work is a significant contribution to the development of a novel whole genome analysis system which has the potential to generate DNA sequence composition profiles with higher resolution (around 550 bp) than other proposals, and thus complement genome assemblies of sequencing reads to help discover genomic structural variations.

6.2. Future work

As part of the work that is currently being carried out at the LMCG, researchers are using the algorithms proposed in this thesis for testing whether it is possible to classify Fscans from DNA with different methylation types. Preliminary results indicate that methylation classes can be identified in Fscans extracted from concatemers by using relatively simple machine learning algorithms with more than 90% accuracy. Since our MC simulation algorithm incorporates different methylation conditions (aka: differential binding), there is potential for it to continue guiding experiments centered on the capability of Fluoroscanning to identify different methylation sites or other kinds of underlying phenomena. Moreover, the development of a Python version of the MC Fluoroscan will enable us to easily incorporate more parameters and noise sources if required in the future.

On the other hand, given that we will generate synthetic Fluoroscans based on the human genome, there is room for testing modern signal alignment algorithms on alignment tasks of varying difficulty. In particular, we want to explore the potential of data-intensive deep learning algorithms, which we were unable to employ in this work due to the limited availability of experimental data. With the availability of genomes from many kinds of organisms, one interesting line of experimentation would be to test whether models trained on the human genome would be able to generalize to the genomes of other organisms without major performance drops. Finally, we also expect the quality of Fluoroscans extracted from laboratory experiments to increase as the protocols continue to get refined and more advanced microscope sensors become available.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2008). *Molecular Biology of the Cell* (M. Anderson & S. Granum, Eds.; 5th). Garland Science.

Aston, C., Hiort, C., & Schwartz, D. C. (1999). Optical Mapping: An Approach for Fine Mapping. *Methods Enzymol.*, *303*, 55–73. https://doi.org/10.1016/S0076-6879(99)03006-2

Bahadar, K., Khaliq, A. A., & Shahid, M. (2016). A morphological hessian based approach for retinal blood vessels segmentation and denoising using region based otsu thresholding. *PLoS One*, *11*(7), 1–19. https://doi.org/10.1371/journal.pone.0158996

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804779

Bennink, M., Schärer, O., Kanaar, R., Sakata-Sogawa, K., Schins, J., Kanger, J., de Grooth, B., & Greve, J. (1999). Single-molecule manipulation of double-stranded dna using optical tweezers: Interaction studies of dna with reca and yoyo-1. *Cytometry*, *36*(3), 200–208. https://doi.org/10.1002/(sici)1097-0320(19990701)36:3&lt;200::aid-cyto9&gt;3.0.co;2-t

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nat Neurosci*, *2*(11), 947–957. https://doi.org/10.1038/14731

Brady, E. (1992). Real-time data compression using a FFT digital signal processor. https://doi.org/10.2172/7275570

Cao, H., Hastie, A. R., Cao, D., Lam, E. T., Sun, Y., Huang, H., Liu, X., Lin, L., Andrews, W., Chan, S., Huang, S., Tong, X., Requa, M., Anantharaman, T., Krogh, A., Yang, H., Cao, H., & Xu, X. (2014). Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience*, *3*(1), 1–11. https://doi.org/10.1186/2047-217X-3-34

Chan, E. K., Cameron, D. L., Petersen, D. C., Lyons, R. J., Baldi, B. F., Papenfuss, A. T., Thomas, D. M., & Hayes, V. M. (2018). Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. *Genome Research*, *28*(5), 726–738. https://doi.org/10.1101/gr.227975.117

Chandra, R., Dagum, L., Kohr, D., Menon, R., Maydan, D., & McDonald, J. (2001). *Parallel programming in OpenMP* (D. E. Penrose & E. Wade, Eds.; 1st). Morgan Kaufmann Publishers.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen,

G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., . . . Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to Algorithms* (2nd). The MIT Press; McGraw-Hill Book Company.

DeGroot, M. H., Schervish, M. J., & Sheet, C. (2011). *Probability and Statistics*. Addison Wesley. https://doi.org/0321709705

Deligeorgiev, T., Kaloyanova, S., & Vaquero, J. (2010). Intercalating cyanine dyes for nucleic acid detection. *Recent Patents on Materials Science*, *2*, 1–26. https://doi.org/10.2174/1874465610902010001

Dimalanta, E. T., Lim, A., Runnheim, R., Lamers, C., Churas, C., Forrest, D. K., Pablo, J. J. D., Graham, M. D., Coppersmith, S. N., Goldstein, S., & Schwartz, D. C. (2004). A Microfluidic System for Large DNA Molecule Arrays. *Anal. Chem.*, *76*(18), 5293–5301. https://doi.org/10.1021/ac0496401

Duarte, M. (2021). *Detecta: A python module to detect events in data* (Version v0.0.5). Zenodo. https://doi.org/10.5281/zenodo.4598962

Dvirnas, A., Pichler, C., Stewart, C. L., Quaderi, S., Nyberg, L. K., Müller, V., Bikkarolla, S. K., Kristiansson, E., Sandegren, L., Westerlund, F., & Ambjörnsson, T. (2018). Facilitated sequence assembly using densely labeled optical DNA barcodes: A combinatorial auction approach. *PLOS ONE*, *13*(3), e0193900. https://doi.org/10.1371/journal.pone.0193900

Gonzalez, R. C., & Woods, R. E. (2008). *Digital image processing* (3rd). Prentice Hall. https://www.imageprocessingplace.com

Guennebaud, G., & Jacob, B. (2010). Eigen v3 [software library]. http://eigen.tuxfamily.org

Guizar-Sicairos, M., Thurman, S. T., & Fienup, J. R. (2008). Efficient subpixel image registration algorithms. *Opt. Lett.*, *33*(2), 156–158. https://doi.org/10.1364/OL.33.000156

Günther, K., Mertig, M., & Seidel, R. (2010). Mechanical and structural properties of YOYO-1 complexed DNA. *Nucleic Acids Res.*, *38*(19), 6526–6532. https://doi.org/10.1093/nar/gkq434

Gupta, A., Kounovsky-Shafer, K. L., Ravindran, P., & Schwartz, D. C. (2016). Optical mapping and nanocoding approaches to whole-genome analysis. *Microfluid. Nanofluidics*, *20*(3), 1–14. https://doi.org/10.1007/s10404-015-1685-y

Gupta, A., Place, M., Goldstein, S., Sarkar, D., Zhou, S., Potamousis, K., Kim, J., Flanagan, C., Li, Y., Newton, M. A., Callander, N. S., Hematti, P., Bresnick, E. H., Ma, J., Asimakopoulos, F., & Schwartz, D. C. (2015). Single-molecule analysis reveals widespread structural variation in multiple myeloma. *Proc. Natl. Acad. Sci.*, *112*(25), 7689–7694. https://doi.org/10.1073/pnas.1418577112

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

International Organization for Standardization. (2012). *ISO/IEC 14882:2011 Information technology — Programming languages — C++*. Geneva, Switzerland, International Organization for Standardization. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50372

Jo, K., Schramm, T. M., & Schwartz, D. C. (2009). A single-molecule barcoding system using nanoslits for DNA analysis : nanocoding. *Methods Mol. Biol.*, *544*(8), 29–42. https://doi.org/10.1007/978-1-59745-483-4_3

Johansen, F., & Jacobsen, J. P. (1998). 1H NMR studies of the bis-intercalation of a homodimeric oxazole yellow dye in DNA oligonucleotides. *J Biomol Struct Dyn*, *16*(2), 205–222.

Johnson, I. (2010). *Molecular probes handbook: A guide to fluorescent probes and labeling technologies*. Life Technologies Corporation. https://books.google.com/books?id=djuacQAACAAJ

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Sci.*, *23*(12), 1650–1666. https://doi.org/10.1002/pro.2552

Kounovsky-Shafer, K. L., Hernandez-Ortiz, J. P., Potamousis, K., Tsvid, G., Place, M., Ravindran, P., Jo, K., Zhou, S., Odijk, T., de Pablo, J. J., & Schwartz, D. C. (2017). Electrostatic confinement and manipulation of DNA molecules for genome analysis. *Proc. Natl. Acad. Sci.*, (January), 13400–13405. https://doi.org/10.1073/pnas.1711069114

Larsson, A., Carlsson, C., Jonsson, M., & Albinsson, B. (1994). Characterization of the Binding of the Fluorescent Dyes YO and YOYO to DNA by Polarized Light Spectroscopy. *J. Am. Chem. Soc.*, *116*(19), 8459–8465. https://doi.org/10.1021/ja00098a004

Lee, S., & Jo, K. (2016). Visualization of Surface-tethered Large DNA Molecules with a Fluorescent Protein DNA Binding Peptide. *Journal of Visualized Experiments: JoVE*, (112). https://doi.org/10.3791/54141

Lee, S., Lee, Y., Kim, Y., Wang, C., Park, J., Jung, G. Y., Chen, Y.-L., Chang, R., Ikeda, S., Sugiyama, H., & Jo, K. (2018). Nanochannel-Confined TAMRA-Polypyrrole Stained DNA Stretching by Varying the Ionic Strength from Micromolar to Millimolar Concentrations. *Polymers*, *11*(1), 15. https://doi.org/10.3390/polym11010015

Lesho, E., Clifford, R., Onmus-Leone, F., Appalla, L., Snesrud, E., Kwak, Y., Ong, A., Maybank, R., Waterman, P., Rohrbeck, P., Julius, M., Roth, A., Martinez, J., Nielsen, L., Steele, E., McGann, P., & Hinkle, M. (2016). The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. Department of Defense. *PLoS One*, *11*(5), 1–12. https://doi.org/10.1371/journal.pone.0155770

Leung, A. K. Y., Kwok, T. P., Wan, R., Xiao, M., Kwok, P. Y., Yip, K. Y., & Chan, T. F. (2017). OMBlast: Alignment tool for optical mapping using a seed-and-extend approach. *Bioinformatics*, *33*(3), 311–319. https://doi.org/10.1093/bioinformatics/btw620

Li, Y., Zhou, S., Schwartz, D. C., & Ma, J. (2016). Allele-Specific Quantification of Structural Variations in Cancer Genomes. *Cell Systems*, *3*(1), 21–34. https://doi.org/10.1016/j.cels.2016.05.007

Louie, E., Ott, J., & Majewski, J. (2003). Nucleotide Frequency Variation Across Human Genes. *Genome Res.*, 2594–2601. https://doi.org/10.1101/gr.1317703.

Majewski, J., Majewski, J., Ott, J., & Ott, J. (2002). Distribution and characterization of regulatory elements in the human genome. *Genome Res.*, *12*(212), 1827–1836. https://doi.org/10.1101/gr.606402.12

Marie, R., Pedersen, J. N., Bauer, D. L., Rasmussen, K. H., Yusuf, M., Volpi, E., Flyvbjerg, H., Kristensen, A., & Mir, K. U. (2013). Integrated view of genome structure and sequence of a single DNA molecule in a nanofluidic device. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(13), 4893–4898. https://doi.org/10.1073/pnas.1214570110

Marie, R., Pedersen, J. N., Bærlocher, L., Koprowska, K., Pødenphant, M., Sabatel, C., Zalkovskij, M., Mironov, A., Bilenberg, B., Ashley, N., Flyvbjerg, H., Bodmer, W. F., Kristensen, A., & Mir, K. U. (2018). Single-molecule DNA-mapping and whole-genome sequencing of individual cells. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44), 11192–11197. https://doi.org/10.1073/pnas.1804194115

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. on Modeling and Computer Simulation*, *8*(1), 3–30.

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Brief. Bioinform.*, *18*(5), arXiv 1603.06430, 851–869. https://doi.org/10.1093/bib/bbw068

Müller, V., Dvirnas, A., Andersson, J., Singh, V., KK, S., Johansson, P., Ebenstein, Y., Ambjörnsson, T., & Westerlund, F. (2019). Enzyme-free optical DNA mapping of the human genome using competitive binding. *Nucleic Acids Research*, *47*(15), e89. https://doi.org/10.1093/nar/gkz489

Nagarajan, N., Read, T. D., & Pop, M. (2008). Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, *24*(10), 1229–1235. https://doi.org/10.1093/bioinformatics/btn102

Nandi, S. (2017). *Statistical Learning Methods for Fluroroscanning* (Ph.D. Thesis). University of Wisconsin-Madison.

Netzel, T. L., Nafisi, K., Zhao, M., Lenhard, J. R., & Johnson, I. (1995). Base-Content Dependence of Emission Enhancements, Quantum Yields, and Lifetimes for Cyanine Dyes Bound to Double-Strand DNA: Photophysical Properties of Monomeric and Bichromomphoric DNA Stains. *J. Phys. Chem.*, *99*(51), 17936–17947. https://doi.org/10.1021/j100051a019

Nyberg, L., Persson, F., Åkerman, B., & Westerlund, F. (2013). Heterogeneous staining: a tool for studies of how fluorescent dyes affect the physical properties of DNA. *Nucleic Acids Research*, *41*(19), e184–e184. https://doi.org/10.1093/nar/gkt755

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Park, J., Lee, S., Won, N., Shin, E., Kim, S.-H., Chun, M.-Y., Gu, J., Jung, G.-Y., Lim, K.-I., & Jo, K. (2019). Single-molecule DNA visualization using AT-specific red and non-specific green DNA-binding fluorescent proteins. *Analyst*, *144*(3), 921–927. https://doi.org/10.1039/C8AN01426D

Pereira, R., Couto, M., Ribeiro, F., Rua, R., Cunha, J., Fernandes, J. P., & Saraiva, J. (2021). Ranking programming languages by energy efficiency. *Science of Computer Programming*, *205*, 102609. https://doi.org/https://doi.org/10.1016/j.scico.2021.102609

Precision Medicine Initiative (PMI) Working Group. (2015). *The precision medicine initiative cohort program – building a research foundation for 21st century medicine* (tech. rep.). http://www.nih.gov/precisionmedicine/

Ravindran, P., & Gupta, A. (2015). Image processing for optical mapping. *Gigascience*, *4*(1), 1–8. https://doi.org/10.1186/s13742-015-0096-z

Reisner, W., Larsen, N. B., Silahtaroglu, A., Kristensen, A., Tommerup, N., Tegenfeldt, J. O., & Flyvbjerg, H. (2010). Single-molecule denaturation mapping of DNA in nanofluidic channels. *Proceedings of the National Academy of Sciences*, *107*(30), 13294–13299. https://doi.org/10.1073/pnas.1007081107

Roy, A., Diao, Y., Evani, U., Abhyankar, A., Howarth, C., Le Priol, R., & Bloom, T. (2017). Massively Parallel Processing of Whole Genome Sequence Data, In *Proc. 2017 acm int. conf. manag. data - sigmod '17*. https://doi.org/10.1145/3035918.3064048

Rye, H. S., Yue, S., Wemmer, D. E., Quesada, M. A., Haugland, R. P., Mathies, R. A., & Glazer, A. N. (1992). Stable fluorescent complexes of double-stranded DNA with bis-intercalating asymmetric cyanine dyes: Properties and applications. *Nucleic Acids Research*, *20*(11), 2803–2812.

Schwartz, D., Li, X., Hernandez, L., Ramnarain, S., Huff, E., & Wang, Y. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science (80-.).*, *262*(5130), 110–114. https://doi.org/10.1126/science.8211116

Shapiro, H. (2004). Excitation and emission spectra of common dyes. *Current Protocols in Cytometry, Chapter 1*, Unit 1.19. https://doi.org/10.1002/0471142956.cy0119s26

Shiguo, Z., Herscheleb, J., & Schwartz, D. C. (2007). A single molecule system for whole genome analysis., In *New high throughput technol. dna seq. genomics*.

Shit, S., Paetzold, J. C., Sekuboyina, A., Zhylka, A., Ezhov, I., Unger, A., Pluim, J. P. W., Tetteh, G., & Menze, B. H. (2020). clDice – a Topology-Preserving Loss Function for Tubular Structure Segmentation, arXiv 2003.07311, 1–23. http://arxiv.org/abs/2003.07311

Spielmann, H. P., Wemmer, D. E., & Jacobsen, J. P. (1995). Solution structure of a DNA complex with the fluorescent bis-intercalator TOTO determined by NMR spectroscopy. *Biochemistry*, *34*(27), 8542–8553.

Tang, H., Lyons, E., & Town, C. D. (2015). Optical mapping in plant comparative genomics. *Gigascience*, *4*(1), 1–6. https://doi.org/10.1186/s13742-015-0044-y

Teague, B., Waterman, M. S., Goldstein, S., Potamousis, K., Zhou, S., Reslewic, S., Sarkar, D., Valouev, A., Churas, C., Kidd, J. M., Kohn, S., Runnheim, R., Lamers, C., Forrest, D., Newton, M. A., Eichler, E. E., Kent-First, M., Surti, U., Livny, M., & Schwartz, D. C. (2010). High-resolution human genome structure by single-molecule analysis. *Proc. Natl. Acad. Sci.*, *107*(24), 10848–10853. https://doi.org/10.1073/pnas.0914638107

Valouev, A., Schwartz, D. C., Zhou, S., & Waterman, M. S. (2006). An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc. Natl. Acad. Sci.*, *103*(43), 15770–15775. https://doi.org/10.1073/pnas.0604040103

Valouev, A., Li, L., Liu, Y.-C., Schwartz, D. C., Yang, Y., Zhang, Y., & Waterman, M. S. (2006). Alignment of Optical Maps. *J. Comput. Biol.*, *13*(2), 442–462. https://doi.org/10.1089/cmb.2006.13.442

Valouev, A., Zhang, Y., Schwartz, D. C., & Waterman, M. S. (2006). Refinement of optical map assemblies. *Bioinformatics*, *22*(10), 1217–1224. https://doi.org/10.1093/bioinformatics/btl063

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in python. *PeerJ*, *2*, e453.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Voigtländer, B. (2015). Data Representation and Image Processing (P. Avouris, B. Bhushan, D. Bimberg, H. Sakaki, K. von Klitzing, & R. Wiesendanger, Eds.; 1st ed.). In P. Avouris, B. Bhushan, D. Bimberg, H. Sakaki, K. von Klitzing, & R. Wiesendanger (Eds.), *Scanning probe microsc. at. force microsc. scanning tunneling microsc.* (1st ed.). Berlin, Heidelberg, Springer-Verlag GmbH Berlin Heidelberg. https://doi.org/10.1016/B978-0-12-814182-3.00005-5

Zhou, S., & Schwartz, D. C. (2004). The Optical Mapping of Microbial Genomes. *ASM News*, *70*(7), 323–330.