Stimulated emission depletion microscopy for super-resolution optical DNA mapping

Louis, Boris

2016

KU LEUVEN

FACULTY OF SCIENCE

Stimulated emission depletion microscopy for super-resolution optical DNA mapping

Boris LOUIS

Supervisor: Prof. J. Hofkens
KU Leuven

Co-supervisor: Prof. B. Leyh
Université de Liège

Mentor: Dr. K. Janssen
KU Leuven

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Chemistry

Academic year 2015-2016

Preface

Firstly, I would like to thank my promotor, Johan Hofkens for giving me the opportunity to work on this interesting project. I also want to thank him supporting me from the beginning to the end. I would also like to thank my co-supervisor Bernard Leyh for accepting the co-promotion of this project, for supporting me and for fruitful discussions about the results.

Secondly, I would like to thank my mentor Kris Janssen. Before coming to Leuven I was a bit worried because Ii was entering in a new environment with many unknown people but he made my adaptation much easier. I would also thank him for trusting me and letting me work in a very independent way while being available whenever it was needed. I want also to thank him for the many contribution he made by improving the programs that were used during this project.

Thirdly, I would like to thank the different persons from the group that scientifically contributed to this project. Su Wang and Sven Van Snick for their preparation of the enzyme and the cofactor respectively. Jia Su for teaching me how to label DNA and many discussions about the sample preparation and the imaging. Christian Steuwe for teaching me a significant part of what I know about optics, setup alignment and building as well as many discussion about STED and microscopy in general. Christian also contributed a lot in making my adaptation easier. Rafael Camacho for many advices on programming and many discussion about the relevance of different parameters in the data analysis.

Finally, I would like to thank all the persons that participated indirectly to the fullfilment of this project. Either by supporting me or giving me the opportunity to escape from the research and the lab for some time. In particular, my friends, my family and my lover .

Boris Louis

Glossary

APD	avalanche photon detector.
aSseq	automated Sanger sequencing.
bp	base pair.
CE	capillary electrophoresis.
CuAAC	copper(I)-catalyzed alkyne-azide cycloaddition.
DNA	deoxyribonucleic acid.
FOV	field of view.
HGP	the human genome project.
InDel	insertion-deletion.
MES	2-(N-morpholino)ethanesulfonic acid.
mTAG	methyl transferase directed transfer of activated groups.
NGS	next-generation sequencing.
NHS	N-Hydroxysuccinimide.
PBS	phosphate buffered saline.
PCR	polymerase chain reaction.
PMT	photo multiplier tube.
PSF	point spread function.
RNA	ribonucleic acid.
SAM	*S*-adenosyl-L-methionine.
SIM	structured illumination microscopy.
SMseq	single-molecule sequencing.
SOFI	super-resolution optical fluctuation imaging.

SPAAC	strain-promoted alkyne-azide cycloaddition.
SRF	super-resolution fluorescence.
SRFM	super-resolution fluorescence microscopy.
STED	stimulated emission depletion.
TPP	time per pixel.

Contents

Abstract

English:
Deoxyribonucleic acid or DNA is one of the most fundamental molecules of life as it has the power to encode the basic structure of every living thing large or small, including us. Not only is DNA responsible for precisely describing every single aspect that makes us what we are, it also directly affects the world around us, every second of every day. Indeed, by unlocking the genetic code embedded in DNA we were already able to create new diagnostics that allow us to detect certain diseases before we can even detect the first symptoms. It allows us to create new, stronger crops that allow us to feed the worlds ever growing population. However, in spite of this newly acquired power to manipulate the very core of life itself we are ever so often reminded of the fact that mankind is still very much subjected to the ever evolving "source code" of life rather than being in control of it. Indeed, many disease causing pathogens exchange DNA that provides them with the ability to withstand even the most powerful known antibiotics. Furthermore, many aspects of the genetic code still remain obfuscated by its complex nature and are very much out of reach of even the most modern sequencing technologies because these often rely on determining sequence information for a large population of DNA. Therefore, the search for genomic analysis strategies that allow us to investigate the code of life at the single molecule level are the next big frontier scientific research. Here, optical DNA mapping is one of the top contenders to address some of the long standing issues that remain with modern "next-generation" sequencing technologies such as their inability to achieve long readout lengths and difficulties encountered when trying to detect long range structural variations in the genome.. In optical mapping, fluorescent molecules are attached to the DNA in a sequence-specific manner. Through subsequent observation of surface deposited, contiguous DNA molecules with a fluorescent microscope, long range information about the sequence can be retrieved. The information content of such genomic maps is of course, less dense than in the case of sequencing approach. However, genomic DNA maps have already proven their worth by serving as scaffolds for sequencing based reconstructions of complex genomes. Furthermore, if the resolution of the microscopy imaging in mapping could be increased beyond the diffraction limit of 250 nm, which roughly corresponds to a map labeling density of one label every 700 to 800 base pairs, the information density of maps would also be increased drastically. Fortunately, recent years have seen an increasing number of developments in so called super-resolution microscopy methods. The founders of this field were even awarded the Nobel prize in 2014. Stimulated emission depletion microscopy (STED) is one of

such techniques and allows to produce images at resolutions exceeding 100 nm in an almost instantaneous way. The presented work aims to evaluate the applicability of STED for optical DNA mapping with an emphasis on optical map characterisation and differentiation. For this reason, STED based DNA mapping was attempted on reference DNA samples of two viruses, phage T7 and phage Lambda. Intensity profiles from DNA images obtained with STED were extracted and compared to *in silico* generated reference intensity profiles for these species. This work demonstrates that STED is applicable to optical DNA mapping but also that it provides a sufficient amount of information to allow for pattern recognition. Indeed, the correct specie was assessed to samples containing one specie. Furthermore, two populations could be distinguished in a sample composed of the two species showing that STED allows for DNA differentiation.

Nederlands:
Deoxyribonucleïnezuur of DNA is een van de meest fundamentele bouwstenen van het leven. Dit bijzondere molecule bezit de mogelijkheid om de structuur van alle levende dingen in de wereld rondom ons te vatten en op te slaan tot in de kleinste details. In die zin maakt DNA ons niet enkel tot wat we zijn maar oefent het op een ongeziene manier ook een enorme invloed uit op elke aspect van ons leven. Dankzij onze altijd groeiende kennis van de eigenschappen van DNA en de genetische code zijn we bijvoorbeeld in vaak reeds in staat om ernstige ziekten op te sporen, nog voor de eerste symptomen zich voordoen. Door de genetische code aan te passen zijn we in staat om nieuwe gewassen te ontwikkelen welke ons zullen toelaten om onder de immer veranderende klimaatsomstandigheden een steeds groeiende bevolking te blijven voeden. Ondanks al deze ontwikkelingen krijgt de mens op regelmatige basis een stevige herinnering aan het feit dat we nog steeds eerder onderworpen zijn aan de eeuwigdurende evolutie van de genetische codes om ons heen, veeleer dan ze volledig te beheersen. Zo is het bijvoorbeeld geweten dat bepaalde pathogene bacteriën in staat zijn om genetisch materiaal uit te wisselen, hetgeen hen toelaat resistentie op te bouwen tegen zelfs de meest krachtige antibiotica. Daarnaast heeft de genetische code nog tal van eigenschappen en structuren welke nog niet of onvoldoende begrepen worden en welke op dit ogenblik maar moeilijk in kaart gebracht kunnen worden met de modernste sequentie betalingstechnieken, zoals bijvoorbeeld in het geval van genetische aberraties zoals inserties, deleties of grootschalige herschikking van genen. De zoektocht naar nieuwe strategieën om enkelvoudige moleculen DNA te kunnen analyseren is daarom een van de grote nieuwe uitdagingen van het wetenschappelijk onderzoek geworden. Hierbij is optische genoom 'mapping' een van de meest recente methodes welke op termijn moeten toelaten om een aantal van de grootste uitdagingen in sequentiebepaling aan te gaan. In optische mapping worden intacte genetische fragmenten voorzien van fluorescente merkers welke aangebracht kunnen worden op een sequentie specifieke manier. Er wordt als het ware een genetische streepjescode aangemaakt welke uniek is voor een specifiek organisme of bijvoorbeeld een beoogd ziektebeeld. Deze code kan relatief eenvoudig worden uitgelezen met een geschikte microscoop. Een dergelijke streepjescode bevat nooit alle details zoals in geval van een doorgedreven sequentiebepaling. Ze biedt echter het voordeel dat enkelvoudige moleculen over erg grote afstanden uitgelezen kunnen worden. Dit laat toe om genetische streepjescodes te gebruiken als een referentiekader bij de wedersamenstelling van complexe genomen op basis van sequentie data. Bovendien kan de hoeveelheid informatie welke vervat zit in de streepjescode gevoelig worden opgedreven door nieuwe ontwikkelingen op het vlak van zogenaamde super resolutie fluorescentie microscopie. Deze nieuwe vormen van microscopie ontketenden in de laatste jaren een revolutie doorheen alle velden van het wetenschappelijk onderzoek, van biologie tot materiaalkunde. Het hoeft dan ook niet te verwonderen dat de wetenschappers die dit veld in het leven riepen recent beloond werden met een Nobelprijs. Gestimuleerde emissie-deletie microscopie (STED) is een van de vele super resolutie technieken. Ze biedt het voordeel om met hoge snelheid super geresolveerde beelden met resoluties lager dan 100 nm op te nemen. Daarom zal ze in dit werk gebruikt worden voor genetische DNA mapping.

Er werd getracht worden om referentiestalen afkomstig van twee virussen, faag T7 en faag Lambda te analyseren met STED. Hiertoe werden STED intensiteitsprofielen van beide stalen vergeleken met *in silico* gesimuleerde maps op basis van referentie genoom data.In dit werk wordt aangetoond dat STED wel degelijk gebruikt kan worden voor de opname van genetische streepjescodes en ons in staat stelt om op basis van deze codes verschillende soorten virussen van elkaar te onderscheiden.

List of Figures

List of Figures

Part I

Introduction

Chapter 1

Optical DNA mapping

1.1 Introduction

This chapter aims to give an overview about DNA and the extraction of the information contained in it. For this purpose, DNA sequencing is discussed as well as the advantages that optical DNA mapping can bring to complement those methods. Eventually, a brief review of the different approaches for optical DNA mapping are presented including the one used by our group, Fluorocode bar coding. The issues encountered in later approach are also discussed.

1.2 Basic properties of DNA

Friedrich Miescher in 1869 was the first to describe the presence of a phosphate rich compound in the nucleus of white blood cells.[15, 66] The substance, which he managed to isolate and initially referred to as *nuclein*, would later be referred to using the now well known name DNA. At the time, Miescher could not have predicted the significance of his discovery and it was only much later that the key role of DNA as the cornerstone of biological life could be fully elucidated.

New advancements in scientific understanding of the nature of Mieschers nuclein would have to wait until the early 1900's when the Russian biochemist Phoebus Levene would be the first to correctly deduce that DNA chemically consists of three distinct moieties, i.e. a phosphate, a sugar and a base.[66] Not only that but he also managed to correctly identify the sugar as being deoxyribose in the case of DNA and ribose in the case of ribonucleic acid (RNA). He further posited that DNA is a bio-polymer composed of four different types of nucleotides. In DNA, each nucleotide consists of a central 2'-deoxyribose with its nitrogen base or nucleoside, at the 1' position and a 5' phosphate and 3' hydroxyl moiety. The hydroxyl and phosphate moiety on neighboring nucleotides can form a phosphodiester bond which ultimately forms the backbone of the single stranded DNA polymer. Additionally, the specific interaction between the 3' hydroxyl and 5' phosphate is what imparts directionality in the context of single DNA strands which would turn out to be another highly

important aspect of DNA.

Nucleotides in DNA and RNA were shown to only differ in the nature of their nucleoside. In total, five types of nitrogen bases exist: single ring Uracyl (U), Cytosine (C), Thymine (T), i.e. the pyrimidines and fused ring Adenine (A) and Guanine (G) and it was soon accepted that DNA only contains A, G, C or T whereas RNA only contains A, G, C and U.

Through a series of clever experiments, Oswald Avery was the first to unambiguously identify DNA as the single carrier of genetic information by observing how physical traits could be transferred between populations of Type III pneumococci by exposing one population to DNA isolated from another.[3, 66] This inspired Chargaff, a chemist who had just developed a new paper chromatography method to separate minute amounts of organic material, to study the nucleotide composition of DNA across different species of organisms. Although he could not explain why, he found that even though the total DNA composition would vary across species, certain properties would nonetheless by highly preserved. Indeed, he could show how the total amount of adenine would always be similar to the total amount of thymine and similarly, the total amount of pyrimidines (C + T) would also be nearly equal.[66] To this day, this observation is known as Chargaff's rule and it would turn out to be crucially important to help understand the final experiments in the quest to determine the structure of DNA.

Francis Crick, James Watson, Maurice Wilkins and Rosalind Franklin famously elucidated the structure of DNA through X-ray diffraction experiments.[88, 66] The first three would go on to share the Nobel prize in medicine in 1962 for their achievement when the scientific community started to realize that knowing the structure of DNA would eventually allow to find a way to decode the information contained in it.[66] One of the major contributing factors that allowed the scientist to propose a structure for DNA based on the diffraction data was in fact Chargaff's rule. The observation that only A/T or G/C pairs would form led Watson and Crick to infer that different single strands of DNA can interact and associate with each other because of the innate ability of nucleosides to form hydrogen bonds (see Figure 1.1). Specifically, G and C are capable of forming three hydrogen bonds whereas A and T can only form 2, a phenomenon which ultimately results in a much stronger interaction between single strands that are GC rich as opposed to AT rich. Here again, the 5' to 3' directionality of single stranded DNA is important as complementary strands in a DNA duplex will have opposing orientations. The hydrogen bonding between matching nucleotides and stacking of the hydrophobic nucleosides are the two factors ultimately contributing to the helical structure of double stranded DNA as first described by Watson and Crick with following basic properties:[66]

1. DNA is a double-stranded helix. The strands are connected by hydrogen bonding between bases. Adenine bases can only be paired with thymines while

cytosines can only be paired with guanines. This is consistent with Chargaff's rule and explains its nature.

2. Most DNA double helices are right-handed what means that when holding your right hand thumb up the other finger curled around it, your thumb would represent the axis around which the helix is formed and your finger the DNA backbone. Only one type of DNA, called Z-DNA, is left-handed.

3. The DNA double helix is anti-parallel. One strand oriented in 5'3' will thus be paired with a strand oriented in 3'5' (and vice versa). As shown in Figure 1.1, nucleotides are linked to each other by their phosphate groups, which bind the 3' end of one sugar to the 5' end of the next sugar.

4. The outer edges of the nitrogen-containing bases also permit an easy access to other molecules such as proteins via hydrogen bonding. This is a key for replication and expression of DNA in which proteins play a crucial role.

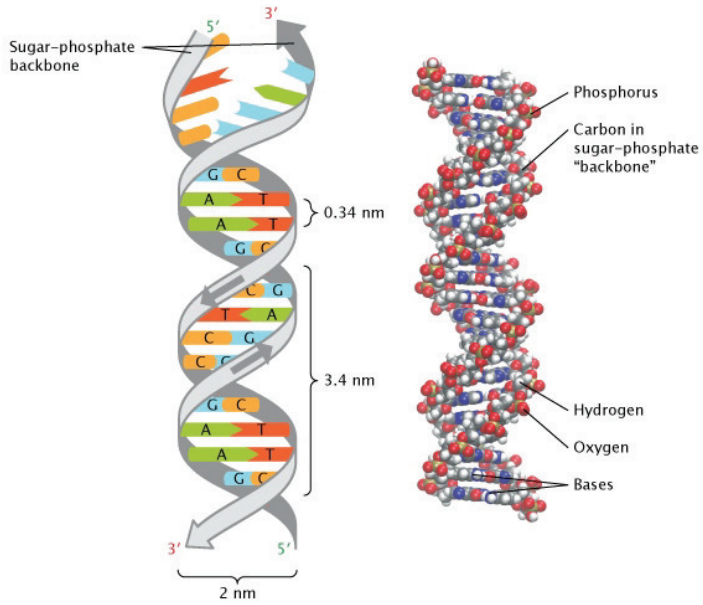

Figure 1.1: The DNA helical structure. Left: the grey bands in represent the sugar-phosphate backbone, showing the anti-parallel orientation of the complementary strands. These strands are held together though base pairing in accordance with Chargaff's rule: guanine (G, shown in blue) binds with cytosine (C, shown in orange) whereas adenine (A, shown in green) binds with the thymine (T, shown in red). The distance between successive base pairs is 0.34 nanometers. The length of one turn of the double-helix is exactly 10 basepairs. Right: space-filling molecular model of DNA. Reproduced from [66]

.

1.3 The genetic code

Knowing the basic structure of DNA, how then is it possible that a biopolymer consisting of just for letters can possibly encode the vast complexity that can be found across all branches of the tree of life? There are 20 naturally occurring amino acids, the monomers which make up proteins that ultimately from everything from enzymes to entire tissues whereas there are only four letters in the DNA alphabet.

George Gamow was the first to tackle this question by postulating that short sequences of nucleotides, i.e. 'codons' are the basic unit of information that is used by all forms of life to encode protein structure. Based on the number of known amino acids, he was able to predict that the length of such a codon would have to be three bases long. Indeed, a length two codon would only be able to code 2^4 or 16 amino acids. After it was established that DNA is transcribed into RNA prior to translation into protein, Nirenberg, Crick, Brenner and others were able to confirm the predictions of Gamow experimentally by incubating artificially prepared monomeric RNA sequences to an isolate containing all the components of the cellular protein synthesis machinery. Using experiments such as these, it soon became possible to fully elucidate the nature of the genetic code:[14]

- Each amino acid is encoded by a *3 base codon.*
- *The genetic code is non-overlapping.* If one considers a code GTCAGC then this code consists of two codons, GTC and AGC and not four codons GTC, TCA, CAG, AGC.
- *The code is not punctuated but rather continuous.* The sequence of bases is always read sequentially from a fixed starting- and endpoint, without interruptions, i.e. all codons are translated into an amino acid.
- *The genetic code is degenerate.* This means that most amino acids can be encoded by more than one codon. Even though there are 64 possible triplets, there are only 20 naturally occurring amino acids. Still, 61 of the 64 triplets nonetheless encode an amino acid, meaning that each amino acid has multiple triplet representations, except for methionine and tryptophan. The other three triplets were shown to be stop codons.

A gene (e.g. code for a protein) is composed of a serie of codons which starts with a starting codon and ends with an ending codon. As DNA can not go out of the nucleus of the cell, the gene needs to be translated in order to be employed by the organism. For this purpose, DNA is translated into a RNA molecule that is complementary of the DNA strand copied. RNA will later go out of the nucleus where it will be read and expressed into proteins. The expression of a gene takes place in two steps, the translation of DNA to RNA and the expression to protein. The first step is reversible whereas the second is irreversible what was described as the central dogma in molecular biology.[13] The full DNA sequence of a complex

organism (e.g eukaryotic cells) is found in the nucleus of the cell under the form of chromosomes. A chromosome is a super-structure of DNA and histones, a protein which ensures the formation and the maintaining of the whole structure. Unlike complex organism, prokaryotic cells does not have histones, the structure of the DNA is therefore compact and localized.

1.4 Reading the genetic code

1.4.1 DNA sequencing, a brief overview

In the early 90's, the the human genome project (HGP) was started. At a cost of nearly 2.7 billion dollars, the decade long effort culminated in what to this day, can be considered as one of the most significant achievements of modern biotechnology: the completion of the first full map of the human genome. The project relied heavily on automated Sanger sequencing (aSseq). Here, fragmented genomic DNA is extended by DNA polymerase in the presence of fluorescently labeled nucleotides. Whenever such a fluorescently labeled nucleotide is incorporated by the enzyme, further extension is no longer possible, thus resulting in a collection of randomly sized fragments. These fragments can be size separated in a process called capillary electrophoresis (CE) where fragments of different size will pass the fluorescence detector at different times. Because each of the four nucleotides carries a differently colored label, it is thus possible to infer the identity of the base at any given position.[12]

Even though the HGP was a success and aSseq proved to be a very robust sequencing technology that allowed for very long contiguous readouts, the applied strategies used within the HGP would not be applicable on a routine basis as they are very labor intensive, costly and ultimately only offered very limited throughput.[53] This realization caused the development of new, so called next-generation sequencing (NGS) technologies which would ultimately need to enable the '1000 dollar' genome.

The term NGS covers a large group of related sequencing technologies, providing an exhaustive overview of all of them would be well beyond the scope of this work but in general, all NGS technologies share three basic characteristics [23] (Figure 1.2):

1. Genomic fragments for sequencing are not prepared through bacterial cloning
2. Orders of magnitude more sequencing reactions in parallel compared to aSseq
3. No CE is needed for readout because bases are read in parallel and cyclically

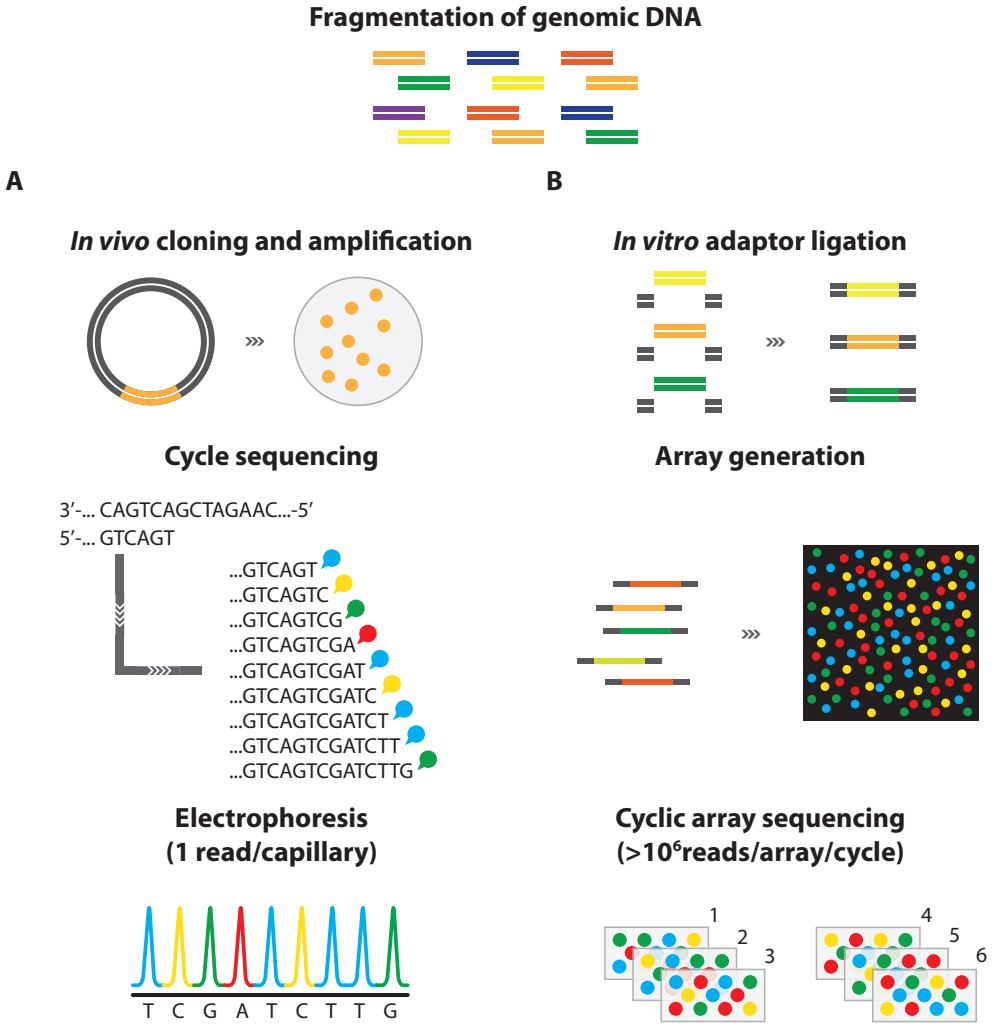

Figure 1.2: Comparison of the Sanger sequencing method and the so-called next-generation sequencing. A. Sanger sequencing: Bacterial cloning generation of DNA fragment is followed by their extension using DNA polymerase. The extension is stopped by the addition of a label that different for each bases. Eventually, the fragment are size separated using electrophoresis, using the information given by the labels and the size, the identity of the base can be tracked down. B. Next generation sequencing technologies : Unlike Sanger sequencing method, generation of genomic fragment is performed in vitro. This approach allows to have more sequencing reactions and a readout in parallel rendering it cheaper and faster than Sanger sequencing.

NGS is currently well established and indeed widely used in genomic analysis and diagnostic applications and the 1000 dollar genome has indeed become a reality. Nonetheless, significant drawbacks still remain, even with the most recent NGS technologies. [23] Indeed, in contrast with aSseq, NGS generally offers much shorter

contiguous readout lengths. This puts high requirements on computational methods that are used to reconstruct the full genomic sequence by analyzing the overlap regions between all these short read fragments.[80] This is particularly true when targeting repetitive regions within the genome such as e.g. when studying copy number variations of genes or tandem repeats.[80] Similarly, NGS also struggles to identify insertion-deletion (InDel) genomic variations.[1] Finally, NGS technologies still rely on polymerase chain reaction (PCR) based amplification of the material to be sequenced. For some genomes, e.g. genomes with AT or GC rich regions such as those of some known pathogens, this might result in bias and ultimately uneven coverage of the genome.[64]
To address these shortcomings of NGS, recent years have seen the development of so called single-molecule sequencing (SMseq) that can help to remove the requirement for amplification based library preparation.[79] A notably simplified sample preparation by the elimination of DNA amplification lead to a significantly higher all-round consistency compared to previous techniques. This makes single-molecule sequencing suitable for a wide range of applications including diagnostic, clinical [54] and epigenetic studies [48].

1.5 Optical DNA mapping

Optical DNA mapping is one of the members of a larger family of single molecule sequencing technologies, together with sequencing by synthesis and nanopore sequencing technologies.[79]. In optical mapping, it is not the aim to extract genomic information at the highest possible density, i.e. at single base resolution. Rather, very long DNA fragments are somehow deposited on a suitable surface or otherwise immobilized, e.g. in microfluidic or even nanofluidic channels. Care is taken to ensure that the DNA will remain intact over lengths of several hundreds of kilobases. Next, sequence information can be extracted by treatment of the DNA using specific restriction enzymes or nicking enzymes, for which the site of action is well known, in conjuction with suitable labeling methods to visualize generated fragments on a microscopy system. This way, approaches similar to those applied in gel based forensic restriction analysis can be used to e.g. identify organisms or the presence of specific genomic sequences in a large mixture of genetic material.[79] Furthermore, highly repetitive genomes or genomes featuring rearrangements or duplications of genes can be more successfully charted, something which is exceedingly difficult with traditional sequencing. This way, single molecule optical mapping has already contributed to provide a scaffold for complete sequencing of the e.g. Maize genome, a genome that is notorious for its large size and highly repetitive nature.[94] Although optical mapping cannot provide very detailed sequence information offered by other approaches.

1.5.1 Optical mapping *via* restriction enzymes

Restriction enzymes are able to cut the DNA at very specific sites (typically 6 or 8 bases long). They are involved in a defense mechanism (e.g. in bacterias) that degrades "foreignDNA " (e.g. viruses). Their uses for DNA mapping was first demonstrated by the Schwartz laboratory in 1993 and ultimately became the most famous approach for single molecule DNA mapping.[74][18] The mapping method is based on the imaging and the determination of the size of the DNA fragment generated by the restriction enzyme on surface-deposited or stretched DNA. The sizing is performed according to the the fluorescence intensity and the apparent contour length of the fragments. The position of the fragments on the gene is then tracked down using the known restriction sites together with the size information of the fragments. Its principle was further developed and the staining with an intercalating dye was demonstrated to allow to size accurately fragment as small as 800 bases.[52][10] DNA as large as 360 kb DNA could also be mapped successfully.[72] A notable contribution of this method is the mapping of methylation status. Methylation is an important genome regulation mechamism in eukaryotic cells. This methylation blocks the activity of the restriction enzyme if one of the base of its recognition site is methylated. By comparing an *in-silico* restriction map with no methylation and the experimental map obtained in presence of methylation one can consequently retrieve information about the methylated sites.[2]

Even though this method has found a lot of applications in genomics, it also has some serious drawbacks. For instance, it is limited in term of the size of the targeted sequence and in resolution. Indeed, small fragments will easily detach from the surface what limits the application of the method. In addition, the size of the fragment is determined by the intensity of an intercalating dye which is sensitive to conformational changes in DNA and variation in binding coefficient limiting its accuracy.[18]

1.5.2 Optical mapping *via* nicking enzymes

Another approach for optical mapping is to use enzymes to label the DNA with a high sequence specificity instead of cutting it. Nicking endonuclease represents one type of enzymes allowing for such approach. Unlike restriction enzymes, nicking enzymes only cleave one strands at the recognition site. DNA polymerase is then used to synthesized a new strand at the "nicked" site and a label can be subsequently incorporated. This principle was first reported in 1977 by Rigby et al. but only applied in 2008 for optical DNA mapping purposes by Xiao et al.[70][93][18] Xiao et al. labeled DNA at nicking sites with a green label and the backbone with a blue label. The DNA was stretched and then observed with total internal reflection fluorescence. Using a statistical analysis, they could extract the most probable position of the sites with a resolution of 800 bases (250 nm). The main limitation in their work was, of course, the resolution of the optical instrument.

The main advantage of this method is the covalent bonding of the label to the DNA which increases the specificity of the labeling as compared to intercalating approaches (non-covalent bonding). This type of labeling is therefore not limited to immobilized observation (surface deposition or stretching) but can also be applied to other approaches such as nanofluidic systems (described below). The main drawbacks of the method is the non specific labeling of the strand by the DNA polymerase and the fact that high density labeling can hardly be used due to the undesired deformation of the DNA structure that it would cause.

1.5.3 Optical mapping in nanofluidic devices

In the previous methods, deposition or stretching of DNA was used prior to image it. A limitation of those methods is that many images have to be taken to get a sufficient statistic increasing tremendously the duration of an experiment. Nanofluidic devices presents an alternative that allows to extend the DNA up to 60-70% of its crystallographic length while getting hundreds of molecules through the field of view in a wide-field microscope. Rapid mapping in parallel is a crucial advantage of this method that other methods should find solutions to.[65][47][18]

Mapping of restriction sites using nanochannels was demonstrated by Riehn et al.[69] Even though the demonstration is remarkable, the use of this approach is rather limited and complicated in nanodevices because one needs to ensure that the DNA is not digested by the enzyme before entering the nanochannels. Therefore, other approaches such as the previously described nicking enzymes are more suited in nanodevices. There use in nanofluidic devices was first reported by Jo et al.[37] In their work, they could determine the position of the nicks on the gene with a standard deviation of 3.3 kb. This is rather low compared to what was obtained in the previous approaches. Another alternative is to use methyltransferase enzymes. It is a family of enzymes that transfers a methyl group to one of the bases of a recognize site (4-6 bases) on the DNA. This type of enzymes is present in many bacterias and allow to protect their own DNA from the activity of restriction enzymes. Methyltransferase approaches have the advantage that it does not affect the structure of the DNA, further details are given in Subsection 1.5.4. The use of this approach in nanofluidic devices was first shown by Kim et al. [40] In their work, they used a six bases recognition site methyltransferase enzyme to label T7 phage DNA at three distinct sites. The enzyme transfered a cofactor linked to a biotin to which a quantum dot was finally attached. Although, the bouding efficiency of the quantum dot was 40% thus the three sites were not always labeled. It was then exploited by Grunwald et al. in which they use mTaqI enzyme (4 bases recognition site) to label T7 and λ-phage DNA with a high density (>100 sites). The molecules were stretched in nanochannels, intensity profiles were then extracted and ultimately matched to in-silico maps by cross-correlation. They could identify the correct DNA from a library of 10 DNA sequences.[25]

The disadvantage of nanofluidic is the random diffusive processes that occur during

the imaging. Corrections for these effect have to be apply.[62] In addition, the use of super-resolution method is render more difficult due to the motion of the molecule.

1.5.4 Optical mapping *via* DNA Fluorocodes

In DNA fluorocode mapping, methyltransferase enzymes are to apply fluorescent labels to contiguous DNA fragments in a sequence specific manner. Next, super-resolution fluorescence microscopy (SRFM) is used to locate these fluorescent labels on surface deposited DNA with high precision (Figure 1.3). The approach is also referred to as methyl transferase directed transfer of activated groups (mTAG), orignally reported by Klimasauskas and Weinhold [43, 50] and later extensively studied in the Hofkens group [60, 61] and is based on the conjunction of a number of technological aspects, which will be discussed next.

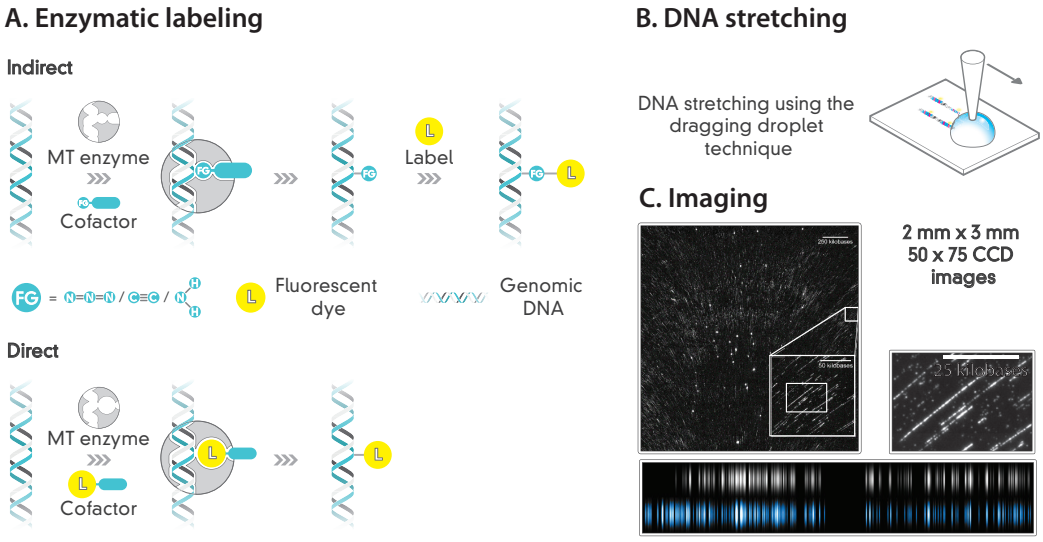

Figure 1.3: The Fluorocode approach for optical DNA mapping : A. High density labeling using enzymatic labeling. The labeling can be carried out indirectly by attaching the cofactor to the DNA and subsequently add the label to the cofactor or directly by labeling the cofactor prior to its incorporation onto the DNA. B. Stretching of the DNA using the rolling droplet method (described below). C. DNA imaging using super-resolution method. Photoactivated localization microscopy is the method that was preferred until now.

Enzyme assisted labeling

In mTAG, either naturally occuring or mutated variants of methyltransferases can be used to transfer either a chemically reactive group or even a full label to DNA at their recognition site.[43, 50, 60, 61]. Natively, methyl transferases transfer a methyl group from their natural cofactor, *S*-adenosyl-L-methionine (SAM) to DNA in a sequence specific manner as they recognize and act on short, palindromic recognition sites, typically from four to six base pairs (bps) in length. However, since methyl-

transferases are a particularly flexible class of enzymes, they can also be used in conjunction with artificial analogs of SAM to transfer other chemical groups to the DNA.[60, 61]

This opens up a host of potential chemistries that can be used to link fluorescent labels to the DNA (Figure 1.4):

1. *Copper(I)-catalyzed alkyne-azide cycloaddition (CuAAC)*: An alkyne cofactor is transferred to the DNA, allowing for subsequent labeling using an azide functionalized dye moiety and well known click chemistry. Although this coupling reaction is highly efficient, the presence of Cu(I) can induce damage in the DNA.

2. *Strain-promoted alkyne-azide cycloaddition (SPAAC)*: Here, an azide cofactor is transferred to the DNA after which a cyclooctyne functionalized dye can spontaneously react with the azide to achieve a 'copper free' click reaction.

3. *N-Hydroxysuccinimide (NHS) coupling*: Here, an amine cofactor is introduced which allows for coupling with carboxylated dyes.

4. *Direct labeling*: Here, a cofactor is prepared that already incorporates the fluorescent label itself.

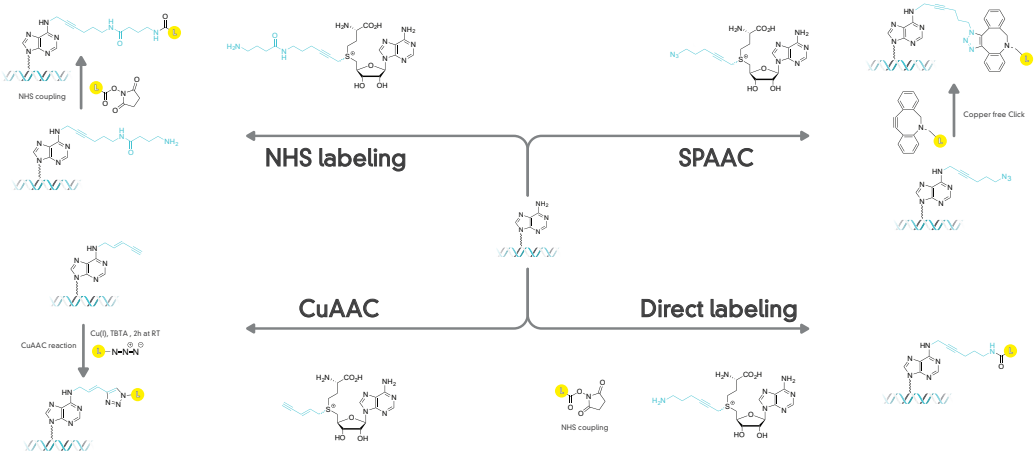

Figure 1.4: Comparison of different approach to attach the label to the DNA. Methyltransferase are a class of enzyme that offers a certain flexibility in the cofactor they can add to the DNA what leads to the development of a panel of method to achieve the desired labeling

.

Whereas CuAAC is a highly efficient reaction that can readily run to completion, Cu(I) induced DNA damage might occur whereas the cyclo octyne reagents used

for the copper free reaction are not readily available. Therefore, to date, the direct labeling and NHS approaches are preferred. Between these two methods, direct labeling might be considered more demanding for the enzyme since it will need to accommodate the presence of an often bulky fluorescent labeling in or near its cofactor binding site. Therefore, in some cases, labeling efficiencies might be reduced. For this reason, withn Hofkens group, the NHS is by far the most used approach.

DNA linearization and molecular combing

To allow proper imaging of labeled DNA molecules, large size DNA fragments need to be deposited on a suitable substrate. Bensimon *et al.* were among the first to report a method for 'molecular combing' where a droplet containg a suspension of genomic DNA would be left to evaporate on a glass substrate that was treated with a silane compound featuring vinyl ($-CH{=}CH_2$) to become hydrophobic.[4, 5] The receding meniscus, in conjunction with the tendency of DNA to adhere to the hydrophobic surface would cause individual DNA molecules to become linearized on the surface. This way, the authors managed to deposit large fragments of λ-phage DNA molecules which, upon deposition, featured lengths of 21.5 ± 0.5 μms. Not only did the authors demonstrate the ability to image large contiguous sequences of DNA but it was also demonstrated how even minute quantities of DNA could be detected, down to a quantity of 10^3 molecules in a single sample. As such, molecular combing was rightly posited as "a way for a faster physical mapping of the genome and for the detection of small quantities of target DNA from a population of molecules".
The technique was further extended by Deen *et al.* who employed a combination of polymer treated glass substrate and so called 'rolling droplets' (Figure 1.6).[19] By dragging a single droplet of suspended DNA over the surface, the flow inside the moving droplet will ensure an enrichment of DNA at the interface between the droplet and the surface, ultimately resulting in much higher densities of DNA on the surface, thus addressing one of the major shortcomings of previously reported optical mapping approaches where DNA was deposited at much lower density, which would limit the amount of information that could be extracted.

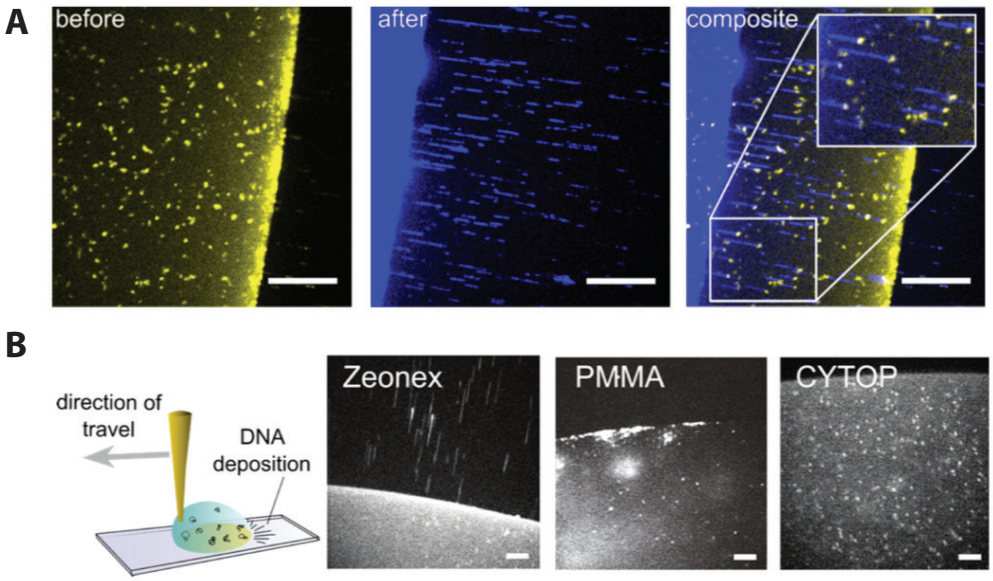

Figure 1.5: Principle of the rolling droplet for DNA stretching. A. In yellow, before the droplet start to move, DNA is deposited on the surface. In blue, the movement of the droplet has stretched the DNA on the surface as the front line moved over them. B. Basic scheme of the method. Additional images demonstrates that zeonex is the coating that allows the most efficient stretching. The images were extracted from[19]

.

Imaging

As will be discussed in detail in the following chapter, optical imaging is inherently subjected to the diffraction limit which states that the maximum resolution that can be obtained, i.e. the minimum separation needed between two features to still be distinguishable, is governed by the wavelenght of the light and the optical properties of the microscope. For a regular optical microscope and visible light, this diffraction limit is in the order of 250-350 nm, a value which corresponds to roughly 700 to 800 bps which is well above the average density of a four base restriction site which can be expected to occur every $4^4 = 256$ bps.[18]

To address this issue, SRFM approaches are thus needed. Whereas many such modalities exist (detailed in Chapter 2), to date a stepwise bleaching method was most frequently used.[60, 61, 18] Indeed, when a sample containing many emitters at relative distances smaller than the diffraction limit are subjected to intense illumination light sources, the fluorophores can be expected to accumulate photodamage and stochastically enter a dark state, i.e. be 'photobleached'.[67, 17] When the sample is thus observed for an extended period of time, with frames being recorded at regular intervals, the final frame can be expected to only contain a few isolated emitters.

with knowledge of the instrument point spread function (PSF), one can determine the positions of these emitters with sub-diffraction limit accuracy by fitting it with a 2D Gaussian profile. Proceeding to the second to last frame, these fitted profiles can then be substracted from that frame, allowing a new round of localization of isolated emitters. This process is repeated until the beginning of the time series is reached and all emitters are localized. Although the approach offers the advantage of extreme simplicity compared to other SRFM modalities, the process of repeated PSF substraction can be expected to induce significant artefacts if the amount of labels increases substantially.[73] Therefore, more suitable SRFM approaches might be evaluated in the future and in fact, this is also the aim of the presented work.

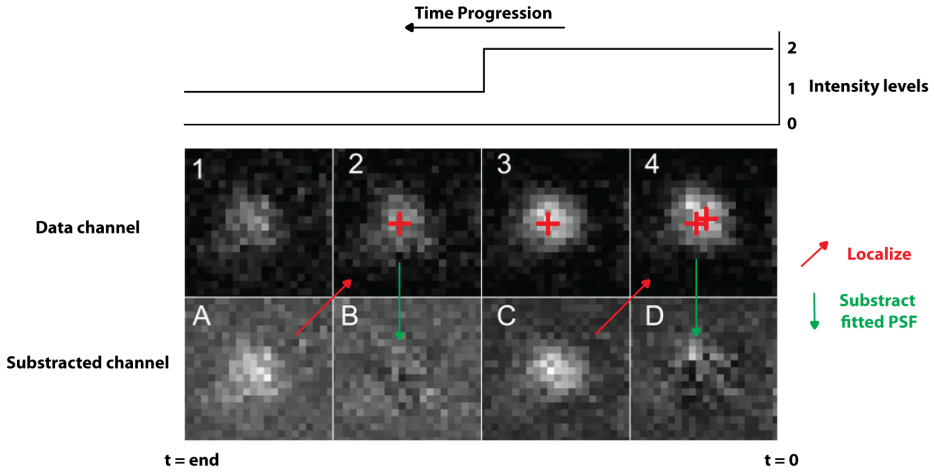

Figure 1.6: Principle and drawback of the bleaching analysis. 1. Last frame. The molecules is localized by a Gaussian fit which is subsequently subtracted from the previous frame(2), B shows that after subtraction, some signal is still left. 3. The previous frame has another molecule on it which is localized and subtracted as before. D shows that two subsequent subtraction can lead to a significant leftover of signal that can induce localization error.

.

Fluorocode matching

After label positions are extracted using SRFM , experimentally obtained sequence maps can be matched to reference genomic data, e.g. to identify the nature of the analyzed DNA sequence.

For this purpose Deen [18] applied the well known Smith-Waterman algorithm. This algorithm operates on the principle that segments, i.e. the linear distance between two successive labels, of the experimental map can be length matched to a corresponding segment in the reference map. As such, the algorithm is similar to frequently used full sequence alignment algorithms it's applicability to restriction

map analysis has already been demonstrated.[18]

Smith-Waterman is actually applies a dynamic programming approach to map matching. The term dynamic programming is an algorithmic strategy that can be applied when recursion would be inefficient approach to solve a particular problem because similar tasks would have to be repeated many times over. This can be readily understood when considering the well known Fibonacci sequence (0, 1, 1, 2, 3, 5, 8, 13).[1] Whereas the first and second number in the series are defined to be 0 and 1 and hence do not need to be calculated, the nth number is defined the sum of the two preceding numbers in the series. So, one can calculate the nth Fibonacci number with a recursive function:

```
fibonacci(n) {
   if (n == 0) {
      return 0;
   } else if (n == 1) {
      return 1;
   } else {
      return fibonacci(n - 1) + fibonacci(n - 2);
   }
}
```

However, this approach would be inefficient as, with increasing n, the function 'fibonacci(n)' would be called many times over for values <n. It would indeed be much more efficient to abandon this recursive, top-down approach in favor of a bottom-up approach where 'fibonacci(n)' is calculated once for low values of n with the outcome stored for re-use to calculate the outcomes for large values of n:

```
fibonacci2(n) {
   table = array of size n + 1;
   for (int i = 0; i < table.length; i++) {
      if (i == 0) {
         table[i] = 0;
      } else if (i == 1) {
         table[i] = 1;
      } else {
         table[i] = table[i - 2] + table[i - 1];
      }
   }

   return table[n];
}
```

Calculating the outcome of the function for large values of n thus becomes more efficient because the outcomes of previous calculations are stored for re-use. Although this approach will ultimately use more memory to store these values, the speed gains can be significant for complex problems as repeatedly calculating the same values is avoided.

[1] $http://www.ibm.com/developerworks/library/j-seqalign/$(accessed on the 15th May 2016)

A similar dynamic programming approach is applied in the Smith-Waterman algorithm when trying to match an experimental map of size m with a reference map of size n:

1. A scoring grid of size $m \cdot n$ is created. This scoring matrix will be filled from the top left towards the right and down.
2. For each position (m,n) in the matrix, a score is assigned which is the maximum of:
   a) The score of element (m-1, n-1) + the independent score for element (m,n)
   b) The score of element (m-1, n) - the gap penalty
   c) The score of element (m, n-1) - the gap penalty
   d) 0
3. After filling the entire matrix in this way, the best total match can be found by tracing the matrix diagonally left and up starting from the highest scoring matrix element. In case of a tie, up-left is favored over up and up is favored over left.

Using this approach, the longest common match between two maps will always be found.
It is worth mentioning a recent implementation of this method by Valouev et. al. [84] In their work, the matching score is described as the logarithm of a likelihood ratio that the segments studied are similar. The likelihood ratio allows to account for issues of restriction mapping such as false cuts, sizing error, missing fragments and distribution of sites when calculating the probability that the fragment are identical by the use of a statistical model. The reader is reffered to Mendelowitz and Pop review for more alignment procedures.[51] An adapted version of this algorithm which account for the experimental issues of the fluorocode approach was succesfully applied for fluorocode map alignment.[18] This approach was shown to be more robust than the cross-correlation of two intensity profiles because of its lower sensitivity to stretching variation. However, this method is not suited for imaging methods such as stimulated emission depletion (STED), super-resolution optical fluctuation imaging (SOFI), structured illumination microscopy (SIM) where the output is a super-resolution image rather than coordinates. Moreover, the two methods were only compared on simulations, never experimentally.

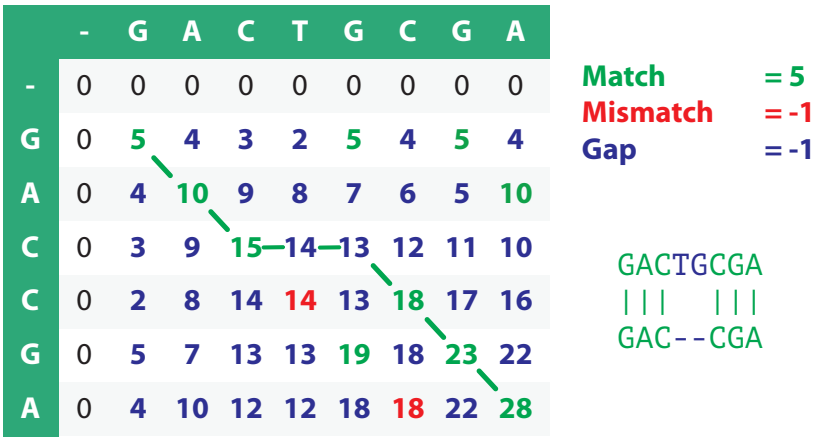

Figure 1.7: Illustration of the scoring assessment used in Smith-Waterman approach. The scoring grid is filled from the top left to the right down. Each position (m,n) in the matrix (grid) is assigned the maximum value of the following : The score of element (m-1,n-1) + independent score of the element, the score of element (m-1,n)- the gap penalty, the score of element(m,n-1) - the gap penalty or 0. Tracing the matrix diagonally left from the highest score element allow to find the best total match

.

Chapter 2

Super-resolution fluorescence

2.1 Introduction

The main disadvantage of optical DNA mapping was, until recently, the limit in resolution of the optical instrument used. This chapter present how optical instruments have evolved from the microscope to the nanoscope. Stimulated emission depletion microscopy is detailed and the aim of the thesis are presented.

2.2 The optical microscope

Hans and Zacharias Janssen, father and son, developed one of the earliest known examples of an optical microscope at the turn of the 16th century. However, it was only designed and used for scientific purposes in the 17th century by Antony Van Leeuwenhoek. At that time, only one single lens was used. A bit later, Antony Van Leeuwenhoek added a second lens that would magnified the image created by the first lens. The main components were then light, an objective and an eyepiece.[1] The later two are lenses. Figure 2.1 below presents the basic principle of a standard light microscope.

The resulting imaginary image is thus a perfect copy of the object but magnified. The magnification depends mainly on the characteristic of the lenses. Ever since it was invented, microscopy has enabled to discover a huge amount of features in most field of sciences but, more particularly, in biology and material sciences. In order to exploit the potential of this invention to its maximum, scientists started to improve the properties of the optical elements composing it.

[1] $http://www.visioneng.com/resources/history-of-the-microscope$(accessed on the 20th February 2016)

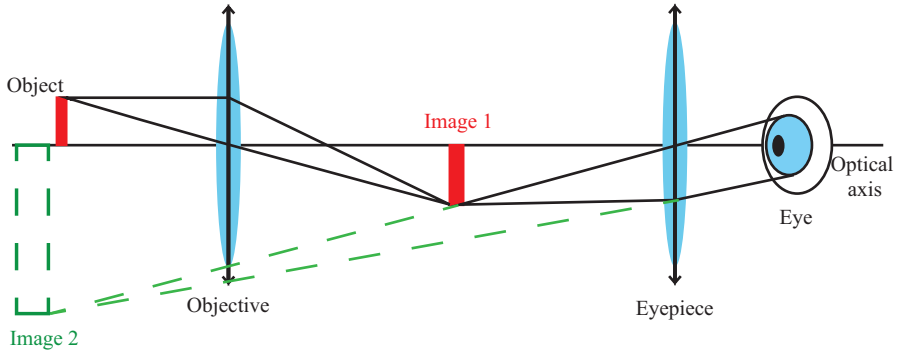

Figure 2.1: Standard Light microscope : The light is shown onto an object, the transmitted light is collected through the objective which creates a real image reversed and magnified that is used by the eyepiece to create an imaginary image further magnified

In the beginning, the main goal was to increase the magnification power which would allow to zoom in onto smaller features. Then, the resolution which can be described as the ability to distinguish two objects that are close in space was found to be a crucial feature of a microscope as well as the contrast which describes how well the image can be distinguish from the background. These three characteristics are the most important in light microscopy. There relative significance depends on the application even though, for any application, the highest the performances of a microscope the better.

However, a limit to the improvements we can make to the microscope's components was reached. Indeed, being able to zoom in infinitely would be useless without a proper resolution and a proper contrast and the resolution was shown to be limited by the diffraction, firstly described by Abbé and Rayleigh in the 19th century.[90, 16, 28] As we will see later, the resolution limit depends on two main factors, the wavelength (shorter the better) and the numerical aperture of the lens (larger the better). Yet, both factors are limited. Indeed, the first one is limited because cell and bio-sample imaging are not compatible with wavelength smaller than 350 nm and the second one is limited because the half aperture angle collected by the lens is technically limited to 70° [31]. This is why, in parallel to the development of new components, scientist started to elaborate other methods by adding and changing the components of the microscope. For instance, fluorescence microscopy brought a high specificity whereas confocal microscopy allowed to improve the resolution and the contrast of the images (more explanation in Subsection 2.3.2). However, the diffraction limit remained unbroken. Methods based on other sources than light were also developed (e.g. electron microscopy and scanning prob microscopy). Despite the advancement in electron microscopy and scanning probe microscopy, most of the life sciences application are still investigated with conventional microscopes.

This is why, in the continuum of the elaboration of new methods, scientist started to develop ways to circumvent the limit in resolution.[28] For now, all the so-called

super-resolution methods are advanced variant of fluorescence microscopy. In the first part of this introduction, we will focus our interest onto fluorescence microscopy, the improvement that confocal microscopy brought to the field and the allowance for dynamic studies that fluorescence correlation spectroscopy (FCS) brought. Finally, i will detail one of the super-resolution method : Stimulated emission depletion (STED) and presents its advantages in comparison to other super-resolution method.

2.3 Fluorescence microscopy

As it can be seen from the scheme of a wide-field fluorescence microscope (Figure 2.2), a fluorescence microscope is not so different from a light microscope.

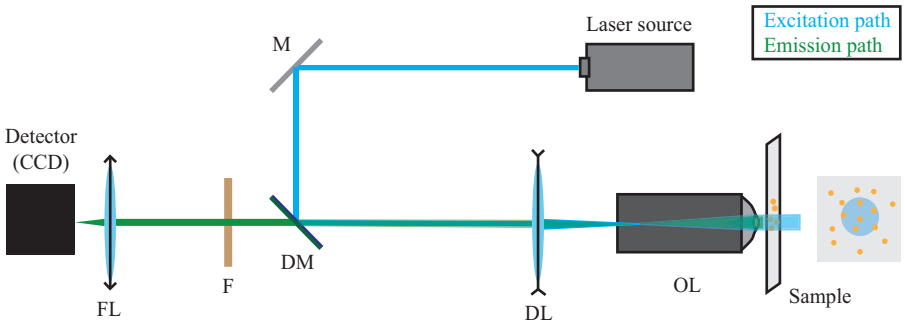

Figure 2.2: Wide-field fluorescence microscope : M Mirror, DM Dichroic mirror, DL Dispersion lens, OL Objective lens, F Filter, FL Focusing lens, Excitation path (blue), Emission path (green)

The main differences are the fact that the objective is placed below the sample and is therefore used in both excitation and emission paths, the source has a narrow wavelength bandwidth therefore, lasers are the most commonly used. The presence of a filter that allow to extract only the emission of the sample (not the reflection/backscattering etc...) and of a dispersion lens that allows to illuminate a great part of the sample (field of view (FOV)) are also two notable differences. The last difference would be the detector used, in wide-field it is usually a CCD camera. However, recent light microscope can also use similar detectors, it is thus not really a difference. In the next section, the principle used in fluorescence microscopy will be discussed.

To explain fluorescence, one should describe the phenomenon that usually comes just before: absorption. When a photon with an adequate energy interacts with a molecule, it can be absorbed. Having an adequate energy, means that the energy should be equal to the difference of energy between two energy levels (vibration, rotation, or electronic) of the molecule for which the transition is allowed. In standard conditions, most of the molecules are found in their electronic ground state and the major part of the population is found in the vibrational ground state. The absorption

will therefore mainly occur from the lowest level (v_0 in S_0 in Figure 2.3). The energy of the absorbed photon will cause the promotion of an electron to a higher energy level (v_x of S_1 where x will depend on the energy of the photon). Absorption is ruled by Beer-Lambert law. Considering a light source of intensity I_0 targeted toward a cell of length l containing a solution of concentration c in solute, Beer-lambert law is written as follow.[82, 83, 81, 44]

$$log(\frac{I_0}{I}) = \epsilon cl \tag{2.1}$$

Where I is the intensity of the light passing through the cell (outcoming beam) and ϵ is the extinction coefficient. The extinction coefficient is a measure of how well the molecule will absorb, it is thus proportional to the absorption cross section. After absorption, the excited state created is not stable. Therefore, the molecule will decay following different possible paths as shown on the simplified Jablonski diagram in Figure 2.3 below.

One can see from Figure 2.3 that there is two main types of decay path : radiative paths and non-radiative paths. The difference between the two is that in the first case the energy is lost by the emission of a photon, in the second case, the energy is lost by friction or distribution to the environment.

As it can be seen from the diagram, the emission is also mainly occurring from the lowest vibrational level from S_1 even though the absorption can cause the vibrational level to be populated. This is because there is a vibrational relaxation. This non-radiative relaxation involves collisions with other molecules and distributing energy to its environment. It happens in a very short time scale, particularly in liquids due to the important number of collisions. Most of the molecules are therefore relaxed to the lowest vibrational level of the first electronic excited states. This is why emission mostly (only) occurs from that level. While on the lowest level of the excited states, different decay paths are shown on the Jablonski diagram (Figure 2.3).[75, 82, 83, 81, 44] The next paragraphs are dedicated to their description starting from the radiative decay paths to the non-radiative decay paths.

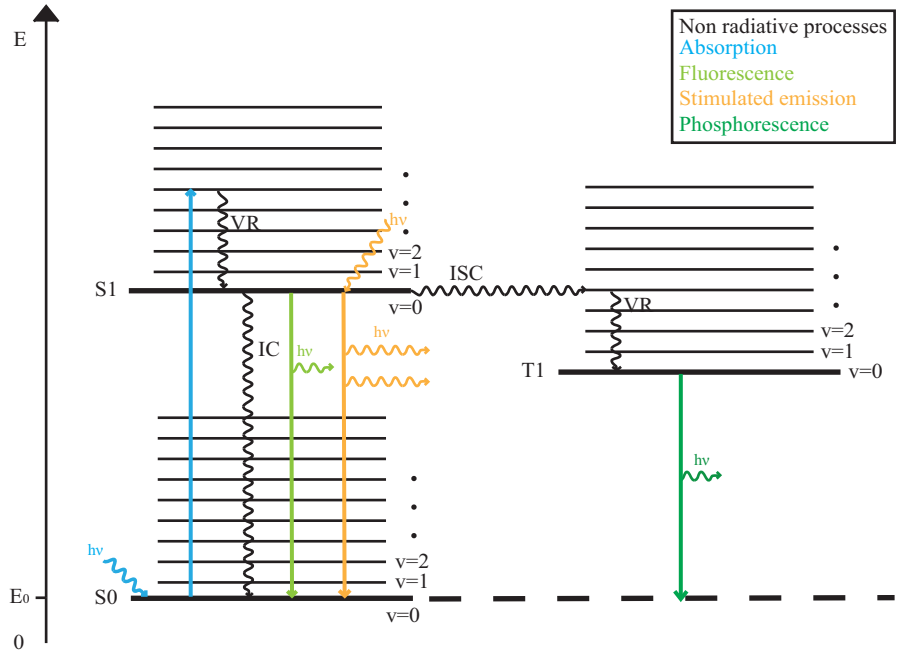

Figure 2.3: Jablonski diagram. IC: Internal conversion, VR: Vibrational relaxation, ISC: Inter-system crossing, v: denote the vibrational levels, S denotes the singlet electronic state, T denotes the triplet electronic state. The straight lines represent the radiative transitions whereas the curvy lines represent non-radiative transition [90, 82, 83, 81, 44]

2.3.1 Fluorescence

Fluorescence is one of the possible decay path and is characterized by the spontaneous emission of a photon, it is thus a radiative path. As the fluorescence only occurs from the lowest vibrational level of S_1 and the decay path can end up on an excited vibrational level, the photon emitted is red-shifted in comparison to the excitation. The fluorescence rate depends on how unstable the excited state is and thus on the lifetime of the state. The fluorescence rate k_{fluo} is equal to $\frac{1}{\tau_{fluo}}$ where τ_{fluo} is the fluorescence lifetime of the state or the lifetime of the state if fluorescence was the only possible decay path. The order of magnitude for τ_{fluo} is usually a few ns to a few hundreds ns. The lifetime of the state (τ) is thus shorter than the fluorescence lifetime. The more unstable the state, the shortest the lifetime. The proportion of each decay path depends on their own lifetime in comparison to the lifetime of the state. Indeed, if $\tau \approx \tau_{fluo}$, the fastest process is fluorescence and thus it is the main decay path. A parameter that is important here is the quantum yield. It can be defined as the number of photons emitted by fluorescence divided by the number of photons that have excited the sample (that were absorbed). It therefore informs us about the proportion of excited molecules that used spontaneous emission as a decay path. This is of course dependent on the molecules and on its environment.[82, 83, 81, 44]

Stimulated emission

The second radiative decay path is called stimulated emission. Even though, it might not be the most important here, it is important to know the concept to understand the STED effect later. Unlike fluorescence, Stimulated emission (as its name stands for), is not spontaneous and needs a external stimulation. This stimulation is usually achieved by a photon. In that case, the emitted photon has the same characteristic (direction, phase, wavelength,...) as the photon that was used to trigger the emission. The stimulated emission is the main principle hidden in laser sources. When illuminating a sample, there is usually both spontaneous and non spontaneous emission. The relative significance of them depends mainly on the wavelength used. Equation 2.2 presents the parameters influencing the stimulated emission cross section.[82, 83, 81, 44, 45]

$$\sigma_s = \frac{\lambda^4 E(\lambda)}{8\pi \cdot c \cdot n^2} \tag{2.2}$$

Where

- σ_s is the stimulated emission cross section in cm^2
- λ is the wavelength in nm
- $E(\lambda)$ is the fluorescence intensity normalize to the quantum yield
- h is the plank constant in J.s
- c is the speed of light in m/s
- n is the refractive index of the media

Equation 2.2 shows that the stimulated emission cross section is proportional to λ^4. Therefore, it varies a lot with the wavelength and is favored at high wavelength (or low frequencies/energies). This is one of the reason why making a laser in the infrared is much easier than in the UV region. Moreover, it is also proportional to the fluorescence and inversely proportional to the quantum yield. The later one can be explained by the definition of the quantum yield. Indeed, as mentioned earlier, the highest the quantum yield, the highest the proportion of molecule taking the fluorescence decay path is. Therefore, it is easy to understand why stimulated emission will be higher for systems having a low quantum yield. The ratio of the fluorescence intensity divided by the quantum yield actually represents the total number of excited molecules. As Stimulated emission acts on excited molecules, it is also understandable that its cross section depends on this ratio. The stimulated

emission rate k_s is equal to $\sigma_s I_{exc} \lambda / hc$.[82, 83, 81, 44]

Phosphorescence

The last radiative path is phosphorescence. It is a phenomenon similar to fluorescence but is much slower. This is because the ISC that allows the electrons to go from the singlet state to the triplet state is very slow because such a transition is not favored. Moreover, phosphorescence is also a transition from a triplet to a singlet state what makes it very slow. When fluorescence happens in a nanosecond scale, phosphorescence happens in a millisecond or even second scale. The rate of phosphorescence depend both on τ_{ISC} and τ_{phos}.[82, 83, 81, 44]

Internal conversion

The first non-radiative decay path is called Internal conversion (IC). It is a dissipation of energy occurring by an "horizontal" transition between the lower vibrational level of the excited state (S_1) with a high energy vibrational state of the ground state that have the same energy. The name internal arises because it occurs between the molecules own levels without the need for an intervention of the environment or other molecules. After this "horizontal transition", the molecule goes back to the ground state by vibrational relaxation. Internal conversion can only happen between state of the same spin (singlet to singlet or triplet to triplet).[82, 83, 81, 44]

Internal system crossing

The last decay path is the intersystem crossing or internal spin conversion (ISC). This conversion is also a kind of Horizontal transition but the particularity is that there is a change in the spin. Indeed, the transition occurs between a singlet state (S_1 in that case) and a triplet state(T_1) or vice versa. This transition is, in most systems not favored because of this change in Spin. In some condition, it can nonetheless become important and then phosphorescence can become an important decay channel. Indeed, once the ISC occurred, Vibrational relaxation is much probable than the inverse ISC, therefore, if ISC is very fast in comparison to the other processes, phosphorescence can become a important decay path.[82, 83, 81, 44]

The different decay paths have all their own timescale that depends on the molecule studied, its environment and the conditions used. Therefore, most of the time, less than 100% of the molecules excited actually fluoresce. Moreover, the Fluorescence is emitted in 4 π which makes the photon collection not so efficient. These different factors explained why lasers are widely used sources in fluorescence microscopy. Indeed, lasers are able to provide a high number of photons per time unit what allow to excite a great part of the sample and thus to retrieve a decent fluorescent intensity. Moreover, lasers usually have a quite narrow wavelength distribution what permit to excite only the specific molecules of interest in a complex matrix what gives to this

technique a high specificity.[82, 83, 81, 44]

However, fluorescence microscopy has some limitation. Indeed, a lot of molecules do not absorb in the visible or near-UV range and therefore, do not emit in the visible range. Moreover, UV light is not suitable for life science studies because it can trigger reactions and degradations in bio-sample. A way to overcome this problem is the use of labels even though, it is not always convenient to work with labels because it adds steps to the sample preparation, it can increase the specificity of the technique a step further. It is also limited in terms of resolution, as it is the case for standard light microscopy.

In essence, the resolution of an optical instrument such as a microscope is the minimal distance between two observed objects such that they can still be observed as to distinct entities. However, this definition is still distinctly qualitative. In an effort to provide a more quantitative description, many scientist have attempted to formalize the concept of resolution such as e.g. Abbé, Rayleigh, Sparrow and Houston. While each of these, Abbé formulation is probably the most regarded and is shown in Equation 2.3.

$$d = \frac{\lambda}{2n \cdot sin(\alpha)} \tag{2.3}$$

Where d is the resolution and alpha is the half aperture angle collected by the optical element (lens). Before detailing other definitions of resolution, one should talk about the main actors responsible for its limitation. The main actor in the resolution limit is the diffraction. It is an optical effect which can be seen as an interaction between the light and an optical element. This effect is mainly observed when the size of the element approaches the wavelength of the light used. The most common example to explain it, is an optical system with light passing through a small slit leading to what is called an Airy pattern (see Figure 2.4). In a microscope, light will be diffracted by the different optical element that compose it leading to a limit in the resolution power.[90][16]

The Airy-pattern is marked in black on the figure to highlight its shape. Depending on the angular distance from the center, a succession of minima and maxima will appear in the pattern. The highest intensity is found for the 0 order (0° deviation from the center). The full width half maximum of the pattern will depend on the wavelength of the light used and the numerical aperture of the objective, thus leading to a limit in resolution.[90][16]

The second actor is thus the numerical aperture NA of the objective. It is a factor appearing in all the resolution formulation (cf. Abbé). This is a characteristic that describes the amount of light collected by an optical element based on the refractive index of the medium n and the full aperture angle collected by the lens θ (see Figure 2.5).

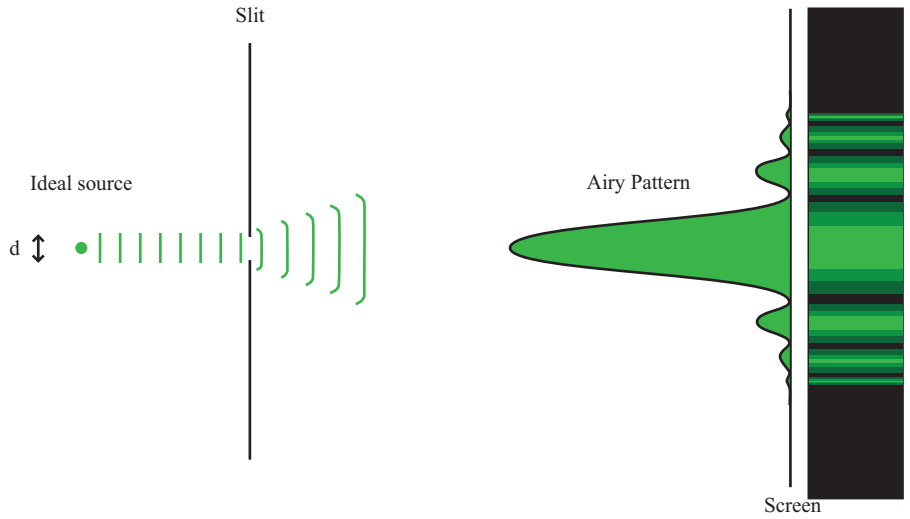

Figure 2.4: Interaction of light with an optical element (in this case slit) result in an Airy pattern which is composed of a high intensity first order maximum can be observed next to first order minima . The right of the figure represents the projection of the Airy pattern on a screen. Diffraction is only observed when the size of the slit is similar or smaller than the wavelength used.[75]

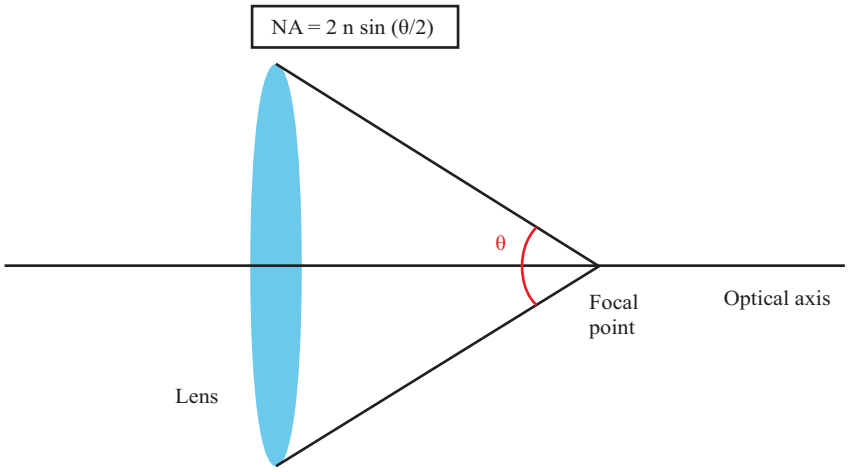

Figure 2.5: The numerical aperture of a lens describes the amount of light it is able to collect. It depends on the refractive index of the medium and the full aperture full aperture angle θ. The bigger the numerical aperture the higher the resolution. However, it is technically limited because the range of media usable is relatively narrow and the full aperture angle is technically limited to 140°.[75]

The full aperture angle is technically limited to 140°, it is thus not possible anymore to play with this parameter to increase the numerical aperture [31]. Nevertheless, one parameter is left : the refractive index of the medium. Scientist developed immersion oil objective because oil has a higher refractive index than air or water thus causing

the numerical aperture to increase. The drawback of this improvement is that brutal changes in refractive index (oil-glass-air) can cause deformation in the image. The dropcast of an index matching solution onto the sample can usually overcome this problem. Unfortunately, the gain is thus also limited giving a numerical aperture equal to 1.4. This is almost twice as much as with dry objective (0.8) what is already a good enhancement. Moreover, these objective can usually have high magnification (x60 to x100).[90][16]

As it was mentioned earlier, various definitions of resolution exist. However, the most widely used is probably the Rayleigh Criterion and thus, we will detail it here. The Houston criterion will also be explained as we will use it later to determine the resolution of the STED setup. The Rayleigh criterion uses the overlap of the Airy function to determine the distance needed between the object to be able to resolve them.[38] *Figure 2.6* illustrates the Rayleigh criterion.

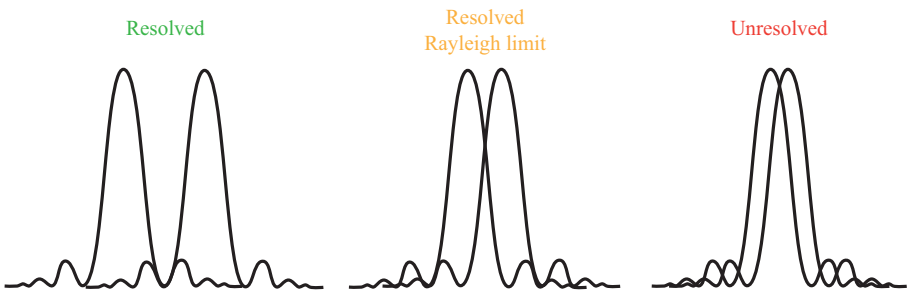

Figure 2.6: Rayleigh criterion: two Airy patterns close by will be resolved if the first minimum of the pattern coincide or is further away from the maximum of the other pattern [75]

This criterion is also used to describe the resolution in fluorescence microscopy. *Equation 2.4* allows us to have an idea about the lateral resolution we can reach for a Standard Wide-field microscope.

$$\Delta r = \frac{0.61 \cdot \lambda}{NA} \tag{2.4}$$

To give a typical number, when an oil immersion objective (NA=1.4) is used with a 500 nm light, the resolution is limited to 220 nm. It is surely sufficient for a range of application but with the not-so-recent tendency study the nanoworld it is not sufficient at all. The Houston criterion simply uses the full-width half maximum to represent the resolution. It is the width of the pattern at the half of the maximum. It is a very convenient way of representing the resolution because by imaging a object that is smaller than the resolution of the instrument, the FWHM gives information on the reachable resolution. We will use this definition later during STED measurement.[38] The Houston criterion is presented in Figure 2.7.

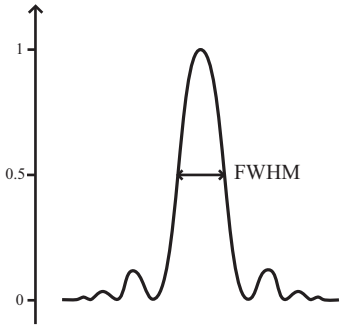

Figure 2.7: The Houston criterion assesses the resolution of an optical instrument by the use of the full width half maximum of the Airy pattern

As mentioned earlier, the contrast is also an important feature of a microscope. It can be describe as how well we are able to distinguish the feature of interest on the image from the background (or noise). The signal-to-noise ratio $\frac{S}{N}$ can thus give a good idea about the contrast of the image.[90]

2.3.2 Confocal fluorescence microscopy

Confocal Fluorescence microscopy is a more advanced technique of fluorescence microscopy. It is more advance in the sense that a few components are added or changed : the excitation is focused onto a spot, a point detector is used and a pinhole. The pinhole also implies to add two lenses. Finally, the fact that the excitation is focused implies the need for a scanning system. The basic layout of a confocal microscope is presented in Figure 2.8.

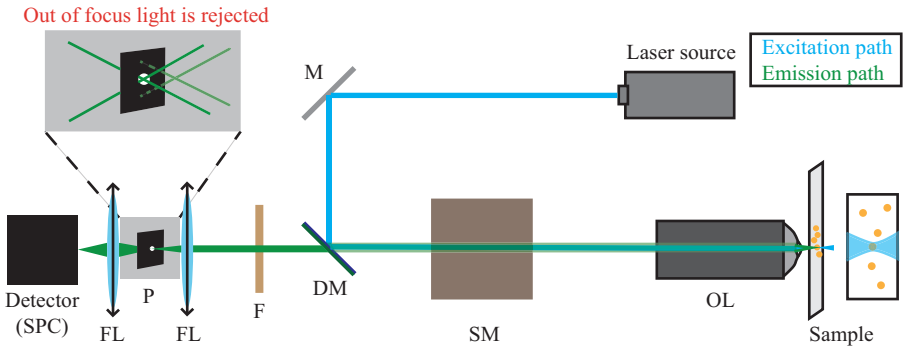

Figure 2.8: Confocal microscope setup : M Mirror, DM Dichroic mirror, SM Scanning module, OL Objective lens, F Filter, FL focusing lens P Pinhole, SPC Single photon counting

The confocal microscope was invented in 1955 and patented in 1961 by Marvin Minsky.[55, 56] Figure 2.8 shows that the light is focused onto the pinhole and then focused onto the detector thus allowing to remove a great part of the out-of-focus light. This accounts for the increase in resolution obtained as stated by Equation

2.5. This equation is valid for infinitesimal pinhole and thus, describes the best resolution one could ever achieve in confocal microscopy. Moreover, the contrast is also enhanced as we eliminate the fluorescent background from out-of-focus planes and thus the signal-to-noise ratio is increased. In addition, it allows to "cut" the sample into slices (optical sectioning) that can be reused to reconstruct a 3D image of the sample.[75, 89]. One of the disadvantages is that due to the high photon density (focus on one point), one can more easily have photo-bleaching effect. One often has to make a compromise between high resolution (require high exposure time) and risk of degradation. This is particularly true when biological sample are concerned. A second disadvantage would be that the source used is focused onto a point and thus the sample has to be scanned what can make data acquisition longer. Finally, the field of view is usually much smaller than in a wide-field microscope. Thanks to the fast scanning now available, the speed disadvantage is reduced and multi-color confocal microscopy is now achievable in a fairly short time-frame [78, 75].

$$\Delta r = \frac{0.37 \cdot \lambda}{NA} \tag{2.5}$$

In the same condition as before, the resolution limit using Equation 2.5 is now around 130 nm. It is much better but still diffraction limited. However, one has to recall that it is the best resolution achievable, practical problems such as making infinitesimal pinhole, signal and alignment issues will not allow to reach this limit. Indeed, even if we had infinitesimal pinhole, the alignment would become very difficult and a lot of signal would be lost. A more realistic coefficient would be close to 0.51 [75]. In a practical way, 200-250 nm is often consider as the resolution limit of a diffraction limited instrument.

2.3.3 Circumventing the diffraction limit

In order to improve the resolution scientists invented other microscopy methods based on other type of measurement or other type of sources. Few examples are Electron microscopy (EM), Atomic force microscopy (AFM), scanning tunneling microscopy (STM). Even though a technique such as EM allows to go much below the resolution limit of a light microscope thanks to the use of the electrons (as oppose to light), the electron beam and the treatments of the sample necessary (e.g. gold coating) are not compatible with bio-samples. Moreover, the treatments makes impossible the use of dynamic sample and thus EM can mostly furnish structural information and not functional ones. STM is only applicable on sample that are conducting what is therefore not interesting for bio-sample. AFM uses the interaction between the sample and a nanometric tip to get information on the sample. However, most of the mode available in AFM implies contact between the probe and the sample and therefore, risk to damage the sample. In addition, AFM has no depth penetration, it can only probe the surface of the object what is, once again, not always useful to

study bio-samples (e.g. inside of a cell).

These reasons explain why most life science measurement are still carried out on light microscope and thus need for the field to go one step further.[28] Indeed, light microscopy and high resolution combined would solve most, if not all the problems previously cited in some of the current high resolution techniques. Scientists thus started to think about ways to circumvent this diffraction law. Different super-resolution method started to emerged in the 90's. Most of them are based on switching the emitters on and off while separating their emission in time and/or space. Among others, let us mention Structured illumination microscopy (SIM), Localization microscopy (PALM and STORM) and Stimulated emission depletion (STED).[36] The next sections will be dedicated to their description. More details will be given for STED as it is the technique of interest here.

2.4 Stimulated Emission Depletion (STED)

2.4.1 Principle

Stimulated emission depletion is a method that was proposed by Stephan W. Hell in 1994 [33]. It is the most elementary RESOLFT super-resolution methods. RESOLFT stands for REversible Saturable Optical Linear Fluorescence Transition.[28] STED is a point spread function engineering technique what basically means that the increase of resolution is obtained by affecting the point spread function (PSF). The point spread function in an imaging system, is the response obtained to a point source (e.g confocal microscopy). The idea Stephan Hell presented was to use a second beam that would envelop the first one in order to inhibit the fluorescence in the outer region of the sample (=size of the PSF reduced). This second beam is different in term of shape, wavelength and intensity (see Figure 2.9). Hence, by discarding some information we can get the desired increase in resolution [75]. The inhibition of the fluorescence is achieved by putting the outer part of the sample in an off state, a state where it can not emit. The phenomenon exploited for this inhibition is called stimulated emission and was already explained in Section 2.3.1. The second beam aim therefore to bring back the molecules from their excited state to their ground state where no emission can be observed [33]. To decrease the size of the PSF, the beam has to have a zero intensity in the center and a high intensity around the center (Doughnut shape). Moreover, recalling from Section 2.3.1, stimulated emission is more efficient when using a low energy beam, this is why the STED beam is red shifted in comparison to the excitation beam. As stimulated emission is the key for the increase of resolution, confocal detection is not essential for resolution enhancement. However, STED is often assimilated with confocal-type method because it also uses a focused excitation spot.[31, 28, 42, 29, 41] Figure 2.9 shows the shape of the two beams and how the increase in resolution is obtained.

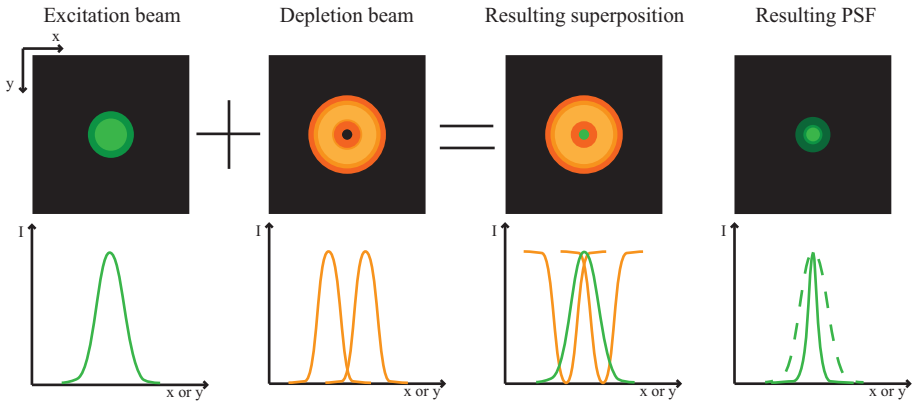

Figure 2.9: The superposition of a doughnut-shaped beam onto the excitation source allows to inhibite the fluorescence in the outer part of the sample consequently reducing the size of the PSF. [46][32]

One can directly infer from the Figure 2.9 that the resolution increase will be a function of the STED beam intensity. Indeed, it will depend on the ability of the STED beam to inhibit the fluorescence. The saturation intensity is the intensity of STED necessary to deplete 50 % of the fluorescence intensity of a fluorophore. One should always uses a STED intensity higher than the saturation intensity.[27, 28] It is however more difficult to guess how the shape of the STED beam (Doughnut) is obtained. It is achieved by the use of a beam shaping device . For instance, placing a phase plate with a coating of local differences in thickness would induce phase change without changing the intensity. Another approach would be the use of birefringent crystal, thus playing with polarisation. The most commonly used 2D shaping devices combines chromaticity of qwartz phase plate with segmentation (see Figure 2.10). This allow to place the phase plate on the common path (exc. and STED beam) while only affecting the STED beam.[68] Figure 2.10 shows how a circularly polarized beam becomes when passing through a segmented waveplate.

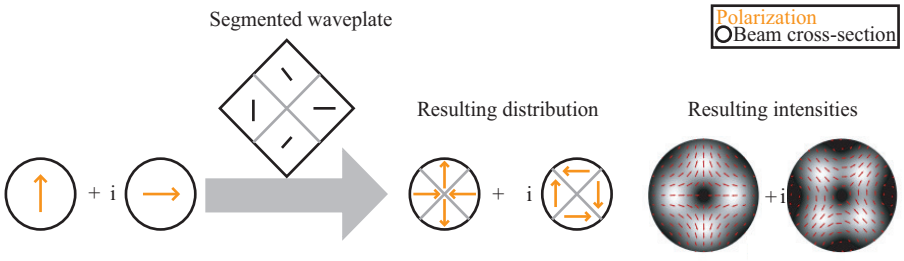

Figure 2.10: A circularly polarized beam passing through a segmented waveplate sees its polarization affected as shown above leading to the resulting intensity distribution shown on the right of the figure and thus to the desired doughnut shape.[68]

The four segments were chosen for manufacturing purposes while the orientation of the optical axis in each segment had to be chosen in order to get a null intensity in

the center and an equal depletion in all positions [68].

In the beginning of the technique, both excitation and STED beam were pulsed [42]. However, for high resolution, high power was needed and the pulse duration of the STED beam was commonly in picosecond/femtosecond. This would of course cause a high load of photon to be received in a short amount of time what can cause damages in biological sample and photobleaching in general. Moreover, this type of pulsed lasers are quite expensive what made STED primarily a technique of fundamental interest. It is with the development of cheaper continuous laser STED that it started to spread more widely[92, 58]. A typical STED setup is presented in Figure 2.11.

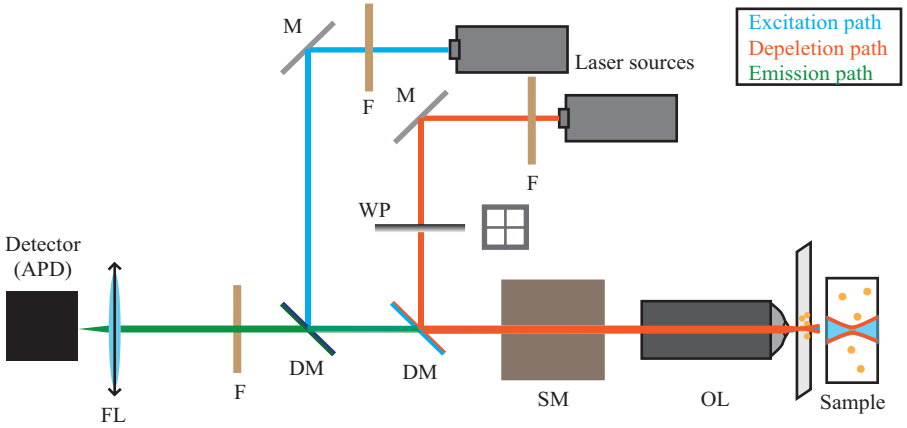

Figure 2.11: Standard STED setup : F Filter, M Mirror, WP Waveplate, DM Dichroic mirror, SM Scanning module, OL Objective lens, FL Focusing lens, F, APD Avanlanche photodiodes. The two beams are combined thank to the use of a dichroic mirror.[68]

The setup we used is an homemade setup and is presented and explained in material and methods. The lateral resolution obtained by STED can be estimated thanks to Equation 2.6 described by Stephan Hell [32, 27, 58].

$$\Delta r \approx \frac{\Delta r_{CM}}{\sqrt{1 + I/I_s}} = \frac{\lambda}{2 \cdot NA \cdot \sqrt{1 + I/I_s}} \tag{2.6}$$

Where Δr_{CM} is the resolution on the confocal image, I is the peak intensity for the STED laser and I_s is the saturation intensity for the fluorophore. Here we can understand why the saturation intensity matter so much. One should be careful, here we talk about confocal image, once again this term is used abusively, it does not always mean than the setup is confocal, it is just assimilated to confocal because of the focused source. It is also likely that i use this term in the results and discussion part, it will not mean confocal, it will just mean that the STED laser was not used. Resolution progressed quite fast in STED, it started with around 150 nm resolution to typical resolution of 35-65.[41][58] For certain organic (resp. Inorganic) fluorophores,

20nm (resp. 5.8nm) have been reported.[68] However, they probably used time-gated detection.

2.4.2 Further resolution improvement by time gating

As mention earlier, the excitation source in a STED setup is pulsed, whereas the STED beam is continuous. The use of continuous STED beam also implies that the intensity is lower and thus that a non-negligible part of the sample has not been exposed sufficiently to be inhibited. It is particularly true for the region close by to the zero of the doughnuts .[86] However, the STED is still effective in this region and thus furnish another decay path to the molecules what leads to a decrease in its lifetime. Thus the lifetime of the molecules depends on the local intensity of the STED beam. Therefore, after an excitation pulse the first photons arriving to the detector are more likely to come from the part of the sample where the STED intensity was high (outer part of the molecules). Similarly, the latest photon to arrive are more likely to come from the part of the molecule where the STED intensity was low (center of the fluorophore). By Time-correlated single photon counting (TCSPC) what is basically single photon counting while noting the arrival time of each photon, one can therefore get spacial information. Time-gating consists in discarding the photons that arrived before a certain time T_g, thus collecting only the photon coming from the center of the fluorophore. This is how one can get a further improvement in resolution. Time gating is thus a further discard of information that allows a further improvement of resolution. One usually do the time gating by post processing the data. Indeed, if TCSPC is used, fitting lifetime curve pixel by pixel on the STED affected PSF allow to have the lifetime distribution of the PSF. Therefore, by cutting off a range of lifetime inferior to a certain value, one can get the resolution enhancement expected from time gating detection. Moreover, the very center of the PSF is almost not affected (STED intensity close to zero) and thus, do not lose many photons. A contrast improvement is also often obtained thank to this technique. Figure2.12 below shows a lifetime distribution on a theoretical beads sample and the resolution enhancement obtained after applying time gating [86, 85].

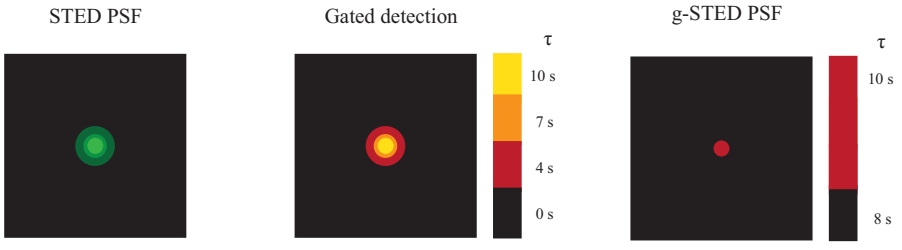

Figure 2.12: Principle of gated detection: from the time arrival of the photons, the corresponding fluorescence lifetime can be extracted. By filtering off the short lifetime (the ones affected by STED), one can obtain the further resolution enhancement presented on the figure.

The only disadvantage of this methods is that you need at first, to have a sufficient

number of photons. Even though, not so many photon are lost in the very center, there are still some losses what can be dramatic if the signal is too weak.

2.4.3 Limitation

STED can be an amazing tool with a broad range of application from material science to biology.[7][11] However, the difficulty of building and aligning in addition to its cost made it, at first, mostly of fundamental interest. Moreover, similarly to fluorescence microscopy, it is limited to fluorophore or fluorophore labeled molecules. In addition, the high intensity of the STED beam often cause photo-bleaching what makes the samples often difficult or even impossible to observe during a long time. Scientist developed buffer or other type of solution to overcome this problem [39]. We also have a limitation in the spectrum of the fluorophores we are looking at, some molecules might absorb the STED beam what would annealed the benefit we get from it. This problem can be overcome by having several sources and depletion lasers but it would be rather costy. As we already discussed, the dynamic that can be studied is limited by the scanning speed on one hand and by the amount of signal we can get from the sample on the other hand. Indeed, when the scanning speed is increased, the integration time is consequently decreased what will cause the signal to decrease and thus can limit the usable speeds. With the fast scanner now available (Galvo and resonance scanner) the limit resides more in the quality of the fluorophore used and the amount of fluorescence they can give. Unfortunately, it often limits the application to relatively slow dynamic. One could imagine to scan very fast while adding several frame on top of each other. It is often not possible because the STED beam would probably make the molecule bleach very fast. Moreover, in term of dynamic studies, having to add several frames is usually not very convenient nor precise. Other approach such as STED combined with fluorescence correlation spectroscopy were shown to be more suitable for dynamic studies.citeHonigmann2014 Finally, such a setup has to be well thought to be convenient for the users. The commercials setup are often less flexible and more expensive than homemade ones. However, a homemade setup can be difficult to align and to use for someone that is not used to it.

2.5 Localization based super-resolution microscopy

Localization super-resolution methods are wide-field techniques as opposed to STED which is a point scan method. They all rely on switching the fluorophore on/off and the accumulation of a huge number of snapshots in order to isolate the emission of the fluorophore in time and space. Two famous localization based super-resolution microscopy are PALM and STORM. The localization precision of these method can be approximated as follow: [36]

$$\Delta_{loc} \approx \frac{\Delta}{\sqrt{N}} \tag{2.7}$$

Where Δ_{loc} is the localization precision, Δ is the size of the PSF and N is the number of photons. One can directly see that if N$\rightarrow \infty$ the localization precision drop to 0. Hence, there is only a practical limit to the localization precision.

PALM stands for Photoactivated localization microscopy. It uses two laser sources, one for activation and one for excitation and photobleaching. The first source is applied in order to bring the molecule in a state in which they can get excited (usually 405 nm). It is applied in a way than only a small fraction of the sample is activated, hence, the molecules can be resolved separately. The excitation source is applied continuously in order to retrieve fluorescence from the activated molecules and then bleach them. Ideally, this cycle (Activation-excitation-bleaching) is pursued until all the molecules are imaged. The molecules isolated can then be localized by post processing the data. The signal of a molecule isolated from the others is summed up over all the frames where it appears and then fitted with a 2D Gaussian. The final super-resolution image can be created by adding a Gaussian signal to the coordinate where molecules were localized. The Gaussian width is chosen to represent the localization uncertainty. Resolution as low as 10 nm was obtained.[28] The problem with this approach is that it necessitates the accumulation of many frames (10^4-10^5) leading in quite long acquisition time (2 to 12 hours for the first publication [6]). Sample and focus drifting might thus be problematic and might affect the resolution enhancement targeted. Moreover, spontaneous blinking of dyes is a problem because it is difficult to distinguish between a different molecules that were activated/deactivated during cycles and the same molecule that is blinking. Research about dye ON/OFF state, blinking and brightness was performed in order to reduce the acquisition time needed as well as the quality of the images. images.[36][46][6][35][30][20]

STORM stands for Stochastic optical reconstruction microscopy. In STORM, emission of the fluorophores is separated in time and space using stochastic blinking of the dye. The blinking can be obtained by the switching of the dye between a fluorescent state and a dark state. One way to do this is to first turn all the dye dark with a red laser, then, turning only part of the dye on using a green laser. Another variant of STORM called dSTORM (direct-STORM), uses the spontaneous blinking of the dye to turn it ON and OFF. The experimental conditions are chosen in order to get, on each frame, most of the molecules in the OFF state and only some in the ON state. This is achieved by making the OFF state last longer than the ON state. Hence, each dye does not have any overlap with its neighbor and can be localized accurately. The localization is performed in the similar way as for PALM. Each dye is localized once per switch cycle, the dispersion of the localization point over many switch cycle will give the precision of the localization. The width of the Gaussian used to reconstruct the super-resolution image represents the dispersion in the localization Resolution as low as 20 nm can be obtained. The resolution is limited by the amount of photon one can retrieve from a single switch cycle. The number of switch cycle before bleaching is also important because the more measurement the more representative the localization will be. The drawback of this approach is that the acquisition is quite long leading to sample and focus drifting. Moreover, the bleaching can be parasitic

because it is not sure that you can observe the dye in an ON state sufficiently before it bleaches leading to some inaccuracies in the localization. [28][36][46][30][71][49]

2.6 Non-linear Structured Illumination Microscopy

Saturated Structured Illumination Microscopy is a super-resolution method which uses a structured illumination rather than an homogeneous light field. The non-linearity arises, in this case, from the saturation of the excited state. The multiplication of the local density of fluorescent dye and the local intensity in excitation generate Moire fringes by frequency mixing. The structured pattern is shifted and rotated onto the sample generating many Moire fringes which can be used to reconstruct an image by post-processing. The data processing is beyond the scope of this thesis. It is worth mentionning that Saturated Structured Illumination Microscopy (SSIM) has provided resolution as low as 50 nm.[**STEDreview**][**SIM**]

2.7 Super-resolution Optical Fluctuation Imaging (SOFI)

SOFI is a super-resolution method that exploits the temporal fluorescence fluctuation of the emitters. It does not need any additional feature on the microscope (Wide-field with CCD camera will do) and it only needs to acquire a short movie. The only conditions are that the fluorescent molecules have to have at least 2 states optically distinguishable between which emitters switch repeatedly and independently and the pixel size should be smaller than the diffraction limit. The signal on a single pixel is due to the superposition of the fluorescence of nearby emitters. The analysis in SOFI uses nth order cumulant (related to the nth order correlation) to filter the signal and let only the highly correlated fluctuations. Hence, if there is an emitter on the pixel, the fluctuation will be highly correlated leading in a strong signal in SOFI analysis. However, on a pixel nearby molecules but not containing any, the fluorescence fluctuations will be poorly correlated as they will be coming from different emitters what gives a low signal in the SOFI final image. Once again, the mathematical tools to understand this method are way beyond this thesis and are let to the reader's curiosity. The resolution enhancement scales with the square root of the cumulant order. Hence, with a 25th order cumulant, 55nm resolution was demonstrated. The advantages of SOFI are that it can be readily used for 3D measurement and the acquisition time is quite decent (few second). Moreover, it can furnish a resolution as low as 10 nm.[21][22]

2.8 Comparison of the super-resolution methods

The previous sections were dedicated to the succinct description of some of the trending super-resolution method. What one should remember about these methods

is that they all have their range of application. For instance, PALM, STORM and SSIM are relatively slow and can only be used to image fixed sample. Their resolution is limited by the sample and focus drifting. Moreover, in the case of PALM and STORM, they are also limited by the properties of the dye used. However, these limitations are not so important because scientist found way to correct for sample drifting and worked extensively on dye properties.[20][71][49]

In the case of SOFI and STED that are faster methods, one can hope to capture dynamic. Indeed, STED has already been demonstrated to achieve video rate. [91] SOFI, on the other hand only needs a few seconds to get a super-resolution images and could thus be used to capture slower dynamics.[21] These two methods would be, in principle, more useful in dynamic studies. However, speed is a non negligible argument when talking about research and thus they could easily supplant their slower fellow for static imaging. The disadvantage of STED is that the implementation of such a technique is quite heavy and need modification to a standard confocal setup. Moreover, one should chose carefully the dye that he needs to use because it should match the spectral range of the lasers used and be sufficiently resistant to high power while giving a great amount of photon.[21] The later disadvantage can be reduces by the use of buffer that enhance the stability and the brightness of the dye.[39] The scanning aspect of STED limits the size of the field of view what is not ideal for certain application. Nevertheless, STED remains faster than SOFI and post-processing free.

2.9 Aim of the thesis

The aim of the work presented here is to **evaluate the applicability of STED nanoscopy as a tool for DNA barcode mapping**. Unlike localization based super-resolution modalities, the resolution enhancement in STED is instantaneous and does not require the accumulation of large numbers of individual images or computationally intensive post-processing. Therefore, STED can potentially be used to acquire super-resolution DNA barcodes much faster than is the case when using other modalities such as e.g. STORM. A **STED system will be designed** and its ability to provide the desired resolution enhancement will be quantified. Subsequently, the feasibility of **STED based imaging of fluorescently labeled DNA will be evaluated**. Indeed, STED requires exposing the sample to be imaged to high power illumination and this might affect the DNA in undesirable ways or rapid photo bleaching of the used fluorescence labels might prevent successful observation. Standardized samples of viral DNA such as phage T7 genomic DNA, Lambda phage genomic DNA and mixture of both will be used for the purpose. The DNA samples will be enzymatically labeled in a sequence specific manner, linearized on a suitable substrate and observed using STED. Next, **barcodes will be extracted as an intensity profile along the axis of surface deposited single DNA molecules**. To allow the use of such DNA barcodes as an identifier for the presence of a certain species in a complex sample, various **mathematical approaches** will be evaluated

with the aim of **quantifying the similarity between experimentally obtained and *in silico* generated reference profiles**. Additionally, simulation of DNA barcode intensity maps will be used to to gain insight in the parameters that might affect matching performance.

Part II

Materials and Methods

Chapter 3

Materials & Methods

3.1 Introduction

A part of this chapter, together with Chapter4 aims to serve as a guide for future users of the STED setup.

3.2 STED microscope

The setup we used is presented in Figure 3.1. An oil objective from Olympus which has a numerical aperture of 1.4 and the magnification x100 is mounted on a pifoc objective piezo scanner (P-721.CLQ, PI) and a stand alone tube-lens (U-DP1XC, Olympus). The galvo scanner is a Yanus IV (Till photonics, Chromaphor). The microscope body is a IX71 from Olympus.

As excitation source, a 485 nm pulsed diode laser is employed (PicoQuant pulsed diode laser LDH-D-C-485). It is cleaned up by a 485±10 nm filter and is coupled into a single mode polarization PANDA fiber (PMC-460Si, Schäfter and Kirchoff), allowing a Gaussian beam profile to be obtained. Two mirrors are placed before to allow the beam to be correctly aligned to the optical axis of the setup. An optical density filter allows to regulate the power arriving to the sample. After the fiber, the beam is collimated by a collimating lens (Linos Qioptica : G052006000, D=10mm, F=20mm) and goes through 2 polarizers ($\lambda/2$ and $\lambda/4$) before being combined with the STED laser on the Dichroic mirror (DM1). The STED solid state depletion laser is a 592 nm continuous wave laser (Genesis MX-590-500 STM, Coherent). It is also cleaned up by a 589 ±10 nm filter before being coupled-in. After coupling through the fiber, polarizers will influence the beam polarization before being combined on DM1. The role of the polarizers is to ensure circular polarization that is a key for doughnut generation by the waveplate later on the optical path. The use of optical fibers to couple lasers into the downstream system allows for easier exchange of lasers should this need arise.

On the common path, the beams travel through a pellicle mirror (equivalent to a beamsplitter 50-50), the pellicle is only there for alignment purpose and can therefore

be removed during the measurements. Then, the beams are both reflected on a second dichroic mirror (DM2). Before being sent into the galvo scanner and the microscope. The doughnut is created just before the objective to ensure that a minimal number of optical element are placed after its formation. This is to ensure that it does not get deformed or changed before going to the sample. The light emitted by the sample can be split into two different paths, the reflection/scattering path and the emission path.

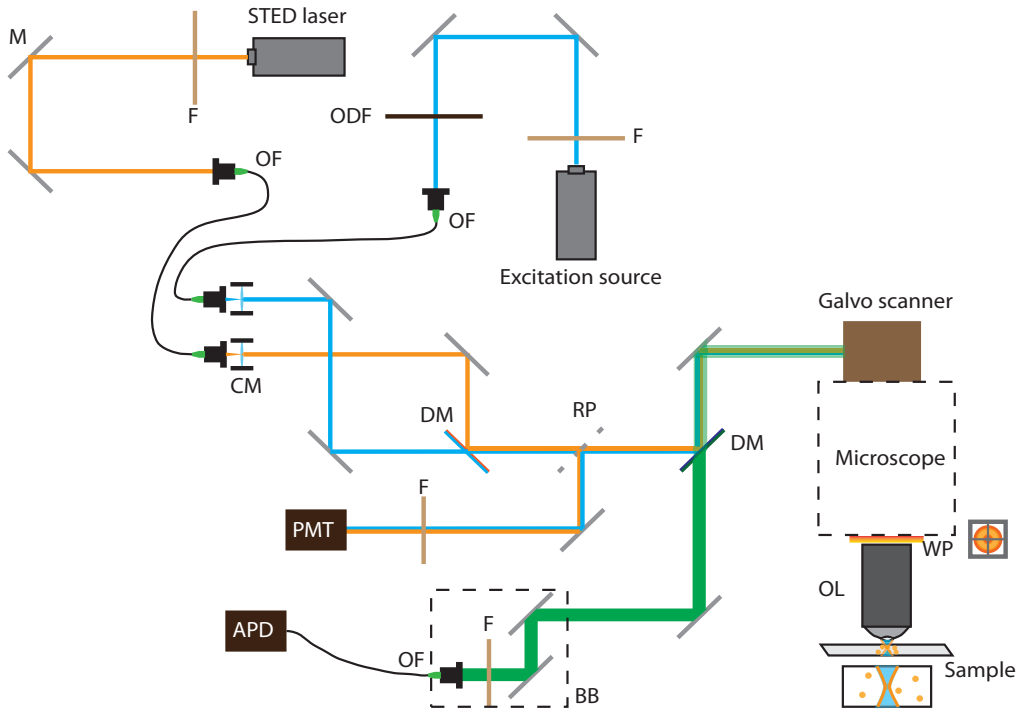

Figure 3.1: Scheme of our STED setup. F: Filter, M: Mirror, OF: Optical fiber, ODF: Optical density filter, CM: Collimating lens, DM: Dichroic Mirror, RP: Removable pellicle mirror, WP: Waveplate, OL: Objective lens, BB: Light-proof enclosure , PMT: Photo-multiplier detector, APD: Avalanche photon detector

The reflection/scattering path is reflected on DM2, then, partially reflected by RP and sent to the photo multiplier tube (PMT). Filters are placed in front to decrease the amount of signal and to remove a part of the noise. This path is only used for the Gold bead sample that is our reference sample for alignment (detailed later). The emission path passes through DM2 and is thus sent to the avalanche photon detector (APD), this path is the one used for the measurement. Once again, filters are used to remove the stray and reflected light.

All the mirrors, polarizers, pellicle were bought from Thorlabs. A brief description

of the electronics used in our setup is presented below
A Pico-quant PDL 800-B box controls the excitation source and its pulse-rate. It also sends information about the rate to the hydra-harp detection system and to the national instrument box 1 for the triggers.

The Coherent power supply (Genesis MX-590-500 STM) controls the power output of the STED laser, with a maximum power of 575 mW.

Two national instrument boxes (USB 6343 X-series, multi function DAQ) control respectively the galvo-scanner (x-y position) and the piezo-scanner (z position). It is necessary to have two of these boxes to be able to do these movements independently and thus simultaneously.

The hydra harp box from PicoQuant (Multi-channel Picosecond event timer) receives the signal from the detector and records the time arrival of the photons. Everything is synchronized by the pulse rate of the excitation laser. Each photon can therefore be assigned to a certain pulse. Moreover, the pulse rate is chosen so the time between to pulse match approximately the fluorescence lifetime of the fluorophore.

The scanner is controlled by a Till photonics scanner power unit and a scanner control unit.

Everything is controlled by a computer which runs a home-made software written by Dr. Janssen. This software was coded on Visual studio 2015 and allows the user to set different acquisition parameters such as the size of the image in pixel , in nm, the exposure time (msec/pixel) and the scanning mode. There are two scanning modes: bi-directional and uni-directional. Finally, it allows the user to move onto the sample and to change the focus. This is done by moving the piezo-scanner. This feature explain the need for 2 national instrument boxes.

3.3 Setup characterization

3.3.1 Laser intensity measurement

The measurement of the output power of the excitation source and the STED laser were measure. For this purpose, a Thorlabs power-meter was used and the power was measured at the sample holder after removal of the objective.

For excitation, the power output was measured for different optical densities ranging from 0 to 4.6. The power output of STED was measured for a range from 0 to 575 mW indicated on the monitor. The power output before and after the fiber coupling was also measured in order to calculate the in-coupling efficiency.

3.3.2 Pixel size determination

To determine pixel sizes a resolution chart was employed. A resolution chart is a glass piece where lines are drawn by means of litography (Thorlabs, R2L2S1P1 Positive, High-Frequency USAF resolution chart). It contains different series of lines, the width of the line and the width of the space between two lines are equal, a single line and a single space compose what is called a cycle. The smallest grid on the resolution chart (228 line/mm = 2.19 µm per line or space) was measured for different scan settings (pixel size, exposure time, scanning mode,...). This was performed to verify that the pixel size set in the program correspond to the pixel size in the final image and that it scales correctly when changing the parameters.

500 × 500 and 1000 × 1000 images were acquired with a set field of view of 10 × 10 µm and 20 × 20 µm. These measurement were performed for both uni-directional and bi-directional scanning mode and for two different exposure time (0.05 and 0.1 ms/pixel).

3.3.3 Resolution assessment

To assess the resolution reachable with our setup, 20 nm fluorescent bead samples were employed. The 20 nm fluorescent beads were provided by Dr. Wouter Sempels as stock solution 2 wt%/v in water. The bead samples were prepared by diluting this stock solution 10000 times before dropcasting it onto a coverslide and let to dry for 25 min in air. All the coverslides employed were Menzel-glaser 22x22 mm and were heat-treated for one day before use in a Nabertherm "More than heat" oven (LE 4/11). Thermoscientific pipette were used with Multi-guard barrier top from Sorenson. The MilliQ water was obtained via a synergy UV millipore purification system.

The measurement were performed using 500 × 500 pixels and 2000 × 2000 nm field of view. The power used was set to a value comprised between 300 and 500 mW on the monitor. The resolution was measured by comparing the intensity profile of the beads on the confocal image, on the STED image and on the g-STED image in both x and y directions. To assess the resolution from the full-width half maximum of the intensity profile, Houston criterion was applied.

3.3.4 Saturation curve

Saturation curves were usually measured before starting an experiment. The wave-plate was first removed what makes the STED beam have a Gaussian shape. Then, the alignment was performed but in this case we overlay two Gaussian beams. Images were then acquired according to what is explained below. For each STED intensity, one measurement is performed with only excitation on, followed by a second measurement with both laser on. Then, by calculating the ratio of the photons obtained,

one can extract the amount of fluorescence that was depleted.

For the bead samples, 500x500 pixels/2000x2000 nm frame were acquired with a time per pixel (TPP) of 0.1 ms. The experiment was performed for 3 different OD filters in front of the excitation source (2, 1.5 and 1 resp. 0.25, 0.73 and 2.5 μW). A range of STED intensity from 0 to 300 mW on the monitor (0-90 mW before the objective). For this experiment the excitation source was a PicoQuant LDH-P-C 470 class III laser with a maximum average power of 5 mW. This source was later changed to the one described in the setup which allows the use of higher power output.

For Atto488, the depletion was measured directly on labeled DNA. 500x500 pixels/5000x5000 nm frame were acquired with a TPP of 0.05 ms. The optical density was set to 2 (13 μW) and a range of STED intensity between 0 and 400 mW (0 to 150 mW before the objective) was used.

3.4 DNA simulation

3.4.1 Programming

For simulations a homemade Matlab program written by Dr. Janssen was employed. The first step was to get information about the targeted genome (T7 and Lambda DNA). Matlab has a toolbox called bioinformatics toolbox which is particularly suitable for such case. Indeed, it allows to go to NCBI website and get the reference for the genome sequence of our choice (in our case T7 and lambda). A function called "getgenbank" allows to retrieve the full sequence of the genome when entering its reference found on NCBI. In addition, a function called "rebasecut" which takes as an input a genome sequence and an enzyme and gives as an output the places on the genome were the restriction enzyme will actually act. As methyltransferase is used here, the place where the corresponding restriction enzyme acts is the place where the label is supposed to be placed. Hence, a perfect map is obtained with the exact position of the label for both T7 and λ-phage DNA.

However, it is know that the data will never be perfect for many reasons and the simulation code also had to account for that. The next step was thus to simulate the different problems that can occur during the sample preparation (fragmentation of the DNA, labeling mistakes, labeling efficiency). The only part that was not simulated is the stretching inhomogeneities. The reason why it was not simulated is that it would have been quite time consuming and also, as no proper model that describe this effect was known, the simulation would have hardly been realistic.

From the sequences obtained in step 1, the program generates random fragments, the size of the fragment and the width of the distribution can be set by the user. Then, for the fragments generated, some labeled positions will randomly be deleted off (80% of the position are kept by default but it can be set to other values). Finally, the

program will add a random numbers of false positive (labeling mistakes) according to Poisson distribution.[19] The rate set per default is 0.1 false positive per kilobase.

The result of this is a matlab "structure" that contains information about the size of the full genome, the length, the starting position and the positions of the labels of the fragments created.

The second part of the program was written by myself and is dedicated to the simulation of the intensity profile from the simulation obtained in first place. First a perfect one dimensional Gaussian is created using resolution input by the user to assess the width of the Gaussian. Then, the perfect Gaussian is sampled (sampling according to a Gaussian distribution) a certain number of time to simulate the fact that we have a limited number of photons coming to the detector (according to Dr. Camacho procedure). Finally, the Gaussian is added to the position where the label are supposed to be which were obtained in the first part of the program.

Now that the program is able to simulate intensity profiles, it can be used to simulate data in order to get information out of them. To be able to extract information out of the simulated intensity profile, comparison with a reference is needed. Hence, another program dedicated to this was written by myself (described below in Subsection 3.4.2).

In order to be able to write and test the algorithm of the program, a 20 kb fragment of T7 DNA was simulated using the default setting of the simulation tool. Intensity profiles for different resolution ranging from 15 nm to 350 nm were compared with T7 and λ-phage DNA reference with the corresponding resolution. The matching scores obtained for different algorithms were then compared (more details in Subsection3.4.2).

Once the more performing algorithm was known, T7 fragment of 5, 10 and 20 kb were simulated using the default settings. Then, intensity profiles were simulated an compared in the same way as described above. Hence, better knowledge the dependency on the size and the resolution on the matching score obtained would be obtained. The final experiment to test the algorithm was to input T7 and Lambda reference on E.Coli and to use the resulting genome as a reference to try to match a 5 kb fragment of T7. This was performed for 250 and 60 nm resolution.

To learn more about the program and the parameters of the system that affect the matching score, larger scale simulations were performed. 500 fragment of T7 and 500 fragment of Lambda were simulated for labeling efficiencies ranging from 10 to 100%. The size was fixed to 20 kb and the false positive rate was kept equal for all simulations. Similar simulations were performed for a fixed labeling efficiency (70%) and a fixed number of false positive ranging from 0 to 5 to investigate the influence of the false positives on the matching score. Finally, similar simulations were performed for 100 % labeling efficiency and no false positive to investigate the influence of the position of the fragment on the DNA on the matching score.

To verify that the program would be able to assess the stretching factor without any bias, simulation were performed onto 100 fragment with a random stretching factor between 1.4 and 1.8. This experiment was performed for two distinct labeling efficiencies (80% and 50%).

3.4.2 Data Analysis

Data analysis was achieved using a homemade program written by myself. It compares one dimensional intensity profile of a sample (simulated or not) to different genome references to assess which one matches best. The way it works is that the fragment intensity profile is compared to a windowed reference that has the same size. The window is then shifted step by step and a matching score is calculated for each reference windows. To calculate this matching score, different algorithms were tested: curve division, overlap integral, cross-correlation and least-square solution. They were compared in their performances to match simulated fragment to their reference. Least-square methods was the one that performed the best and thus, it was employed for the rest of the analysis.

3.5 DNA imaging

3.5.1 Zeonex coating of the coverslides

Zeonex coating is required in order to render the glass surface originally hydrophilic, hydrophobic. This is necessary to achieve the stretching via the rolling droplet method. In addition, Zeonex was shown to bind DNA more efficiently than other polymers allowing for better stretching. [19].

All the Zeonex solutions were prepared according to the procedure described in [19]. 1.5 g of Zeonex (Zeon chemical company) were weighed on a Mettler Toledo ML204 and dissolved in 100 ml of Toluene (Sigma Aldrich, analytical grade >99.5% pure) in order to obtain a concentration of 1.5 wt%/v. The solution was finally sonicated for an hour on a 2510 Branson sonicator (Bransonic ultrasonic cleaner). The solution was then stored in the fridge.

The Spin-coating was performed on oven treated 22 × 22 Menzel-glaser coverslides using a Laurell Technologies spin coater (model :WS-650Mz-23NPPB). The quality and the thickness of the film was believed to affect the STED doughnut. Different spinning rate were used ranging from 2000 rpm to 10000rpm to investigate the effect of the thickness of the film on the quality of the STED imaging. A few droplets (3-4) were deposited before starting the spinning in all the cases and the sample was always spun for 1 minute.

3.5.2 DNA labeling

For all the labeling, the same procedure was used. This procedure was taught to me by Dr. Su who participated in its development. It comports two main steps : the first one is the linking of the Atto488-NHS dye to the MTC6-cofactor in PBS buffer, the second one is the MTaqI-assited labeling of the DNA with the cofactor labeled in step 1. A cutsmart buffer is used to end the reaction in the first step and a proteinase K is added to end the reaction in the second step by digesting the MTaqI. T7, λ-phage DNA and a mixture (1:2) of those two species were prepared to be imaged later on.

The MTC6-cofactor was synthesized by Dr. Van Snick, the mTaqI enzyme was prepared by Dr. Wang and the Atto488-NHS-ester dye was provided by ATTO-TEC. phosphate buffered saline (PBS) buffer was made from Sigma-Aldrich packet (one pouch prepares one liter of 0.01 M PBS, PH=7.4). The concentration was chosen to be 20 times higher than indicated on the package aiming for the concentration to be 0.01 in the final DNA solution to ensure a pH of 7.4. The DNA labeling was carried out in 0.5 mL Eppendorf DNA lobind tubes and centrifugation step were performed on a Microcentrifuge 5418 of the same company. The cutsmart buffer was composed of 500 mM potassium acetate, 100 mM magnesium acetate, 200 mM Tris-acetate and 1 mg/mL of Bovine serum albumin (BSA). These compounds were provided by Sigma-Alrich. Incubation was performed on an Eppendorf Thermomixer Comfort. The Proteinase K was provided by New England Biolab (800 unit/mL, molecular biology grade). Pipettes and milliQ water used were from the same source as described in Subsection 3.3.3

For the first step, all the components were first placed in an ice box to preserve them. Then, in a 1.5 mL DNA Lobind tube also placed in the ice-box, the components were added in the following order : 3.2 µl of MTC6 cofactor (C= 1.5 mM) with 0.6 µl of Atto488 dye (C=20 mM) , 1 µl 20×PBS buffer and 3.2 µl of milliQ water. Between each addition, the sample was centrifuged a few seconds to ensure a good mixing of the different component. The solution was then let in the ice-box for 25 minutes. Finally, 1µl of Cut smart buffer was added to stop the reaction, the solution is ready for the next step 1 min later.

For the second step, an incubator was first set to 60°C and proteinase K was incubated at 50°C. Then, in the same tube that was utilized for the first step the components were added in the following order : 4 µl of mTaqI enzyme (C=1.6 mg/mL), 4 µl of milliQ water and 2 µl of DNA (500ng/µl). The DNA being either T7, λ-phage DNA or a mixture of them. The solution is then let to incubate for 20 minutes at 60°C. Between each additions, the solution was centrifuged for a few second. However, after addition of DNA, it was not done anymore to avoid the shearing of DNA. Moreover, for the same reason, the sharp part of the tip was cut each time a solution containing DNA was handled. Finally, 1 µl proteinase K was added to the mixture to stop the reaction and the final solution was incubated for

30 min.

3.5.3 DNA purification

The DNA is now labeled, the final volume is 20 µl, the next step is the purification. The solution contains free protein/enzyme/cofactor and even free labels that are not suited for the imaging later.

The purification was performed with a "Genomic DNA clean and concentrator-10" kit from Zymo-research according to what is advised by the company. This kits contains spin columns that are composed of one tube with a filter at the bottom that is attached to a collection tube. It also contain a binding buffer that specifically attaches DNA to the filter, a washing buffer that is used to wash over the other components that may have been attached with the DNA and an elution buffer that is employed to recuperate the cleaned DNA from the filter of the spin column.

Twice the volume of the DNA solution of binding buffer was added to the solution (40 µl) before being transferred to the spin column. The solution was then centrifuged at full speed for 30 sec. After that, the solution was washed with 200 µl washing buffer and centrifuged at full speed for another 30 sec (16000 rpm), the washing step was performed twice in total. At this stage, the DNA is still attached to the filter and the tube containing the washing solution can be discarded. A new 1.5ml Ependorf LoBind tube was attached below the filter. Finally, 10 µl of elution buffer were added onto the filter and the tube were centrifuged full speed to get the DNA back in solution. The elution step was repeated once more in order to collect most of the DNA. The whole purification process is repeated once again to ensure a "spectroscopicly" clean DNA. The final concentration of the solution is checked with a Biodrop µlite and is usually below 50 ng/µl. Indeed, the initial stock DNA is 500 ng/µl, is diluted around 10 times in the labeling process and a part of the DNA is usually lost in the purification step. The solution is then stored in the freezer at -20 °C and is ready for further use.

3.5.4 DNA Stretching

Prior to DNA stretching, the solution had to be diluted about 10 times and 2-(N-morpholino)ethanesulfonic acid (2-(N-morpholino)ethanesulfonic acid (MES) buffer) was added to it. The dilution is needed to get a good stretching and to ensure that single molecules can be distinguish on the coverslides. The MES buffer controls the pH allowing an easier rolling of the droplet and hence, a better stretching.

MES buffer was prepared with 500 mM potassium acetate, 100 mM magnesium acetate, 200 mM MES and 1mg/mL BSA, the pH was adjusted to pH=5.7. All the compounds were provided by Sigma-Aldrich.

The dilution was performed using the elution buffer from the Zymo research kit in order to get a concentration comprised between 4 and 5 ng/µl. 2 µl of DNA solution were diluted with the appropriate amount of elution buffer, the DNA concentration was then verified and adjusted if needed. Then, 1 to 2µl of MES buffer were added in order to get 5-10% in volume of the buffer in the final solution. 2 µl of the resulting solution were dropcasted on a zeonex coated slide near border. Then, the motorized tip is placed just above the droplet according to the procedure described in [19]. The stretching was always performed from up to down at a relatively slow speed (about 4 mm/min). A line from a border of the slide to the opposite border is thus obtained. Finally, the slides is dried under vacuum for at least one hour. The drier was covered with aluminum foil to avoid photo-bleaching of the sample.

3.6 Setup alignment

The alignment has to be performed daily before the start of an experiment. Moreover, in case of long duration experiment, it is always preferable to check the alignment again after few hours. The alignment consists of several step : adjusting the excitation beam through the objective, alignment of the detection paths, visual overlay of the beam, centering of the waveplate with the STED beam and check the overlay with a gold bead standard sample.

Adjustments of the excitation beam was performed using the two mirrors that send the beam to the galvo scanner. While moving in and out of focus, the beam should be nicely round and the intensity equally distributed. The scattering path was aligned using the reflection from a coverslide optimizing the signal the detector receives. The APD detection path was aligned by dropcasting a standardized fluoresceine dye solution (0.1 nM) onto a coverslide and optimizing the signal received by the APD. The overlay of the two beams was realized by removing the objective and looking at the setup-roof. Turning the lasers on and off alternatively allows to see if they are correctly overlaid or not. If it was not the case, DM1 was used to realign the STED beam. Indeed, the excitation beam was aligned in first position so it should not be moved afterwards. The centering of the waveplate was also performed by looking at the setup-roof, a cross can be seen (due to the waveplate) in the beams, this cross should be placed in the center of the STED beam (as accurately as possible). Finally, the overlay of the beam was performed by imaging a gold bead sample using the scattering of the beams on the sample. The imaging was performed on a relatively small field of view (500 × 500 nm) and a high scanning speed (0.02-0.05 ms/pixel) to ensure a high accuracy. By turning the beam alternatively on/off and by marking the center of the excitation PSF, one can align the STED so the marking ends up in the center of the doughnut. This alignment ensures a good overlay in the x-y plane. However, a good overlay in the z direction is also needed. *Figure 3.2* shows how the overlay is performed in the different planes.

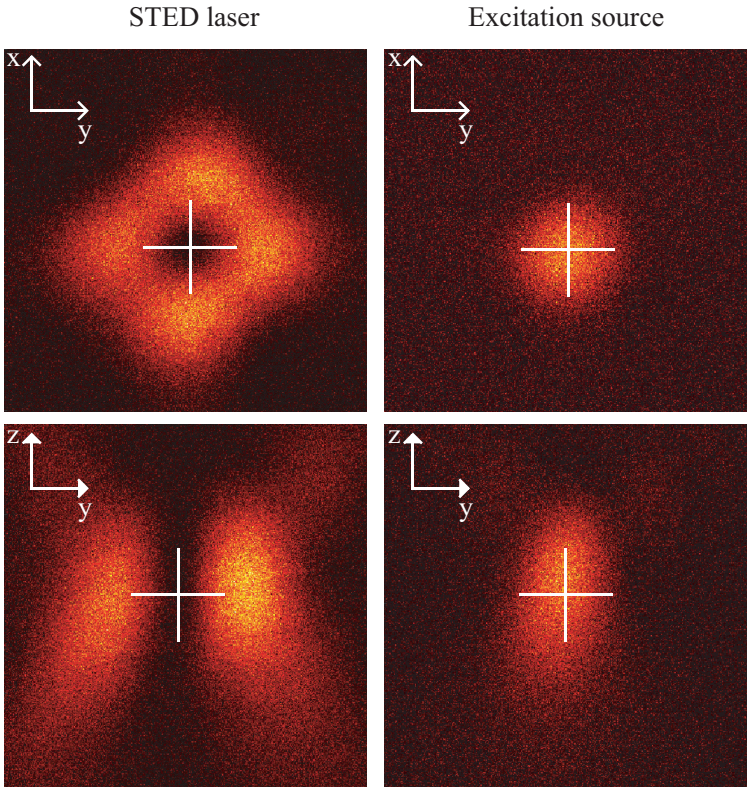

Figure 3.2: Overlay of STED laser and the excitation source. The upper part represents how the overlay is performed in the x-y plane. The cross is first placed in the center of the excitation PSF created by scattering on gold bead, then the excitation is turned off and the STED on. The STED is then moved to match the cross to its center. By successive repetition of these step, good overlay can be ensured. The lower part represent the same alignment performed in the z-y plane, similar images can be obtained for the z-x plane.

A similar image can be obtained for the z-x plane but is not shown here. The fiber in-coupling also need to be verified but not as regularly as the rest. Indeed, it is relatively stable and a slight loss of power will not have a significant effect on the setup performance. Nevertheless, it is important to check it from time to time to ensure optimal performances.

3.6.1 DNA imaging

DNA imaging was an issue in the first place when using the STED laser. The first issue was the photobleaching. To fix this problem, an imaging buffer was used. This buffer is based on glucose Oxidase and allow to stabilize the DNA while avoiding oxygen to enter in contact with the sample. A dilution buffer is first prepared : 50mM Tris-HCl (>99.8 % pure) buffer, 10 wt% D-(+)-glucose (99.5 % pure), 10 mM

NaCl (99.84 % pure). These compounds were provided by Sigma-Aldrich. The final imaging buffer is prepared by adding 5 µl of glucose oxidase in 495 µl of Dilution buffer in order to get 1 to 2% of glucose oxidase in the final solution.
DNA observation was performed setting a large field of view (20 × 20 µm) with a small pixel size 1000 × 1000 pixels images 20 nm/pixel . The time per pixel was 0.05 or 0.1 ms/pixel depending on the amount of signal obtained from the sample. This was also chosen in order to avoid setting the acquisition time too long as we believe STED's main advantage is its speed. Using such settings the acquisition time is typically 2 minutes per frame for unidirectional scanning and 1 minute per frame for bi-directional scanning. Once the problem of STED and zeonex was solved, images were acquired for T7, Lambda and a mixture of both DNA using the same settings.

3.6.2 Data analysis

For each DNA strand on an image, 3 line profiles were extracted with a distances of 0.5 pixels between them. The three profiles were multiplied to reduce the uncorrelated noise, then the 3rd root of the resulting profile was taken.

Different analyses were then performed on the data. Depending on the quality of the intensity profile extracted (in term of background and noise), different approaches could be used. The least-square methods offers to add a second shape for the fitting that can be set as a straight line (e.g. y=1). This line allows to account for the background and will be referred to as the background line. Data can also be pre-processed by the program, a pre-process that will be refer to as denoising. The denoising is performed by smoothing the experimental data and calculating the residual between the smoothing and the data. Then, everything that is smaller than the residual is set to 0. In doing so, we hope to do kill the noize while keeping the information.

On T7 and λ-phage DNA 4 analysis were performed. With or without denoising, each possibility with or without the background line. For the mixture, the denoising and the background line were used.

3.7 Health, safety and environment

In any laboratory work, health, safety and environment are topics of primer importance. A particular care was therefore addressed to them throughout the work performed for in this thesis.
The work in a spectroscopy lab using homemade setups implies the need of an awareness about the risk that one is laying itself open to. In the case of this thesis, the use of a class 4 laser (STED laser) is probably the most risky work that was performed. OD=4 goggles at 590 nm were worn during experiments. In addition, appropriate covering of the open laser sources was used to avoid the light going anywhere but the location where it was needed, automated shutters and warning light at the entrance were also used for a maximal safety.

All the chemicals used during the sample preparation were handled wearing a lab coat, safety goggles and gloves. Most of the chemicals used in the daily work were relatively safe, the list below presents the chemical that were not and thus for which special care was addressed (e.g. work under fumehood, appropriate waste disposal).

- **Toluene:** Flammable, can cause skin/eye irritation, toxic when inhaled or swallowed, carcinogen
- **Ethanol absolute:** Flammable, can cause skin/eye irritation, toxic when inhaled or swallowed
- **Binding buffer:** Contain guanidinium chloride which can cause skin/eye irritation and is toxic when swallowed

All the other chemicals are classified as non hazardous and not dangerous. Nevertheless, lab coat, gloves and goggles were worn to ensure an optimal safety in any condition. The sample containing glass were thrown in the glass waste bin and the vials containing biological waste (Enzyme, DNA, protein,...) were thrown in the contaminated solid waste.

Part III

Results and Discussion

Chapter 4

STED setup performance evaluation

4.1 Introduction

Prior to do any measurement on "real" samples, it was crucial to assess the performances of the setup. In addition, it allowed me to get used to running the setup. Different characterization were performed using Fluorescent bead and gold bead standardized samples. The setup was characterized in term of the power output of its laser sources, the size of the field of view, the shape of the doughnut it can generate and the resolution it can reached. Additionally, its ability to deplete fluorescent beads was also tested. Finally, the correspondence between the pixel size input by the user and the pixel size on the ultimate image was verified. This chapter together with Chapter 3 aims to be used as a guide for future users of the STED setup.

4.2 Evaluation of laser intensity

The power output available of the laser sources of our setup was assessed by measuring the intensity right after the waveplate for different power set (more details in Chapter 3). The results of the power output measurement on the laser sources are presented in Figure 4.1.

Knowledge of those curves will allow to retrieve the power output of the lasers by a simple measurement of the in-coupling efficiency at the fiber followed by short calculation. The huge difference between the in-coupling efficiency of the lasers is explained by the quality of the beams prior to entering the fiber. The STED beam provides a much nicer Gaussian shape that the excitation source accounting for this difference.

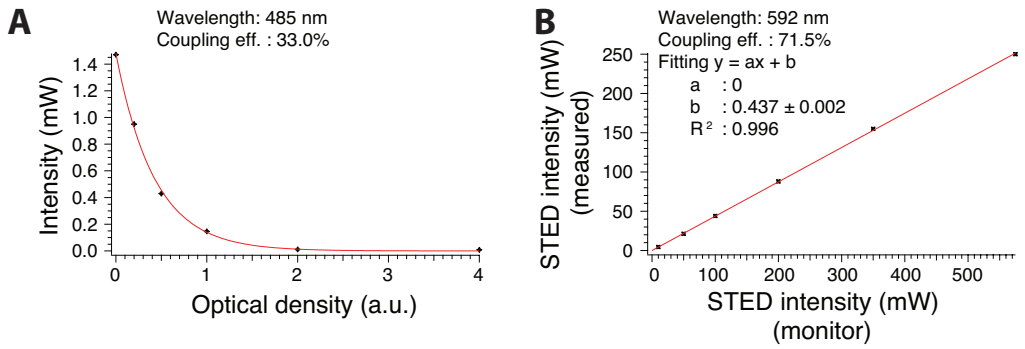

Figure 4.1: A. Measurement of the power output of the excitation source for different optical density filter placed on its optical path. The exponential behavior observed is a consequence of Beer-Lambert law. B. Comparison between the STED power set on the monitor and the STED power measured right after the waveplate.

4.3 Pixel size

The pixel size of the setup was checked by imaging a resolution chart reference sample. Figure 4.2 below shows the lines of the resolution chart and an intensity profile extracted from them. The settings employed for this image were: 500 x 500 pixels and 10 µm x 10 µm field of view. The expected size of a pixel is therefore 20 nm.

The whole intensity profile is 320 pixels long. It corresponds to 4 cycles on the resolution chart, thus, 17.6 µm. The pixel size is therefore 55 nm what is 2.75 times bigger than we thought. This factor will have to be taken into account in the future experiment. Over the different setting tried, the scaling of the pixel size remained consistent.

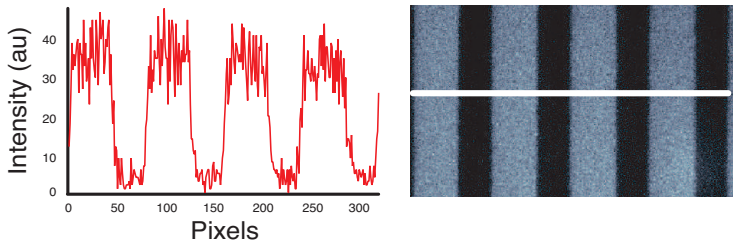

Figure 4.2: Intensity profile extracted from a resolution chart image (up), image of the resolution chart, 4 cycles can be observed (down). Each bright or dark line is 2.19 µm

4.4 Doughnut

Due to some issues with the polarizers they were removed for a certain time, the shape of the corresponding doughnut is shown on the left in Figure 4.3. However, by putting the polarizers back and tuning them carefully while imaging the goldbeads i could get the much better shape that is shown on the left in Figure 4.3.

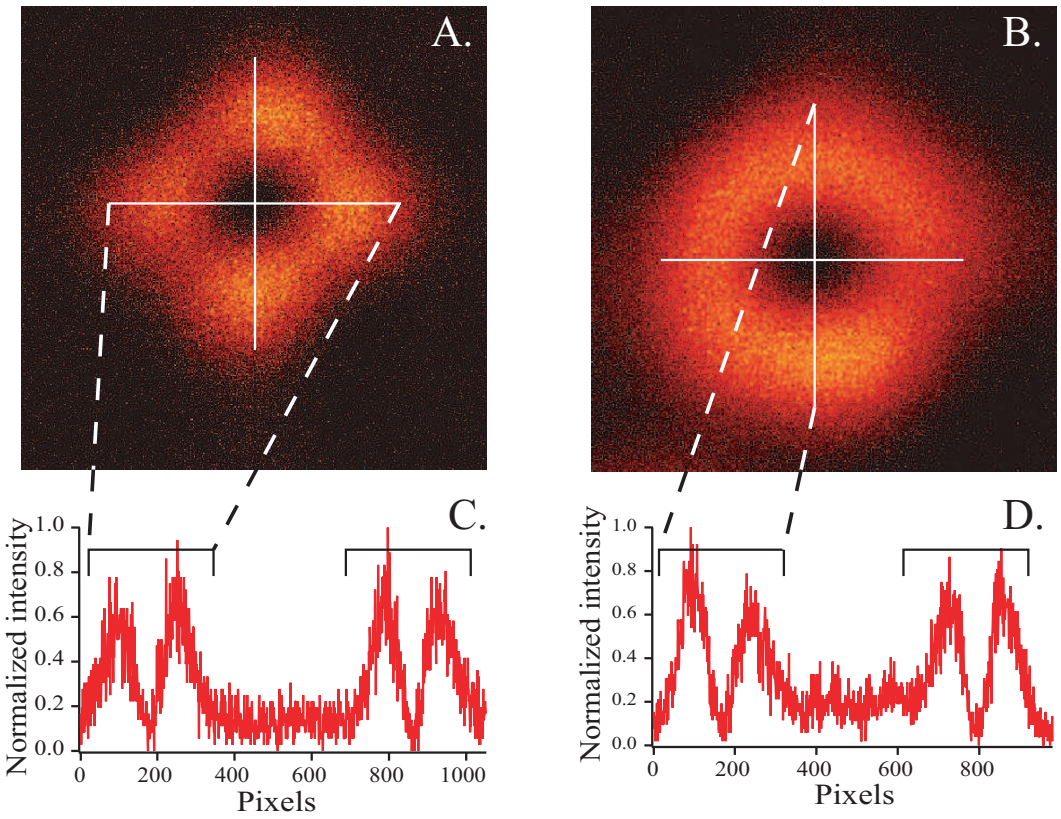

Figure 4.3: A. Unpolarized STED beam doughnut B. Circularly polarized STED beam doughnut C. Intensity profiles in x and y direction for unpolarized doughnut D. Intensity profiles in x and y for polarized doughnut The intensity is much more uniform for the polarized doughnut whereas the doughnut minimum is better for unpolarized doughnut

The two shapes work fine for STED imaging. However, the roundish doughnut has a more uniform intensity distribution than the other one. It can be rather difficult to obtain the proper intensity distribution without the polarizers. Intensity distribution is an issue because in some cases it can leads to differences in resolution along y and x and thus a distorted image. On the other hand, the null of the doughnut tended to be better without the polarizers and is more difficult to improve with the polarized doughnut. The quality of the null is also a key in the resolution enhancement provided by STED. On a daily basis the polarization should be tweaked

in order to get the best compromise between null and intensity distribution in order to get the best performances. The size of the center depends mainly on the intensity and can therefore not be discussed here.

4.5 Resolution

The resolution achievable was tested onto a standard sample (23 nm fluobeads). Figure 4.4 shows the confocal, the STED and the gated-STED image of a beads as well as their respective intensity profile.

One can see that the best achievable resolution is around 68 nm using time-gating. However, this experiment was carried out on a beads sample with very high STED power, it is likely to be the best resolution we can achieve on the setup. By fitting the intensity profile by a gaussian, one can obtained the superposition in Figure 4.5. This figure emphasizes the resolution improvement offered by both STED and g-STED.

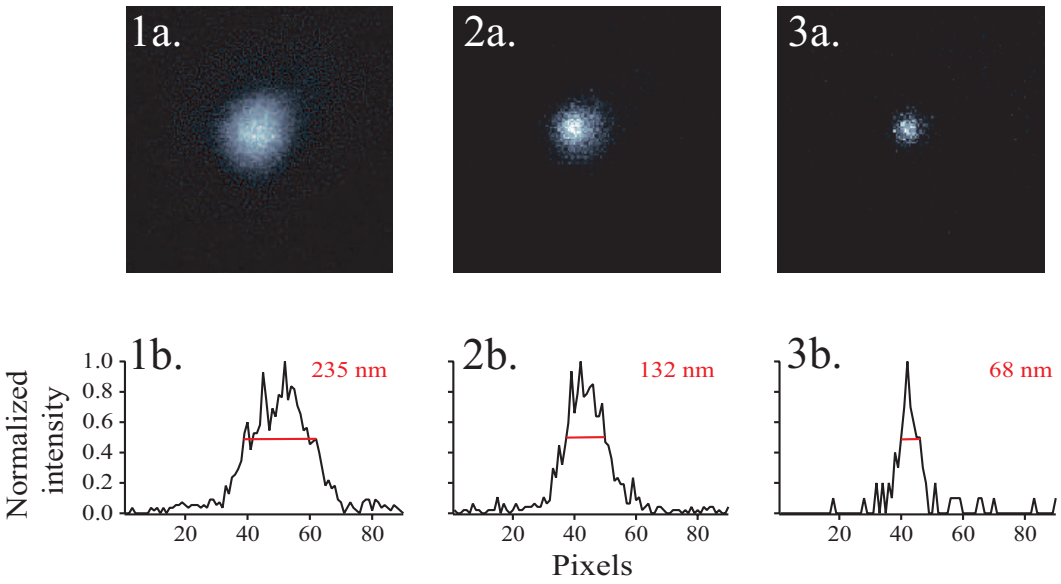

Figure 4.4: Achievable resolution for confocal, STED and g-STED imaging. 1a. Confocal image of a fluorescent bead 1b. Resulting intensity profile 2a. STED image of the same bead 2b. Resulting intensity profile 3a. g-STED image of the same bead 3b. Resulting intensity profile. STED allows for a resolution enhancement of a factor 2 compared to confocal while g-STED allows for a further resolution enhancement (factor 4 compared to confocal)

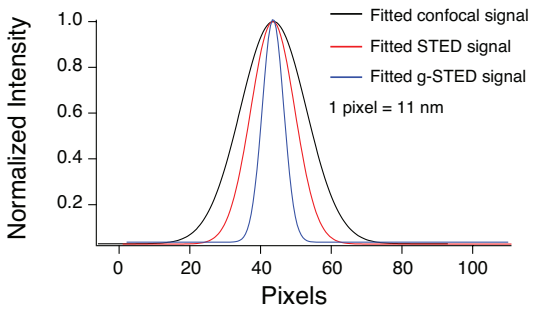

Figure 4.5: Comparison of the Gaussian fits of the intensity profiles obtained from confocal, STED and g-STED image of a fluorescent bead. This picture illustrates more clearly the resolution enhancement provided by STED and g-STED respectively.

4.6 Field of view

Knowing that we are using a scanning method, it was important to know the size of the field of view that we can observe. The maximum size we can ever image is about 165 x 165 µm taking into account the factor determined earlier. This is the limitation of our scanning system but this is more than sufficient for single molecule applications. However, it does not tell us whether the doughnut behave equally well in every places on the image when scanning large areas. The scanning module is placed in a position which limits the movement of the laser on the waveplate but, we need to make sure that while the scanner moves, the cross of the waveplate is still in the center of the doughnut. By imaging the doughnut on gold beads at different positions, the range where it remains efficient could be identify. Figure 4.6 shows the shape of the doughnut for different distances away from the center of the scanning system.
Hence, if one scan a squared area with the central position in the middle, the biggest area one can scan in efficient depletion conditions is 55 x 55 µm (thus 20 x 20 µm in the imaging parameters).

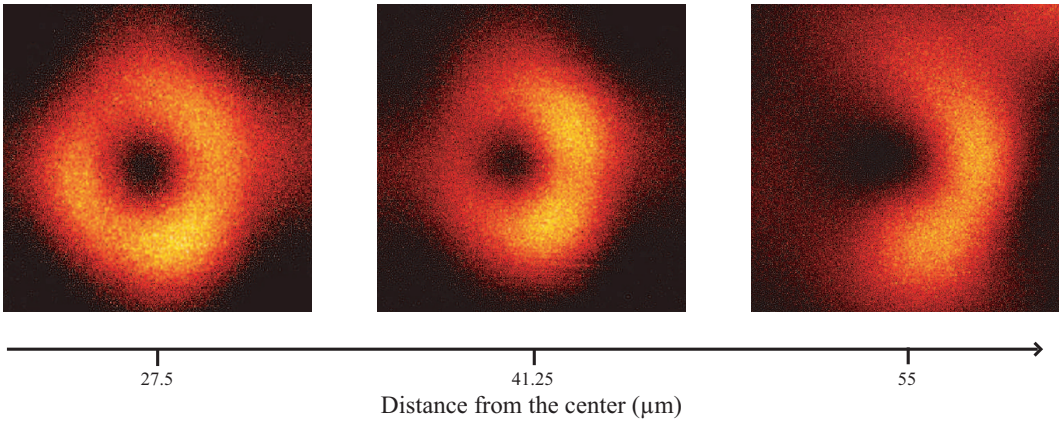

Figure 4.6: Comparison of the quality of the doughnut shape for different distances away from the center of the scanning system. The doughnut's shape is distorted for distances larger than 27.5 µm and unusable for distances greater than 41.25 µm

4.7 Saturation curve

Saturation curves were measured for the fluobeads used for dynamic experiment. The main point was to have an idea about how well the setup was actually depleting and what was the influence of the excitation power. *Figure 4.7* shows the results obtained.

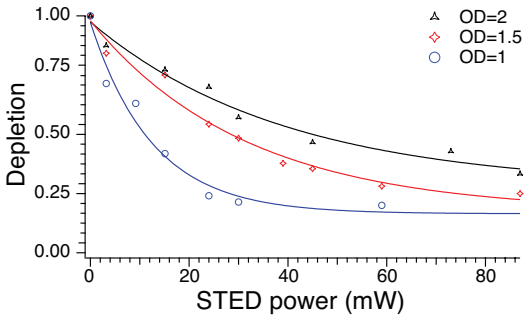

Figure 4.7: The saturation curves show that increasing the excitation power tends to enhance the depletion. This is consistent with what was mentioned in Chapter 2 that the cross section of stimulated emission depend on the number of excited molecules and thus the excitation source. Multiple measurement would have improved the match of the data to the trending line.

It can be seen that the data point are not perfectly placed on the trending line. The first reason is because ideally, the measurements have to be performed many times for each power whereas it was only performed 3 times here.

Nevertheless, one can easily observe that the saturation intensity (intensity for 50 % depletion) decreases when the excitation power increases. This is consistent with what we saw in the introduction, the stimulated emission cross-section is dependent on the number of excited molecule and thus, on the excitation power. It is particularly interesting to verify this feature in a practical way because we now know that increasing the excitation power should make the resolution enhancement higher for a fixed STED power.

4.8 Conclusions

Thank to the use of two simple standardized samples, the performances and the limits of the setup could be assessed successfully. The main results are summarized below :

- The maximal power of the laser sources before the objective is 1.4 mW for the excitation source and 250 mW for the STED laser for a in-coupling of 33% and 71.5% respectively.
- The pixel size was found to be 2.75 bigger than the pixel size set by the user.
- Two different doughnut shapes can be obtained by tweaking the polariztation. Unpolarized beam provides a better null but a less homogeneous intensity distribution while the perfectly circularly polarized beam provides an homogeneous intensity distribution with a lower null quality. The user should therefore aim for a compromise between these two extreme cases
- The resolution achievable with our setup is about 120 nm for STED and 70 nm using time-gating detection
- The size of the field of view was found not to be limited by the capability of the scanning system but rather by the preserving of the doughnut shape which can not be ensured for scanning positions further away from the center than 55 µm.

Chapter 5

DNA imaging

5.1 Introduction

Prior to try to match intensity pattern for species identification and differentiation, images had to be acquired. This chapter presents an overview of the problems that were encountered while attempting STED imaging on labeled DNA and the solutions found to solve them.
The two sequences of DNA that were studied are originated from two phages T7 and Lambda. The reason why these are used in this study is because these two DNA have been extensively studied and are therefore well known and well characterized. Before starting to image them, some basics were collected.

T7-DNA is a 39937 base pairs (bp) genome, its size is about 13.6 μm when considering 3.4 Å/bp. However, if one considers the stretching factor of the rolling droplet method (around 1.6-1.7 [19]), the expected length of T7-DNA is about 20 μm.

λ-DNA is a 48502 bp genome, its size is about 16.3 μm. The expected length considering the stretching factor is about 26 μm.

The TaqI enzyme used recognizes 5'-TCGA or 3'-AGCT and cut the sequence T–CGA or AGC–T. In MTaqI, the label is inserted at that particular position. Assuming a labeling efficiency of 100%, T7 DNA should have 111 sites. Lambda DNA on the other hand, should have 121 sites labeled.

5.2 DNA imaging

We already know beforehand that the high power required in continuous STED setups will limit the choice of the dye we can use for labeling. In addition, it is possible that it causes undesired damages to the DNA. Figure 5.1 below shows the first attempt of imaging DNA using STED. The sample was prepared by Dr.Su using Alexa 488 dye.

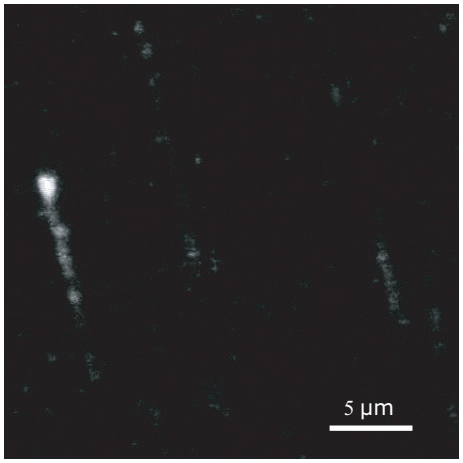

Figure 5.1: First measurement on DNA using STED (27.5 μm x 27.5 μm frame). The amount of details that can be observed is rather small due the signal to noise ratio that is rather low. In addition, fast photobleaching was observed and no resolution enhancement could be observed.

On the image above, only a few DNA molecules can be observed and only one is appearing clearly. It does not look like STED has improved anything because the main spot is still relatively big. It could be because the DNA was not stretched properly. A global observation from the first experiment session is that the photo-bleaching was quite significant, even when performing confocal measurements. A Solution had to be found.

The first one that came to mind was to change the dye to a dye that is well known for its particularly high photostability : Atto dye. Atto 488 is the one that matches best the laser sources of our setup.

Additionally, "imaging buffer" are often used to do super-resolution experiment (mostly for PALM/STORM etc...). Different types of imaging buffer exist, they are usually composed of a reducing agent that avoid the oxygen to be in contact with the dye by reacting with it.[39] The buffer used here is based on glucose oxidase (GLOX) and is described in Material and method.

Once the change of dye was performed, a saturation curve was recorded on Atto 488 attached to the DNA to check if it could indeed perform well. Figure 5.2 presents the saturation curve obtained.

One can see from the figure below that the Atto488 dyes has very good properties for STED. Indeed, its fluorescence can be quickly and efficiently reduced. The saturation intensity is only 10.5 mW and after 60 mW, it is maximally reduced.

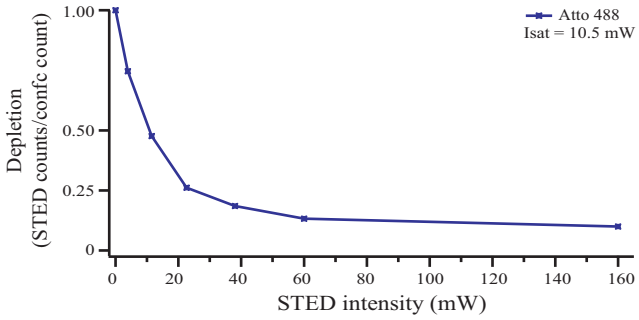

Figure 5.2: Saturation curve of Atto 488 dye shows good properties for STED microscopy. The saturation intensity is 10.5 mW what is rather low and 60 mW is already sufficient to deplete the fluorescence to its maximum.

The dye being changed and the buffer being ready another attempt onto DNA labeled by Dr. Su but Stretched and dried by myself was performed. Figure 5.3 shows the best STED image i could get out of it.

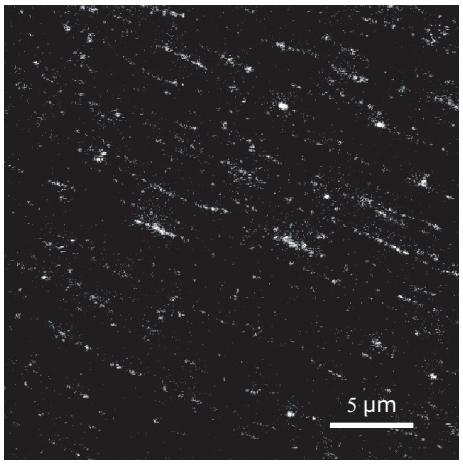

Figure 5.3: First high resolution images obtained with STED on T7-DNA atto488-labeled (27.5 μm x 27.5 μm frame). The amount of details is much higher than on the previous image. No significant photobleaching was observed. However, the stretching is not good because a great part of the molecules is curvy. Moreover, the DNA molecules are quite short indicating a high degree of fragmentation.

One can directly see that there is actually much more to see on this frame. However, the length of the DNA molecules seen on the picture is quite short considering the size of the image. They are probably closer to 5 μm than 20 μm. This problem can be due to a poor labeling efficiency or a bad handling of the DNA. Indeed, if the labeling efficiency is poor, a part of the sites are not labeled making one molecule appears as several. Strand can be sheared and/or broken due to bad handling. Several step in the sample preparation process can induce shearing : pipetting,

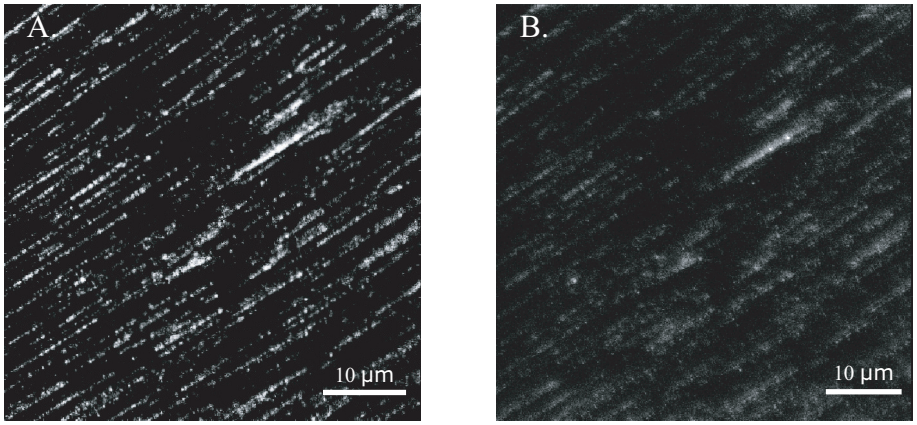

Figure 5.4: Comparison of images obtain for Confocal (A) and STED (B) measurements on the same area of the same sample. The confocal image looks rather good and shows great improvement in the stretching and the fragmentation of the DNA. The STED image, on the other hand, is completely blurry. The signal-to-noise as well as the resolution decreased.

centrifuge, stretching,... An operator should be particularly careful with those step (e.g. cutting the sharpest part of the tip, minimizing the number and the duration of the centrifuge). Finally, the DNA strand does not look stretched properly, they look slightly curvy. This could be due to irregularities of the surface or in the stretching. Indeed, if the surface is not perfectly flat, the DNA will get curvy to follow the shape of the surface when getting stretched on it. If the stretching speed is too slow, it can also be that the DNA does not get stretched properly. The stretching is relatively standardized, it is not very likely that the issue comes from there but the influence of the speed on the stretching quality was never deeply investigated.

Nevertheless, an improvement in the quality of the image compared to the first attempt can be clearly observed. Moreover, the STED seems to work because the feature are really relatively thin. In order to really check whether STED does improve the image or not, comparison between confocal and STED is needed.

Figure 5.4 shows a comparison of a confocal image and the corresponding STED image for the first sample entirely prepared by myself (from the spin coating of the slides to the stretching of the DNA and from the labeling to the purification of the DNA).

It can be clearly seen from the confocal image above that the sample looks already much better in terms of length (here the image is twice as large as the previous ones) and straightness. However, the STED image does not show any improvement and even make the image worst. It seems like STED completely bleaches the sample. In addition, it seems that the STED laser was absorbed in some places because signal appeared where there was no signal before. In a general way, the image got

more blurry when using the STED.It is difficult to know exactly what is happening here. However, it is likely that using high exposure time on STED after a high exposure time on confocal is not the best way to get good STED images. Indeed, confocal has already a trend for photobleaching due to the huge load of photon focused onto one spot, thus using STED right afterward on the same place is not ideal.

In order to find out the source of the problem, i tried different power of STED, i improved the alignment of the setup as much as i could and i made new labeled DNA but nothing seemed to work or even to show any slight improvement. Because the STED laser was giving signal even with the excitation source off, i suspected that it was absorbed by something on the sample. As people had regularly doubts about contamination of the common Zeonex solution at that time, i started to think that it might be the issue. The spin-coater was also suspected of contamination. In order to figure out where the problem was coming from, I made 2 new zeonex solutions (from an old batch of zeonex and a recently opened one). Then i spincoated 4 slides using the following scheme:

- New batch solution, other spin-coater, normal spin-coating program
- New batch solution, normal spin-coater, normal spin-coating program
- old batch solution, other spin-coater, normal spin-coating program
- new batch solution, other spin-coater, other program (6000rpm)

Then, DNA was stretched onto all of these surfaces and imaged in confocal and in STED. In such a way, the influence of different parameter should be determined. Figure 5.1 shows a good STED image obtained during this experiment.

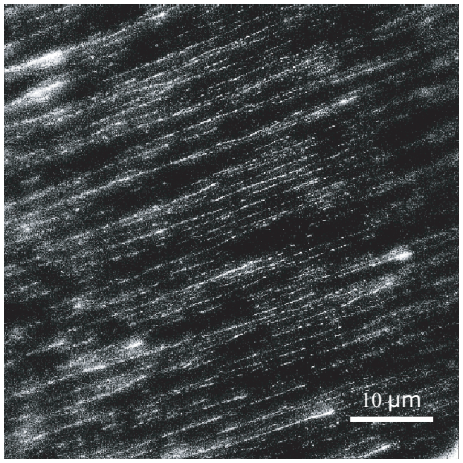

Figure 5.5: STED image (55μm x 55μm) from the new batch, the new spin-coater and the normal program. The sides of the image are out of focus. This is likely due to a misplacement of the holder. In the center of the image, high straightness of the strands as well as high resolution can be observed. The zeonex solution was indeed an issue

In the central region of the image, we can clearly see that STED worked or at least that it did not cause the trouble it was causing before. It is suspected that the slides was slightly misplaced on the holder explaining the blurry effect on the sides (out-of-focus). Once again, a huge improvement is observed when comparing with the first attempt what shows that we are on the right track. However, out of the 4 samples prepared, it was not so clear which one was better than the other, a global improvement was rather observed. It was therefore difficult to evaluate the influence of the different parameters. It is thus likely that the Zeonex solution was the main problem. The only parameter that was not really fixed yet was the spin-coating (amount, duration, timing). Due to problem of stretching in the very beginning, the way i was spin-coated was changed many times and in the end, never standardized. The way the spin-coating is performed could actually matter because the shape of the STED beam is a very important feature for the resolution enhancement and it could be affected by the milieu it goes through. Indeed, it is produced with polarization tuning and a waveplate, the results is that the shape of the doughnut is very sensitive to change in refractive index.[24] Therefore, irregularities on the surface as well as the thickness of the zeonex film could actually induce some changes in the polarization/shape of the doughnut and thus not give the resolution enhancement expected. I then decided to standardize the way i was spin-coating (a few drops before the program starts) while trying different spin-coating speed (2000, 4000, 6000, 8000, 10000). Figure 5.6 presents the outcome of this experiment for different spin-coating speed. The size of the images are the same (27.5 x 27.5 μm).

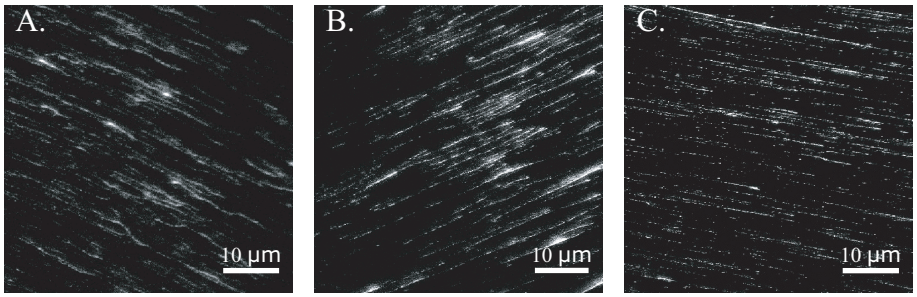

Figure 5.6: Effect of the spin coating rate on the quality of the STED imaging: A. 2000rpm, B. 4000rpm, C. 10000rpm. The series of images shows that the quality of a STED images is dependent on the spin-coating rate. In addition, the straightness of the molecules is also improved. This is explained by the change in thickness and surface roughness occurring with the change in spin-coating speed.

In a global way, all the images look quite good compared to most of the images shown before. This is likely due to the replacement of the zeonex solution and the standardization of the spin-coating. If one looks more carefully to the images, it can be seen that the one a 2000 rpm is the worst one. Indeed, the image looks kind of blurry at some places (similar to what was observed before) and the DNA does not look properly stretched. Indeed, the DNA are not straight at all. The one at 4000 rpm is clearly better than the one at 2000 rpm, particularly in term of straightness

of the DNA. This seems to indicate that the speed of the spin coating is important. Finally, 10000 rpm is by far, the image where the larger amount of details can be seen and the lower amount of blurry effect is observed and where the straightness of the strands is the highest.

This experiment demonstrated that the efficiency of STED is actually influenced by the quality and/or the thickness of the film. It was shown that the thickness of the film is inversely proportional to the spin-coating speed. [26, 57] This would mean that, in our case, the thinner the film, the better. This is consistent because if the change of refractive index affect, the thinner the zone in which it occurs, the smaller should be the effect. However, it also seems like the molecules are better stretched on the surface. This could indicate that the surface roughness has diminished with the increase in spin-coating speed.

The surface roughness is a characterization of the deviation of the surface from its ideal shape. In our case, we would like the surface to be as flat as possible, thus the roughness would represent how imperfectly flat our surface is. Hence, a rough zeonex film would wiggle a lot from its supposedly flat position leading to poorly stretched DNA as it is observed on the image of the surface spin-coated at 2000 rpm. The surface roughness was also shown to affect the optical properties of certain materials.[34, 76, 63] It is therefore reasonable to think that it could affect the STED doughnut as it is known to be sensitive to changes in its optical paths. In that case, it would confirm that the surface roughness decreased with the spin-coating speed because the image improved with the increase in spin-coating speed. However, it was difficult to find information to confirm this trend for the surface roughness in the literrature. Indeed, the surface roughness depends on the material studied and on many other parameters. It was indeed shown, for instance, that the solvent used to dissolve the polymer or the concentration of this one matters. In term of the solvent, the more volatile the solvent the thicker and rougher the film.[26] About the concentration, the higher the concentration, the rougher the surface.[26, 57, 59] In our case, we use a relatively low concentration and relatively non-volatile solvent. The surface roughness was demonstrated to depend non linearly on the spin-coating speed for ZnO whereas it was not depending on the speed for PMMA.[87, 57] However, in the later case, the range of speed studied was much smaller. Anyhow, in our conditions, it seems reasonable to assume that the film would get more homogeneous as the spin coating speed increases. Moreover, all the experimental observations seem to go in that direction. It would be nevertheless interesting to confirm by measuring the surface roughness of zeonex film for different spin-coating speed (e.g. AFM).

During further experiment, the freshness of the Zeonex solution also happened to matter. I suspect that it is because polymers molecules "like" polymer molecules. Therefore, they tend to get together after a while what is not suited for later spin coating. Indeed, having small cluster of polymer could affect greatly the homogeneity of the film and the resulting surface roughness.

We have now a better idea about the parameters that influence the quality of the STED imaging and which condition are more suitable for this particular purpose. Hence, images of T7, lambda and a mixture of both DNA could be acquired. Figure 5.7 shows a good images obtained using STED on labeled DNA using the experimental condition determined by the previous experiment.

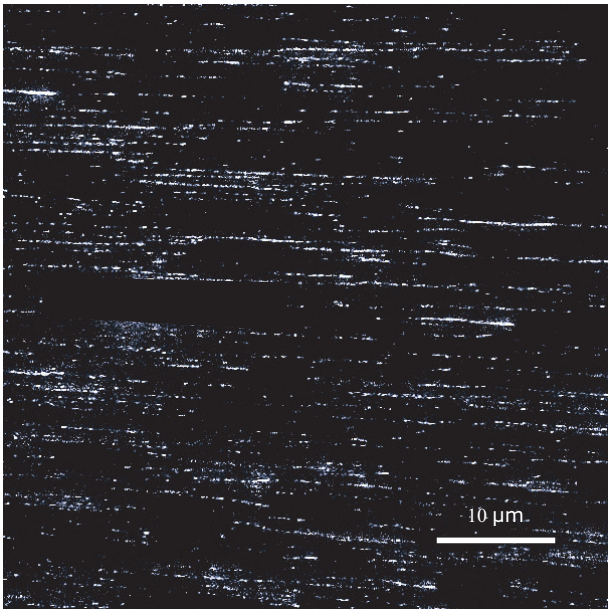

Figure 5.7: STED image (55μm x 55μm). This image shows relatively high density in DNA molecules. Additionally, the resolution enhancement obtained as well as the stretching look particularly good.

5.3 Conclusions

High resolution images of sequence-specific labeled DNA could be obtained using Stimulated emission depletion. To do so, we had to improve the way the surface was coated because the thickness of the film as well as the roughness of the surface have a direct impact on the imaging quality of STED. These experiments also allowed us to get an insight in the practical limitation of STED imaging. The optimal conditions for STED imaging are listed below :

- Atto 488 dye labeling in combination to glucose oxidase based imaging buffer should be used to limit the photobleaching and enhance the brightness of the dye
- The spin-coating of the slide should be performed at 10000 rpm for optimal stretching and optimal STED resolution enhancement

- 55μm x 55μm and 27.5μm x 27.5μm field of views seem a good compromise between getting a sufficiently small pixel size while getting a large number of DNA molecule in
- Exposure time of 0.1 ms/pixel is ideal because it allows to get high quality images with a decent measurement duration
- STED power of 300 mW on the monitor is a good compromise between high resolution and low photobleaching

Chapter 6

Species differentiation

6.1 Introduction

Now that data are accumulated, information can be extracted from them. Prior to the final differentiation of two species, more information about the system studied and the issues that might be encountered in the data analysis were needed. Simulation were used for this purpose. Matching of intensity profiles was then performed in solution composed of only one of the species studied to assess the performances of the program. Finally, a mixture of the two species was investigated.

6.2 Towards species identification

6.2.1 Simulations

In order to benchmark the program that would allow us to analyze the intensity pattern extracted from the images, we simulated artificial ones. More information can be found on the way it was achieved in Chapter 3.

Before discussing simulations, it is interesting to have a look at the position of the label on the DNA we are going to study. In addition, we could have information about how the density of information is distributed along the DNA. For this purpose, perfect intensity profile of T7 and Lambda were simulated. T7 genome was divided into four sections and lambda into five. The size of the sections was similar for both DNA. Then, the integral of a single peak (one isolated dye) was calculated and compared to the integral of each regions . This procedure allowed me to get information about the amount of information contained in each sections. This was realized simulating a relatively the pattern at high resolution (60 nm) for the pattern. In that way, we can easier compare the shape of the profile and the characteristics of the different zones. Here, it is interesting to note that T7 has a higher information content than Lambda. Indeed, T7 is 39 kb and has 112 sites labeled (2.8 sites per kb) whereas Lambda is 48 kb and has 121 sites (2.52 sites per kb). Figure 6.1 shows the distribution of the information density over T7 and Lambda DNA.

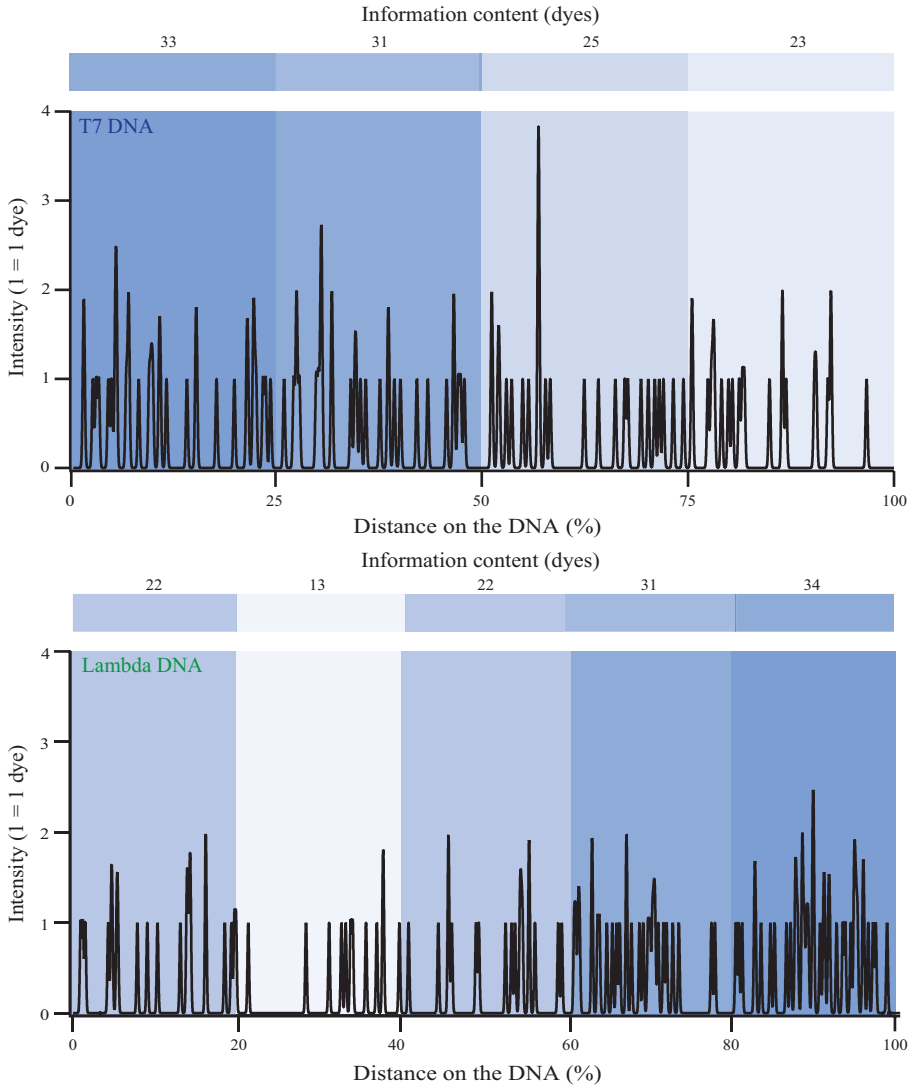

Figure 6.1: Distribution of the information density on T7 and Lambda DNA. The color scale from the dark blue to the light blue represents respectively high and low information density regions. It can be observed that Lambda DNA as much more variation in its density distribution than T7 DNA.

It can be seen from the figure above that the density of information is much more regularly dispersed over T7 DNA than Lambda. Indeed, T7 could actually be splat into two halves of similar information density whereas has large variation in its information density. This is confirmed by the fact that the ratio of the biggest density to the smallest is about 1.4 for T7 and is 2.6 for Lambda.

High density zone are interesting because they carry more information. However, with a resolution that is poorer than 60 nm, what is our case, the high density zone might not be fully resolved (density is too high). In particular, the first region of T7 and

the last region of Lambda will likely look very similar. On the other hand, the lower density zone carry less information but will look more distinct even at lower resolution. High density zone might also cause trouble in the fitting. Indeed, they are more likely to give high scores in fitting even though it might not be the right place or even the right reference. The reason for this is that it will always be easier to fit with more information on the reference than with less. This is particularly true when it comes to relatively noisy data because a noisy signal will seem information dense for the program. Hence, in our future simulation, we will also pay attention on where the simulated data are actually matching on the right and wrong reference to see if it verifies.

It can also be seen that, as a consequence of its higher density of information and its more uniform distribution, T7 DNA give rise to higher peaks (e.g. the one at 55%). This is likely due to multiple dyes that are so close that they cannot be resolved even with a 60 nm resolution. That central feature is very interesting for pattern recognition using STED. Indeed, the intensity of the peaks in STED can be related to the number of dye which is not the case in method like STORM. Hence, such a high peak with smaller one on its side should be very characteristic of T7.

In order to find the more suitable algorithm for the matching of pattern, we compared different method :

- Overlap integral
- Least-square method
- Curve division
- Cross-correlation

They were compared in their performances to match a 20 kb fragments. While doing this analysis, it occurred that comparing the method only base on the score obtained by the algorithm was meaningless. It was much more efficient and meaningful to compare the signal to noise ratio of the algorithm output. For instance, for the Overlap integral, taking the signal to noise ratio of the plot of the integral as a function of the fragment position on the reference. This way of doing allows a better comparison of the different method but also, give an idea about how significant the matching peak compared to the other position attempted. The method that appears to perform best in that respect was the Least-square method. The advantage of this method is that it was most of the time providing one peak very intense for the right match leading to higher scores than the other method where the match was not as clear.

The least-square method was thus kept for the rest of the analysis. For this method, the value compared is the mean residual of the fit from which i subtract the minimum residual and the result is divide by the standard deviation. This leads to a signal-to-noise ratio.

The performance of the least-square method were then assessed in terms of the size of the fragment it was able to match and the resolution it would need to match it. In addition, it would provide us an idea about the scaling of the matching score with the size and the resolution. The performances of the least-square method for 5, 10 and 20 kb T7 simulated intensity profiles are summarized in Figure 6.2.

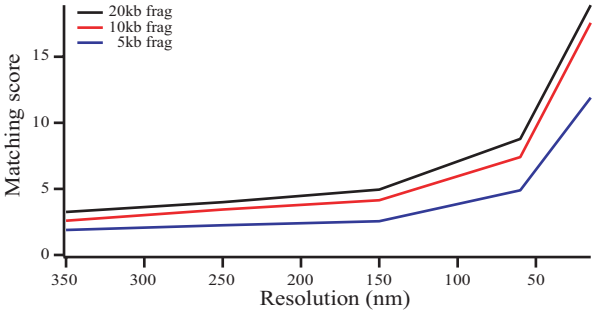

Figure 6.2: Matching score obtained for simulated T7 fragments (5, 10 and 20 kb) compared with T7 reference for different resolutions of intensity pattern. Scaling of the matching score with the size of the fragment as well as with the resolution improvement can be observed.

One can directly notice that for the correct reference, the matching score increases with the size and the resolution of the intensity profile. This was expected because as the length of the DNA increases, the amount of information also increases. The same conclusion can be drawn for the resolution. However, the increase of the matching score is not linear along with the resolution. Indeed, between 350 and 150 nm resolution the matching score is actually not much affected by a change in resolution. It might be that in that range, the resolution is not sufficient to get a significant increase in the amount of information. The reason for this could be that the density of information on the DNA is too high and necessitate a relatively high resolution to get additional information. If one compares the number of dyes on the DNA to its size, one can find that the space between two dyes should be, on average, about 125 nm for T7 and 132 nm for lambda when assuming no stretching (conditions in which we are with our simulations). However, these numbers consider a labeling efficiency of 100%. With the labeling efficiency of our simulation (80%) the average distance between two dyes goes to 156 and 165 nm for T7 and Lambda respectively.

If one considers the presence of false positives this value should again decrease a bit. In any case, the fact that the average distance between two dyes is close to 150 nm could explain why the score does not change much for lower resolutions. The resolution increase does actually not augment significantly the amount of information because the molecules are too close to be resolved. However, we have seen that the density of information is not uniformly distributed what explains that even with a resolution poorer than 150 nm the matching score scales, to some extent, with the resolution.
The trends discussed above were also observed when comparing with the wrong

reference. However, there were not followed as nicely (e.g. 10 kb fragments would give better scores than 20 kb ones for a range of resolutions). This is of course because the information is not specific to this reference. Moreover, the increase of resolution was also causing changes in the place where the fragment matched. As the matching was not operate at the same place, the correlation with the resolution should not be as good as when it is the case. The matching scores were also much lower but still relatively high (between 1.5 and 4). This is because our program is implemented to find the best match, therefore, it will always find a place where a fragment match, to some extent. Hence, it is already clear that we will not be able to get a threshold value above which we can consider a match and below which a mismatch. The attribution of the matching should be performed by comparison of the score obtained for different references. As we will focus our interest onto two species, a matching ratio can be used for that purpose. Figure 6.3 compares the matching score by taking the ratio between the one obtained for T7 reference and Lambda reference.

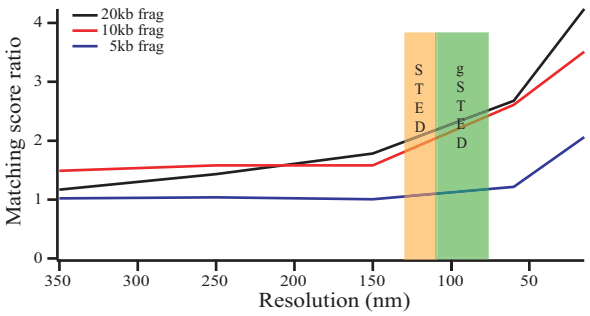

Figure 6.3: Matching score ratio obtained for simulated T7 fragment (5, 10, 20 kb) for different resolutions. STED matches the region where the scaling of the matching score with the resolution start to be more pronounced justifying its use.

The figure above demonstrates clearly that for a 5 kb fragment, association to the the correct reference is hardly achieved unless the resolution is high (<150 nm). This is when considering one as a threshold for the matching ratio what is really the minimal threshold one can consider. However, for longer fragment, a resolution comprised between 100-150 would already be sufficient to differentiate them. Hence, we see that our STED setup fit exactly the zone where it start to be interesting to use super-resolution methods. This justify the use of STED rather than confocal or Wide-field. Moreover, we also see that we should not expect high differentiation score. Indeed, with our resolution, the differentiation score is about 1.5-2 for 10-20 kb fragment. In addition to this, experimental data will likely have noisier profiles. We also know that there are stretching variation between DNA strand and also stretching inhomogeneities within a strands (this is discussed below). Moreover, it could also be that the labeling efficiency assumed in the simulations (80%) and the false positive rate (0.1/kb, poisson distributed) are respectively overestimated/underestimated

what could lead to lower scores.

In order to test the robustness of the program, T7 and Lambda genomes were inserted in a larger database (E.Coli genome). Labeling of the resulting genome was simulated using our simulation routine. Then, i use the matching program to match a very small fragment (5kb) using the EColi as the reference. In order to limit the size of the array used, the genome was converted in nm using a stretching factor of 1 (0.34 nm/bp). Figure 6.4 shows the residual obtained from the attempt of matching. The T7 fragment is expected to match around 680000 nm, where T7 genome starts.

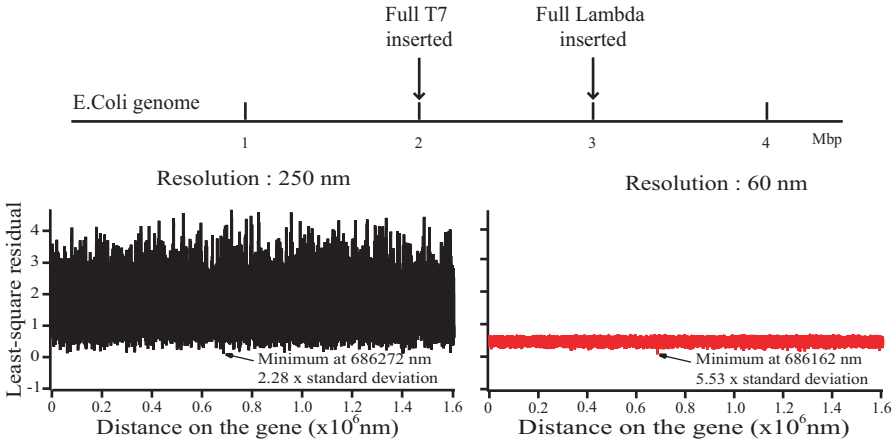

Figure 6.4: Residual of the least-square method when matching a 5 kb T7 fragment onto T7 reference inserted in E.Coli genome. Matching in a larger database necessitates a higher resolution to get a sufficient degree of certainty on the matching.

One can see that with a low resolution, a match at a right place is obtained but with a very low certainty. Indeed, the signal-to-noise ratio (matching score) is very low and there is actually no obvious matching when looking at the residuals plot. However, with a high resolution, a clear match is obtained at the exact spot where the 5 kb fragment was supposed to be. The reason why the matching position is so different between the two is due to a shift in the position due to the way the program was written. After correction for this shift the match for 250 nm is 13 nm away from the exact value and thus from the match for 60 nm resolution. Nevertheless, the same conclusion as before is obtained, we need a high resolution to get a high matching score and to differentiate the real match from the rest. This is particularly true when considering a very small fragment in a very large database.

This small experiment shows that the program and the way we use it is actually quite good because it could manage to match a 5 kb fragment (12.5% of the full genome) to a database that is about 115 times larger than the full T7 genome. On top of this, one should not forget that we removed 20 % of the information, added some false information and added some noise. The result is a fragment that contains

less than 10 % of the information contained in the full genome that could be match successfully in a very large database.

Now that we know that the program is relatively robust and that we have an idea about the effect of the resolution and the size of the fragment on the matching score, we can run simulation in similar condition as the experimental one. Indeed, we know the resolution of our setup and the pixel size we used during the imaging. As the size of the fragment was already investigated a bit, we will use a fixed size to investigate the other parameters. Investigating the size more deeply is not so interesting considering that we do not simulate the stretching. Hence, we expect the matching score to increase with the size. Moreover, the size of the fragment is not something we have a real control on, we can try to reduce the amount of fragmentation by careful handling but not control it. Two parameters that are interesting and that we can control a bit more, is the labeling efficiency and the number of false positives. Hence, it would be interesting to know which one matter more. Is it better to have a high labeling efficiency but with a lot of false positives or to have a low labeling efficiency but few false positives, for instance. We will also try to see how wide is the distribution of the matching scores for T7 and Lambda. We will also try to find whether the matching on the wrong reference follows a trend or not. Indeed, we have seen that the density of information on the strands is not uniformly distributed and we could expect the matching to give higher score when matched in places where the information content is high. Therefore, it could also be that wrong matching tend to happen in high information content zone. We can also hope to identify places that are similar on the DNA. To answer these questions, 500 T7 and 500 lambda fragment were simulated for different labeling efficiency keeping the size (20 kb) and the false positive rate (0.1/kb) constant. Analyzing the population will allow us to understand better the system.

To make the understanding of the figure easier, a convention for the color was chosen : All the figures and graphs concerning T7 DNA will be in Blue whereas the ones concerning Lambda DNA will be in green. This convention will be kept for the experimental part as well. Figure 6.5 presents the distribution of the matching score obtained for T7 DNA (blue, A) and Lambda DNA (Green B) when compared with the corresponding reference. The labeling efficiency is 80%.
The shape of the distribution is rather difficult to interpret. Indeed, the matching score can depend on many parameters : The size of the fragment, the labeling efficiency, the number of false positives, the place on the DNA where it was created and matched. In this experiment, the size of the fragment and the labeling efficiency were fixed. Hence, we expect the distribution to depend mostly on the number of false positives and the place where the DNA was created. The place where it was created can influence the matching score because, as we discussed before, the density of information on the DNA reference is not uniformly distributed. Hence, as an example, a 20 kb fragment generated from T7 which start at the beginning (High information density region) or at the half of the genome (lower information

density region) will have relatively different information content and thus different matching scores. One should also to consider that, even at fixed labeling efficiency or false positives, the dyes that were deleted/added in the process are chosen randomly. Hence, even with all the parameters fixed, for a given place on the DNA, we would still have a certain distribution of matching scores.

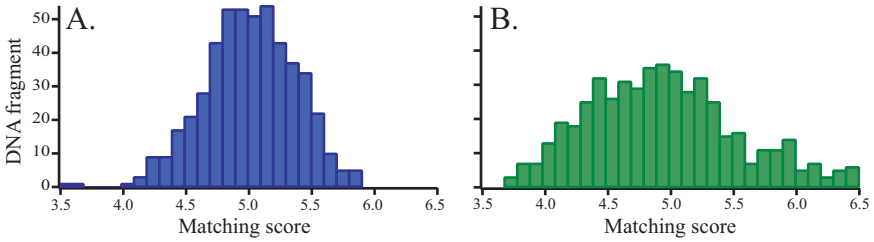

Figure 6.5: Distribution of the matching score for A. T7 fragments matched with T7 reference B. Lambda fragments matched on Lambda reference. For T7 DNA, the maximum of the distribution is slightly shifted towards high values whereas it is shifted towards low values for Lambda DNA. The distribution of the matching scores is also broader for Lambda.

Predicting the matching score is thus rather challenging. In addition to the effect of the simulated fragment itself, it is likely that the program will tend to fit better on high density information region of the reference. It is indeed, easier to fit with more information than not enough. However, what we are actually looking at, is the signal to noise ratio on the residual and thus, how significant the match is compare to the other position attempted. It could therefore be that matching in high density region will give a lower signal to noise ratio because the matching score will be high at different placed of the high density region. On the other hand, fragment simulated from a low density region will likely not carry enough information to get high score in high density zones, the matching will likely be more specific and thus give a high signal to noise ratio. We are here in presence of two trends that are opposite showing that the program we use will also participate in the complexity of the distribution.

The shape of the distributions resemble an asymmetric Gaussian distribution. For T7 the asymmetry is slightly in favor of higher scores whereas it is in favor of lower scores for Lambda. This also gives an average score slightly higher for T7 than Lambda. This could be an indication that high information density fragments will indeed tend to give higher score. Indeed, T7 has globally a higher information density and most of it is contained in the first 55-60% what would account for the shift of the maximum towards higher score whereas Lambda has about 40% of high information density region and 60% of low information density region accounting for the shift towards lower matching score.

One can also see that the distribution is much broader for T7 than Lambda DNA. This could be explained by the fact that more positions are attempted for Lambda

DNA than T7 DNA. Indeed, 20 kb represents about 50% of the T7 genome while it only represents 40% of the lambda genome. Hence, there are more possibilities to make different DNA strands from Lambda than for T7 resulting in a broader distribution. In addition to this, the information content varies much more on Lambda than T7 what could also participate in the broadening of the distribution.

It is interesting to note that the shape of the distribution remained the same even for lower labeling efficiency (e.g. 50%). Two main differences were observed. The first one, is a shift of the distribution towards lower scores. This was expected because the amount of information is lower. The second difference is that the distribution broadened. The reason for this broadening is likely to be the fact that as the labeling efficiency diminish, the number of possibility for the dye position increases leading to a broader distribution of matching scores.

Here, we have only talked about the matching on the right reference. However, the matching ratio is also very important because it will be the value we will look at later in the experiment. In that case, one should not forget that some places on the DNA will contain information that are more specific meaning that they will not resemble a place on the other reference. This is likely not dependent on information density and will nevertheless affect the matching on the wrong reference and thus the matching ratio.

No evidence of dependency of the matching score with the matching position was found when looking at the data. It is likely that 500 molecules is not enough to find a trend without fixing more parameters. This will be investigated later in this chapter. For now, we will look at the distribution of the matching positions. Figure 6.6 compares the matching position of T7 and lambda on the right and the wrong reference.

The matching position on the right reference (left) was always corresponding with the starting position of the generated fragment. This was expected because we know that for 20 kb fragment the program is able to match in 100% of the cases on simulations. It can be seen from the figure above that the matching position on the right reference is randomly distributed. This of course due to the implementation of our simulations.

On the other hand, the matching position on the wrong reference is not randomized at all. In the case of T7, most of the fragment were matched at the end of Lambda fragment. This is consistent because the end of Lambda DNA is the part that is the most information dense. Hence, we clearly see that T7 that has globally a higher information density, tend to match on the high density zone of Lambda. This shows that the program tries to match zone of similar density. However, it could also be that matching in higher density information zone is just favored.

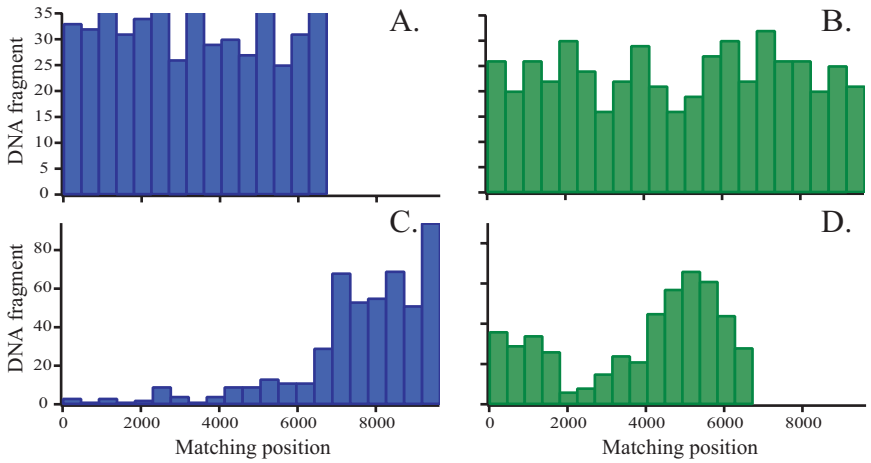

Figure 6.6: Distribution of the matching position for A. T7 fragments matched on T7 reference B. Lambda fragments matched on Lambda reference C. T7 fragments matched on Lambda reference D. Lambda fragments matched on T7 reference. When the matched fragment corresponds to the reference, the position are randomly distributed while it is not the case when the fragment does not correspond to the reference.

In the case of Lambda, the distribution of the matching position is different. Indeed, a great part is matched in the lower information content zone of T7 whereas a small part is matched on the high information content zone. This seems to confirm the idea that places of similar information content tend to match. Indeed, from Figure 6.1 we know that only a small part of the 20 kb fragment will be generated from the high density zone and thus likely to match in the beginning of T7 (High information density zone). Similar justification can explain the proportion of matches in the lower density region of T7.

In our previous justification we forgot that no matter the density of information (within a certain limit) places can look similar at our resolution. This phenomenon is likely due to the lack of resolution. Indeed, when one look at the reference map presented above, it is difficult to see lot of similarities. In any case, places of similar information content are more likely to resemble than places of different information density thus our justification was still accounting a bit for similarities.

It is important to note that the shape of the distribution of the matching on the wrong reference did not change much with the labeling efficiency. This tends to confirm that there are actually some similarities between the references. Indeed, at very low labeling efficiency the argument of information content is not so valid anymore but the distribution still did not change. On one hands, this is a good news, the fitting is actually matching area that looks alike. On the other hands, it means that at the resolution we are working with, the maps have a certain degree of similarities. This probably account for the ratio of the matching score being so low even for the relatively perfect data that our simulations are.

The reason why no clear trend could be found between the matching position and the matching score is because many parameters have to be taken into account. Indeed, the number of false positives and the labeling efficiency makes the number of possibilities for a given position relatively high. Hence, to be able to extract a trend on this we should either fix more parameters, either get a larger statistic (10000 fragments or so). To avoid unnecessarily long computation time. 500 fragments with 100% labeling efficiency and 0 false positive were simulated to investigate the influence of the matching position. In addition we will try to get an insight in the similarities of the map by comparing matching position on right and wrong references. The matching on the right reference is actually the place where the fragment was generated on the genome.Figure 6.7 presents a plot of the matching score as a function of the matching position and a plot of the matching position on the wrong reference as a function of the matching position on the right reference (position where it was generated).

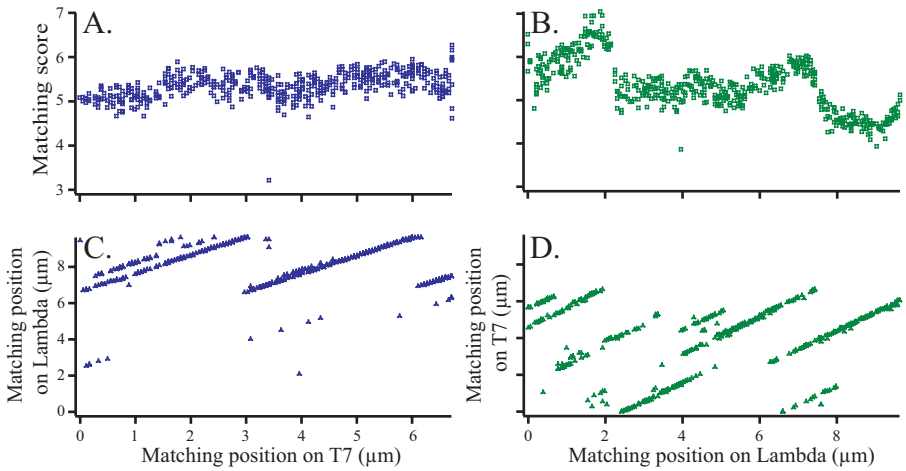

Figure 6.7: Influence of the position where the fragment was generated on A. the matching score for T7 fragments matched on T7 reference B. The matching score for Lambda fragments matched on Lambda reference C. The matching position of T7 fragments on Lambda reference D. The matching position of Lambda fragment on T7 reference. By comparing with Figure 6.1, it can be seen from A and B that fragments that are generated from high density region tend to give lower score (and vice versa). Figure C and D shows that there is a certain level of similarity between the two intensity patterns.

In the first place, we will only consider the top part of the figure. It represents the influence of the place where the fragment was generated on the matching score.

As mentioned before, the program should tend to give better fitting onto high density fragment in high density zone. However, as we are looking at the signal to noise ratio, a second trend also had to be taken into account: a fragment generated from a low density zone is more likely to give a high signal to noise ratio in the residual because

it will only match there whereas fragment generated from a high density zone will give high score for different shift in the high density region. Until now, we only had a few indication about those trend and it was difficult to determine which one tends to be stronger than the other. However, from the figure above it can clearly be seen that the fragment generated from the high density zone (in the beginning for T7, in the end for Lambda) actually give lower matching scores. This seems to indicate that fragments generated from low density indeed yields better signal to noise ratios. The fact that the highest matching scores are obtained in the beginning for Lambda DNA completely confirmed this observation because the fragment generated there contains the very low density region of Lambda. Therefore, our justification that T7 has a higher density of information to explain why it has, on average, higher matching score appears not to hold to. This can be explained by considering that 20 kb represents 50% of T7 whereas it only represents 40% of Lambda. It is thus normal that the matching gives lower score, 20 kb represent a smaller proportion of the genome for Lambda. It is likely that for fragments representing the same proportion, Lambda one would give higher score because of its low information density zones.

One can also observed that Lambda score is much more influenced by the matching position than T7. Two reasons can explain this. Firstly, there is more possibilities to generate 20 kb fragment on Lambda than on T7 what will tend to cause more variation. Secondly, the information density is much more homogeneous on T7 than Lambda DNA.

We will now look at the two figures on the right. It presents the influence of the place where the fragment was generated on the matching position on the wrong reference. It can be seen that for one position generated, there is usually more than one match possible. This is normal because the information is not supposed to be specific to that reference. Although, it can be observed that there is a certain degree of similarity. Indeed, we can see some straight line on the profile. A straight line means that when the place where the fragment is generated is shifted, the place where it matches is also shifted. This clearly shows that they are similarities between the reference. It is particularly true for T7 which has almost a 1 to 1 correspondence between the generated and the matching position on lambda. The relatively high level of similarities is likely due to the resolution that is not that high compared to the density of information. We also see that for lambda, the fragment that are matched onto the beginning of T7, are not only the ones generated from the end of Lambda. This indicates that similarities in the pattern has a stronger influence than similarity in the density of information.

The influence of the labeling efficiency on the matching ratio was also investigated. The average value of matching score for both reference and for the matching ratio was extracted for different labeling efficiencies ranging from 10 to 100%. Figure 6.8 shows the evolution of the matching score on the right and the wrong simulation for both DNA.

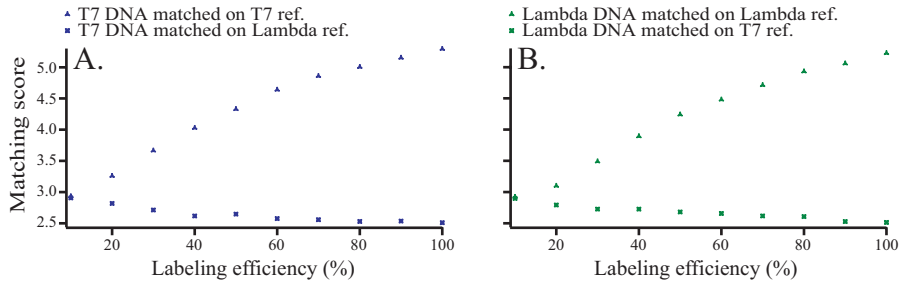

Figure 6.8: Evolution of the matching scores for the correct and the wrong reference as a function of the labeling efficiency for A. T7 DNA simulated fragments B. Lambda DNA simulated fragments. The matching score is greatly increased for the correct matches whereas it decrease slowly for the wrong matches. The two matching curves converge to the same value at very low labeling, no distinction is possible between correct and wrong matches.

It can be seen from the figure above that the matching score for the right reference increases with the labeling efficiency. This behavior was expected because as the labeling efficiency increase, the amount of information that matches on the fragment also increases giving it more similarities to its kind. However, it can also be seen that the increase tend to slow down at high labeling efficiency. This would mean that at a certain point, the amount of information does not increase much anymore. This trend is likely to be proportional to the resolution. Indeed, it could be that at high labeling efficiencies, the information added does not increase much the matching score because, our method is not able to distinguish the information added from the rich information already present on the fragment.

It can also be observed from the figure that the matching score for the wrong reference tend to decrease with the labeling efficiency. It is simply because we add information that are not specific to that reference, it thus become more difficult for the program to find good matches. However, it is also much less sensitive to the labeling efficiency than the matching score for the right reference. The reason why is because the information is non-specific but not anti-specific. What i mean by anti-specific is that all the information added would render the matching worst as opposed to specific where all the information added improves the matching score. Non-specific information are, from the point of view of the wrong reference just random info, thus some of them might make the score better while other might not. In a global way the score decreases because the likelihood that the added help for matching on the wrong reference is low but this explains why it is less sensitive.

The final observation that can be made is that the two curves tend to merge to a common value at low labeling efficiencies. This is because as the labeling efficiency diminish, the content in specific information will decrease and thus the fragment will be able to match any DNA. In the limit case where there is no label on the DNA, all DNA will have the same matching score, we are not able to distinguish them we do not have information.

In this case, it will be particularly interesting to look at the behavior of the matching ratio with the labeling efficiency because from the two graphs here it is difficult to see any difference between the trend for Lambda and T7. Figure 6.9 shows the evolution of the matching ratio with the labeling efficiency.

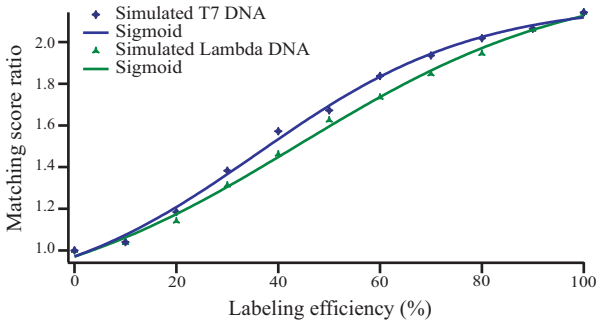

Figure 6.9: Comparison of the evolution of the matching ratio with the labeling efficiency for T7 and Lambda DNA. The evolution of the matching score is very similar to a sigmoid, it evolves slowly for low labeling efficiency, the information is not specific. For mid-labeling efficiency the matching ratio scales increase rapidly with the labeling efficiency. For high labeling efficiency the increase in matching ratio slow down, saturation of information is observed.

It can be seen that both lambda and T7 DNA have a similar evolution along with the labeling efficiency. It has a slight sigmoidal trend. The sigmoidal line was added mostly to guide the eyes, it will not be used to extract any parameters. We will now take the evolution of the curve step by step.

At low efficiencies, the ratio does not evolve much. This is because wrong and right matches tend to the same value because the information content is not sufficient to be significantly more specific to one or the other. Then, when a certain threshold is reached, specific information start to appear and the matching score tend to increase faster. This threshold seems to be smaller for T7 than Lambda DNA. This is probably because the density of information on T7 is slightly higher, hence for the same increase in labeling efficiency, more labels are actually added on the strand. This might appear contradictory to the previous observation that low density regions tend to give higher score but it is not. Indeed, we are talking here about a threshold and it is reasonable to think that a higher density region will more quickly have enough information to be differentiate than a low density region. This does not forbid the low density region to give higher scores at high labeling efficiency.

When a second threshold is passed, the matching ratio evolve more slowly towards its maximal value. This slow down in the evolution of the matching ratio is likely due to the fact that our resolution does not allow us to resolve the additional information that we get because it is lost in the information that is already there. We will call

this effect the saturation of information. One can also observe that T7 tend to saturate faster than Lambda. This is because of the higher information density of T7. Hence, for a fixed resolution, the saturation will appear faster if the density of information on the DNA is high what is fully consistent with our explanation. This is particularly interesting because it means that the labeling efficiency targeted should depend on the resolution we have. With a poorer resolution, less label are needed because otherwise you saturate your information anyway. However, a poorer resolution will also result in a lower matching score what should of course, be taken into account.

The last experimental parameters will be investigated with simulations is the influence of the number of false positives. For this, simulations were performed for both DNA, at 70% labeling efficiency for 0, 1, 2, 3 , 4 and 5 false positives. The number were chosen to cover most of the range obtained by the generated Poisson distribution when using the typical rate used in the other simulations. Getting the influence of the false positives is slightly trickier because it is likely that the effect of adding false positives will depend on the amount of information already present. Hence we will use the ratio wrong info/total info in our analysis. Figure 6.10 shows the results obtained for T7 and Lambda simulation in term of the matching score on both reference and the matching ratio.

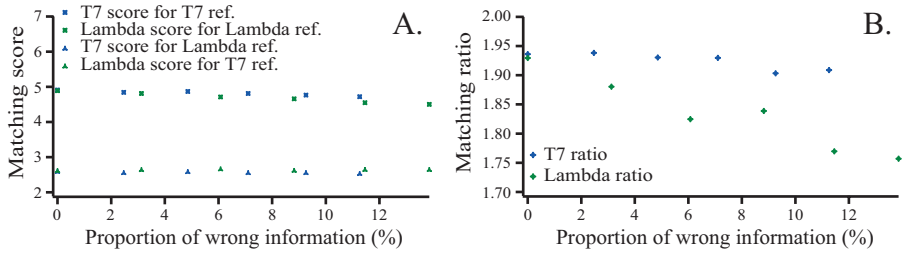

Figure 6.10: Influence of the proportion of wrong information on A. The matching score for simulated fragment of both DNA for both reference B. The matching ratio for both sets of simulated fragments

The fact that the trend are not as clear as for the previous simulations is likely due to the fact that we should have taken a higher statistic in this case. Indeed, the total number of dye can vary a lot, hence, the ratio wrong/total information will also vary a lot, getting a representative average would need more statistic to get clearer trend. We will first focus our interest on the figure on the left.

The first thing that strikes on this figure is the mirror symmetry that is present between the data points for the matching on the right reference (above) and the data points for the matching on the wrong reference (below). However, this is actually mostly due to the relative position of the point for Lambda and T7. It is also due to the fact that matching on the wrong reference tend to have an opposite trend to matching on the right reference. Indeed, the matching score for the right reference tend to diminish with the number of false positives whereas the one for the wrong

reference tend to increase. For matching of T7 fragment on Lambda however, a slight decrease was observed in the matching score. In any case, those trends are very small and it does not seem like the number of false positives affects greatly the matching scores. The explanation here is rather similar to the reason why increasing the labeling efficiency was affecting less the matching score on the wrong reference. Indeed, once again, the information added is just randomly placed and is thus non specific. Hence, the effect is relatively small because a certain part of the false positives might contribute in a positive way to the matching score and another part might contribute in a negative way. In addition, the number of false positives is relatively small and thus most of the information remains correct. It is likely that for smaller labeling efficiency, the effect would be more pronounced. This seems to be confirmed by the fact that Lambda DNA score is slightly more affected by the number of false positives. Indeed, Lambda DNA has a lower information content and is thus, more affected.

The density of information of the reference can also be used to explain the trend observed. In the case of T7, its high labeling density protect it in a way from the effect of false positives. We have here again a manifestation of the saturation of information, the information added is wrong but is lost in the rest and thus has a very small effect on the matching score. As T7 only matched the high density region of Lambda, the same effect is observed and the decrease of the matching score for Lambda reference is very small. As a consequence of the two decreases the matching ratio (figure on the right) is relatively stable. In the case of Lambda, the matching on the right reference is more affected because of its poorer information density. Hence, the false positives are fully resolved and tend to diminish the matching score. On the other hand, the matching on the wrong reference increases with the number of false positives because the density of information tend to approach the one of T7 reference with the added false positives. Both effect are very small but as they are going in opposite directions. As a consequence, the ratio is much more affected for Lambda DNA.

The results from the simulations emphasized the complexity of the system studied. Indeed, we could see, thanks to the simulations, the numerous parameters influencing the matching and their interconnection. The program shows consistent behavior with the different changes performed in the simulations. We also showed that our STED setup is in the appropriate range of resolution to get a good differentiation. These demonstrations are encouraging for our experiment and we have great hope that our method could identify a DNA from its intensity pattern.

However, several parameters were not taken into account in the simulations. Indeed, we did not consider the stretching variations nor the orientation of the DNA in our simulations. Two types of stretching variations are found, stretching variations among the different strands and stretching consistency along a single strand. The reason why we did not simulated them is that we only have a succinct idea about the stretching variation between the strands and basically no idea about the stretching inconsistencies. Hence, it was not possible to simulate these effects in a realistic way.

To account for the variation in stretching among different DNA strands, the experimental profiles will be compared to references generated with different stretching factor. The problem of this approach is that the same experimental profile could have a strong match for a certain factor with T7, for instance, but also a strong match with Lambda reference with another stretching factor. The range of stretching factor should then not be too large to limit such problems and to be realistic anyhow. From the knowledge of the group, in the typical stretching conditions, an average of 1.6 was found. We will thus confront our experimental data with stretching factor going from 1.2 to 2. An additional consideration could help us to figure out whether the match makes sense or not. If we consider the DNA as a spring, longer DNA should tend to be stretched more than shorter ones. Hence, if we get, for a long DNA a strong match with lambda with a low stretching factor and a strong match with T7 with a high stretching factor, T7 is more likely to be the right reference. We could even correct the matching scores with a factor of probability depending on the size and the stretching factor distribution. However, to use this consideration, we would need to have a better knowledge of the stretching factor.

Accounting for stretching inhomogeneities is very difficult. As we have no idea about how strong is the phenomenon, it is difficult to predict its importance. If the variation are relatively small, we can may be think that the resolution that we have with STED will not really be able to resolve these inhomogeneities. Hence, in that case, a certain lack of resolution could actually make our profile having an inherent robustness to stretching inhomogeneities. Moreover, if we assume that it is a random process, we could assume that the effect would tend to cancel out over the strand. If the variation are relatively strong, it is likely that we will not be able to match properly our data. In that case we would have a sort of paradoxe. Indeed, long DNA are more likely to have more information but also more likely to have more significant stretching inhomogeneities whereas small fragment will have less stretching inhomogeneities but also less information. As this problem will likely be the one giving us the most trouble, other analysis route which could allow to account, to some extent for this effect, will be discussed as a perspective.

To account for the direction of the DNA, experimental data will be compared to the reference oriented in both direction.

6.2.2 Proof of concept

Now that we could match simulated data and learn a lot about how complex the system studied is, we can try to tackle our problem with experiment. We will first ensure that STED measurements indeed allow to get additional information compared to confocal measurements. Hence, a line profile was drew along a DNA on both STED and confocal image allowing to compare there profile. An additional line was drew perpendicularly to the strand to get an idea about the resolution obtained

for both methods. Figure 6.11 presents an Confocal and a STED image as well as the resulting intensity profiles.

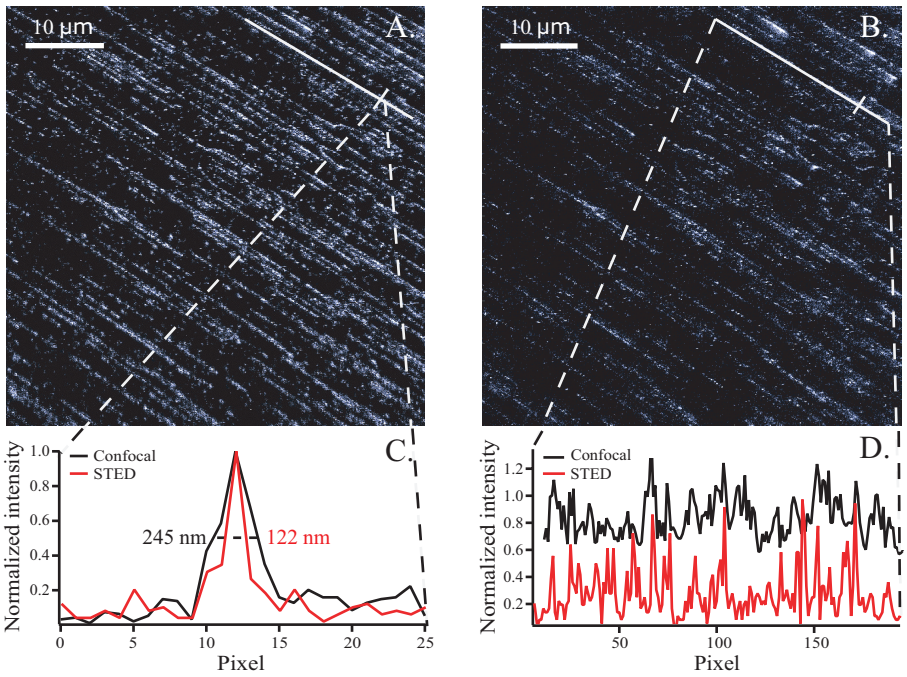

Figure 6.11: Comparison confocal and STED measurement: A. Confocal image B. STED image C. Comparison of the resolutions D. Comparison of the shape of the intensity profiles

From the figure above, one can directly see that STED allows to get a better resolution. However, on this picture, it also looks like it bleaches the sample a lot. The main reason for this (Except from the tendency of STED to make dyes bleach very fast), is that STED scan was run with a relatively high exposure time after a confocal scan also run with a high exposure time. Hence, it is perhaps not the best condition to compare them but it is the more straightforward because we can compare similar information. Despite those conditions, the resolution of STED is 2 times better than in confocal. As a direct consequence of this resolution improvement, the intensity pattern extracted is much more detailed. Indeed, the peaks look much sharper and the signal to noise got enhanced. However, it can also be seen that some feature are absent on the STED pattern but this is probably due to the double scanning. For future measurement, STED and confocal will be taken onto different spots on the sample to avoid these undesirable effects. The use of STED in the context of DNA mapping is therefore fully justified because we can extract more detailed pattern from its images than we would with diffraction limited tools.

Before pursuing, a remark should be made about getting intensity pattern from an image. One should always try to orient the DNA sample parallel or perpendicular to the scanning line and avoid diagonal orientation. The reason for this is that when

extracting intensity pattern, the pixel size you get when extracting in diagonal is larger than the pixel size you actually set on your image. It is larger by a factor which depends on the orientation angle of the line drew with respect to the horizontal or vertical and, in the worst case, this factor is $\sqrt{2}$. This is because when taking a DNA strand at 45 degree from the scanning line, the number of pixel that you get on the profile is correct but the size of the pixels is given by their diagonals rather than their sides. Hence one would get a profile more poorly resolved for DNA taken in diagonal. Additionally, having the DNA parallel to the scanning line will diminish the scanning time per strand what will diminish the risk of sample and focus drifting. This advantage is rather small giving that STED is already supposedly fast.

The next step was to extract a few pattern and check if we could see similar features in those patterns. Figure 6.12 shows two patterns extracted from two different T7 DNA on the same frame that are well matched.

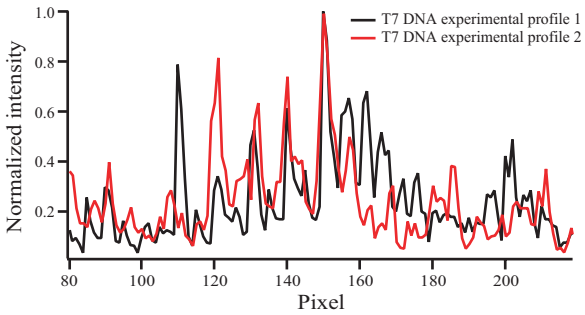

Figure 6.12: The two experimental intensity profiles extracted from different images of T7 DNA are very similar and the main features could be matched successfully. Secondary features also shows similarities in shape but the peak are slightly shifted from each other and could not be matched. This might be a first concrete evidence of stretching inhomogeneities

It can be seen from the figure above that the two profiles are quite well matched. Even though the intensity are different the shapes of the patterns are very similar on many locations. However, one can see that on the side of the pattern, the peaks are not matched as efficiently as in the central part. They may of course have different missing label and false positives. In addition, the two molecules may have a stretching factor that differ slightly. What is also possible is that the stretching may not be consistent over the full length of the DNA. Hence, it we could well be in presence of two very similar pattern that have stretching inconsistency which would explain that the shapes are very similar but some peak are shifted. Another explanation could be that when extracting the pattern the line was not perfectly parallel to the DNA. These two patterns illustrates quite well the increased complexity of the system when considering all the possible bias from the preparation of the sample to the imaging and data extraction. It also illustrates how complex it can be for the program to match two patterns. Indeed, even though they look the same in some part and by

eyes we would say that they represent the same region on the DNA, the differences, no matter their origins could affect greatly the outcome of the matching.

What is likely to be crucial is the stretching inhomogeneities. Indeed, it can be easily understood that if the pattern are very similar in term of shape but the peak are displaced because of stretching inconsistencies, it will be difficult for the program to match it to a perfect reference where these effect are not taken into account. It might be considered as one of the disadvantages of our analysis method, it is relatively rigid and is not supposed to be so robust when facing such issues.[18] Only experiment will tell us the importance of this effect.

The fact that we are able to find similar patterns on different DNA is already promising for the data analysis. Indeed, if we can see by eyes that they are similar, we can hope that the program will be able to do so as well. The fact that similar patterns can be found also shows that the labeling is sequence specific. This was of course thought before but this is a clear evidence. However, the fact that we can find similar patterns proof that the labeling is indeed sequence specific but does not proof that it should match the reference. Indeed, if the mechanism is slightly different that what we think, it could be that the profile does not match the reference because it has a different pattern on it. Figure 6.13 shows a matching obtained for T7 DNA onto T7 and Lambda reference using our procedure.

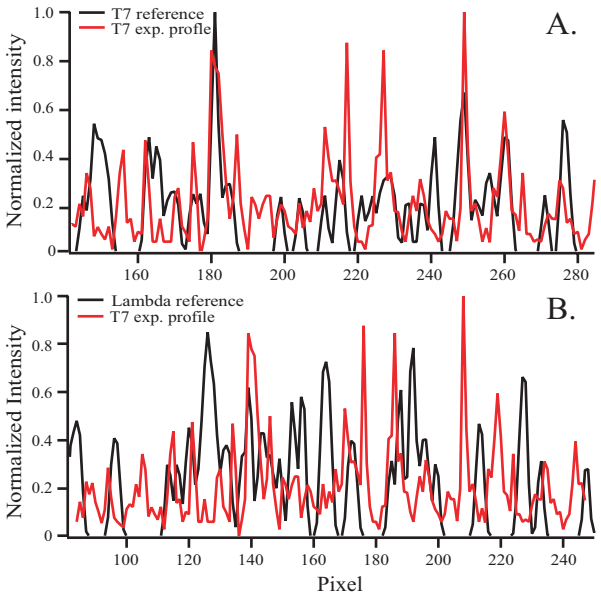

Figure 6.13: Matching of an experimental profile extracted from T7 DNA images onto A. T7 reference B. Lambda reference. The matching on T7 reference is clearly better because the main features of the profiles are matched whereas it is not the case for matching on Lambda reference.

From the image above, it can be seen that the experimental profile is much better matched onto T7 reference than Lambda. Indeed, most of the features on the reference are matched while it is not the case on Lambda matching. However, it also look like some feature are at places where they should not be and some are missing. Both phenomenon can have 2 different explanations.

Missing features are most likely due to the labeling efficiency that is known to be below 100 %, hence some dyes are missing. A feature misplaced on the other side, is most likely due to errors of labeling or "false positives". Both of these explanations are known phenomenon that occurs during the sample preparation. The last explanation is common to both phenomenon. A missing feature somewhere could actually be missing because it is slightly shifted from its position. This would explain both a missing feature and a additional one, phenomenon that were treated separately before. The explanation for this shift in position, would be stretching inhomogeneities. Indeed, if there are variations in stretching, features that should have been matched will not be matched anymore as the reference has a fixed stretching factor. This phenomenon is likely the cause of the mismatched of the two peaks a bit before 220 pixels. Indeed, we can clearly see that the profile and the reference have the same shapes but the profile is slightly shifted while being matched at other places. The fact that it is matching some feature but not other even though they have the same shape can only be explained by stretching inhomogeneities.

One can also see that the labeling efficiency should be relatively high because the amount of feature on the reference and the amount of feature on the experimental data are relatively similar. This is a good news for future analysis because high labeling efficiencies are supposed to give higher scores.

Hence, this matching example illustrates perfectly the different issues that we will have to face during the data analysis. Indeed, in this case, it can even be seen by eyes that the matching is better on T7 but in some cases, with missing labels, false positives and stretching inhomogeneities the information can be greatly deformed and thus misinterpreted by the program.

6.2.3 Fine tuning of the program

Due to the presence of relatively important noise and background on some of the intensity profile, a question came to mind, should we analyze the profiles in a more complicated way. The least-square method offers the possibility to account for background by adding a fitting shape (straight line y=1) that it can use to level during the fitting. A second thought was to denoise the data by setting every feature that was lower than a certain value to 0. We tried 4 different possibilities : Not denoised no background line, not denoised with background line, denoised no background line and denoised with background line. When comparing the number of fragment correctly matched (the outcome was known beforehand) and the matching scores,

we could find out that the simplest form actually worked better (no denoising and no background line) for our T7 data. It is likely that our denoising method was not perfect, leading to some loss of information. The background line did improve in some cases but it turns out that in most of the cases the matching got worst(less than 10% of the matching score were better with the line). The reason for this is that the addition of the line renders the system more complex what can lead to errors, particularly when it is not needed. The fact that denoising and background line were not needed is likely linked to the quality of the data. The T7 samples imaged were quite good in term of labeling efficiency and the quality of the images were quite good as well. Hence, we should keep in mind the possibilities that we have for analyzing the data in case we do not have as good data as these ones.

Then, for the best matches, i checked the distribution of the stretching factor that was coming out from it. It turns out that there was a strong bias towards high values such as 2. Simulation of fragment with a random stretching factor were then performed to check whether the program was able to find the correct stretching factor. This was tested for 80 and 50% labeling efficiency and in 100% of the cases, it could find the proper match. This proofed that the bias was not coming from the program.

From the literature, it turns out that a stretching factor higher than 1.7-1.8 starts to greatly affect the structure of the DNA. Actually, stretching in that range is already a state called overstretching of DNA. After an overstretching plateau, the force necessitates to elongate the DNA more increases and would cause further distortion in its structure. [9][8][77] Therefore, the program should not be able to find matches that make sense on a reference with a fully distorted strand. For this reason, the range of the reference tested was restricted to 1.4-1.8.

6.3 Species identification

Before starting to attempt to investigate a complex mixture (two species or more), species should be investigated individually. Indeed, doing so will allow to measure as if it was an unknown sample but knowing the outcome of the program. Therefore, we will be able to assess how well it can actually match, but also the proportion of high scores, the proportion of undetermined and the proportion of wrong matches. It will also allow us to know what are the typical scores we can expect and if the program performs in a consistent way.

6.3.1 T7 DNA

73 intensity profiles were extracted from the STED images acquired on T7. They were analyzed with our procedure and the matching ratio was extracted. 76% of the fragments analyzed gave a score higher than one and thus in favor of T7. However 1 is not a very precise delimitation because the program is not that accurate, thus, a

score of 0.99 or 1.01 have probably the same chances to be T7 than Lambda. As we do not have any data or ideas about how the matching score should behave, the threshold will be arbitrarily set to 1.1 for a good match and 0.9 for a mistake. Figure 6.14 below shows the distribution of the matching scores, the red zone is the zone where the matching score is below 0.9 (valid match towards lambda), the orange zone is the zone where the matching score do not allow to assign a DNA, the green zone is the zone where the fragments are correctly assigned to T7.

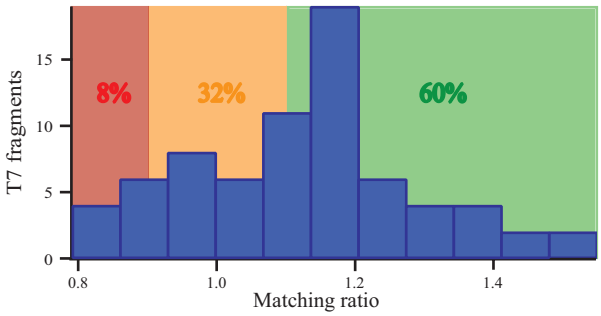

Figure 6.14: Distribution of the matching score ratio for experimental profiles extracted from T7 DNA images. Green : profiles assigned to the correct reference, Orange : unassigned profiles, Red : profiles assigned to the wrong reference. The figure shows that a great proportion of the profiles analyzed could be matched to the correct reference while only few mistakes occurred.

Before even considering the threshold, it can be seen that the maximum of the distribution is largely in favor of T7 DNA. As a consequence, 60% of the DNA are matched with a score higher than 1.1. Those results are a very good sign that the program is able to recognize a DNA based on intensity profiles. In addition we can also see that even with a threshold that is not very strict, we make only 8% of mistakes. By adjusting the criterion, one could make the amount of mistakes smaller but it would also decrease the proportion of the fragment correctly matched and increase the width of the uncertainty zone. As an example, one could set the threshold to be 0.8-1.2 for the uncertainty zone. In that case, we would make no mistakes but we would only match correctly 15-20% of our data. This criterion will thus depend on what the user prefers : matching a lot with a few mistakes or doing less matching but no mistakes. To determine a more meaningful threshold, one would need a much higher statistic, hence, we will keep this one for the remaining of the work.

If we now compare those results with the one obtained in the simulation, we can see that the matching ratio are relatively low. Indeed, for simulation of 20 kb fragments value comprise between 1.5 and 2 were obtained considering 80% labeling efficiency. The first reason that could explain this is that the labeling efficiency of 80% may have been overestimated. Although, if we only consider the labeling efficiency, the scores are still quite low because even at 50% labeling efficiency the score are higher in the simulations (see Figure 6.9). In addition, the matching shown above (Figure

6.13) shows clearly that the labeling efficiency is not that bad. The number of false positives was shown to have almost no influence on the matching score so it is not likely to contribute to the observed effect. The fact that we also used the map in both direction could partially explain this. Indeed, using the flipped reference gives more chances for the wrong match to get a high score and thus will decrease the ratio from a certain amount. However, it is most likely that the noise on the profile, the variation in intensity (does not corresponds to the reference as well as a simulation) and the stretching inhomogeneities are the factor that mainly played a role in these low scores. It is not to be excluded that artefacts could have been introduced from the semi automated profile extraction . The last explanation is that it could also be that we actually studied fragment that were smaller than 20 kb. To be able to verify the latest explanation, we will need the stretching factors extracted from the program. If we now combined all the explanation from above, it is completely understandable that the matching ratios are lower than the one obtained from the simulations.

For the rest of the analysis we will only use the part that gave proper matching scores (green zone). The percentage of uses of flipped and unflipped reference for those data was very close to 50-50 what is a side proof that the program work consistently. We will now check the distribution of the stretching factor to see we can count on our program to extract useful values. The distribution of the stretching factors extracted from the program is shown in Figure 6.15.

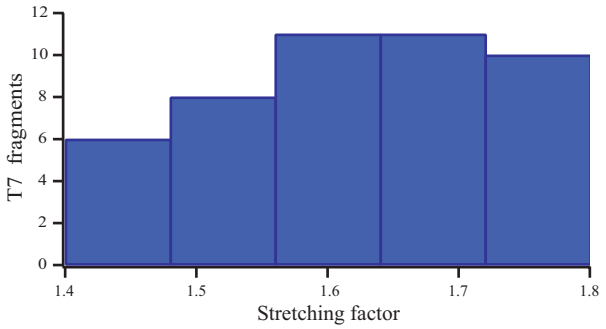

Figure 6.15: Distribution of the stretching factor extracted from the program after matching T7 DNA intensity profiles to T7 reference

The average value of the stretching factors extracted is about 1.63. This is pretty close to the supposed 1.67 factor of the method [19]. The study was performed on Lambda DNA so it could actually be that T7 which is shorter on average tend to be more difficult to stretch. This is a reasonable explanation if we assume a Hookean behavior. [9] Of course, the sample here is rather small (42 fragments), it is thus difficult to get reliable statistics out of it what could explain the difference.

If one looks at the shape of the distribution, it seems like there is a slight trend

towards overstretching. As mentioned before, the all the factor tested here are in the overstretching regime of DNA. This regimes end up with a high increase in force needed to further extent the DNA . It could thus be that the stretching force was not sufficient to break the bound of the DNA and thus stopped where this wall starts (1.7-1.8).[9][8][77] It is therefore more likely to have overstretching than the opposite, in that respect, the distribution makes sense.

The last explanation is that the program does not match a fragment to a reference that is smaller that the fragment. This condition makes sense because a fragment should not be longer than the reference. However, some fragments were quite long and were only tested for 1.6, 1.7 and 1.8, for instance. Such situation can lead to a bias because for 1.6 stretching factor we might try 15 positions, for 1.7, 60 and for 1.8, 100 or so. Hence, when calculating the signal to noise ratio, the match at 1.6 has a disadvantage because the matching peak will weigh more in the standard deviation and the mean of the residual. In this case, no strong bias is observed so it is likely that this is not an issue.

Therefore, the stretching factors extracted seem more than reasonable in term of values and distribution what indicates that the program is able to extract such information. We can then try to use them to get the distribution of sizes of the fragments expressed in base pair. Figure 6.16 shows this distribution.

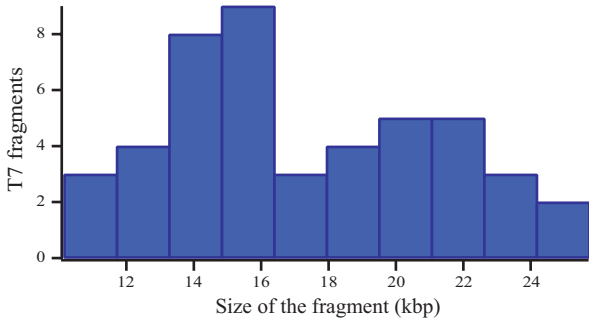

Figure 6.16: Distribution of size for the T7 fragments correctly matched. Bias introduced by the semi-automated extraction of the profiles affected greatly the size distribution of the sample.

The distribution of the sizes does not follow a Gaussian distribution as we could expect the random breaking of the chain to follow. It even seems like there are two populations, one populated with shorter fragments and one populated with longer ones.

The main reason for this is the way the data were extracted. Indeed, getting a nice size distribution was not the purpose here, the aim was rather to get high quality profiles. Hence, a lot of the DNA molecules were not taken in the analysis. The biggest pieces were often split into two pieces to avoid getting too much stretch-

ing inhomogeneities, the shorter ones were avoided because of their probable too poor information content. Moreover, to demonstrate that small and intermediate fragments can be matched with a good accuracy, they were prioritized during the analysis. All these selections induced bias in the size distribution. Yet, what we are really interested in here,is the actual numbers.

We can see that a great part of the DNA studied is actually below 20 kb what can participate in explaining the relatively low matching ratio scores.

After analyzing the distribution of the sizes, the extraction of a few trend was attempted : stretching vs length of the fragment, matching ratio vs length of the fragment, matching ratio vs integration of the fragment. These values are supposedly linked to some extent. It is interesting to check whether we can access these information from our matching or not.

The stretching did not show any clear dependency on the length of the fragment. We could have expected that the smaller DNA fragment would tend to have a lower stretching factor and the longer a higher one. However, the number of strand is probably not sufficient to get such data out.

The matching ratio was depending only slightly on the length of the fragment. The average matching ratio for fragment<16 kb was 1.18 while it was 1.25 for fragment above this value. The stretching inhomogeneities is likely to be the reason why the difference is so small. Indeed, longer DNA will carry more information but takes longer time to elongate on the slide leading to a higher likelihood for stretching variations. On the other hand, shorter DNA will carry less information but the information will be less distorted by stretching variations.

The matching ratio was not depending on the integral of the fragment. The integral extracted was normalized onto the size of the fragment and thus, supposedly representing the amount of information contained in it. However, noise and background variations were probably the issue that made this information not useful at all.

To summarize, it could be demonstrated how the program was able to match properly a great part of our sample. We can thus identify a specie from another one by matching intensity pattern extracted from STED images. These results look very promising because the program we use for data analysis is rather simple, the fact that it performs so well on experimental data gives good hope about what could be done with some improvements. From this matching, we are also able to extract the stretching factor what could be useful to investigate that part of the sample preparation a bit more deeply. Indeed, the influence of the stretching speed should be thoroughly investigated because it is likely that high speed will lead to a more homogeneous stretching but on the other hand, might risk to distort its structure. However, it was never deeply investigated because getting information about the stretching factor is not so easy. If this method would allow us to do it,

it could lead to a better knowledge of the influence of the stretching factor on the resulting pattern matching. We could also extract a size distribution on the sample but one needs to be careful not to induce any bias in the measurement in that case.

Before pursuing with Lambda DNA, it was important to definitely check that STED is indeed performing better than confocal for pattern recognition. Hence, pattern were extracted from confocal images on T7, the resulting matching score distribution is presented in Figure 6.17.

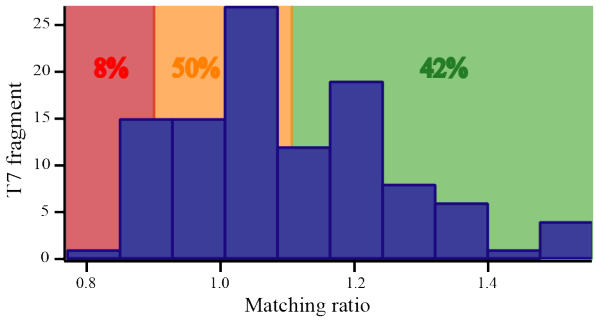

Figure 6.17: Distribution of the matching ratios obtained from intensity profiles extracted from confocal image of T7 DNA. Green : profiles assigned to the correct reference, Orange : unassigned profiles, Red : profiles assigned to the wrong reference. The matching for confocal measurement could assign 42% to the correct reference while not making more mistakes than with STED.

It can be seen from the figure above that a decent matching is obtained using confocal. If one compares with STED, the larger peak of the distribution is in the uncertainty zone whereas it was in the correct matching zone for STED. As a consequence 42% of the strand are matched using confocal whereas 60 % were matched with STED. The improvement that STED bring is not that high, we matched about 1.5 times more fragment thanks to STED. This was however expected from the simulation. Indeed, if one recalls the figures about matching and resolution (Figure 6.2 and Figure 6.3), our setup is actually in the beginning of the zone where super-resolution methods start to bring additional information. This explains why the increase is not so high. However, it still matches better and it totally worth to use STED in this context. The fact that the trend observed accords well with the simulation is a good sign that the way we are treating and analyzing the data is the way to do it and that the program actually kept its consistency even when it came to experimental data.

One can also observed that the amount of mistakes made remained about the same. This is actually a good sign because it seems to indicate that the mistakes we are doing are likely due to errors in data extraction rather than bad pattern recognition. Indeed, if it was due to bad pattern recognition, it is likely that the wrong matches would have increase with the resolution being lower. What is observed here is that the exchange operated between the correct matches and the uncertainty zone

what is a good indication that the pattern are still somewhat recognized but with a lower certainty. This is exactly what we could expect from having a lower resolution.

The fact that the program managed to identify the correct DNA on two completely different sets of data shows that it is fully able to recognize and match a pattern.

One should not forget, when comparing STED and confocal, that we did not exploit the time gating available in STED. This time gating is supposed to improve the resolution of STED by almost a factor two, leading to a resolution comprised between 60 and 80 nm. Hence, STED with time gating is likely to give much better matching than confocal.

Time gating will not be exploited in this thesis because the data did not permit it. Indeed, a part of the data did not have a pixel size small enough to satisfy the Nyquist criterion after gating. Another part of the data did not have enough signal to allow a further discard of photons. Hence, there was too few images that would have allowed to extract intensity profiles and i lacked of time to do additional measurements. It is a bit unfortunate and i have to admit that i am quite curious about what it could lead to.

6.3.2 Lambda DNA

Before to go to the results of the analysis performed on Lambda DNA images, it is important to notify that the images obtained on Lambda DNA were not as good in quality as the one for T7. Even after several attempts, the labeling was not excellent and the background relatively high. It was difficult to track the source of the problem because the labeling depends on many parameters and many chemicals that are prepared by the people of the lab. In addition to this, the null of the doughnut was not as good it could have been even after careful alignment. Decent images could nevertheless be obtained but it is likely that we will need to change the way we analyze the data. For instance, using the background line from the program and/or to denoise the map beforehand. It turned out that denoising and the background line were needed to get a decent identification. This emphasize the need to get as good data as possible beforehand.

Figure 6.18 shows the distribution of matching ratios (Lambda score / T7 score) obtained from Lambda DNA. The same delimitation for the zone of uncertainty was used as for T7 (0.9 to 1.1).
It can be seen from Figure 6.18 that the program was also able to recognize lambda DNA to some extent. Indeed, there is much more DNA on the right part of the distribution than on the left. However, in this case, the amount of data that is uncertainly or wrongly identified is much than for T7. Moreover, we see that the uncertain zone is much more populated for scores below one. It is likely due to the quality of the data that were knew as being inferior to the one obtained for T7. Indeed, noisy data may appear, for the program, as more information dense and

thus could tend to be associated to T7 that is more information dense than Lambda. This could explain why there is a relatively strong bias towards T7 matching. In addition, we know from simulation that for a given size, Lambda DNA tend to have a wider distribution and lower matching scores. This explains partially the smaller proportion of DNA correctly matched. Once again, the size of the fragment was not checked beforehand and could also explain the low scores as well as all the extraction issues described before.

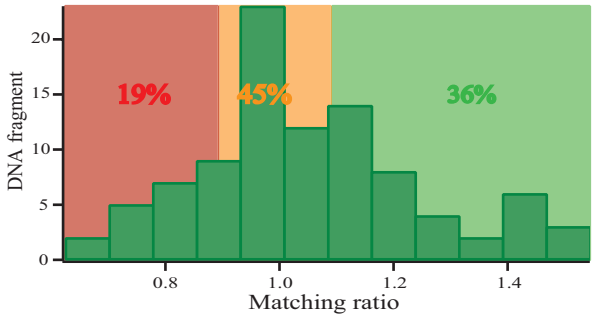

Figure 6.18: Distribution of the matching ratios for intensity profiles extracted from Lambda DNA STED images. Green : profiles assigned to the correct reference, Orange : Unassigned profiles, Red : profiles assigned to the wrong reference. Only 36% were matched to the correct reference. Additionally, the number of mistakes made increased significantly.

It is now clear that STED displays relevance for DNA recognition. Indeed, if one recalls how the confocal was performing, it is likely that in this case, we would get about as many DNA correctly matched as the ones wrongly matched with a large zone of uncertain matches. This would mean that we would not actually be able to identify the specie. However, STED also participated together with the sample in making the images not as good as before.

36% of the molecules could be associated to the right reference and we will keep this part of the sample for the rest of our analysis.Figure 6.19 presents the distribution of the stretching factor obtained for the best matches.

In this case, the distribution is more strongly oriented towards overstretching than it was for T7. However, the average is 1.67 what corresponds to the expected value of the method.[19] The trend for overstretching is stronger than for T7 what makes sense if we consider a Hookean spring.[9] The fact that the distribution tend to increase for 1.7-1.8 could match the idea that it is difficult to stretch more than this (force needed increase a lot) hence more stretching can hardly occur. The quality of the data may also have played a role in the assignment of the stretching factor giving some mistakes. The same explanation as for T7 that Long DNA tend to give poorer signal to noise ratio for small stretching factor than for big ones due to the amount of data point can be used here as well. Finally, the amount of molecules left is

even lower here than it was for T7, it is thus difficult to obtain good statistics out of it.

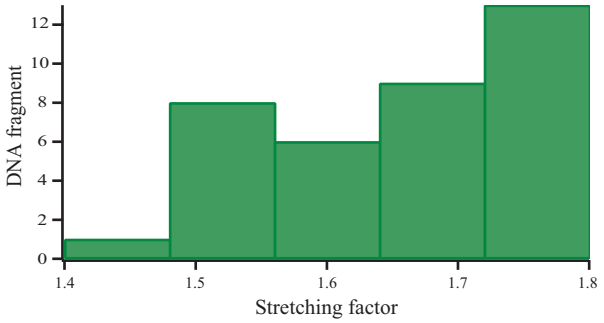

Figure 6.19: Distribution of the stretching factor obtained for Lambda DNA fragments. The distribution is strongly oriented towards overstretching.

We will now use the stretching factor to calculate the size of the fragment in base pair. Figure 6.20 shows the distribution of the sizes of the Lambda DNA properly matched.

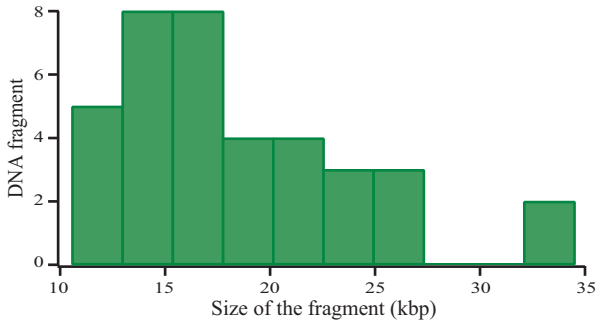

Figure 6.20: Distribution of the sizes obtained from Lambda DNA fragments. The shape of the size distribution resemble a Poisson distribution. This is due to the bias induced by the semi-automated extraction of the data.

It can be seen from the figure above that the maximum of the distribution is around the same position as for T7. We could have expected that Lambda DNA would have a maximum value shifted towards higher value when compared to T7. It is particularly true because the stretching factor distribution tended more towards over stretching. What could explain this is the quality of the sample which was much more fragmented than the one for T7.

In this case, the shape of the distribution look much nicer than the one for T7. The fact that very short pieces are not present in the distribution is due to the fact that they were avoided when analyzing the data what introduces a bias in a distribution supposedly normal leading to a shape closer to a Poisson distribution.

From the results obtained for T7 and Lambda, we can now say that our method performs relatively well when it comes to DNA identification. It can also give a relatively good idea about the size and the stretching factor obtained in the method even though some bias were enlightened. However, it was shown to be very sensitive to the quality of the data. A solution to this issue would be to be able to choose whether we want to save the profile or not what is not possible right now. This would allow the user to filter off the profiles that look too noisy or that have a large amount of background. Moreover, one could therefore analyze all the strands on the image and keep the best looking profiles for the analysis instead of trying to filter off the DNA according to the way they look. Thus, the filtration will be based on how the profile actually look which is likely to be more accurate.

We also learned that the method as it was applied gave a certain amount of mistakes. One could enlarge the zone of uncertainty in order to do less mistakes but it will also decrease the amount of DNA that are matched to the right reference. It also seemed that noisier data tend to be attributed to T7 due to its higher information content. This could explain together with the quality of the data that the identification of T7 works better than for Lambda DNA.

With what we learn from the previous experiment, we can try to study a mixture of those DNA. The method, in the state it is now, is probably not accurate enough to be able to assess the proportion of each DNA due to the amount of mistakes we are doing. Hence, the main purpose will be to verify whether we are able to distinguish two populations or not. It will also allow us to compare the distribution of stretching factors and sizes in presence of the same experimental bias. Indeed, in this case, both DNA are treated in the same way because i did not know which one was which beforehand.

6.4 Species differentiation

A mixture of T7 and Lambda DNA was prepared in a 1:2 ratio. This ratio was in favor of e.g a 1:1 ratio to ensure that the difference in concentration would still be significant after all sample preparation procedure which might introduce biais(e.g. purification). Hence, once again, the complexity of the sample was increased but a known parameter remains allowing to verify the meaning of the results. However, as mentioned earlier, it is clear that the quantification will be difficult and that improvement to the method should be brought before.

This mixture was prepared the same day as the Lambda DNA and measured in similar conditions. Hence, we will keep the same parameter for the data analysis (Background line and denoising).

For this part, we will define the matching ratio in a different way, it will always

be T7 score/ Lambda score. Hence, a score superior to 1.1 will be considered as a match for T7, a score below 0.9 will be considered as a match for lambda. Figure 6.21 shows the distribution of the matching scores obtained (Same color convention, green for Lambda blue for T7).

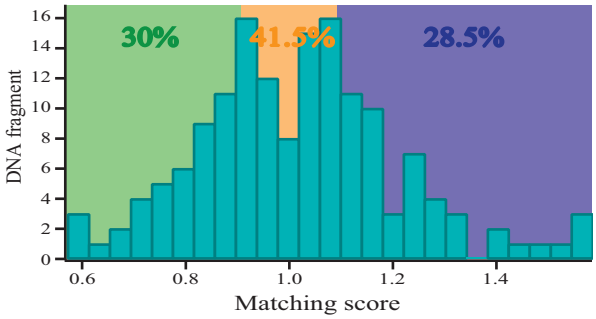

Figure 6.21: Distribution of the matching scores from the intensity profile extracted from STED image of a 1:2 T7:Lambda mixture. Green: profiles assigned to Lambda DNA, Orange: Unassigned profiles, Blue: profiles assigned to T7 DNA. Two population distributions can be distinguished, Lambda DNA is slightly more represented than T7 DNA.

It can be directly seen that the shape of the distribution is much different than before and that it is very likely to corresponds to two populations. Indeed, we have two maximums, a first around 0.90-0.95 and a second about 1.05-1.1. Those maximum respectively corresponds to what would be expected for Lambda and T7. Moreover, it can be seen that the proportion of the matches for T7 and lambda are very similar what is very different to the proportion we were getting when only one specie was present. It is a good indication that the method could be applicable for specie differentiation. However, we were expecting to have a 1:2 proportion between Lambda and T7 and it is not really the case. Several explanations can be used to justify this.

The first one is that the purification of the DNA after labeling might have affected the ratio provided initially. It would have probably be better, in a first time, to mix two solutions of DNA labeled and purified separately allowing to have a more precise idea about their respective concentration. The reason why it was not performed that way was to achieve an experiment that would resemble as much as possible the final application where DNA are extracted from an unknown sample (e.g. bacterias, viruses) and labeled together. In this case, a step was jumped in a way but distinguishing two populations was the main purpose here anyhow.

The second reason is that we have seen before that Lambda DNA tends to have a stronger bias towards T7 than the opposite situation. Hence, taking this into account in our calculation would lead to proportion close to the ratio prepared. However, to be able to take this into account we would need larger amount of data in order to do

it in a relevant way.

The third reason is that only 155 strands were studied. It is definitely not enough to be able to characterize the macroscopic proportion based on single molecule observations. Indeed, it could simply be that in the images and the strands analyzed there was about as many T7 than Lambda.

Nevertheless, two populations could be distinguished. Each one of those populations will now be used to compare their stretching factor and their size distribution. It is particularly interesting to do it in this case because the two populations were prepared and extracted in the same way and thus should have the same bias if any. Hence, the comparison here is much more relevant than before. Figure 6.22 compares the stretching factor obtained for the population assigned to Lambda and the one assigned to T7.

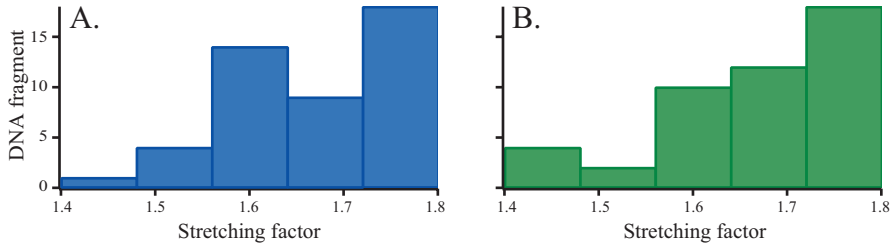

Figure 6.22: Distribution of the stretching factor extracted from the matching of A. T7-assigned population of the mixture B. Lambda-assigned population of the mixture. Both distributions look similar, they have a trend towards overstretching. It is however more pronounced for Lamda what confirms what was already observed before.

The average in both case is about 1.68. Once again, it is quite close to the 1.67 expected from the deposition method.[19] However, both distributions have a clear trend for overstretching. Similarly to what was observed before, Lambda DNA has a more pronounced trend for overstretching than T7. The fact that the trend in the outputs of the program remain consistent for a specie is an indication that the value are not find randomly and that they indeed are meaningful. However, the trend for overstretching of T7 is more pronounced here. It is clear that number of fragment studied does not allow to get really reliable data and that much more should be studied to get the trend out.

The fact that the trend for overstretching was present on all the data means that the program remain consistent over the different sample. This consistency can either mean that there is a bias in the attribution of the stretching factor by the program or that the program is actually measuring the proper stretching factor. In that case, it seems that the explanation for overstretching due to the strong increase of force needed to stretch the DNA more than 1.7-1.8 times its size is the reason why a great

proportion of the DNA studied got stretched as much.[9][8][77]

The overstretching could also be explained by a bias towards overstretching mentioned before. This bias could be corrected with the help of a factor of probability. Indeed, if we know that a certain stretching factor is more probable for a certain size, we could correct the score to account for this. However, one should pay attention because the distribution should not be forced to look like we want it to look. To do this correction in a smart way, we would need statistically significant data about the stretching factor on DNA stained with YOYO dye. Then, a factor could be determined and used to remove the matching on 1.8 when its score is close to the matching to another stretching factor more probable. In addition, it should also improve the number of molecules correctly matched because it is not rare to see a good match on one DNA being "killed" by a better matched on the other DNA with an extreme stretching factor (either too small or too large).

This time we will confront the size distribution obtained in micrometer and the one in kilobase for both populations. Figure 6.23 shows the confrontation of the size distributions for both populations differentiated.

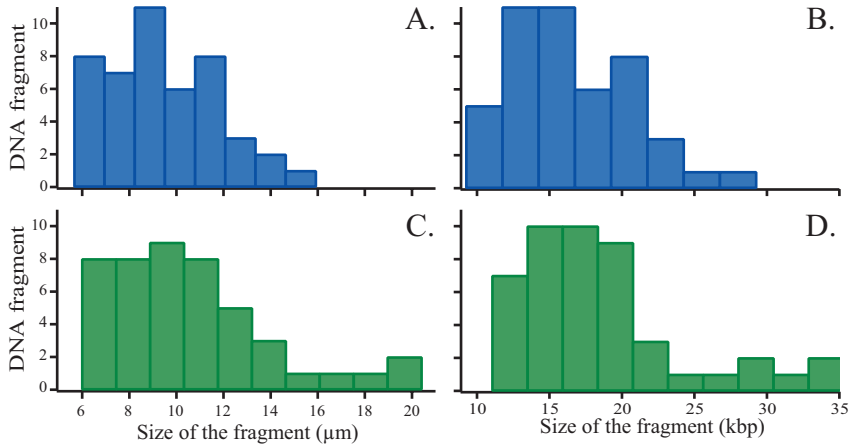

Figure 6.23: Size distribution of DNA molecules for the T7-assigned population of the mixture expressed in micrometer (A) and in kbp (B) and for the Lamda-assigned population in micrometer (C) and in kbp (D). The distribution tend to narrow when being calculating the size in base pairs. It indicates that small fragment tends to have smaller stretching factor and larger fragment tend to have larger stretching factor.

We will first focus onto the two figures on the left. The length distribution of the fragment is rather similar for both DNA. Indeed, most of the DNA have a length comprised between 6 and 12 μm for both T7 and lambda DNA and they have the same cutoff value of about 6 μm. We can clearly see here that the same bias onto small molecules was introduced by myself.

On average, the Lambda molecules are a bit longer than the T7 ones. This was expected because lambda is longer than T7. It can be clearly seen that all the molecules larger than 16 μm were assigned to Lambda. However T7 with a 1.6 stretching factor is about 21 μm and could thus perfectly be 16 μm long. It could imply the presence of a bias from the program due to the fact that long DNA have less attempt on T7 than Lambda what often leads to a poorer signal to noise ratio. It is of course, difficult to say because the number of molecule above this length is too small (4) to really draw conclusions.

From the figure on the right, it can be seen that T7 distribution kept a very similar shape even though a slight narrowing is observed. The same effect but more pronounced is observed for Lambda DNA. The fact that the distribution narrowed is due to the fact that the smaller fragment were on average less stretched than the longer fragment. Indeed, if that is the case, a small fragment with a small stretching factor will carry a bit more information (in bp) that what we could guess from its side whereas a longer fragment with a high stretching factor will carry less information than what we think from its size. Carrying more or less information is considered with respect to a 1.6 stretching factor. Hence, we can clearly emphasize here that the stretching factor are somewhat consistent with the length even though a tendency for overstretching was present.

6.5 Conclusions

Thanks to the use of simulation, a program that is able to match an experimentally obtained intensity pattern to a bank of reference could be optimized. From these simulations, the parameters that influence the matching could be determined and their respective effect quantified. Those parameters are :

- **Size of the fragment:** The longer the fragments, the higher the amount of information and the higher the resulting matching score will be
- **Resolution:** The matching score increases slowly with the resolution for resolution lower than the average distance between two sites. However, it increases dramatically as the resolution pass this threshold
- **Density of information:** fragments generated from low information density regions are more likely to give high scores than the ones generated from high information density regions. Larger variations of the information content on Lambda DNA were shown to cause its matching score distribution to be broader. Eventually, the region with similar information density tended to resemble. This effect was more pronounced for high density regions
- **Labeling efficiency:** a threshold value labeling efficiency is needed to get specific information. This threshold depend on the density of information on

the DNA and the resolution of the pattern. When this minimum is reached, the matching score scales linearly with the labeling efficiency. Eventually, a second threshold is reached and the matching score evolves slowly towards a maximal value. This second threshold is due to a saturation of information, it follows similar dependencies as the first threshold discussed

- **False positive:** The number of false positive was shown to have almost no influence on the matching score. The effect was nevertheless more pronounced on lower information density regions (e.g. Lambda vs T7). The influence is not pronounced because adding non-specific information does not mean adding anti-specific information. Hence, the effect of the numerous labeled correctly placed was not balanced by the wrongly placed ones

Thanks to the use of the program developed, analysis of experimental intensity pattern extracted from STED images could be performed. STED was shown to yield more detailed intensity patterns than confocal imaging. Similar patterns could be found among different DNA molecules and could subsequently matched to the correct reference. These facts were used as a proof that the presented method can work to identify a specie from its labeled DNA intensity pattern. On images containing only one specie, identification of the specie was achieved by comparing the intensity pattern to a bank of references. The quality of the images and the pattern extracted was shown to be crucial for the pattern recognition. In addition, patterns extracted from confocal data were demonstrated to yield lower pattern recognition than with STED. From the matching, information about the stretching factor and the size distribution of the sample could also be extracted. For these values to be reliable, it is believed that more measurement are needed as well as a side verification.

Finally, on a sample prepared from two different DNA, differentiation of two population could be successfully achieved. The amount of molecule in the study was however not sufficient to allow to track the macroscopic proportions of the sample.

Part IV

Conclusions & Perspectives

A STED setup was successfully built and its performances assessed:

- Laser sources can provide a power output of 1.4 mW and 250 mW for the excitation and STED laser respectively
- Maximal field of view is limited to 55 um due to the deformation of the doughnut for scanning position beyond this distance
- Saturation intensities measured were comprised between 10-50 mW depending on the dye and the excitation power used
- Achievable resolution is 120 nm for STED and 70 nm for time-gated detection

These performances could be reached on a daily basis showing a good overall stability of the setup. Effort could be directed towards a thorough characterization of the ability of the setup to provide dynamic information using normal scanning mode and compare the performances obtained with the one obtained with other dynamic approaches such as STED-FCS. Indeed, the speed of STED is much higher than other resolution methods (e.g PALM, STORM) that can not achieve a high frame rate. Studying the ability of the setup to provide dynamic information is therefore particularly relevant.

The ability of the setup to perform imaging on sequence-specific labeled DNA was demonstrated under certain conditions:

- The use of dyes highly resistant to photobleaching (e.g. Atto dyes)
- The use of Imaging buffer to increase the dye brightness and/or to reduce the photobleaching
- The use of high spin-coating rates for zeonex to reduce the thickness of the film and increase its homogeneity

Even though high resolution images could be obtained, a deeper investigation would be needed. Indeed, time gated detection was not used and could offer a further gain in resolution what could be crucial for the resulting intensity pattern. In addition, further investigation on the effect of the spin-coating rate should be performed, for instance, by studying the surface thickness and roughness by e.g. AFM for different coating settings. In addition, improvement of the labeling method should be addressed to render parameters such as labeling efficiency more reproducible. This is crucial to ensure a good quality and predictability of the results.

A program to analyze intensity pattern was written and its performances to match an intensity pattern were tested on simulated DNA intensity patterns. The simulations also allowed to get an insight onto the experimental parameters that influence the matching score:

- **Size of the fragment:** The longer the fragments, the higher the amount of information and the higher the resulting matching score will be
- **Resolution:** The matching score increase slowly with the resolution for resolution smaller than the average distance between sites (150 nm) but increases dramatically for better resolutions
- **Density of information:** fragments originated from low density region of the genome tend to give higher score than the ones originated from high density regions. In addition, the later tend to have a higher degree of similarity between the two species
- **Labeling efficiency:** The matching score was greatly influenced by the labeling efficiency. For very low labeling efficiency the information is not specific and matching score ratio is close to one. Then, after a certain threshold is reached, the matching ratio scales almost linearly with the labeling efficiency. Finally, a saturation threshold is reached and the matching ratio increase slowly to its maximal value. Both threshold depends on the density of information on the strand and the resolution
- **False positive:** False positives do not present a significant influence onto the matching score. Their effect is however dependent on the density of information and the resolution and can thus become more significant

The density of information and the resolution remain the parameters than influence the most the matching score. This is because they influence the effect of all the other parameters. Comparison with statistically significant experimental data should be needed to confirm those trends experimentally.
Matching of experimental profiles was successfully performed for T7 DNA, Lambda DNA and a 1:2 mixture of both specie :

- **T7 DNA STED image:** 60 % of the experimental profile extracted from STED images could be associated to the correct reference while doing only 8% mistakes
- **T7 DNA Confocal image:** 42 % of of the experimental profile extracted could be associated to the correct reference while doing only 8% mistakes
- **Lambda DNA STED image:** 40 % of the experimental profile extracted from STED images could be associated to the correct reference while doing only 19% mistakes
- **Mixture STED image:** 30% of the DNA were associated to lambda and 28.5 % were associated to T7, the rest was undetermined. Two populations could be clearly distinguished even though the amount of mistakes is not known

The method used showed a good ability to identify a specie in simple samples as well as more complex ones. Those results demonstrate that intensity profile matching

methods are sufficiently resistant to stretching variation to be applicable to DNA differentiation. This is in opposition to what was postulated in Johem's work. This work opens to DNA differentiation a range of super-resolution techniques such as SIM and SOFI that yield a super resolution image rather than coordinate and for which Smith-Waterman approach is not suited.

However, the relevance of STED for such application would need a further investigation by the use, for instance, of time gating. Indeed, confocal microscopy was not performing significantly less than STED while being much easier to implement. Hence, showing that time gating provides a significant improvement in the differentiation of DNA would bring the final justification of the relevance of STED for this particular application. Other methods such as SOFI should also be tested and compared to what STED can provide. A comparison of the performances of STORM and/or PALM matching with Smith-Waterman algorithm with STED, SOFI and /or SIM using intensity pattern matching would be very interesting to know more precisely on which approach, efforts should be focused on. In addition, larger amount of DNA molecules should be studied and compared to a larger databank to confirm the robustness of the program and its ability to match in conditions closer to the final application.

The program performances were quite good already. However, the algorithm is rather simple and improvement could probably be made easily. By acquiring a deeper knowledge of the stretching factor in routine conditions, we could discriminate some wrong matches by the addition of a probability factor that will represent the likelihood for a certain size of fragment to have a certain stretching factor.

Machine learning is also an approach that could be used to improve the DNA recognition. It consists in giving the program a huge amount of data (e.g. intensity patter) while giving him, in our case, a parameter to which the data should be associated to (e.g. DNA corresponding). By providing the program with a sufficient amount of data, it could be able to assign a new set of data to the correct parameter. However, such implementation is quite heavy and one should first think about the time that such a program would take to build and balance this with the benefits we would retrieve from it.

Bibliography

[1] C. Alkan, S. Sajjadian, and E. E. Eichler. Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1):61–65, 2011. ISSN: 1548-7091.

[2] G. E. Ananiev, S. Goldstein, R. Runnheim, D. K. Forrest, S. Zhou, K. Potamousis, C. P. Churas, V. Bergendahl, J. A. Thomson, and D. C. Schwartz. Optical mapping discerns genome wide DNA methylation profiles. *Bmc molecular biology*, 9:68, 2008. ISSN: 1471-2199.

[3] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *The journal of experimental medicine*, 79(2):137–158, 1944.

[4] A Bensimon, A Simon, A Chiffaudel, V Croquette, F Heslot, and D Bensimon. Alignment and sensitive detection of DNA by a moving interface. *Science*, 265(5181):2096–2098, 1994. ISSN: 0036-8075.

[5] D. Bensimon, A. J. Simon, V. Croquette, and A. Bensimon. Stretching DNA with a Receding Meniscus: Experiments and Models. *Physical review letters*, 74(23):4754–4757, 1995. ISSN: 0031-9007.

[6] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess. Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science*, 313(5793):1642–1645, 2006. ISSN: 0036-8075.

[7] D. Busko, S. Baluschev, D. Crespy, A. Turshatov, and K. Landfester. New possibilities for materials science with STED microscopy. *Micron*, 43(5):583–588, 2012. ISSN: 0968-4328.

[8] C. Bustamante, Z. Bryant, and S. B. Smith. Ten years of tension: single-molecule DNA mechanics. *Nature*, 421(6921):423–427, 2003. ISSN: 0028-0836.

[9] C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith. Single-molecule studies of DNA mechanics. *Current opinion in structural biology*, 10(3):279–285, 2000. ISSN: 0959440X.

[10] W Cai, H Aburatani, V. P. Stanton, D. E. Housman, Y. K. Wang, and D. C. Schwartz. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the national academy of sciences*, 92(11):5164–5168, 1995. ISSN: 0027-8424.

[11] R. Chéreau, J. Tønnesen, and U. V. Nägerl. STED microscopy for nanoscale imaging in living brain slices. *Methods*, 88:57–66, 2015. ISSN: 1046-2023.

[12] H Chial. DNA sequencing technologies key to the Human Genome Project. *Nature education*, 1(1):219, 2008.

[13] F Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, 1970.

[14] F. H. C. Crick, L. Barnett, S Brenner, and R. J. Watts-Tobin. General Nature of the Genetic Code for Proteins. *Nature*, 192(4809):1227–1232, 1961.

[15] R. Dahm. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581, 2008. ISSN: 1432-1203.

[16] M. W. Davidson and M. Abramowitz. Optical Microscopy. In, *Encyclopedia of imaging science and technology*. John Wiley & Sons, Inc., 2002.

[17] P. Dedecker, B. Muls, A. Deres, H. Uji-i, J.-i. Hotta, M. Sliwa, J.-P. Soumillion, K. Müllen, J. Enderlein, and J. Hofkens. Defocused Wide-field Imaging Unravels Structural and Temporal Heterogeneity in Complex Systems. *Advanced materials*, 21(10-11):1079–1090, 2009. ISSN: 09359648.

[18] J. Deen. High resolution DNA mapping for species identification. Doctoral Thesis. KU Leuven, 2016, p. 143.

[19] J. Deen, W. Sempels, R. De Dier, J. Vermant, P. Dedecker, J. Hofkens, and R. K. Neely. Combing of Genomic DNA from Droplets Containing Picograms of Material. *Acs nano*, 9(1):809–816, 2015. ISSN: 1936-0851.

[20] G. T. Dempsey, J. C. Vaughan, K. H. Chen, M. Bates, and X. Zhuang. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat meth*, 8(12):1027–1036, 2011. ISSN: 1548-7091.

[21] T Dertinger, R Colyer, G Iyer, S Weiss, and J Enderlein. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proceedings of the national academy of sciences of the united states of america*, 106(52):22287–22292, 2009. ISSN: 0027-8424.

[22] T. Dertinger, R. Colyer, R. Vogel, J. Enderlein, and S. Weiss. Achieving increased resolution and more pixels with Superresolution Optical Fluctuation Imaging (SOFI). *Optics express*, 18(18):18875–18885, 2010.

[23] E. L. v. Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014. ISSN: 01689525.

[24] J. N. Farahani, M. J. Schibler, and L. A. Bentolila. Stimulated Emission Depletion (STED) Microscopy: from Theory to Practice. *Microscopy: science, technology, applications and education*, 2(4):1539–1547, 2010.

[25] A. Grunwald, M. Dahan, A. Giesbertz, A. Nilsson, L. K. Nyberg, E. Weinhold, T. Ambjörnsson, F. Westerlund, and Y. Ebenstein. Bacteriophage strain typing by rapid single molecule analysis. *Nucleic acids research*, 2015.

[26] D. B. Hall, P. Underhill, and J. M. Torkelson. Spin coating of thin and ultrathin polymer films. *Polymer engineering & science*, 38(12):2039–2045, 1998. ISSN: 0032-3888.

[27] B. Harke, J. Keller, C. K. Ullal, V. Westphal, A. Schönle, and S. W. Hell. Resolution scaling in STED microscopy. *Optics express*, 16(6):4154–4162, 2008.

[28] S. W. Hell. Far-Field Optical Nanoscopy. *Science*, 316(5828):1153–1158, 2007.

[29] S. W. Hell. Increasing the Resolution of Far-Field Fluorescence Light Microscopy by Point-Spread-Function Engineering BT - Topics in Fluorescence Spectroscopy: Volume 5: Nonlinear and Two-Photon-Induced Fluorescence. In J. R. Lakowicz, editor, pp. 361–426. Springer US, Boston, MA, 2002.

[30] S. W. Hell. Microscopy and its focal switch. *Nat meth*, 6(1):24–32, 2009. ISSN: 1548-7091.

[31] S. W. Hell. Toward fluorescence nanoscopy. *Nat biotech*, 21(11):1347–1355, 2003. ISSN: 1087-0156.

[32] S. W. Hell, M. Dyba, and S. Jakobs. Concepts for nanoscale resolution in fluorescence microscopy. *Current opinion in neurobiology*, 14(5):599–609, 2004. ISSN: 0959-4388.

[33] S. W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics letters*, 19(11):780, 1994. ISSN: 0146-9592.

[34] H. Hermann, W. Pitschke, and N. Mattern. Surface Roughness of Porous Materials and Its Characterization by X-Ray Absorption Measurements. *Physica status solidi (a)*, 132(1):103–114, 1992. ISSN: 00318965.

[35] S. T. Hess, T. P. Girirajan, and M. D. Mason. Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. *Biophysical journal*, 91(11):4258–4272, 2006. ISSN: 00063495.

[36] B. Huang, M. Bates, and X. Zhuang. Super resolution fluorescence microscopy. *Annual review of biochemistry*, 78:993–1016, 2009. ISSN: 0066-4154.

[37] K. Jo, D. M. Dhingra, T. Odijk, J. J. d. Pablo, M. D. Graham, R. Runnheim, D. Forrest, and D. C. Schwartz. A single-molecule barcoding system using nanoslits for DNA analysis. *Proceedings of the national academy of sciences of the united states of america*, 104(8):2673–8, 2007. ISSN: 0027-8424.

[38] A. W. Jones, J Bland-Hawthorn, and P. L. Shopbell. Towards a general definition for spectroscopic resolution. In *Astronomical data analysis software and systems iv*. Vol. 77, 1995, p. 503.

[39] R. Kasper, B. Harke, C. Forthmann, P. Tinnefeld, S. W. Hell, and M. Sauer. Single-Molecule STED Microscopy with Photostable Organic Fluorophores. *Small*, 6(13):1379–1384, 2010. ISSN: 1613-6829.

[40] S. Kim, A. Gottfried, R. R. Lin, T. Dertinger, A. S. Kim, S. Chung, R. A. Colyer, E. Weinhold, S. Weiss, and Y. Ebenstein. Enzymatically Incorporated Genomic Tags for Optical Mapping of DNA-Binding Proteins. *Angewandte chemie international edition*, 51(15):3578–3581, 2012. ISSN: 14337851.

[41] T. A. Klar and S. W. Hell. Subdiffraction resolution in far-field fluorescence microscopy. *Optics letters*, 24(14):954–956, 1999.

[42] T. A. Klar, S. Jakobs, M. Dyba, A. Egner, and S. W. Hell. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proceedings of the national academy of sciences*, 97(15):8206–8210, 2000.

[43] S. Klimasauskas and E. Weinhold. A new tool for biotechnology: AdoMet-dependent methyltransferases. *Trends in biotechnology*, 25(3):99–104, 2007. ISSN: 01677799.

[44] J. R. Lakowicz, ed. *Principles of Fluorescence Spectroscopy*. Springer US, Boston, MA, 2006.

[45] G. Laporte and D. Psaltis. STED imaging of green fluorescent nanodiamonds containing nitrogen-vacancy-nitrogen centers. *Biomedical optics express*, 7(1):34–44, 2016. ISSN: 2156-7085.

[46] M. A. Lauterbach and C. Eggeling. Foundations of Sted Microscopy. In, pp. 41–71, 2014.

[47] S. L. Levy and H. G. Craighead. DNA manipulation, sorting, and mapping in nanofluidic systems. *Chemical society reviews*, 39(3):1133–1152, 2010. ISSN: 0306-0012.

[48] M. Levy-Sakin, A. Grunwald, S. Kim, N. R. Gassman, A. Gottfried, J. Antelman, Y. Kim, S. O. Ho, R. Samuel, X. Michalet, R. R. Lin, T. Dertinger, A. S. Kim, S. Chung, R. A. Colyer, E. Weinhold, S. Weiss, and Y. Ebenstein. Toward Single-Molecule Optical Mapping of the Epigenome. *Acs nano*, 8(1):14–26, 2014. ISSN: 1936-0851.

[49] S. v. d. Linde, A. Löschberger, T. Klein, M. Heidbreder, S. Wolter, M. Heilemann, and M. Sauer. Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nature protocols*, 6(7):991–1009, 2011. ISSN: 1750-2799.

[50] G. Lukinavicius, V. Lapiene, Z. Stasevskij, C. Dalhoff, E. Weinhold, and S. Klimasauskas. Targeted Labeling of DNA by Methyltransferase-Directed Transfer of Activated Groups (mTAG). *Journal of the american chemical society*, 129(10):2758–2759, 2007. ISSN: 0002-7863.

[51] L. Mendelowitz and M. Pop. Computational methods for optical mapping. *Gigascience*, 3(1):33, 2014. ISSN: 2047-217X.

[52] X Meng, K Benson, K Chada, E. J. Huff, and D. C. Schwartz. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature genetics*, 9(4):432–8, 1995. ISSN: 1061-4036.

[53] M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews. genetics*, 11(1):31–46, 2010. ISSN: 1471-0064.

[54] P. M. Milos. Emergence of single-molecule sequencing and potential for molecular diagnostic applications. *Expert review of molecular diagnostics*, 9(7):659–666, 2009. ISSN: 1473-7159.

[55] M Minsky. Memoir on inventing the confocal scanning microscope. *Scanning*, 10(4):128–138, 1988. ISSN: 1932-8745.

[56] M Minsky. Microscopy apparatus. 1961.

[57] E. Mohajerani, F. Farajollahi, R. Mahzoon, and S. Baghery. Morphological and thickness analysis for PMMA spin coated films. *Journal of optoelectronics and advanced materials*, 9(12):3901–3906, 2007. ISSN: 14544164.

[58] G. Moneron, R. Medda, B. Hein, A. Giske, V. Westphal, and S. W. Hell. Fast STED microscopy with continuous wave fiber lasers. *Optics express*, 18(2):1302–1309, 2010.

[59] P. Muller-Buschbaum, J. S. Gutmann, M. Wolkenhauer, J. Kraus, M. Stamm, D. Smilgies, and W. Petry. Solvent-Induced Surface Morphology of Thin Polymer Films. *Macromolecules*, 34(5):1369–1375, 2001. ISSN: 0024-9297.

[60] R. K. Neely, P. Dedecker, J.-i. Hotta, G. Urbanaviciute, S. Klimasauskas, and J. Hofkens. DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chemical science*, 1(4):453–460, 2010. ISSN: 2041-6520.

[61] R. K. Neely, J. Deen, and J. Hofkens. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95(5):298–311, 2011. ISSN: 1097-0282.

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

[63] a. I. Oliva, V. Sosa, and J. L. Pena. Effect of indium tin oxide substrate roughness on the morphology , structural and optical properties of CdS thin films. *Applied surface science*:340–346, 2000. ISSN: 01694332.

[64] S. O. Oyola, T. D. Otto, Y. Gu, G. Maslen, M. Manske, S. Campino, D. J. Turner, B. MacInnis, D. P. Kwiatkowski, H. P. Swerdlow, and M. A. Quail. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *Bmc genomics*, 13(1):1, 2012. ISSN: 1471-2164.

[65] F. Persson and J. O. Tegenfeldt. DNA in nanochannels-directly visualizing genomic information. *Chemical society reviews*, 39(3):985–999, 2010. ISSN: 0306-0012.

[66] L Pray. Discovery of DNA structure and function: Watson and Crick. *Nature education*, 1(1), 2008.

[67] X. Qu, D. Wu, L. Mets, and N. F. Scherer. Nanometer-localized multiple single-molecule fluorescence microscopy. *Proceedings of the national academy of sciences*, 101(31):11298–11303, 2004. ISSN: 0027-8424.

[68] M. Reuss. Simpler STED setups. *Dissertation*, 2010.

[69] R. Riehn, M. Lu, Y.-M. Wang, S. F. Lim, E. C. Cox, and R. H. Austin. Restriction mapping in nanofluidic devices. *Proceedings of the national academy of sciences*, 102(29):10012–10016, 2005. ISSN: 0027-8424.

[70] P. W. Rigby, M Dieckmann, C Rhodes, and P Berg. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *Journal of molecular biology*, 113(1):237–51, 1977. ISSN: 0022-2836.

[71] M. J. Rust, M. Bates, and X. Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods*, 3(10):793–795, 2006. ISSN: 1548-7091.

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

[73] I. Schoen. Localization Precision in Stepwise Photobleaching Experiments. *Biophysical journal*, 107(9):2122–2129, 2014. ISSN: 00063495.

[74] D. Schwartz, X Li, L. Hernandez, S. Ramnarain, E. Huff, and Y. Wang. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262(5130):110–114, 1993. ISSN: 0036-8075.

[75] W. Sempels. Fluorescence microscopy for fast and local scale dynamics in suspensions, 2015.

[76] P. F. Smith, I. Chun, G. Liu, D. Dimitrievich, J. Rasburn, and G. J. Vancso. Studies of optical haze and surface morphology of blown polyethylene films using atomic force microscopy. *Polymer engineering & science*, 36(16):2129–2134, 1996. ISSN: 0032-3888.

[77] S. B. Smith, Y. Cui, and C. Bustamante. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science*, 271(5250):795–799, 1996. ISSN: 0036-8075.

[78] C. M. St Croix, S. H. Shand, and S. C. Watkins. Confocal microscopy: comparisons, applications, and problems. *Biotechniques*, 39(6 Suppl), 2005. ISSN: 19409818.

[79] J. F. Thompson and P. M. Milos. The properties and applications of single-molecule DNA sequencing. *Genome biology*, 12(2):1–10, 2011. ISSN: 1474-760X.

[80] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews genetics*, 2011. ISSN: 1471-0056.

[81] B. Valeur and M. N. Berberan-Santos. Environmental Effects on Fluorescence Emission. In, *Molecular fluorescence*, pp. 109–140. Wiley-VCH Verlag GmbH & Co. KGaA, 2012.

[82] B. Valeur and M. N. Berberan-Santos. Introduction. In, *Molecular fluorescence*, pp. 1–30. Wiley-VCH Verlag GmbH & Co. KGaA, 2012.

[83] B. Valeur and M. N. Berberan-Santos. Structural Effects on Fluorescence Emission. In, *Molecular fluorescence*, pp. 75–107. Wiley-VCH Verlag GmbH & Co. KGaA, 2012.

[84] A. Valouev, D. C. Schwartz, S. Zhou, and M. S. Waterman. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the national academy of sciences*, 103(43):15770–15775, 2006. ISSN: 0027-8424.

[85] G. Vicidomini, G. Moneron, K. Y. Han, V. Westphal, H. Ta, M. Reuss, J. Engelhardt, C. Eggeling, and S. W. Hell. Sharper low-power STED nanoscopy by time gating. *Nat meth*, 8(7):571–573, 2011. ISSN: 1548-7091.

[86] G. Vicidomini, A. Schonle, H. Ta, K. Y. Han, G. Moneron, C. Eggeling, and S. W. Hell. STED Nanoscopy with Time-Gated Detection: Theoretical and Experimental Aspects. *Plos one*, 8(1):e54421, 2013.

[87] S. B. W, N. Murthy, M Krishna, and S. C. Sharma. Investigation of Influence of Spin Coating Parameters on the Morphology of ZnO Thin Films by Taguchi Method. *Int. j. thin film sci. tec*, 2(2):143–154, 2013. ISSN: 2090-9519.

[88] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.

[89] R. H. Webb. Confocal optical microscopy. *Reports on progress in physics*, 59(3):427–471, 1999. ISSN: 0034-4885.

[90] R Wegerhoff, O Weidlich, and M Kassens. Basics of light microscopy and imaging. *Imaging & microscopy*:56, 2007.

[91] V. Westphal, S. O. Rizzoli, M. a. Lauterbach, D. Kamin, R. Jahn, and S. W. Hell. Video-rate far-field optical nanoscopy dissects synaptic vesicle movement. *Science (new york, n.y.)*, 320(5873):246–249, 2008. ISSN: 0036-8075.

[92] K. I. Willig, B. Harke, R. Medda, and S. W. Hell. STED microscopy with continuous wave beams. *Nature methods*, 4(11):915–918, 2007. ISSN: 1548-7091.

[93] M. Xiao, A. Phong, C. Ha, T.-F. Chan, D. Cai, L. Leung, E. Wan, A. L. Kistler, J. L. DeRisi, P. R. Selvin, and P.-Y. Kwok. Rapid DNA mapping by fluorescent single molecule detection. *Nucleic acids research*, 35(3):e16–e16, 2007. ISSN: 0305-1048.

[94] S. Zhou, F. Wei, J. Nguyen, M. Bechner, K. Potamousis, S. Goldstein, L. Pape, M. R. Mehan, C. Churas, S. Pasternak, D. K. Forrest, R. Wise, D. Ware, R. A. Wing, M. S. Waterman, M. Livny, and D. C. Schwartz. A Single Molecule Scaffold for the Maize Genome. *Plos genetics*, 5(11):e1000711, 2009. J. R. Ecker, editor. ISSN: 1553-7404.

Molecular Imaging and Photonics
Celestijnenlaan 200F, Bus 2402
3000 LEUVEN, BELGIË
tel. + 32 16 32 74 18
fax + 32 16 32 79 90
www.kuleuven.be