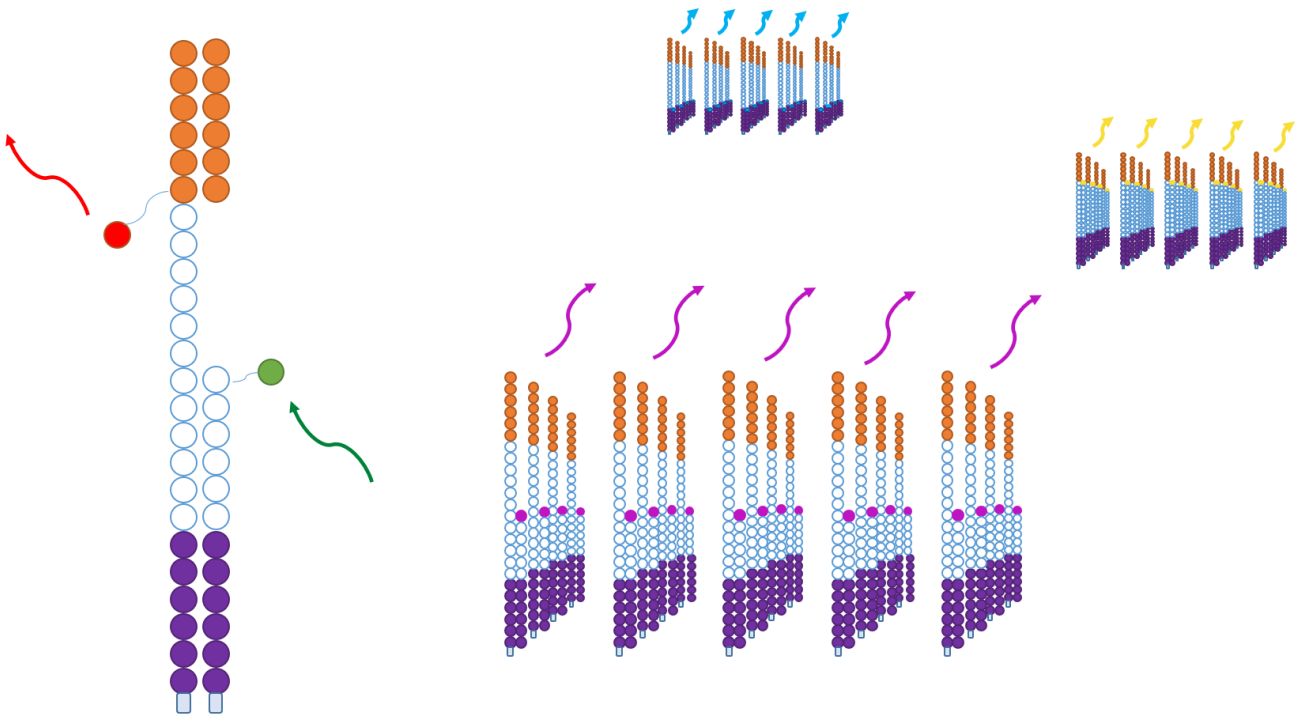
Combining Single-Molecule FRET with High-Throughput Sequencing

A powerful new tool for researching single-stranded DNA

F. S. Brandenburg



TU Delft

Combining Single-Molecule FRET with High-Throughput Sequencing

A powerful new tool for researching single-stranded DNA

by

F. S. Brandenburg

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Tuesday December 7, 2021 at 15:00 PM.

Student number:	4445643	
Project duration:	April 1, 2021 – December 7, 2021	
Daily supervisor:	dr. Sung Hyun Kim	
Thesis committee:	Prof. dr. Chirlmin Joo,	TU Delft, supervisor
	Dr. Martin Depken,	TU Delft
	Prof. dr. ir. S. J. T. van Noort,	Universiteit Leiden

This thesis is confidential and cannot be made public until December 31, 2021.

TU Delft

Contents

Abstract

Single-stranded DNA (ssDNA) is involved in many important cellular processes such as the replication, transcription and repair of our genome. It is also involved in the creation of so called telomeres, end-caps that protect chromosomes from degradation and are linked to aging. ssDNA is also used extensively in modern DNA nano-fabrication. Examples of this include DNA-origami, which can be used to create nanometer scale structures in programmable shapes and aptamers, ssDNA architectures that bind with high affinity and high specificity to a target that show great promise in use as novel therapeutics.

This makes ssDNA structure an interesting topic of study. Such structural assays have been done using various techniques, including but not limited to: FRET, NMR, optical/magnetic tweezers and AFM. A downside to all these techniques is that only a couple of sequences can be measured at a time, making it difficult to sufficiently sample the vastness of sequence space available.

In this thesis I demonstrate a novel technique combining single-molecule FRET combined next-generation high-throughput optical sequencing. I show that the two different measurements can be performed on the same chip and effectively mapped to each other. First, using traditional low-throughput methods, the FRET efficiency was measured for a cy3-cy5 pair separated by a 8 nucleotide piece of single-stranded DNA. This was done for 12 different sequences that were selected to vary in terms of base stacking, bulkiness, hydrogen bonds and other structural factors. The experiment was repeated using the high-throughput platform. The results of the high-throughput method were compared with the results from the low-throughput method, and show a correlation factor of 0.75. These experiments show that this technique can be used as an effective tool in performing FRET measurements on a minimum of 1500 sequences up to possibly 910,000 sequences in one measurement, making it an exciting new tool for structural research into ssDNA.

1

Introduction

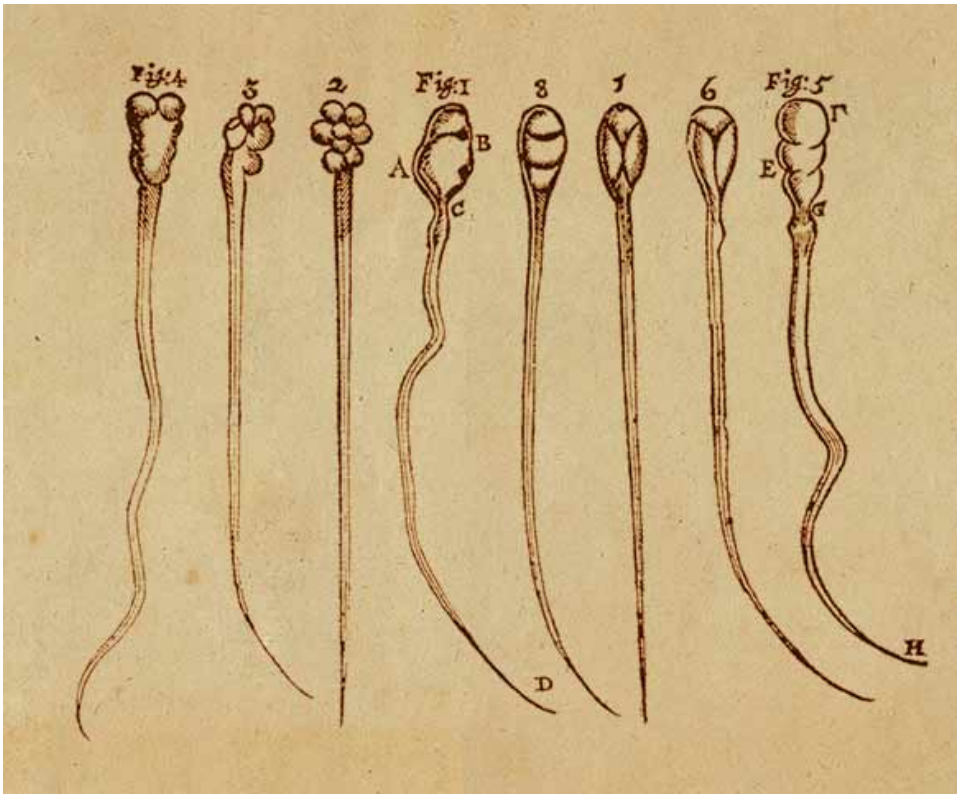
1.1. A short history of DNA

1.1.1. The first steps

Life is as mysterious as it is interesting. Since the dawn of time people have been wondering where life comes from, what we have in common, and what makes us different.

In the mean time scientist continued to tirelessly work on more robust and empirical explanations. A big break through came with the advent of microscopes, early ones designed by the likes of Robert Hooke and Antonie van Leeuwenhoek. In 1677, Antonie van Leeuwenhoek became the first person to ever see a human cell, when he looked through his microscope at a sample of his own sperm. What he saw where hundreds of what he called 'animicules', tiny eel-like creatures, swimming around[16]. Now we now that what he really saw were spermatozoa, also known as sperm cells.

Figure 1.1: Some of van Leeuwenhoek's drawing of this first recorded human cells



After this first discovery of the cell scientist began to see cells in all living things. This eventually led to the development of *Cell Theory* which states that "The cell is the unit of structure, physiology, and organization in all living things". However if we are all made from the same building blocks, how come we all look so different? Put a mouse and an elephant next to each other and the difference is night and day, take a cell of a mouse and a cell of an elephant and the difference is not so clear.

The answer lies in how these cells are organised. A elephant has the same cells as a mouse, just more of them and in different configurations. If that is true, then there must be a blue print somewhere in the cell that explains how to build an organism. Ever since Mendel it was known that some characteristics of a organism were controlled by genes, and that these genes even could be transferred from parent to child. Where this information was stored and how this information was translated into physical characteristics remained a mystery.

1.1.2. The discovery of the double-helix

Then during the spring of 1952 Rosalind Franklin and a graduate student of her were performing x-ray diffraction measurements on *deoxyribonucleic acid*, now colloquially known under the acronym DNA. In 1953 James Watson and Francis Crick developed based on this data the familiar model for DNA that we know and love today: the famous double helix, with a sugar-phosphate backbone on

the outside, and nucleotide bases on the inside. These nucleotide bases come in 4 flavours: adenine (A), thymine (T), cytosine (C), guanine (G). These four bases together encode all information for every organism in existence(!).

The discovery of DNA gave rise to incredible in our understanding of life. DNA became the integral to the so called *Central Dogma of Biology*, which explains how information in the cell gets transcribed, transported and translated into actual function. The discovery of DNA gave us a better understanding on how to grow better crops, making variants that are more drought-tolerant, resistant to diseases, or more nutritious and tasty. We used it better classify all the different species around us, and give us insight on who our closest relatives in the animal kingdom are. DNA analysis is now an essential component in crime forensics, and we used it to solve many crimes. DNA has helped us find long lost family members. And in the medical field, DNA helps us better treat certain cancer patients, and helps us quickly and reliably diagnose diseases through a PCR-test.

1.2. Sometimes less is more

Naturally, a lot of research is focused on the characteristics of the double-stranded DNA helix, since that is form we most often see. However DNA doesn't only exist as the neat double helix that we all learned about in high school. Single-stranded DNA (ssDNA) is also an interesting molecule to study. Since ssDNA isn't paired up with a second strand, it is less rigid then double-stranded DNA (dsDNA). This gives it the ability to create interesting secondary structures, making it an ideal material for the fabrication of nanostructures. This is what makes studying ssDNA so interesting, since we can see it both in all kinds of natural processes in our own cells, as well as in the most some of the world's most advanced nanofabrication labs.

1.2.1. The role of ssDNA in the cell

There are various processes in a cell where double-stranded DNA (dsDNA) unzips and forms two strands of single-stranded DNA (ssDNA). In fact during almost all biochemical reaction involved in DNA replication and repair steps, ssDNA is involved in some capacity.[9]

A particularly interesting interaction is that of ssDNA with DNA-binding proteins. These interaction plays a crucial role in the aforementioned processes. Replication protein A (RPA) for instance interacts with ssDNA to suppress the formation of any secondary structures. This helps DNA-polymerase (itself also an DNA-binding protein) in assembling double-stranded DNA from a single-stranded template. Chromosomal DNA often terminates with a 3'-overhang of ssDNA, which together with another class of DNA-binding proteins form the so called telomeres, an structure that

is essential for chromosomal stability. Consequently, shortening of the telomeres is strongly linked to aging. [21] [4]

Another place where we can find ssDNA is in viruses. ssDNA viruses affect all three domains of life: Archaea, Bacteria and Eukarya. ssDNA infections in humans don't appear to cause any major diseases. Curiously however, ssDNA viruses are very common in plants or cause large losses of agricultural productivity each year. Over one third of viruses found in plants are ssDNA viruses. [28] [17] The effectiveness of viruses is largely determined by their viral packaging, a process in which their genome gets condensed into an protein capsule. The workings of this process are still poorly understood. [27] [26]

1.2.2. The use of ssDNA in nano-fabrication

However nothing is more human to look towards the materials and tools that are given to us by nature, and to think how can we use these to alter this and use them for our own goals and desires? In the case of ssDNA that is as a platform for building nano-structures. ssDNA has a couple of properties that make it uniquely suitable for this cause. Since ssDNA is quite flexible it is relatively easy to create complex shapes. Because DNA been a part of cellular for billions of years, we can make use of the very efficient tools evolution has created for copying, cutting and repairing DNA without having to invent them ourselves. And lastly, the structure of four nucleotide bases consisting of the two pairs that exclusively bind with each other gives a certain degree of programmability that is very conducive to the human design process.

After laying out the fundamental framework in the early 80s, N.C. Seeman created the first DNA nano-structure in 1991: a DNA nano-cube. Since then the field has grown exponentially. Nowadays, researchers can virtually produce any two- or three-dimensional shape that they so desire, using DNA-origami. Examples include: 2-dimensional shapes like stars, hearts or smileys; 3-dimenional shapes or even detailed pictures of the Mona Lisa. [8] [23] [24]

Another exiting development in the field of DNA nanotechnology are aptamers. Aptamers are made out of a single strand of ssDNA (or ssRNA) that folds into a three-dimensional secondary structure. They can be selected to specifically bind to a certain target. Their specificity rivals that of more traditional antibodies, but they also come with a few significant advantages. Since their structure is made out of oligonucleotides, they are relatively easy to manufacture and are therefore cheaper. They are more temperature stable. They are also easier to change and design. They also do not cause an immune response when used inside the body, in contrast to some protein based therapeutics. Currently multiple aptamers are being developed as a cure for disease. [15] [25]

1.3. ssDNA as a subject of study

As we can see, ssDNA plays in a pivotal role in many important processes, such as: cellular processes (transcription and replication), aging (telomeres), modern advanced nano-fabrication (DNA origami) and novel therapeutics (aptamers) to name a few. In many of these cases, it is crucial to know the secondary structure of ssDNA, which of course is dependant on its sequence of nucleotide bases. However, much less is known about the structure-sequence relation of ssDNA compared to dsDNA, especially at short length scales of a few nucleotides. Gaining more insight how the secondary structure of ssDNA comes about can therefore be crucial in developing the next generation of ssDNA applications.

1.3.1. Measuring end-to-end distance using FRET

FRET is a measurement technique in which we make use of a special property of fluorophores. If two fluorophores that have some spectral overlap between their emission and absorption spectra, energy from a one fluorophore can transfer non-radiatively (that is, without photons) to another. The other fluorophore can then emit the photon, but now at a lower wavelength then the original excitation photon. This process happens to be extremely sensitive to distance. FRET can therefore be used as a *molecular ruler*, to measure very short distances of 1-10 nm.

Using FRET, we can measure the average end-to-end distance of short pieces of ssDNA. This end-to-end distance is determined by the relative flexibility of the ssDNA, higher flexibility will give the ability for the ssDNA to more curl up more, which reduces the end-to-end distance. Flexibility of individual stretches of ssDNA is one of the important parameters for a 3-dimensional structure.

In this thesis I will first demonstrate that we can measure the difference in end-to-end distance in short 8 nucleotide sequences of ssDNA. We will also look at the effect of different salt concentrations on the flexibility of ssDNA.

1.3.2. High-throughput FRET measurements using Next Generation Sequencing

However using traditional single-molecule FRET, it is quite cumbersome to measure large quantities of different sequences, since every sequences has to pipetted into a separate observation chamber. That is why we developed a novel technique in which we combine single-molecule FRET with next generation optical sequencing. Using Illumina's optical sequencer we can preform single-molecule experiments on the same chip as sequencing, which gives us the ability to correlate the two. This removes the need for the physical separation of the different sequences. This will drastically increase the amount of sequences that can be assayed at once, making the technique an effective tool for high-throughput measurements of end-to-end distance.

1.4. Physical principles of ssDNA structure

There are many fundamental forces that can influence the structure of DNA. Since DNA is a long sequences of nucleotides, we lend a lot from general polymer theory. However there are also some interactions unique to DNA, such as base stacking.

1.4.1. Bond stretching

Bond stretching describes the interaction between atoms when they move out of the equilibrium distance. The harmonic oscillator is a common model to describe these interactions. In the harmonic oscillator forces are given by the simple Hooke's law, meaning that the resulting force from moving out of equilibrium is opposite to that movement, resulting in movement back to the equilibrium position or oscillation around the equilibrium position. Using the linear Hooke's law results in a quadratic potential, however Lennard-Jones potential and quartic potentials are also used. [2]

1.4.2. Bond angle potential

Analogous to the the fact that translational movements are influenced by the bond stretching potential, rotational movements are influenced by the bond angle potential. The ideal bond angle between two is explained by the electron orbitals. For example a $\pi - bond$ is usually 180 degrees. However neighboring atoms and the environment can cause the actual bond angle to deviate from this ideal angle. This results in a rotational force trying to move the angle back to equilibrium. Again quadratic potentials are most used, however different potentials like the trigonometric metric potential also exist. [3]

1.4.3. Base stacking

The interactions often touted as most influential in DNA structure are the stacking interactions. Because the surface of the nucleotide bases have so few polar groups they are quite hydrophobic, contrast to the phosphate groups also present in DNA. This causes DNA to organise with nucleotide bases in at the center, shielded by the sugar backbone and the phosphate groups that organize on the outside. To further minimise the exposed surface of the nucleotide bases, it is beneficial for the bases to turn around the axis to fit closer together. This is where double stranded DNA gets its characteristic 30 degree tilt. The energy that is released by this base stacking is different for subsequent base pairs.[1]The efficiency of this base stacking is not the same for all bases. In general guanine and cytosine stack better then adenine and thymine. And the differences between base pair stacking energies are conserved over a wide range of ionic conditions. [14]

While the situation for single-stranded DNA (ssDNA) is of course different, stacking still occurs, however often in a lesser amount. ssDNA can still organise in helical structures, but also often contains lengths of unstacked domains. Particularly long tracts of adenine are known to exhibit strong helical stacking behaviour. [19]

1.4.4. Excluded volume

Two parts of a polymer cannot occupy the same volume. In general if we have a monomer of radius R, we cannot place a second monomer of equal size closer then twice that radius. While the volume of the monomer is $\frac{4}{3}R^3\pi$, the excluded volume has radius $2R$, and is 8 times as large at $6R^3\pi$. Imagine that we have two polymers, A and B, of equal length, but polymer A having a larger monomer size then polymer B. In this case, all else being equal, polymer A will be less flexible then polymer B. [12] [20] [22] In the context of DNA, purines (adenine and guanine) contain two carbon rings, while pyrimidines (cytosine and thymine) only have one carbon ring. This makes purines more bulky then pyrimidines.

1.4.5. Electrostatic interactions

Electrostatic interactions concern the repulsive of attractive forces caused by electrical charges within the a molecule or system. DNA itself is quite negatively charge in solution. However this charge is not equally dispersed through the molecule, but is concentrated in the phosphate groups along the backbone. In general strong repulsive electrostatic forces make a polymer more stiff. The negative charges of the phosphate groups can be screened by positively charged ions in solution. When this occurs, the repulsive forces between the monomers are reduced, and as a result the DNA molecule becomes more flexible. [5] [6] [22]

1.4.6. Hydrogen Bonds

Hydrogen bonds are weak bonds that occur between different hydrogen atoms within a molecule or between molecules. For this to happen, the hydrogen atom must be covalently bound to a more electronegative atom, such as oxygen or nitrogen. Since the shared electrons spend more time close to the electronegative atom then to the hydrogen atom, this causes a charge imbalance. The hydrogen atom becomes slightly positively charged, and the other atom becomes slightly negatively charged. The attractive force between these small positive and negative charges are the cause of these hydrogen bonds. Hydrogen bonds are (partially) responsible for the secondary structure of a lot of molecules, including nucleotides and DNA. More specifically, in DNA hydrogen bonds are also responsible for the complementary binding of nucleotide bases. The adenine and thymine a coupled by two hydrogen bonds, while the cytosine and guanine pair are coupled by three hydrogen bonds. Single-stranded DNA obviously has no complementary strand to bind to, however it can form loops within it self. This can give ssDNA some unique secondary structures.

2

Materials & Methods

2.1. Total Internal Reflection Fluorescence Microscopy

Total Internal Reflection Fluorescence (TIRF) microscopy is a technique in which we make use of evanescent waves to only illuminate a small band beneath the surface. This has the advantage that we only illuminate molecules bound to surface and minimise background noise from molecules in suspension.

2.1.1. Working principles

Light passing through a medium with refractive index n_1, encountering a certain different medium with refractive index n_2 at an angle of incidence Θ_i will cause the transmitted light beam to bend with an angle Θ_t as described by Snellius' Law:

$$n_1 sin\Theta_i = n_2 sin\Theta_t \tag{2.1}$$

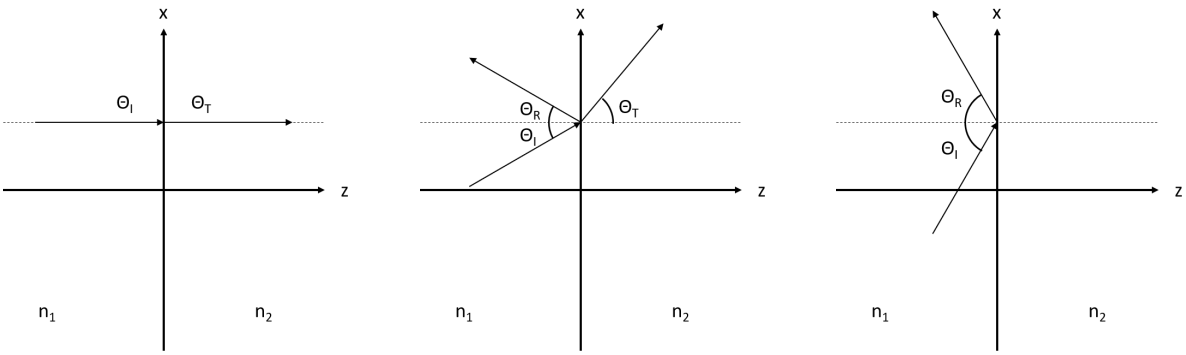
Usually when approaching a new surface, light is both partially reflected and transmitted. However as the refraction angle approaches 90° all light will be reflected. The angle of incidence at which this happens is called the critical angle Θ_c

$$\Theta_c = \Theta_i = sin^{-1}(\frac{n_2}{n_1} sin\Theta_t) \tag{2.2}$$

And since $\Theta_t = 90°$

$$\Theta_c = sin^{-1}(\frac{n_2}{n_1}) \tag{2.3}$$

Figure 2.1: Three cases illustrating: a.) full transmission; b.) partial transmission and partial reflection; c.) full reflection

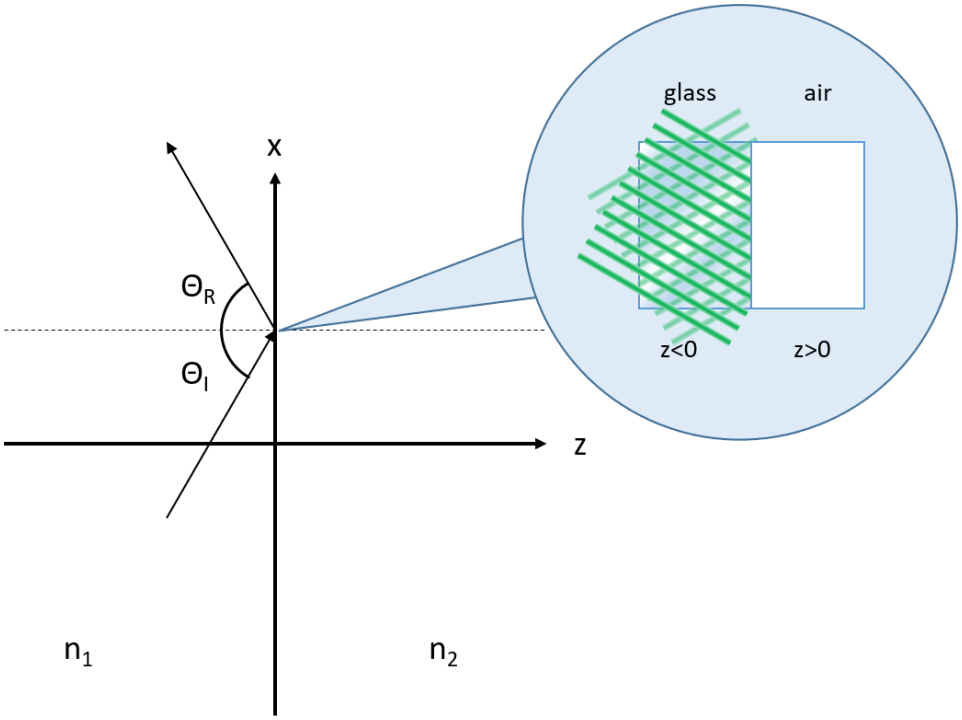


(a) $\Theta_i = 0$ (b) $0 < \Theta_i < \Theta_c$ (c) $\Theta_i > \Theta_c$

This is only possible when $\frac{n2}{n1} < 1$, so when moving from a medium with a higher refractive index to a medium with a lower refractive index. An example of this when light moves through glass ($n \approx 1.6$) towards an interface with air ($n \approx 1$).

When all light is reflected there will be a certain electric field E at one side of the boundary. Since there is no light transmitted through the boundary, one would assume that there also is no electric field. However this would cause a discontinuity error exactly at the boundary.

Figure 2.2: Approaching the boundary from negative z there clearly is a time-dependant electric field at the boundary. However approaching the boundary from positive z, there appears to be no electric field, since there is no transmission



The logical conclusion then is that there is an electric field at the other side of the boundary, and this field is known as the evanescent field. A general description of the electric field of an electromagnetic wave $E(\vec{\mathbf{r}}, t)$, variable in space $\vec{\mathbf{r}}$ and time t.

$$E(\vec{\mathbf{r}}, t) = \mathrm{Re}\left\{E(\vec{\mathbf{r}})e^{i\vec{\mathbf{k}}\cdot\vec{\mathbf{r}}-i\omega t}\right\}, \quad \vec{\mathbf{r}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \vec{\mathbf{k}} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \tag{2.4}$$

with wave vector $\vec{\mathbf{k}}$ and complex field E_0. We know that waves propagate in the direction of the wave vector $\vec{\mathbf{k}}$. Generally, if a wave is transmitted, k_z should be a non-zero, real number. We can express k_z in term of the other wave number k_x and k_y and the angular velocity ω

$$k_z = \sqrt{\frac{\omega^2}{c^2} - (k_x^2 + k_y^2)} \tag{2.5}$$

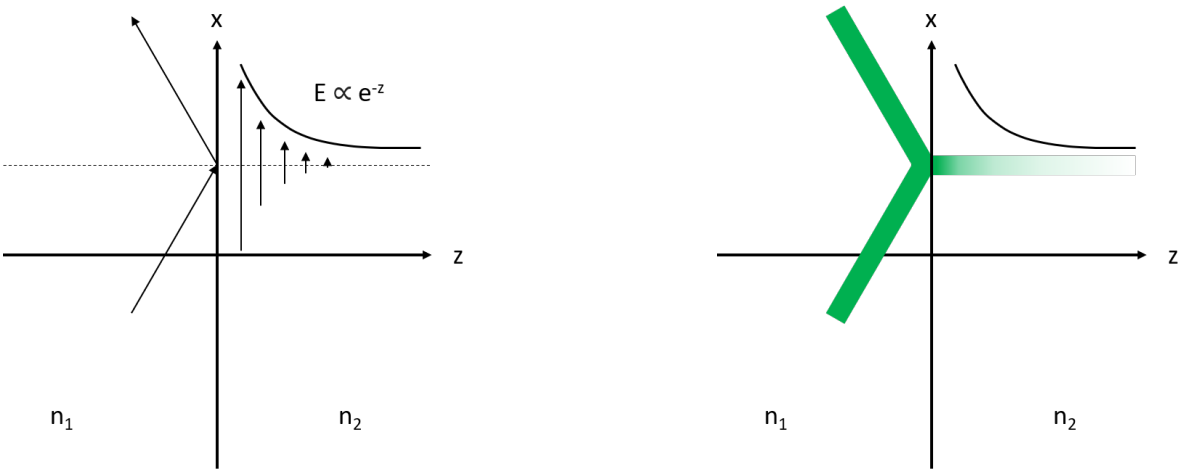
If the wave is transmitted, k_z should be real, and we therefore we can have the constraint that $\frac{\omega^2}{c^2} > k_x^2 + k_y^2$. Conversely, if the wave is reflected, it cannot have a real-valued k_z, but it can still be imaginary-valued: $k_z = \gamma i$ with γ a real-valued, positive constant. We can then rewrite the general formula given in 2.4 as follows:

$$E(\vec{\mathbf{r}},t) = \mathrm{Re}\left\{E(\vec{\mathbf{r}})e^{\pm i(k_x x+k_y y)-i\omega t}e^{\pm i(\gamma i z)}\right\} = \mathrm{Re}\left\{E(\vec{\mathbf{r}})e^{\pm i(k_x x+k_y y)-i\omega t}\right\}e^{\mp\gamma z} = \mathrm{Re}\left\{E(\vec{\mathbf{r}})e^{\pm i(k_x x+k_y y)-i\omega t}\right\}e^{\mp\gamma z} \tag{2.6}$$

This leaves us with with two solutions: one exponentially growing solution and one exponentially decaying solution. However the exponentially growing solution is not physical, since it is at odds with the principle of conservation of energy. That leaves us with one solution that is exponentially decaying:

$$E(\vec{\mathbf{r}},t) = \mathrm{Re}\left\{E(\vec{\mathbf{r}})e^{\pm i(k_x x+k_y y)-i\omega t}\right\}e^{-\gamma z} \quad \text{for } z>0 \tag{2.7}$$

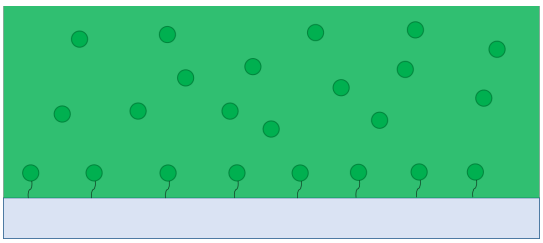
Figure 2.3: Schematic representation of the evanescent field



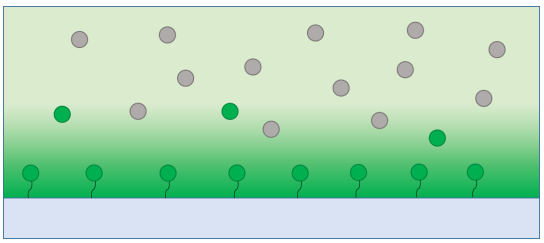
From equation 2.7 we can see that the evanescent field propagates in the xy-plane and exponentially decays in amplitude in the positive z-axis.

As mentioned before, the properties of the evanescent field make it very useful in minimizing background noise from fluorophores moving around in suspension by only illuminating a very small band close to the surface.

Figure 2.4: Difference between a.) direct illumination and b.) illumination by an evanescent field



(a) Direct illumination excites all fluorophores present in the volume



(b) The evanescent field excites only fluorophores close to the surface

As the light intensity of the evanescent is of course continuously decaying it is impossible to define a hard border between the illuminated zone and the non-illuminated zone, however a commonly used length scale is when the intensity of the electric field has decayed a $1/e$

$$d = \frac{\lambda}{4\pi\sqrt{n_2^2 \sin^2(\Theta_i) - n_1^2}} \tag{2.8}$$

2.2. Förster Resonance Energy Transfer

Förster Resonance Energy Transfer (FRET) is a technique in which energy is transferred from a excited fluorophore to another one through nonradiative dipole-dipole interactions. FRET usually occurs on very small length scales, $1-10nm$. The FRET efficiency (the average percentage of energy that gets transferred from one fluorophore to the next) is heavily dependent on the the distance between the fluorophores, scaling with a sixth power. This makes FRET very sensitive to small changes in distance.

2.2.1. Working Principles

When a fluorophore enters an excited state there are a couple of pathways through which this energy can be released. Energy can be released nonradiatively, for example through vibrational relaxation, in which the molecule gains a higher vibrational state, or through collision that release heat energy. In fluorophore energy can also be released through radiative processes such as fluorescence, in which the molecule lowers its energy state through the emission of a photon. FRET gives yet another pathway for a molecule to return to its ground state. Through FRET a donor fluorophore can transfer its energy through dipole-dipole coupling to a acceptor fluorophore, without emitting a photon.

The FRET efficiency E_{FRET} can be defined as the ratio between rate constant for FRET k_{FRET} divided by the rate constant for all possible relaxation mechanisms (in this case through FRET(k_{FRET}), radiative relaxation(k_r), and nonradiative relaxation(k_{nr})).

$$E_{FRET} = \frac{k_{FRET}}{k_{FRET} + k_r + k_{nr}} \tag{2.9}$$

The rate constant for FRET k_{FRET} is inversely related to the distance between the FRET pair to the sixth power

$$k_{FRET} = \frac{1}{\tau_D}\left(\frac{R_0}{r}\right)^6 \tag{2.10}$$

With τ_D being the lifetime of the excited state of the donor, and R_0 is an empirical quantity known as the *Förster radius*, a number that denotes the distance at which the FRET efficiency is 50%. The Förster radius mainly dependent on the quantum efficiency of the donor and the spectral overlap between the donor emission spectrum and the acceptor absorption spectrum.

The excitation lifetime for the donor is inversely proportional to the rate of energy leaking away through other pathways then FRET

$$\tau_D = \frac{1}{k_r + k_{nr}} \tag{2.11}$$

Combining the above equations we can come to the following expression for the FRET efficiency:

$$E_{FRET} = \frac{R_0^6}{r^6 + R_0^6} \tag{2.12}$$

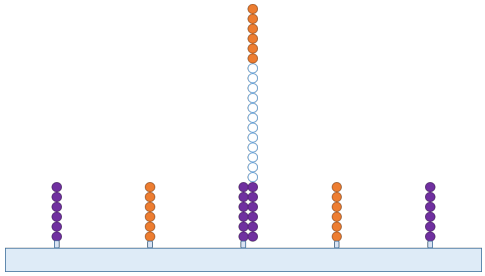
To calculate the FRET efficiency from the measured intensity traces we take the ratio of the intensity of the acceptors I_D (which corresponds to the amount of photons emitted through FRET) and the sum of the donors and acceptors (which corresponds to the total amount of photons absorbed). We can do this separately for each molecule, making it a single-molecule FRET measurement. This means we can assess each molecule individually, making it possible to study dynamics or study different molecules in the same field of view. [10] [7] [13]

2.3. Optical Sequencing

During this project sequencing has been preformed using Illumina optical sequencers. During sequencing the Illumina sequencer is loaded with a chip. This chip is also manufactured by Illumina, and is pre-loaded with two types of short oligonucletides covalently bound to the surface, also called sometimes called *the lawn*.

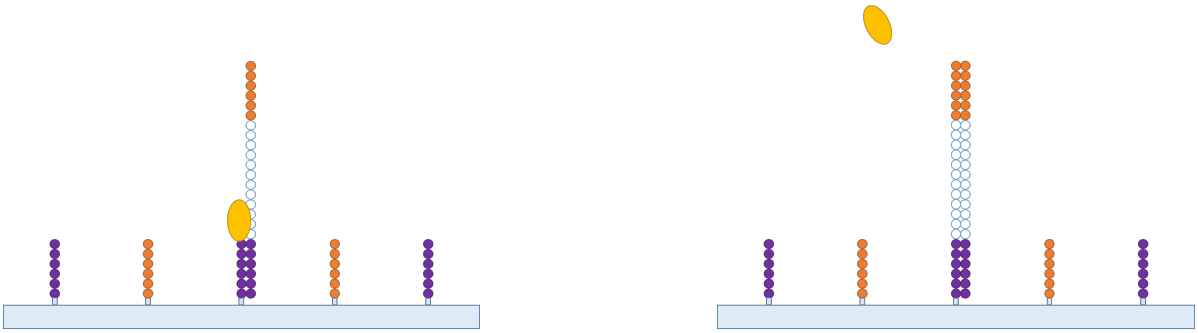
These two oligonucleotides, called p5 primer and p7 primer, are also found on both ends of the DNA that is to be sequenced. The DNA can therefore hybridise to the primers present on the surface.

Figure 2.5: DNA can hybridise to the p7 primers on the surface



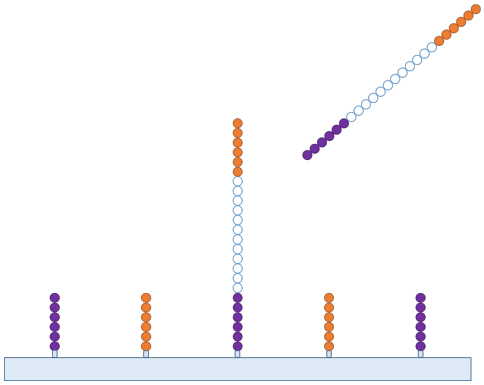
During the first step, DNA polymerases will attach to the single-stranded DNA fragments and polymerise the ssDNA into dsDNA. The complementary strand will now be covalently bound to the p7 primer.

Figure 2.6: DNA polymerase binds to ssDNA and turns it into dsDNA



Next, the DNA will be denatured, washing away the hybridised strand and leaving only the strand covalently bound to the primer.

Figure 2.7: Template strand is washed away, leaving the complementary DNA covalently bound to the surface



Now the sequencer will go through a step called bridge amplification. The opposite end of the DNA fragment contains the complement of the p5 primer, which will now start hybridising to a neighboring, empty p5 primer. DNA polymerase will start turning the ssDNA into dsDNA again. The DNA fragment now forms an arc on the surface, hence the name bridge amplification.

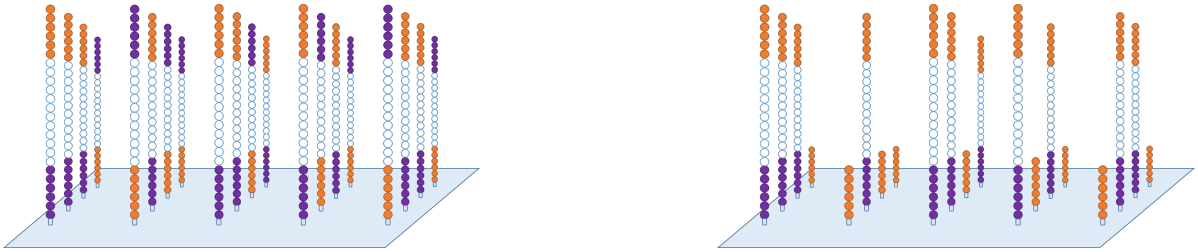
Figure 2.8: DNA arcs over to neighboring p5 primer and polymerises again

Next, there is another denaturing step, in which the dsDNA is turned back into ssDNA. However since the other end of the complementary strand is now covalently bound to the surface, it is not washed away and the amount of DNA fragments on the surface is doubled.

Figure 2.9: DNA is denatured and leaves two strands bound to the surface

This bridge amplification is repeated over and over until *clusters* form. Clusters are dense packings of DNA covalently bound to the surface of the chip. Since the clusters have grown from a single fragment of DNA, they all equal to the original sequence or are complementary to it. After clusters are formed all fragment bound to p5 are washed away so that the remaining fragments now all share only one sequence.

Figure 2.10: DNA fragments bound to p5 are washed away so that only one sequence remains

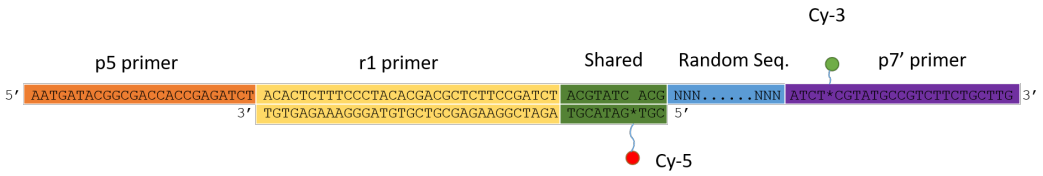


The last step is the actual sequencing. Special nucleotides are introduced that have a fluorophore attached. These fluorophores are too weak to preform single-molecule measurements, but since the bridge amplification has created dense clusters, there will be many fluorophores closely packed that can produce a readable signal. Reading out these flashes per cluster gives out the sequence of the DNA that originally bound there. Using this sequencing technique, many clusters can be read out in parallel, making it an excellent tool in high-throughput research.

2.4. DNA Design

The DNA design for the experiments is build from a couple of building blocks. Starting from the 3'-end it looks as follows. The first is p7' primer, which is complementary to the p7 primer. The p7 primer (and the p5 primer) are short DNA oligos used in Illumina sequencer to attach DNA to the chip. This p7' primer also is labeled with a cy3 fluorophore. Next up is a variable sequence. This could be of variable length, however for the experiments preformed in this thesis 12 sequence of 8 nucleotide length have been chosen. These 12 sequences have been chosen such that they vary as much as in their physical characteristics such as base stacking, bulkiness, purine and pyrimidine content, etc. A length of 8 nucleotides has been chosen to make it sequence space not too large (8 nucleotides already have 65,536 different configurations). After the variable sequence there is a short piece of shared sequence, which contains a cy5 fluorophore. Next up is the r1 primer, which is used in sequencing by the Illumina sequencer. Lastly there is a p5 primer, which is used in to bridge multiplication during sequencing. This piece of DNA is hybridised with the cover stand, which contains the complement to the r1 primer and the shared sequence. The cover strand also contains a cy5 fluorophore. When the DNA strand is hybridised to the p7 primer on the surface, most of the DNA will be double-stranded. The variable sequence will be left single-stranded. With the two fluorophores attached on either side, the fret efficiency can be measured. The p5 primer also has been left single-stranded, for fear if the p5' primer would be added to the design, the DNA could also hybridise to p5 primers on the surface.

Figure 2.11: Sequence of DNA design used to measure FRET efficiencies of ssDNA pieces



2.5. Quartz flow-cell

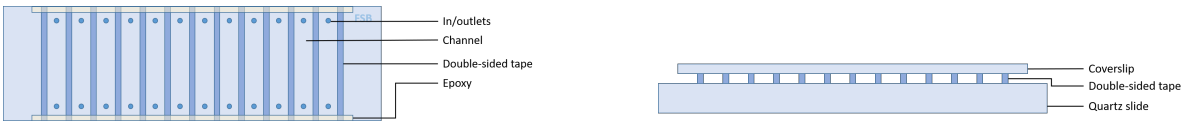
The low-throughput were performed on a quartz flow-cell. The flow-cell consist out of a quartz slide and a cover slip, with channels formed by double-sided tape. To tether DNA to the surface, a couple steps have to be taken. Firstly the quartz slide and cover slip have to properly cleaned. The cleaning protocol consist of 8 steps. First glass slides are cleaned with water and detergent and any macroscopic material is removed. Next the slides are sonicated with detergent for 20 minutes. The detergent is washed away first with tap water and later with Milli-Q water. The slides are sonicated for another 5 minutes in Milli-Q water. The solution is replaced with acetone and the slides are sonicated for another 15 minutes. The solution is replaced with potassium hydroxide at 1M and the slides are sonicated for another 20 minutes. Next the slides will be cleaned of any remaining organics using Piranha (a mixture of 75 mL sulphuric acid and 25 mL of 30% hydrogen peroxide). The slides are incubated for 20 minutes in this mixture. Lastly they are stored in Milli-Q. The cover slips under go the same cleaning steps except for the Piranha etching

After they have been cleaned, they are first PEGylated with a mixture of 40:1 of mPEG-Silane:Biotin-PEG-Silane. This can be done in bulk, and PEGylated quartz slides are kept in the freezer.

2.5.1. Assembling flow-cell

At the day of the experiment, the flow-cell is assembled. The quartz slide is covered with 13 thin strips of double-sided tape, creating 12 channels. Then a cover slip is put on top. The sides of the flow-cell are then sealed with epoxy. This is left to cure for 30 minutes. The DNA cannot directly bind to the PEG or biotin, so a couple more steps have to be taken when the flow-cell is assembled.

Figure 2.12: Top and side profile of the quartz flow-cell

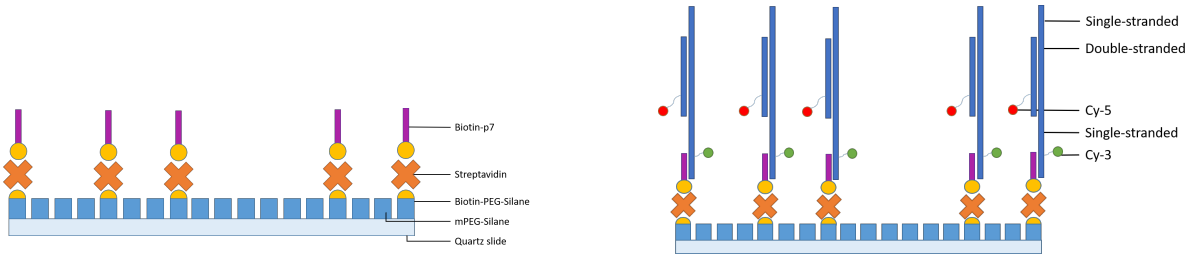


2.5.2. Building tethering surface

With the flow-cell assembled, 12 μL streptavidin 0.5 mg/mL in a buffer of 10 mM TrisHCl at pH 8 and 50 mM of NaCl (also called T50) is flushed through the channels and incubated for 3 minutes. Streptavidin specifically binds to biotin, and has four binding spots. Biotin is present at the surface through the Biotin-PEG-Silane molecules. After the streptavidin has been washed away with 100 μL of T50. 50 μL of biotin-p7 in T50 at 1 nM is flushed through the channels and incubated for 3 minutes. The biotin-p7 will bind to one of leftover streptavidin binding spots. We are now left of a

surface with p7 primers, which the sample can hybridise to. We take these extra steps with the p7 primer instead of directly binding our DNA to the PEG-silane-biotin because we want to make our DNA sample compatible with the Illumina sequencer.

Figure 2.13: Schematic representation of the tethering mechanism

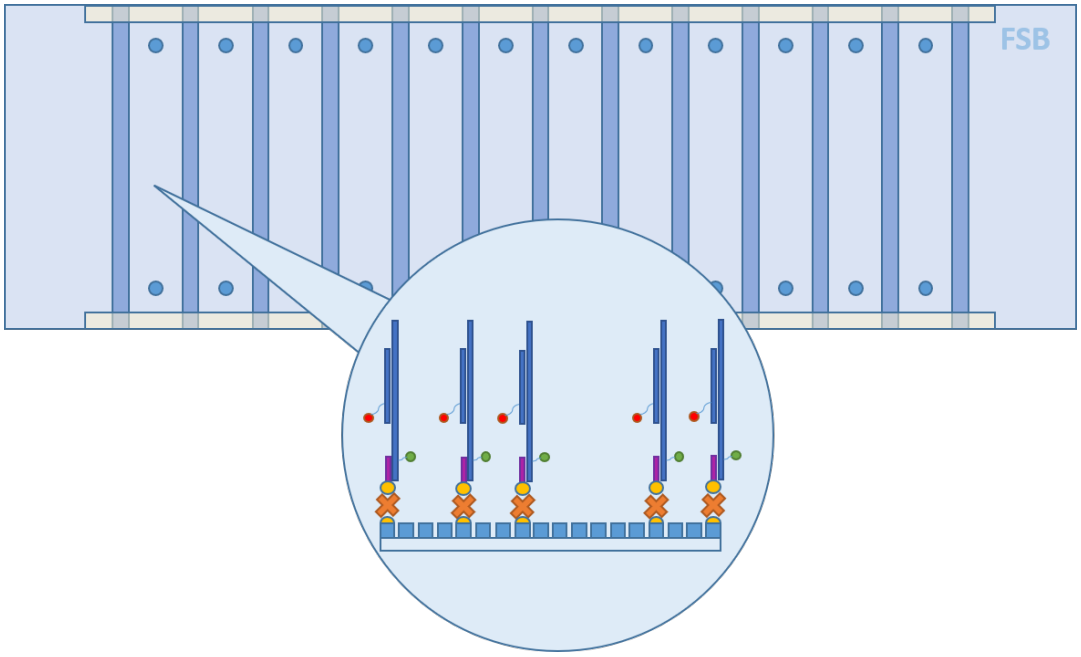


Next we add 50 μL of DNA dissolved in T50 with 10 mM of $MgCl_2$. The concentrations of DNA is determined with trial and error, but is typically around 100 pM. The DNA incubates for 20 minutes. With the DNA hybridised to the surface there is one last step before imaging. To increase fluorophore life-time we add an imaging buffer to the channels. This buffer contains Trolox, which is a triplet quencher. To lower the oxygen content, we also add Protocatechuate decarboxylase (PCD) and Protocatechuic acid (PCA). PCD uses oxygen to convert PCA into 3-carboxy-cis,cis-muconate.

$$3,4-\text{dihydroxybenzoate(PCA)} + O_2 \xrightarrow{\text{PCD}} 3\text{-carboxy}-\text{cis, cis}-\text{muconate} \tag{2.13}$$

We also add the desired NaCl and $MgCl_2$ content. With imaging buffer inside the channels we have can start imaging. Because PCD uses up PCA over time, it's advisable to work quickly and finish imaging within 30 minutes.

Figure 2.14: Quartz flow-cell, zoomed in on the surface



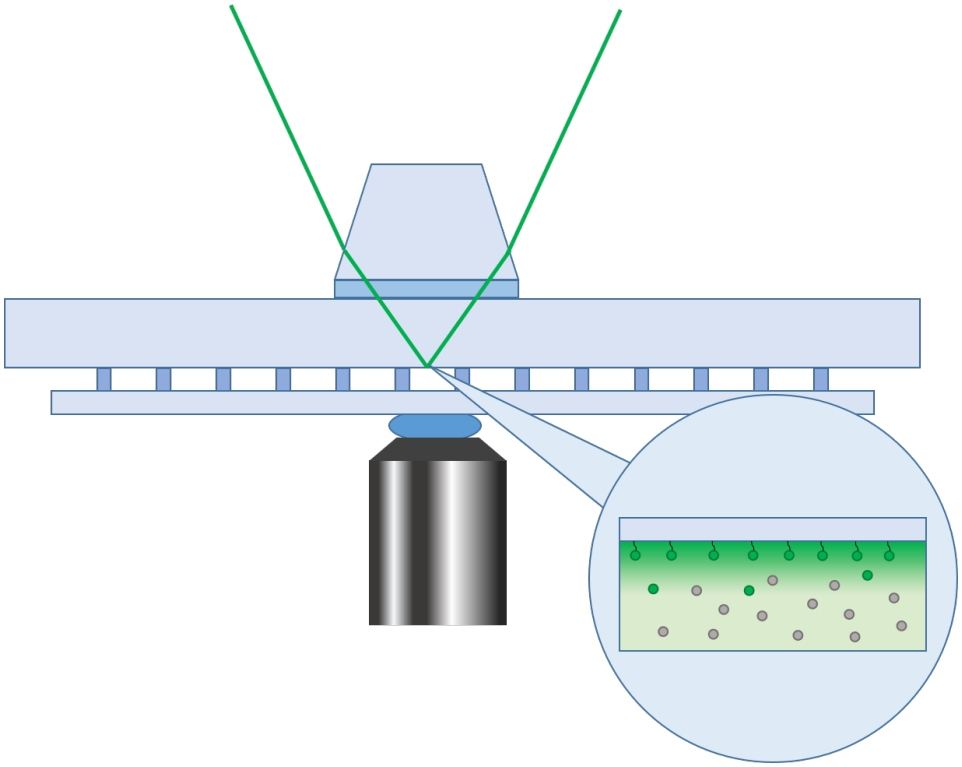
2.6. Microscope setup

During this thesis, two different microscope setups have been used. The low-throughput data was collected using a prism-type TIRF microscope. The high-throughput data was collected using a objective-type TIRF microscope.

2.6.1. Prism-type TIRF Microscope

The prism-type TIRF microscope is a custom-built setup. The setup is build around a Nikon Eclipse Ti2 inverted microscope. The laser light is produced by a 500 mW, 532 nm laser module inside an Oxxius LaserBoxx. A 60x water immersion objective (CFI Plan Apochromat VC 60X WI) magnifies the image, after which the light beam passes through a 390 to 690 nm bandpass filter (ET700sp-2) to block any stray light, but allows the emitted light from the cy3 and cy5 dyes through. A quad-notch filter (NF03-405/488/532/635E-25) then filters out the 532 nm light produced by the excitation laser. After passing through the quad-notch filter the light beam moves on to the emission box.

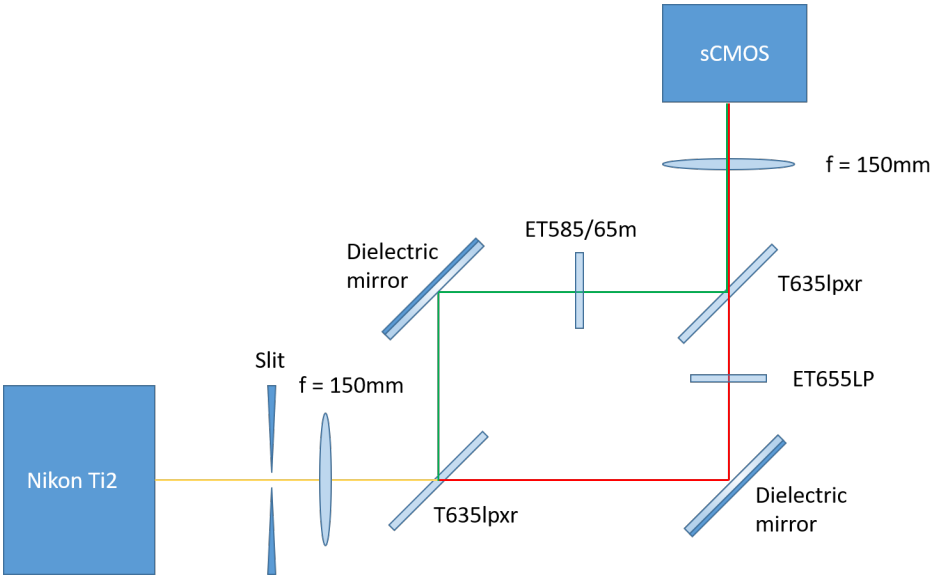
Figure 2.15: Schematic of prism-type TIRF microscopy



Inside the emission box, light will first pass through a narrow slit to narrow the field of view. A lens (f=150mm) will collimate the light beam. A dichroic mirror (T635lpxr) splits the fluorescent light, letting light with a wavelength of 635 or higher through, while reflecting light with a lower wavelength. A 655 nm long pass filter (ET655LP) will filter out any residual green or blue light from

the transmitted beam. A 552 to 618 nm band pass filter removes any residual red or blue light from the reflected beam. Another dichroic mirror (T635lpxr) brings the two beams back together. A final lens (f=150mm) projects the image on the sCMOS camera (Photometrics Prime BSI). The camera has a 13.3 mm sensor size and takes images in resolution of 2048x2048 pixels. The microscope is controlled using NIS-Elements software package.

Figure 2.16: Schematic of optical path inside the emission box

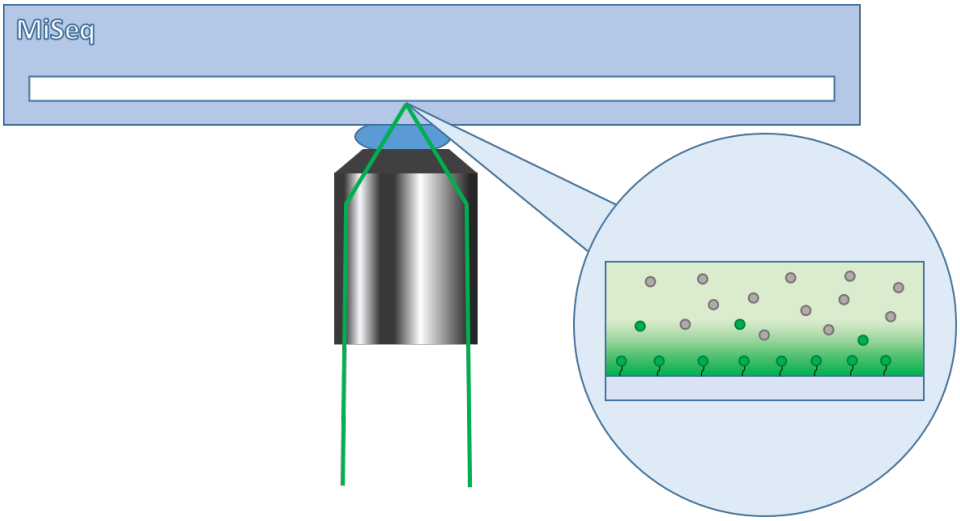


2.6.2. Objective-type TIRF Microscope

Measurements on the Illumina chip were made on a objective-type TIRF Microscope setup. The setup is build on the same Nikon Eclipse Ti2 inverted microscope. A Gataca iLAS 2 azimutal TIRF illumination module provides the 561 nm and 642 nm laser beams. An 100x oil objective lens (Nikon SR Apo TIRF 100x/1.49 NA) magnifies the image. An image splitter (Andor Optosplit II) seperates the cy3 signal from the cy5 signal. The cy3 and cy5 channel are projected next to each other on a CCD camera (Andor iXON Ultra EM-CCD). Data acquisition is done by using the Metamorph software.

The objective-type TIRF setup has a smaller field of view compared to the prism-type TIRF setup, which makes imaging take longer. However, imaging the MiSeq chip on the objective-type TIRF setup gives high background noise in the cy5 channel when excited by a green laser. This may be due the type of glass used in the MiSeq chip, which might contain impurities. The MiSeq chip has a thin side and a thick side. In objective-type TIRF, the laser passes through the thin side. In prism-type TIRF, the laser passes through the thick side, which might explain the increase in background noise.

Figure 2.17: Schematic of objective-type TIRF microscopy



3

Results

3.1. Low-throughput measurements on the quartz slide

Before attempting any high-throughput measurements, we start by preforming low-throughput measurement on a quartz glass flow cell. This is to test if our experimental design works well first in a controlled environment. The MiSeq chip is proprietary technology, and we therefore do not know the exact surface chemistry. Starting on a quartz slide helps us eliminating any confounding factors.

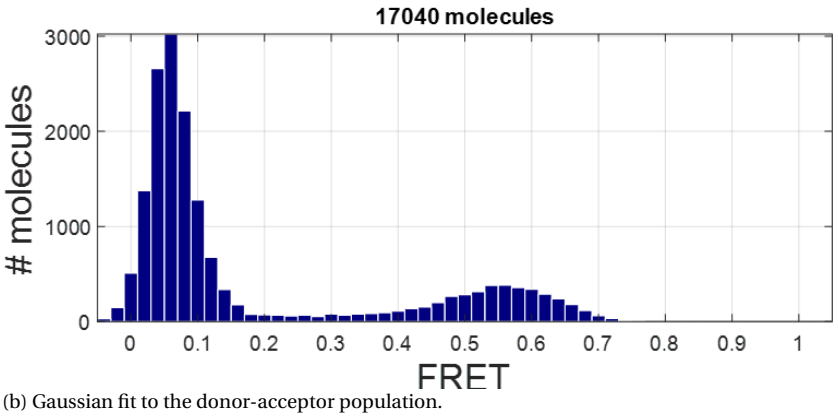
There are three main questions we wanted to answer with the experiments on the quartz slide: Do the different sequences indeed have different end-to-end distances? Are these differences enough to measure using our FRET technique? And does salt concentration play a significant effect?

3.1.1. Determining average FRET efficiency

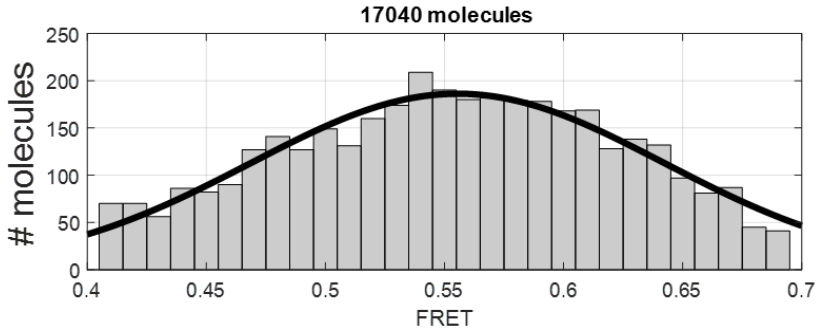
We can trace each molecule individually and calculate the FRET efficiency for each molecule. Due to natural variance, not every molecule will have exactly the same value. If we create a histogram of all these molecules, we can fit a normal distribution and get an average value. If we create this histogram, we can see two distinct peaks, one at approximately FRET efficiency 0.05 and one between FRET efficiency between 0.5 and 0.9 (see figure 3.1a). The first peak is a peak of molecules with only donor fluorophores. Since there is no second fluorphore, there is no energy transfer and the FRET efficiency is 0 (intensity of the cy5 channel is zero). Since there is always some background signal the value appears to be 0.05. The second peak is created by actual FRET transfer between the two

fluorophores. We can fit a Gaussian curve to the second peak to get the mean FRET value (see figure 3.1b).

(a) FRET histogram with two population, a large donor-only population and a smaller donor-acceptor population

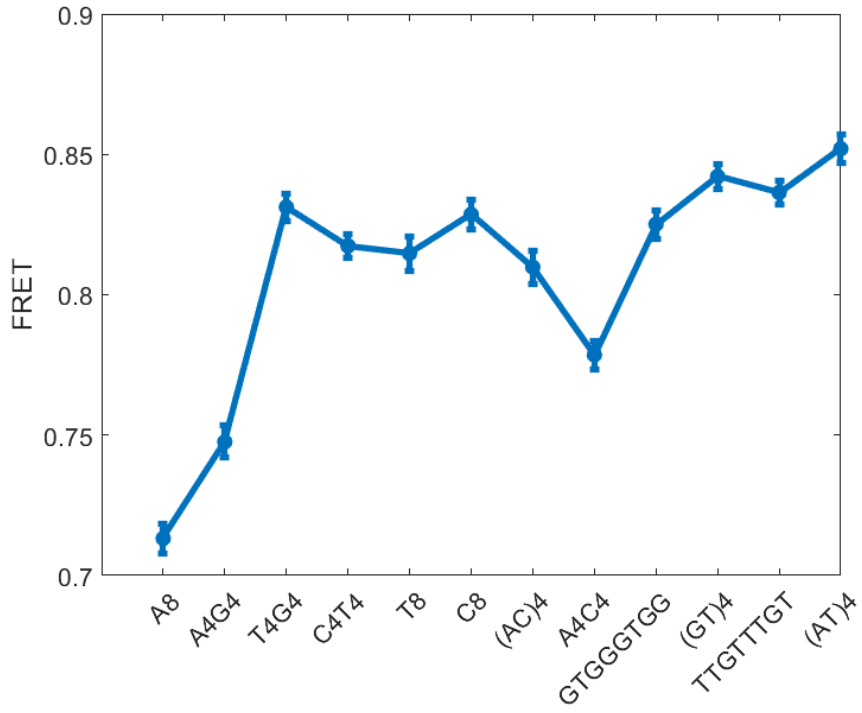


(b) Gaussian fit to the donor-acceptor population.



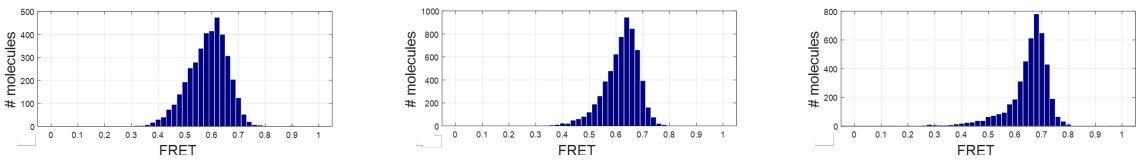
Now to answer our first question, can we actually measure any difference between the different sequences? It turns out we can. For instance at an salt concentration of 5 mM NaCl and 10 mM $MgCl_2$ we can the FRET efficiencies vary between 0.71 and 0.85 (see figure 3.2), a quite noticeable difference. The average standard deviation of the FRET peak is around 0.074. The standard error of the mean at approximately 0.005 is negligible however, due to the large amount of data points.

Figure 3.2: Average FRET efficiency for different sequences at 5 mM NaCl and 10 mM $MgCl_2$



To see if salt has any significant effect we tried out 5 different salt concentrations: 5 mM NaCl, 50 mM NaCl, 500 mM NaCl, 5 mM NaCl + 10 mM $MgCl_2$, and 5 mM NaCl + 100 mM $MgCl_2$. We clearly observe in figure 3.3 that with increasing ionic strength the average FRET values also increase, meaning that the ssDNA has become more flexible resulting in a lower end-to-end distance. This is likely caused by the positive ions neutralising the charge of the phosphate backbone.

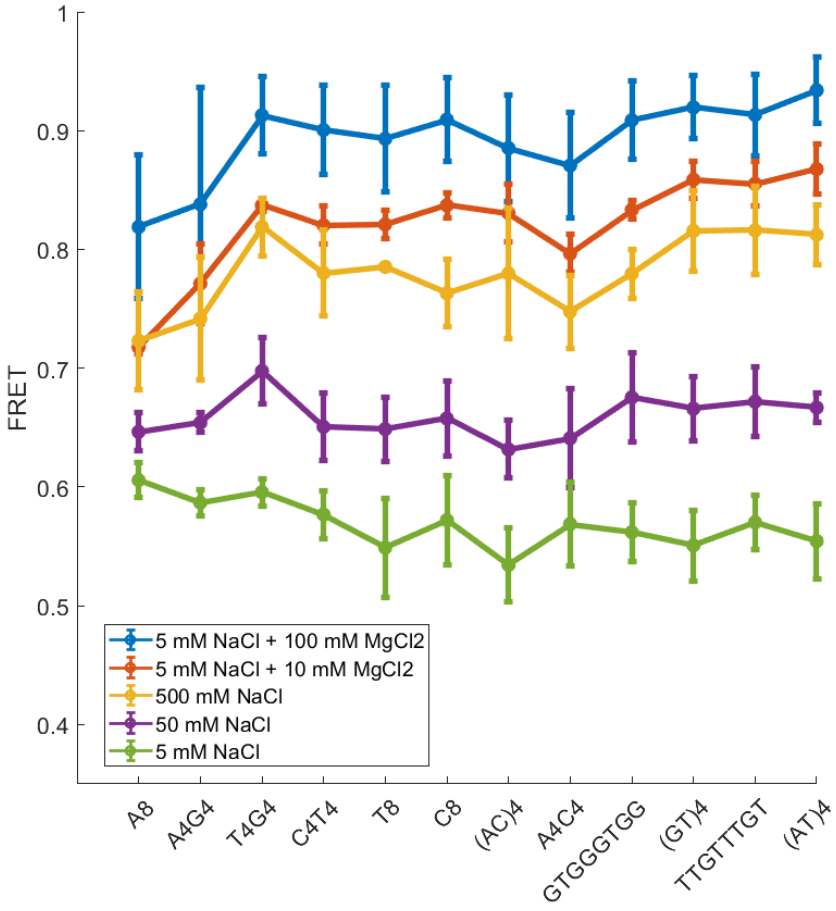
Figure 3.3: FRET histograms for the same sequence at 5, 50, and 500 mM NaCl. Average FRET value increases with ionic strength.



(a) 5 mM NaCl (b) 50 mM NaCl (c) 500 mM NaCl

Figure 3.4 summarises all the obtained FRET efficiencies for the 12 sequences and 5 different salt concentrations that were part of the experiments. Each data point is at least repeated thee different times.

Figure 3.4: Composite image of FRET efficiencies of 12 different sequences over 5 different salt concentrations



The error bars displayed here denote the standard deviation between the average FRET efficiencies over the three different experiments. As mentioned before, the standard error of the mean should be quite low given the large amount of molecules, in the order of 0.001. The error bars in the picture are in the range from 0.005 to 0.044, with some standard deviations being significantly higher then expected. A possible explanation to this could be slight differences in salt concentrations in the buffers between experiments. We can see from the data salt concentration has a strong effect on the measured FRET efficiencies. We can however see that some of the same trends hold true for different measurements. We can observe for instance that T4G4 always seems to be the sequence with the highest FRET efficiency, indicating that it is more flexible. The only exception to this is for the lowest salt concentration of 5 mM NaCl, indicating that at these low salt concentrations electrostatic interaction become more dominant. We expect that at higher salt concentrations charges along the backbone get more effectively screened, diminishing the electrostatic repulsion, making the ssDNA more flexible, resulting in higher FRET values. We also might expect that this

Table 3.1: Table summarising the ionic strength of the different buffers and the average FRET efficiency at these conditions

Buffer	Ionic Strength (mM)	Average FRET efficiency
5 mM NaCl + 100 mM MgCl2	305	0.89
5 mM NaCl + 100 mM MgCl2	35	0.82
500 mM NaCl	500	0.78
50 mM NaCl	50	0.66
5 mM NaCl	5	0.57

scales with the ionic strength. If we look at the ionic strength of the different buffer solutions and compare them with the results from figure 3.4)

We see from table 3.1 that the scaling of FRET efficiency with ionic strength holds true for the monovalent buffers. However it seems that $MgCl_2$ has a larger effect on the FRET efficiency then could be expected on basis of ionic strength. We can therefore assume see divalent ions such as $MgCl_2$ are more efficient at suppressing the negatively charged phosphate groups compared to monovalent ions such as Na^+, since even at a lower ionic strength, the presence of $MgCl_2$ ions causes the FRET efficiencies to be higher then the presence of Na^+ at an higher ionic strength. Since the average FRET efficiency of 500 mM NaCl and 5 mM NaCl and 10 mM $MgCl_2$ are quite similar, we can calculate that $MgCl_2$ ions are at least 15 times as effective at screening the phosphate backbone compared to Na^+ ions. It also seems that the more the charge of DNA is being screened by positive ions, the higher the variance in FRET efficiencies. For instance for the difference between the highest and lowest FRET efficiency at 5 mM of NaCl is only 0.07, while at 5 mM of NaCl and 100 mM of $MgCl_2$ it is almost double that at 0.012. This might be an indication that as electrostatic interactions are not very sequence specific, and as they weaken sequence specific effects become more pronounced.

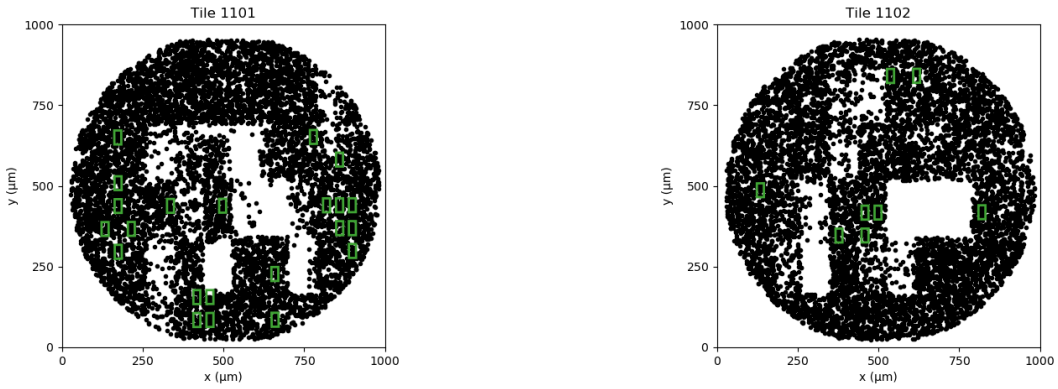
3.2. MiSeq chip

The experiments on the quartz slide gave a proof-of-concept that we can measure sequence dependant end-to-end distances using FRET. But to get a better idea of the base specific interactions, we have to explore more sequence space, and therefore we should move to the high-throughput platform.

But before we widen our net we first need to confirm if the results achieved on the quartz slide could be reproduced on the the high-throughput platform. The experiment was performed a MiSeq nano chip, which is capable of sequencing up to 1 million sequences. The Illumina sequencer doesn't sequence the entire chip, but only inside two round sequencing tiles. We therefore are also only interested in measuring the FRET values in these areas. We use the automatic stage scan the surface of these two tiles grid-wise.

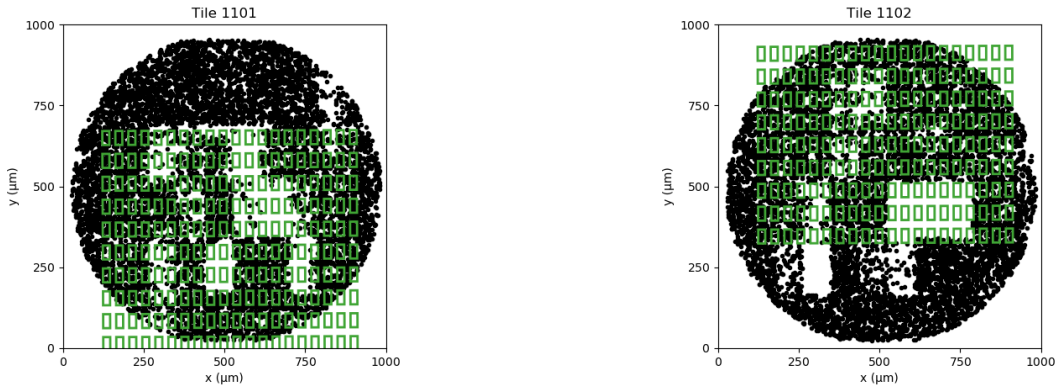
The image below shows the initial mapping between the single-molecule fluorescence data and the sequence data.

Figure 3.5: Image showing the initial matches



As can be seen from figure 3.5, only a few field of views can be found through this process. However since we know the images have been made in a regular grid we can use the stage coordinates to find the rest of the field of views.

Figure 3.6: Image showing all fields of view found using stage coordinates



As can be seen in figure 3.6 the image grid doesn't exactly line up with the sequencing tile. This is because the MiSeq doesn't contain a nicely defined origin point from which a measurement can be aligned. Therefore the microscope must be aligned by hand at the start of the experiment, which of course introduces room for error.

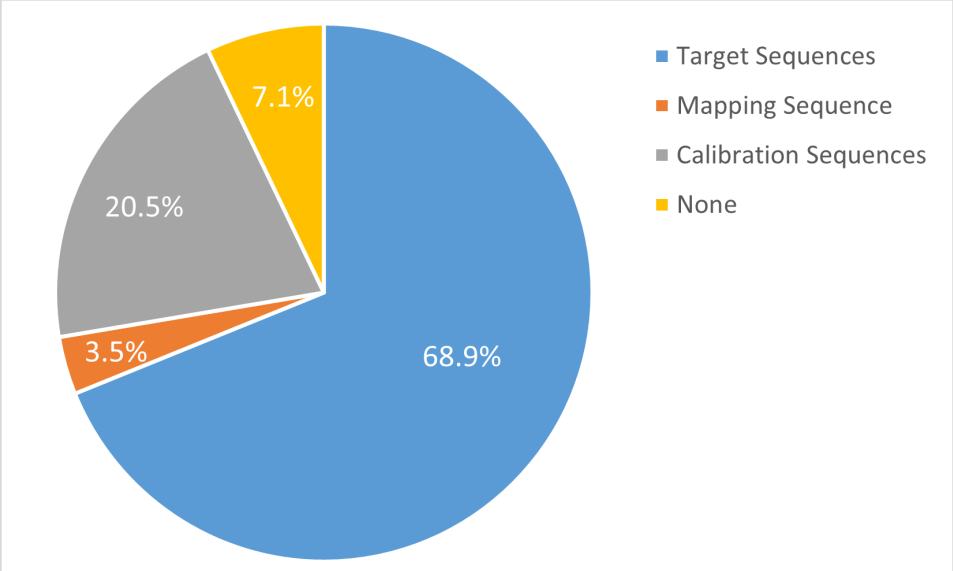
Next the sequence data is characterised. We have 3 different groups of DNA bound to the surface, the mapping sequence, calibration sequences, and target sequences (SKxx). We sequenced for a total length of 42 nucleotides. Sequencing only starts after the r1 primer. The first 42 nucleotides (after the r1 primer) of the different sequences can be found in the table below.

Table 3.2: The 42 first nucleotides after the r1 primer for the various sequences used in the experiment

Calibration sequence	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAATGCCTAGCCG
Mapping sequence	ACTGACTGTAACAACAACAACAATAACAACAACAACAATAAC
SK029	ACGTATCACGAAAAAAAAATCXCGTATGCCGTCTTCTGCTTG
SK030	ACGTATCACGAAAAGGGGATCXCGTATGCCGTCTTCTGCTTG
SK031	ACGTATCACGTTTTGGGGATCXCGTATGCCGTCTTCTGCTTG
SK032	ACGTATCACGCCCCTTTTATCXCGTATGCCGTCTTCTGCTTG
SK033	ACGTATCACGTTTTTTTTATCXCGTATGCCGTCTTCTGCTTG
SK041	ACGTATCACGCCCCCCCCATCXCGTATGCCGTCTTCTGCTTG
SK042	ACGTATCACGACACACACATCXCGTATGCCGTCTTCTGCTTG
SK043	ACGTATCACGAAAACCCCATCXCGTATGCCGTCTTCTGCTTG
SK044	ACGTATCACGGTGGGTGGATCXCGTATGCCGTCTTCTGCTTG
SK045	ACGTATCACGGTGTGTGTATCXCGTATGCCGTCTTCTGCTTG
SK046	ACGTATCACGTTGTTTGTATCXCGTATGCCGTCTTCTGCTTG
SK047	ACGTATCACGATATATATATCXCGTATGCCGTCTTCTGCTTG

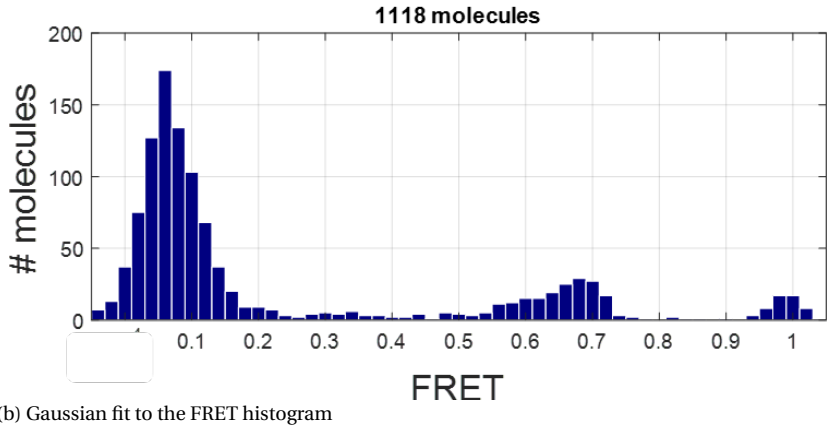
Out of 141,326 sequences, we characterise 97,329 (68.9%) as target sequences, 4,919 (3.5%) as mapping sequences, 29,031 (20.5%) as calibration sequence. This means that 92.9% of all sequences are coupled to a specific sequence and 10,047 (7.1%) are left uncharacterised.

Figure 3.7: Distribution of assigned sequence names

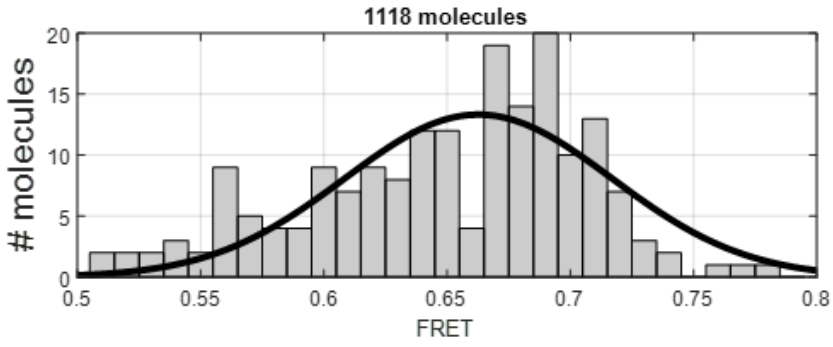


After peak finding we end up with 85,423 found molecules. Over all 12 target sequences, we end up with a total of 10,581 coordinates that have both single-molecule data and sequencing data. Intensity traces of the single-molecule data are then extracted and ordered per sequence. Looking at the traces for a specific sequence, we get a FRET histogram that look as follows:

(a) FRET histogram measured on the MiSeq chip



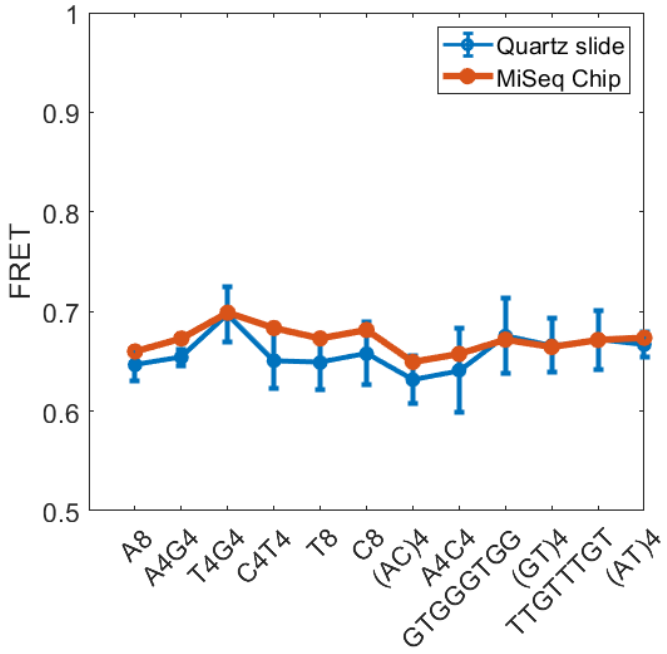
(b) Gaussian fit to the FRET histogram



In this histogram we can see three peaks instead of only two. We still have the donor-only peak around FRET efficiency 0.05 and the FRET peak between FRET efficiency 0.6 and 0.7, but also have an extra acceptor-only peak around FRET efficiency 1. The average standard deviation of the FRET peak is 0.039. Since we could only match 10,581 molecules, our number of molecules per sequence is actually quite much lower then on the quartz slide.

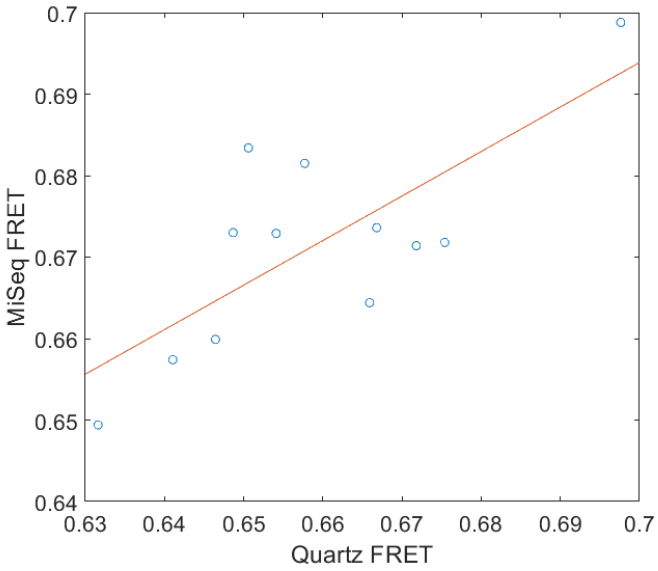
Calculating the FRET efficiencies the same way we did before, we can compare them to the low-throughput data. The results are summarised in figure 3.9.

Figure 3.9: FRET efficiencies for various sequences as measured on a quartz slide (blue) and the MiSeq chip (orange)



As can be seen from figure 3.9, the quartz slide data and the MiSeq data strongly correlate. This proves that the high-throughput pipeline is accurate enough to correctly characterise single-molecule spots and separate them based on sequence. This removes the need for different sequences to be physically separated like in the low-throughput method and opens up the possibility to measure a large amount of sequences on the same chip. If we plot the FRET efficiencies of both methods against each other, we see that they positively correlate with a Pearson correlation coefficient of $\rho = 0.75$

Figure 3.10: The quartz slide data and MiSeq date correlate with $\rho = 0.75$



4

Discussion

The low-throughput experiments we showed that using our DNA design we can measure FRET efficiencies across a short piece ssDNA. It was shown that these FRET values were different depending on the sequence, and our technique was sensitive enough to measure these differences. Next the same experiment was repeated using the high-throughput platform. This was done as a proof of concept for the high-throughput method that has the potential to measure many sequences at once by making use of next generation sequencing. As of now, the technique has not yet reached its full potential. In the following paragraphs possible improvements to the system will be discussed.

4.1. Theoretical versus achieved data points

The Illumina sequencer is able to sequence 1 million unique reads across both tiles. However the FRET efficiencies are calculated on the basis of only 10,581 molecules. What is the reason that the number of usable molecules is so much smaller then the theoretical limit? And can we further improve on the method?

First of all, the sequencer only found a total of 581,205 clusters, of which 479,464 where able to be sequenced successfully, which is about 82%. This means that the cluster density on the chip could be increased to better utilise it's full potential.

Still, for a molecule to be included in the final data set, it needs to have both single-molecule fluorescence data, as well as have sequencing data. Looking at the image below we can see in the sequencing tile roughly outlined by the black dots, and we can see the imaged fields of view as

the green rectangles. As can be clearly seen, the green rectangles don't cover the entire tile, in two separate ways. Firstly, the grid doesn't align perfectly with the sequencing tile. Since the MiSeq chip doesn't have a clearly defined origin point, it's difficult to perfectly center the microscope on the sequencing tile. Secondly, there is also space in between the fields of view. Additionally, during the peak finding process, a certain margin around the image is disregarded. This is to ignore any peaks that are partially cut off by the border.

Firstly, measuring the overlap between the imaging grid and the sequencing tile, we find a percentage of 66% for the first tile and a percentage of 64% for the second tile. The imaged area of the fields of view is 44% of the imaging grid. The fields of view are saved as 512 by 512 pixel images, meaning both the donor and acceptor images are 512 by 256 pixels. Since a border of 20 pixels is excluded from peak finding, this means that only 78% of molecules in a field of view can be found.

Putting this all together, of all available molecules on the surface, we expect only 22% to be found by the peak finding algorithm. Assuming that every cluster has grown from a single piece of DNA, we would expect to find $8.9 \cdot 10^4$ single-molecule spots. The analysis code returns 85,423 molecules, which is around what we expected. Similarly for the amount of sequences present in this region we can expect to find $1.4 \cdot 10^5$ sequences. The analysis code returns 141,326 sequences, which corresponds to the calculated value.

To explain why the amount of molecules is lower then the amount of clusters we can think of a couple of reasons. Firstly, the software might not correctly find all molecules. If two molecules are two close to each other, they don't form a well-defined Gaussian peak, but rather a dumbbell shape, and the software will not find them. Alternatively, there might be other reasons why there some peaks are not found, such as defects on the surface. Fluorophores might also bleach before being imaged. Lastly, some of the DNA attached to the surface might not actually be labeled with a fluorescent dye. This DNA will not show up in the fluorescent image but can still be sequenced. The labelling efficiency has been determined after the labeling process. The average labelling efficiency, weighted for the relative occurrence of each sequence, is 84%. This alone cannot explain the difference between the expected amount of molecules and the measured spots. The rest might be explained by the other factors mentioned before.

Out of these 141,326 sequences and 85,423 single-molecule spots, only 10,581 are successfully mapped, meaning that they have both sequence data and single-molecule fluorescence data. This is low but inline with other experiments.

4.2. The different populations in the FRET histogram

If we look at the FRET histograms, we can distinguish three major groups, molecules that only have the donor fluorophore (FRET efficiency ≈ 0%), molecules that have both donor and acceptor (FRET efficiency between 60% and 70%), and molecules that only have acceptor fluorophores (FRET efficiency of ≈ 100%).

4.2.1. Donor-only molecules

The donor-only molecules make out by far the largest group of molecules. Ideally all molecules would have both donor and acceptor fluorophores, since only then we can measure our desired FRET efficiencies. When looking at the MiSeq data, $7.1 \cdot 10^3$ molecules could be attributed to donor-only molecules, while only $1.7 \cdot 10^3$ could be attributed to molecules with both donor and acceptor fluorophores. This means that only 19% of the molecules are double-labeled. The are multiple reasons why there are so many donor-only molecules, such as: poor labeling efficiency, poor hybridisation efficiency, bleaching of the fluorophores, poor mapping of the cy3 and cy5 channels.

Labelling efficiency

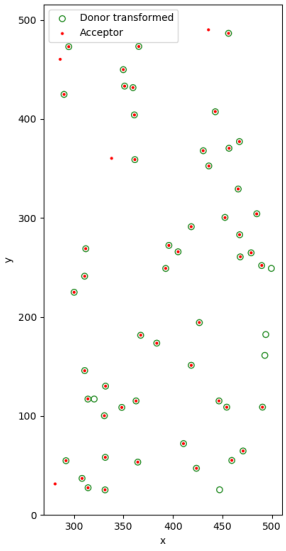
Firstly looking at the labeling efficiency, we know that our sample has an average labeling efficiency of 84% for the cy3 (donor) dye. If the labeling efficiency of the cy5 is much lower, this could explain the perceived difference. However, the labeling efficiency of the cy5 dye is measured to be 96%, which is even higher then for the cy3 dye. This could therefore not be an explanation.

Channel mapping

If the mapping between the cy3 channel and the cy5 channel is very poor, we could end up with a situation in which the donor peaks do have corresponding acceptor peaks, but are erroneously mapped to a random patch of background.

If the mapping between the cy3 and cy5 channel would be poor, we expect to see only a few donors and acceptors being matched. We could also don't expect to see the centers of the donors to consistently line up with the centers of the acceptors. This would give the appearance that are a lot of donor-only molecules. However, if we look at the at the mapping file, we find the opposite. A large majority of the donors are matched to acceptors, and centers align nicely.

Figure 4.1: Image showing the mapping between donors (green circles) and acceptors (red dots)



Bleaching

Another reason for the large donor-only population could be if for some reasons a lot of fluorophores are being bleached before the measurement. This doesn't have to be specific to the acceptor fluorophores, since there might also be a lot of molecules with only a cy5 dye, but we cannot see them since they are not directly illuminated by the green laser. That being said, the acceptor dyes are still more susceptible to bleaching.

Each position is imaged for 5 seconds, if we compare the number of fluorophores in the first and last frame of each movie we can get a idea of how much photo-bleaching has taken place. We find that for the cy3 dyes, $13 \pm 0.92\%$ of the fluorophores have been bleached by the end of the movie, and for the cy5 dyes, $28 \pm 2.7\%$ has been bleached by the end of the movie. This alone is not sufficient to explain the donor-only population, especially because for analysis only frames 3 through 12 have been taken into account, which responds to 0.3 to 1.2 seconds after the start of imaging. At this point even less molecules should be bleached.

Bleaching could however also occur before imaging. Maybe a large portion of fluorophores are being bleached due to the ambient light in the lab, which could explain why we see so many donor-only molecules. Each time a fluorophore absorbs a photon, there is a change of photo-bleaching. There for to get an idea if this would be plausible explanation, we should compare the amount of photons delivered by the environment compared to during the experiment, which is equivalent with the average energy delivered by each source.

First of all, our laser has a max output of $500mW$. This energy is spread over a beam spot of approximately $64\mu m$ in diameter. This gives us, combined with the 5 second imaging time, a energy density of around $7.8 \cdot 10^8 J/m^2$. According to the International Energy Conservation Code (IECC 2021) a professional laboratory should have a recommended lighting power density of $14.3 W/m^2$ [**?**]. I've done a total of 17 experiments with cy3-cy5 pair. Taking an average of 4 hours per experiments, and assuming the sample is continuously left exposed to the light, this gives us a delivered energy density of $3.5 \cdot 10^6 J/m^2$. This is two order of magnitude lower then the energy delivered during the experiment. Combined with the fact that the samples often are inside a covered box of ice, and the fact that the indoor lighting is not of a single wavelength, but actually a broad spectrum, most of which is not even able to excite the fluorophores. This makes it very unlikely that the samples have been significantly bleached before the experiment started by artificial lighting.

Hybridisation efficiency

Lastly, the large amount of donor-only molecules might due to poor hybridisation between the two DNA-strands. If the cover strand is not bound to the primary strand, there will be no acceptor dye. Reasons why the hybridisation efficiency might be low could be a too low concentration of monovalent sodium ions or too steep temperature gradient when cooling down the DNA. To check if this might be the case, a new sample was hybridised, now with NaCl at $200mM$ instead of $50mM$ and using a PCR machine to carefully let the temperature drop over a period of approximately 8 hours. Measuring this new sample still saw only 24% of molecules labeled with both fluorophores, so no significant improvement over the original sample.

If the problem does not lie in the hybridisation conditions, it might be a problem with composition of the primary and cover strand. If for some reason the sequence of some of the molecules deviates from the intended sequence and the bases are not complementary, hybridisation can not occur. However, we know that of the primary, strand upwards of 97% of the sequences are characterised, meaning we can be fairly sure that these sequences are indeed correct. As per design, we don't have sequencing data on the cover strand. More over, during the original labeling of the cover strand, there appeared to be 6 times as much DNA as ordered. The DNA was then appropriately diluted and used as is. However this might be in indication that the ordered DNA sample was not pure, or something in the process had gone wrong. If the sample contained other DNA next to the ordered sequence, this could explain why the hybridisation efficiency is so low. The would not be equal amounts of cover strand and primary strand during hybridisation, resulting in only part of the DNA molecules to be successfully hybridised.

4.3. Recommendations

To improve the pipeline, a couple of changes to the procedure can be implemented. I will discuss them here, roughly in order from easy to hard.

First of all, it's important to change the imaging grid such that there are no more gaps between the fields of view. This makes sure the entire surface actually gets appropriately scanned, and makes better use of the MiSeq chip.

Related to this it might also be a good idea to "oversize" the imaging grid. By making the imaging grid, it makes it more likely to capture the entire sequencing tile. This of course comes at the expensive of more measurement time. However in our case this is not a big issue at 5 seconds a movie and 400 movies, it takes little over half an hour to image everything. Quadrupling the imaging grid makes it way more likely to measure the entire sequencing tile, while still keeping the total measuring time manageable. If longer movies need to be made or a sequencing chip is used that has more tiles, this might not be possible.

Another quick improvement is to increase the cluster density by adding a higher concentration of DNA. The DNA concentration necessary to achieve the ideal cluster density is not trivially determined, and requires some trial and error. DNA concentrations of samples also tends to vary over time as DNA tends to stick to the sides of eppendorfs over multiple freeze and thaw cycles. Care also has to be taken into not overdoing this, since a too high cluster density is also detrimental to the sequencing.

To decrease the donor-only peak, it might be a good idea to repeat all the labelling and hybridisation steps starting from fresh to see if anywhere in the process something went wrong. Preform the hybridisation on the PCR machine instead of the heat block for better temperature control, and using higher salt concentrations in the buffer.

4.4. Improvements on the Illumina platform

As has been shown in this thesis, the Illumina-FRET platform is a powerful tool for the research of ssDNA. However, the experiments as shown could (and have) been performed on a quartz slide as well. The true advantage of the Illumina-FRET platforms shines through when moving to high-throughput, when imaging tens, hundreds or even thousands of sequences simultaneously. It might therefore be interesting to look into what might be some of the theoretical limits of the platform, and what kind of throughput we can expect.

Three different scenarios are evaluated: one conservative scenario in which we can only improve

some minor issues, a more realistic scenario in which we also improve some more major problems, and an optimistic scenario in which we make some big break-trough improvements.

For the conservative estimate no significant improvements the mapping between the single-molecule fluorescent data and the sequencing data can be achieved. Assumed is also that fraction of molecules labeled with both cy3 and cy5 cannot be improve. Some minor issues have been resolved though: full imaging coverage of the sequencing tiles and increased cluster density to recommended levels. These small improvements would give the ability to analyse up to 120 different sequences on the MiSeq Nano, and up to 1500 sequences on the regular sized chip using the v3 reagent kit.

For a realistic improvement scenario labeling process is improved such that the fraction of double-labeled molecules is increased to 80%. Mapping efficiency is also improved to 12.5 %, which is more in line with other experiments. With these improvement we can expect to analyse up to 900 sequences on the MiSeq Nano and up to 11,000 sequences on the regular chip.

For the optimistic scenario we require some major improvements on the mapping of single-molecule fluorescence data to the sequencing data. In an ideal scenario this would improve this to 90%. Reducing the amount molecules per sequence permits to explore more sequence space. This comes at a cost of less accuracy. Still, reducing the amount to just 10 molecules per sequence would theoretically have a standard variation of just 0.012. Making these changes, the high-throughput technique would be able to analyse up to 73,000 molecules on the MiSeq Nano chip and up to 910,000 molecules on the regular sized chip.

5

Conclusion

5.1. Applications

5.1.1. Model verification

One of the possible applications for the technique described in this thesis is to validate potential theoretical models for ssDNA. ssDNA is widely used in all kinds of nano-fabrications, and an improved theoretical basis could help in the development of those. These theoretical models would have certain predictions regarding flexibility, average end-to-end distance, and short-length interactions and could be verified using our technique. Measurements such as FRET, but also other structural imaging techniques such as atomic force microscopy (AFM), optical/magnetic tweezers or X-ray diffraction could be done before, but are usually quite low-throughput and are therefore limited to just a handful of sequences. This of course is less desirable. Using high-throughput FRET, and more significant part of the sequence space can be explored. This could also to an iterative loop in which models are continuously updated and improved, which could lead to a better physical understanding of certain sequence-specific structural interactions.

5.1.2. Aptamer screening

Aptamer a short pieces of ssDNA (typically <100 nucleotides) that can bind specifically to certain targets. Aptamers are quite similar to antibodies in this way, but made from nucleotides instead of aminoacids. Normally these apatamers are developed through a process called Systematic Evolution of Ligands by Exponential Enrichment (SELEX).This is cyclic process which produces aptamers with high affinity and specificity. When a aptamer sequence is known, it is actually quite easy to produce them. They can often be ordered through regular oligo providers. This is in contrast to antibodies, which often need to be harvested from (animal) blood plasma, which make them quite expensive to produce and purify.

Because of their binding properties and affordability, aptamers are an interesting candidate for creating biosensors for disease diagnostics. When creating these sensors, selectivity is of course of high concern. SELEX doesn't generally result in a single aptamer, but rather a family of aptamers that all show high affinity and selectivity binding. Using our high-throughput platform, this final library can be screened for highest affinity and selectivity, to ultimately select the best preforming one.

High-throughput screening methods of aptamers have been developed before using next generation sequencing. These often however take measurements at the cluster level, not the single-molecule level. This makes it difficult to measure some kinetics, such as on and of binding times. [11] [29] [18]

5.1.3. Transient-binding aptamers

SELEX generally creates high-affinity binding aptamers, but this might not always be something that is desired. Sometimes transient-binding, binding for only a couple of seconds, might be more desirable. For instance, in super-resolution microscopy, "blinking" fluorophores are used to move past the diffraction limit. DNA paint is a technique in which this blinking is achieved by two strands of DNA that have transient-binding properties with each other. DNA paint often involves the need for linkers, for example an antibody, which hampers the resolution through the so called linkage error. Antibodies are relatively large at around $15nm$ in size. Aptamers have the advantage that are often quite small at around $2nm$, which would severely decrease the linkage error.

As stated before, SELEX produces high-affinity, high-selectivity aptamers, not transient-binding aptamers. Additionally, on and off binding times can only really be determined when looking at individual molecules. To create a transient-binding aptamer, one could start at high-affinity aptamer and create a library of mutants. Some of these mutations would hopefully lower its affinity to

the target. This resulting library can then be screened for the desired binding characteristics using high-throughput FRET measurements.

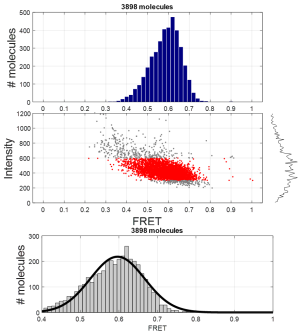
A

Appendix

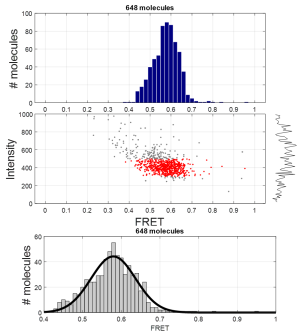
A.1. 22-06-2021 FRET measurements at various salt concentrations

A.1.1. 5mM NaCl

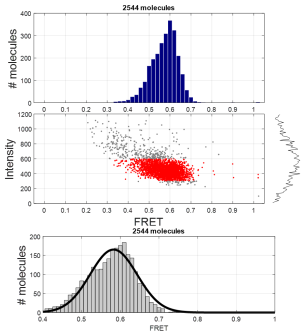
(a) AAAAAAAA



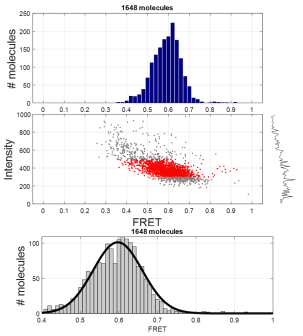
(b) AAAAGGGG



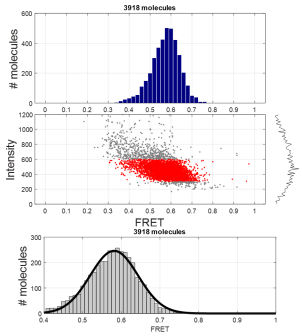
(c) TTTTGGGG



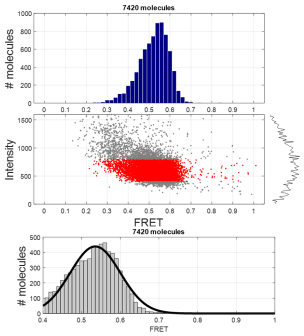
(d) CCCCTTTT



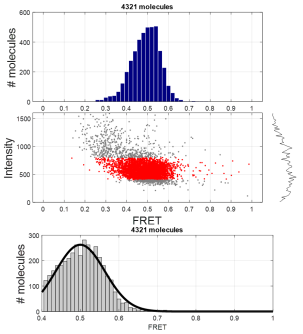
(e) TTTTTTTT



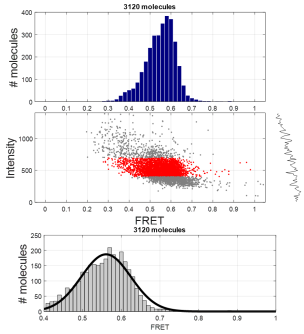
(f) CCCCCCC



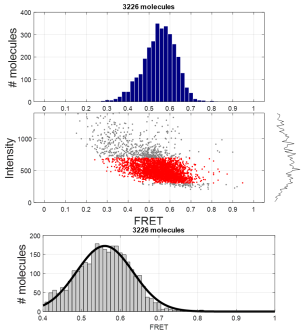
(g) ACACACAC



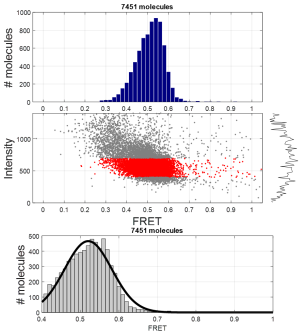
(h) AAAACCCC



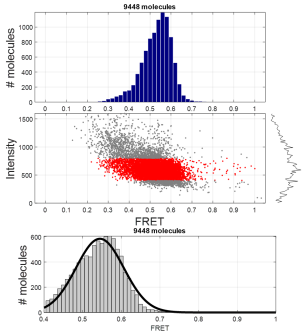
(i) GTGGGTGG



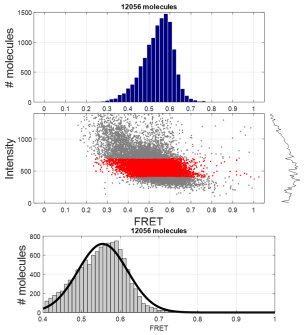
(j) GTGTGTGT



(k) TTGTTTGT

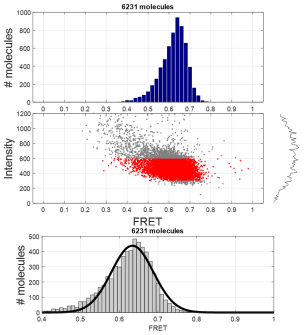


(l) ATATATAT

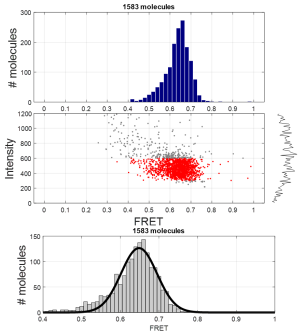


A.1.2. 50 mM NaCl

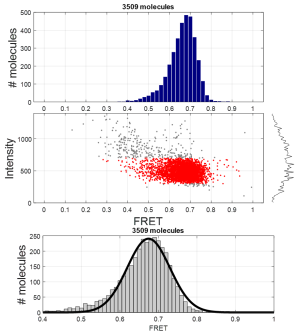
(a) AAAAAAAA



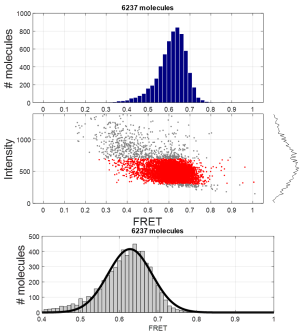
(b) AAAAGGGG



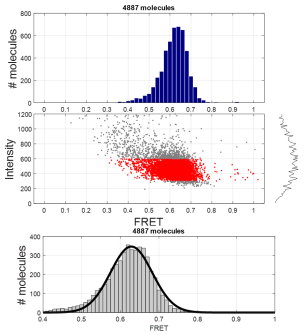
(c) TTTTGGGG



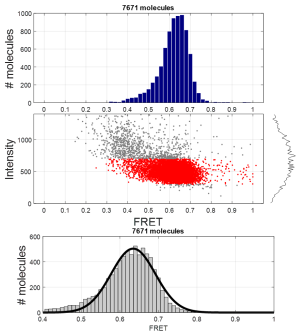
(d) CCCCTTTT



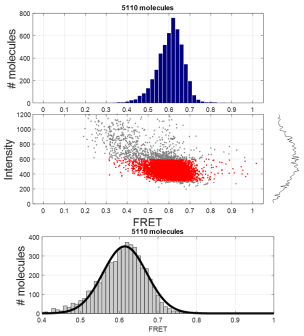
(e) TTTTTTTT



(f) CCCCCCC

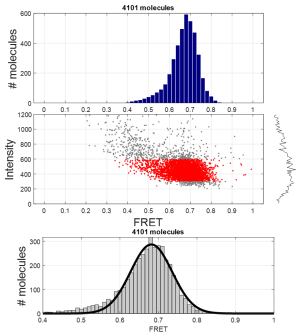


(g) ACACACAC

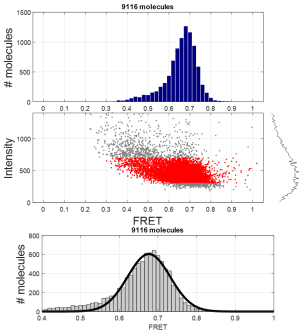


(h) AAAACCCC

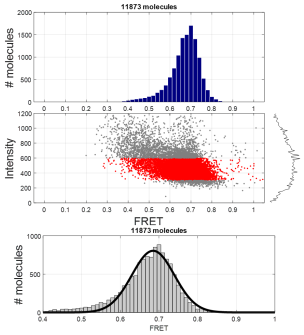
(i) GTGGGTGG



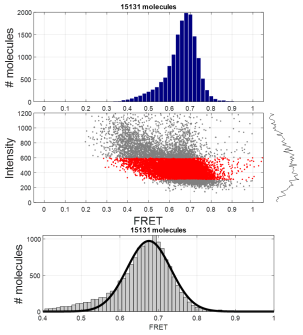
(j) GTGTGTGT



(k) TTGTTTGT

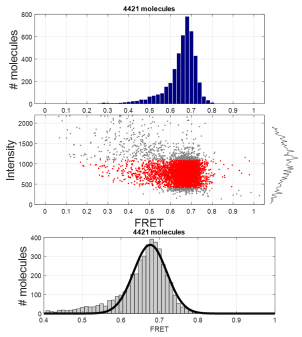


(l) ATATATAT

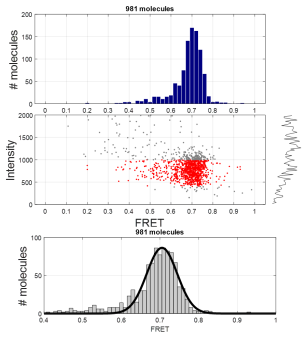


A.1.3. 500 mM NaCl

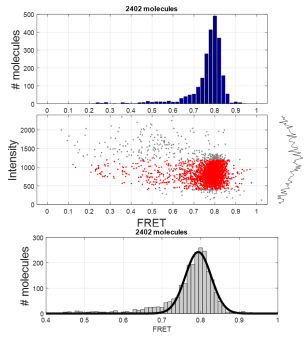
(a) AAAAAAAA



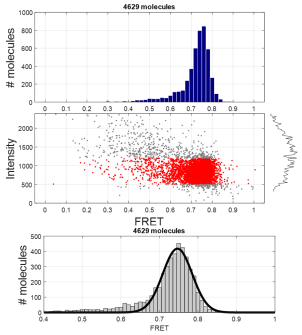
(b) AAAAGGGG



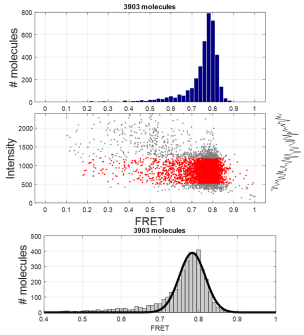
(c) TTTTGGGG



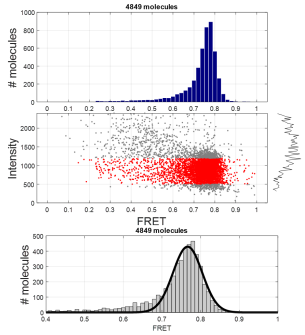
(d) CCCCTTTT



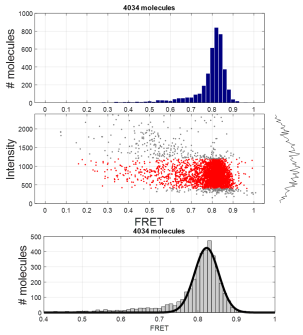
(e) TTTTTTTT



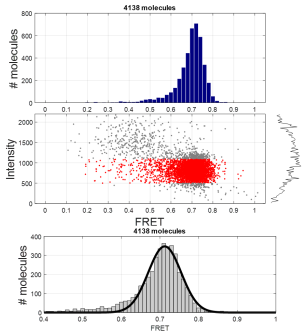
(f) CCCCCCC



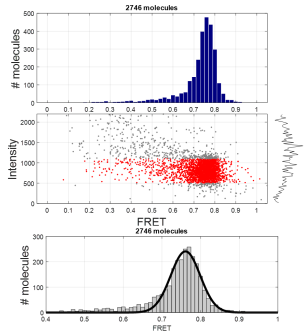
(g) ACACACAC



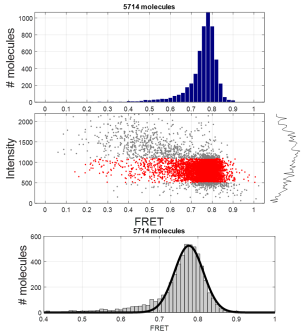
(h) AAAACCCC



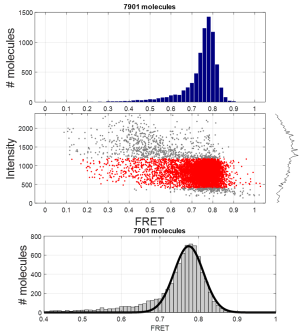
(i) GTGGGTGG



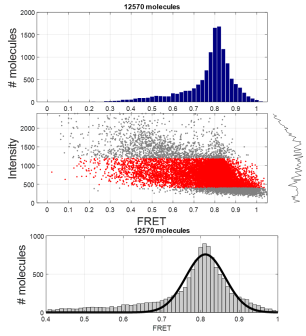
(j) GTGTGTGT



(k) TTGTTTGT

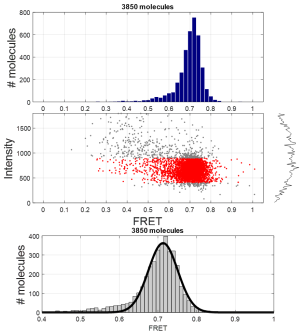


(l) ATATATAT

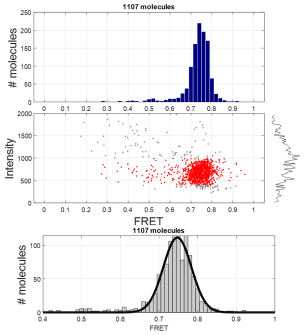


A.1.4. 5mM NaCl + 10 mM MgCl2

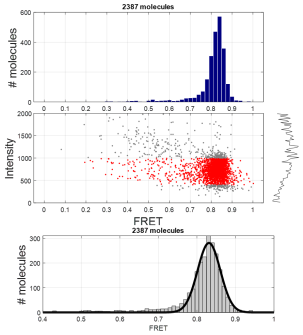
(a) AAAAAAAA



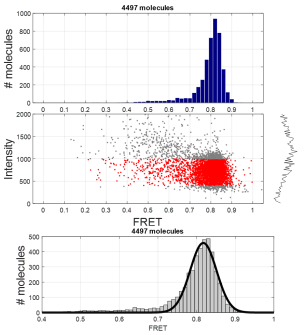
(b) AAAAGGGG



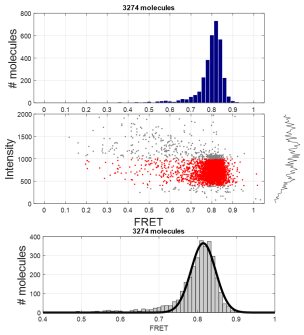
(c) TTTTGGGG



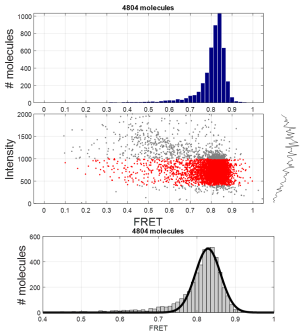
(d) CCCCTTTT



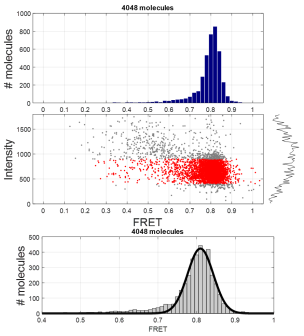
(e) TTTTTTTT



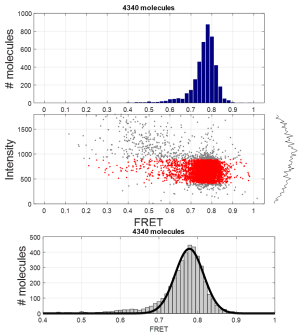
(f) CCCCCCC



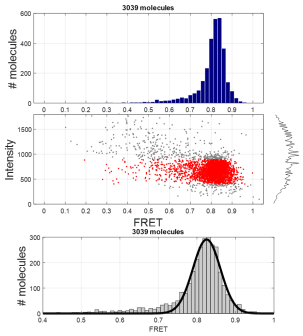
(g) ACACACAC



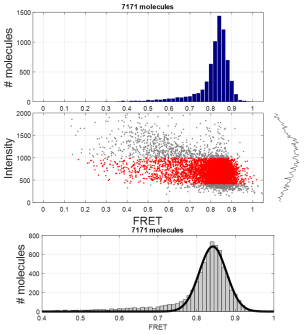
(h) AAAACCCC



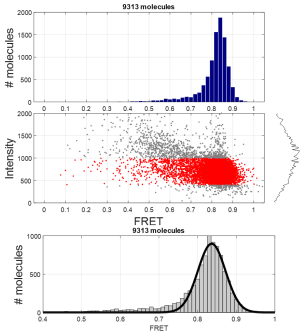
(i) GTGGGTGG



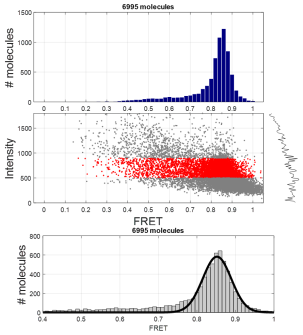
(j) GTGTGTGT



(k) TTGTTTGT

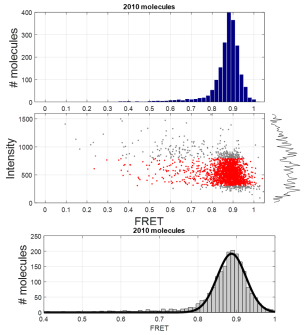


(l) ATATATAT

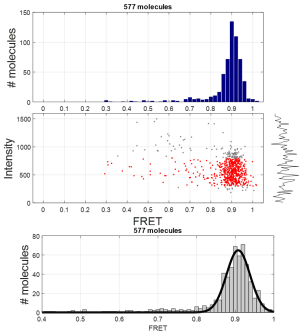


A.1.5. 5mM NaCl + 100 mM MgCl2

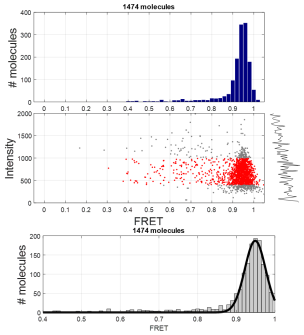
(a) AAAAAAAA



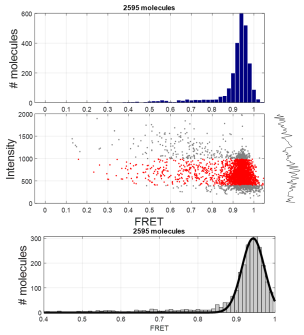
(b) AAAAGGGG



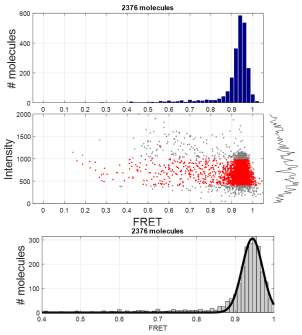
(c) TTTTGGGG



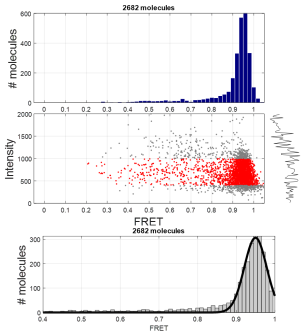
(d) CCCCTTTT



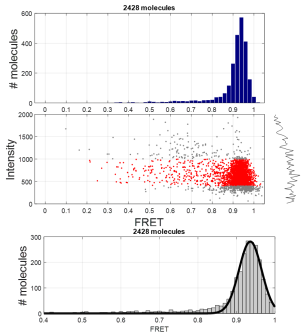
(e) TTTTTTTT



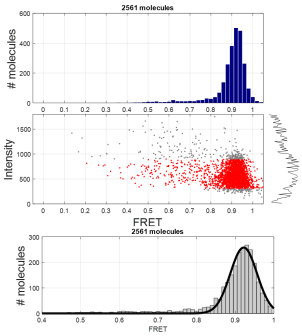
(f) CCCCCCC



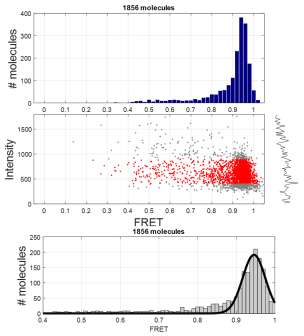
(g) ACACACAC



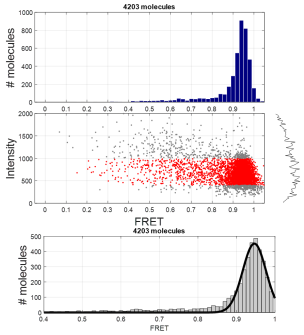
(h) AAAACCCC



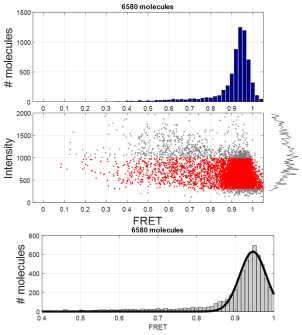
(i) GTGGGTGG



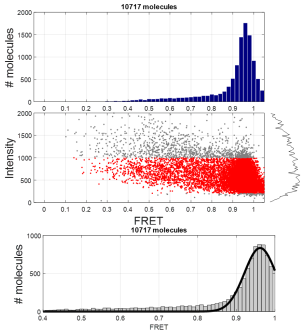
(j) GTGTGTGT



(k) TTGTTTGT



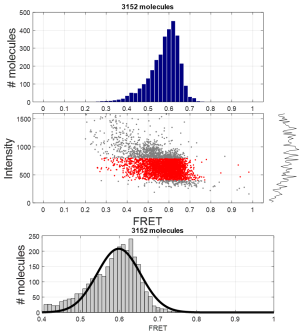
(l) ATATATAT



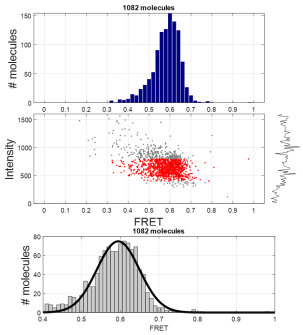
A.2. 23-06-2021 FRET measurements at various salt concentrations

A.2.1. 5mM NaCl

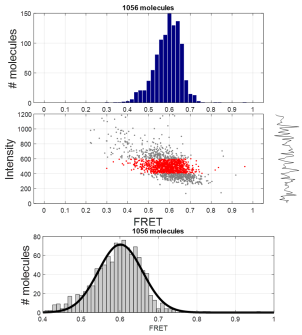
(a) AAAAAAAA



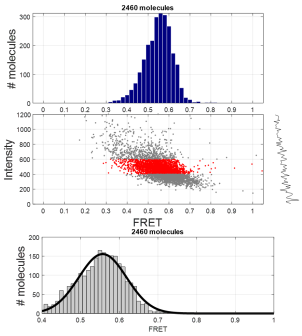
(b) AAAAGGGG



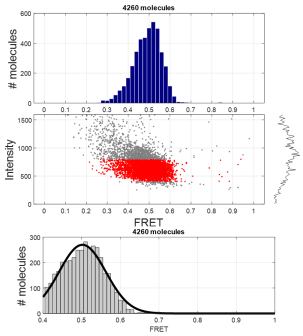
(c) TTTTGGGG



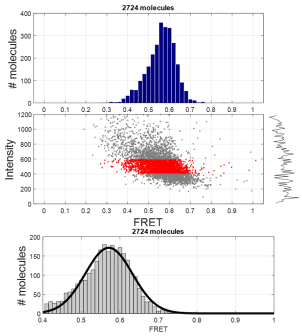
(d) CCCCTTTT



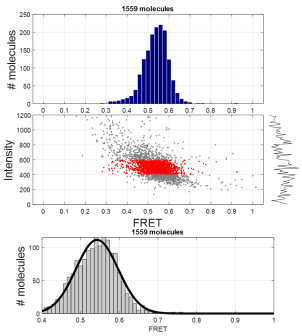
(e) TTTTTTTT



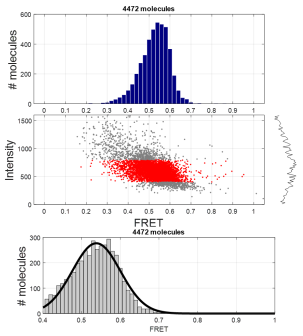
(f) CCCCCCC



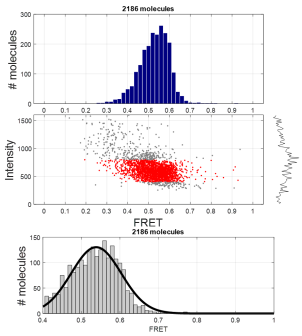
(g) ACACACAC



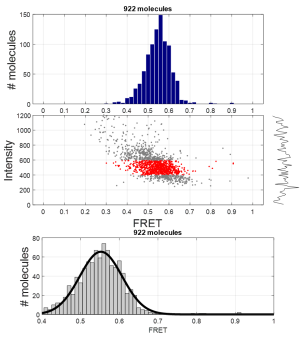
(h) AAAACCCC



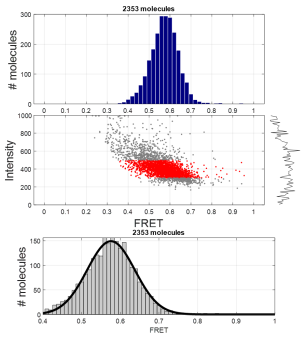
(i) GTGGGTGG



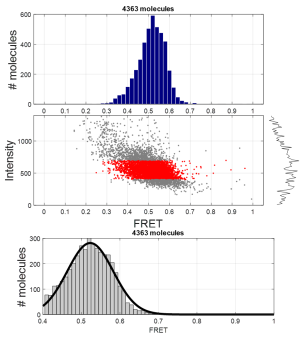
(j) GTGTGTGT



(k) TTGTTTGT

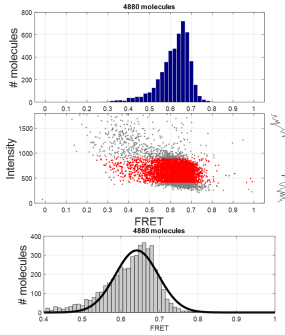


(l) ATATATAT

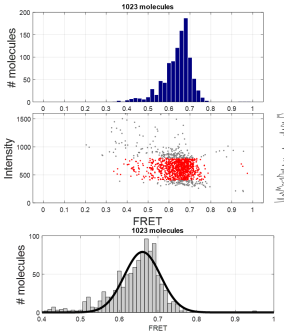


A.2.2. 50 mM NaCl

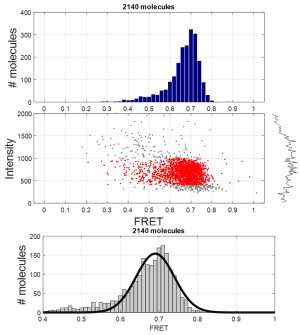
(a) AAAAAAAA



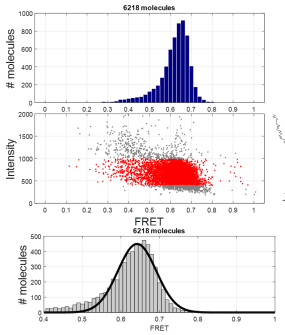
(b) AAAAGGGG



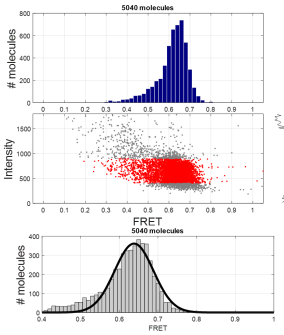
(c) TTTTGGGG



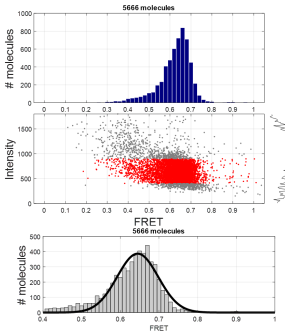
(d) CCCCTTTT



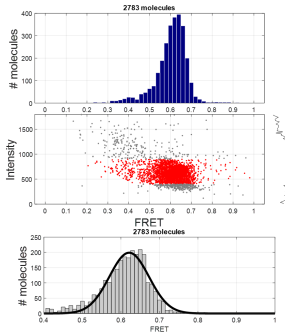
(e) TTTTTTTT



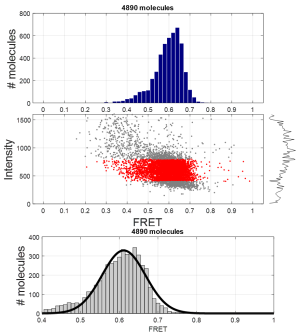
(f) CCCCCCC



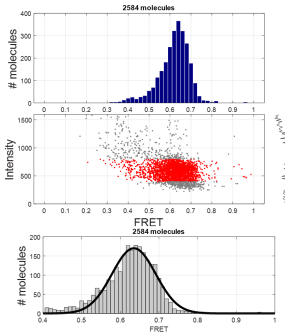
(g) ACACACAC



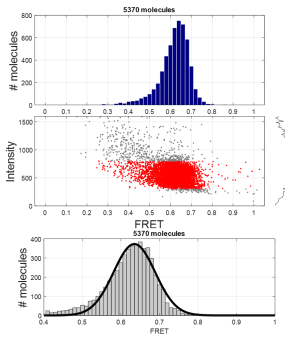
(h) AAAACCCC



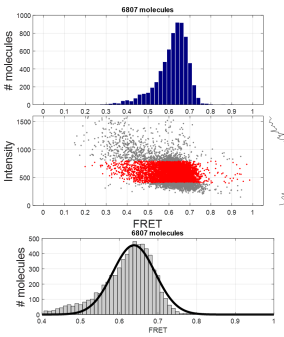
(i) GTGGGTGG



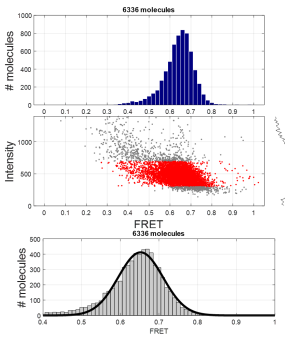
(j) GTGTGTGT



(k) TTGTTTGT

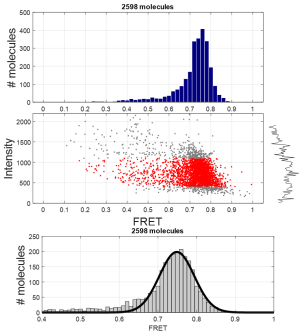


(l) ATATATAT

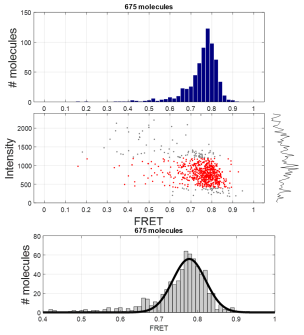


A.2.3. 500 mM NaCl

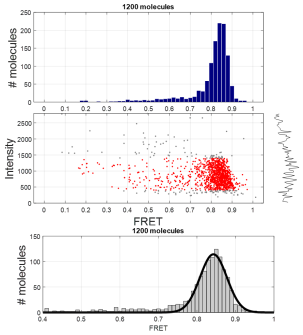
(a) AAAAAAAA



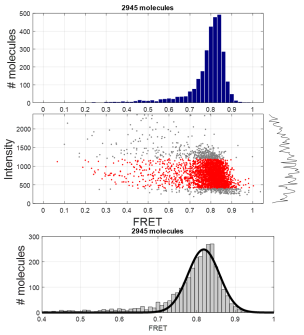
(b) AAAAGGGG



(c) TTTTGGGG

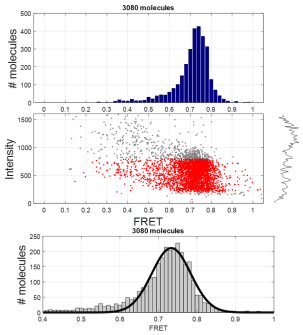


(d) CCCCTTTT

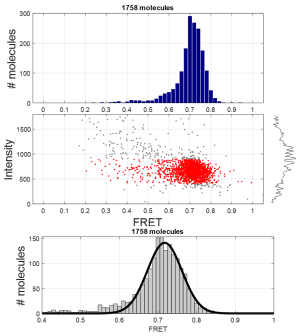


(e) TTTTTTTT

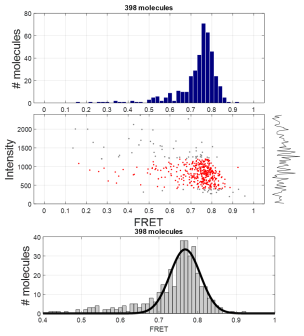
(f) CCCCCCC



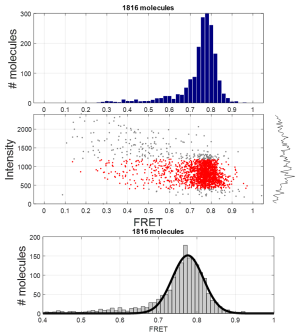
(g) ACACACAC



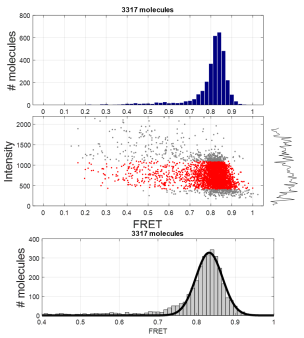
(h) AAAACCCC



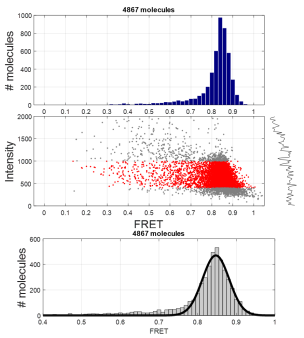
(i) GTGGGTGG



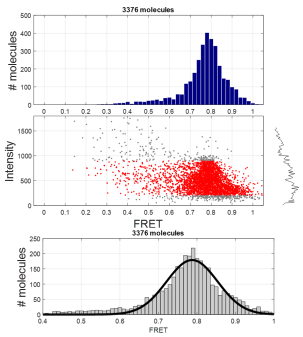
(j) GTGTGTGT



(k) TTGTTTGT

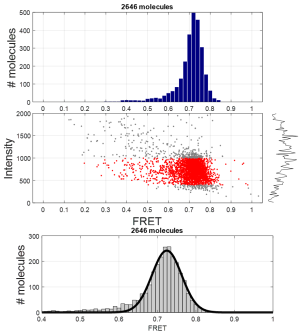


(l) ATATATAT

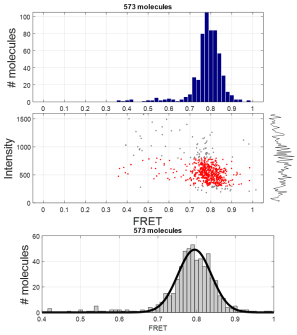


A.2.4. 5mM NaCl + 10 mM MgCl2

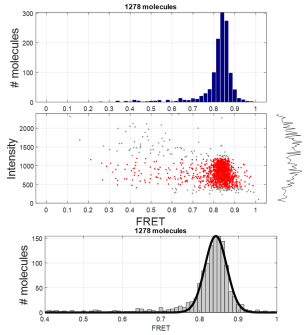
(a) AAAAAAAA



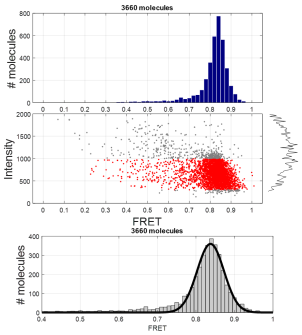
(b) AAAAGGGG



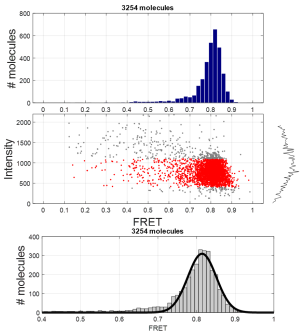
(c) TTTTGGGG



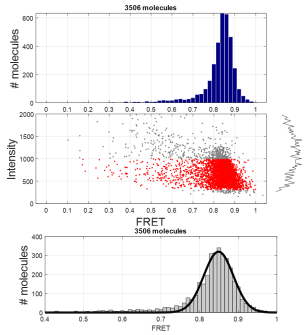
(d) CCCCTTTT



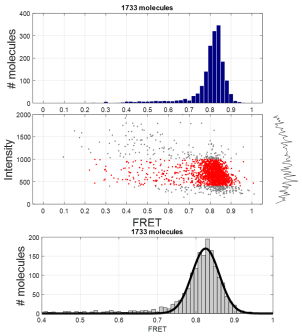
(e) TTTTTTTT



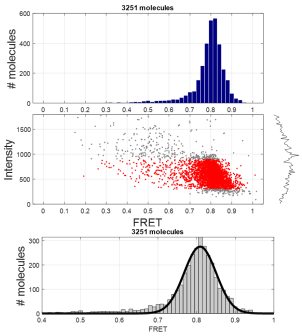
(f) CCCCCCC



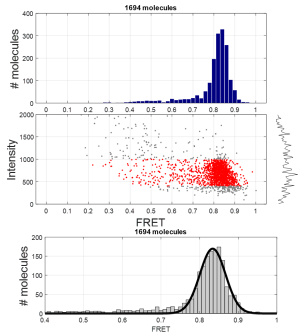
(g) ACACACAC



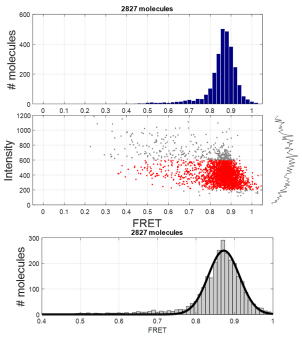
(h) AAAACCCC



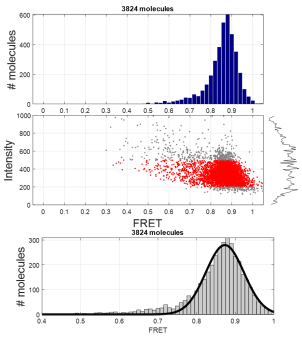
(i) GTGGGTGG



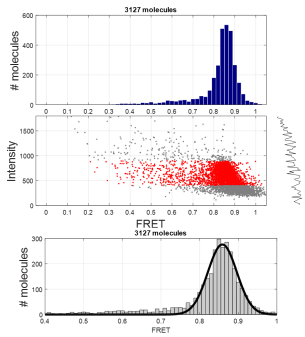
(j) GTGTGTGT



(k) TTGTTTGT

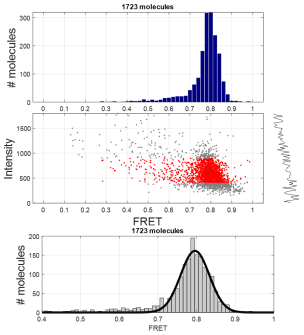


(l) ATATATAT

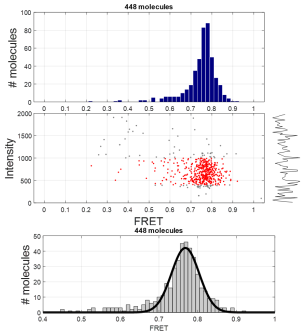


A.2.5. 5mM NaCl + 100 mM MgCl2

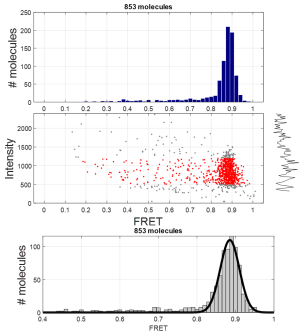
(a) AAAAAAAA



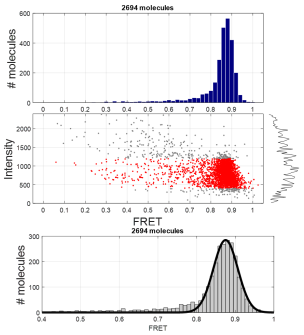
(b) AAAAGGGG



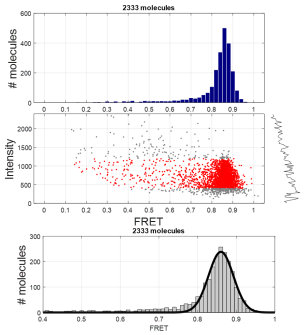
(c) TTTTGGGG



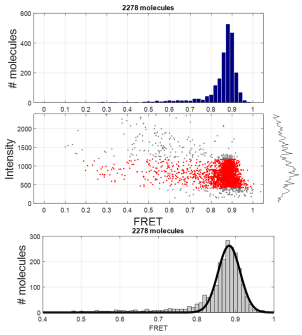
(d) CCCCTTTT



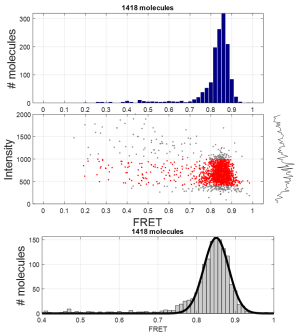
(e) TTTTTTTT



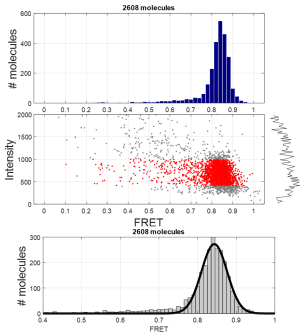
(f) CCCCCCC



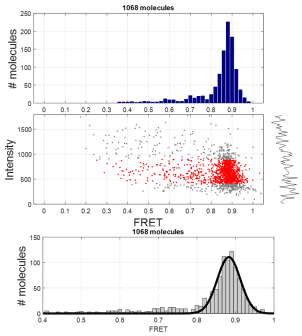
(g) ACACACAC



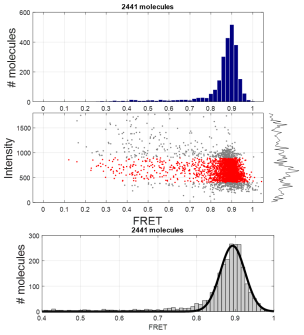
(h) AAAACCCC



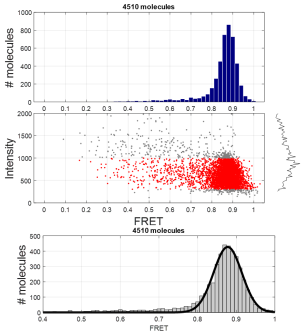
(i) GTGGGTGG



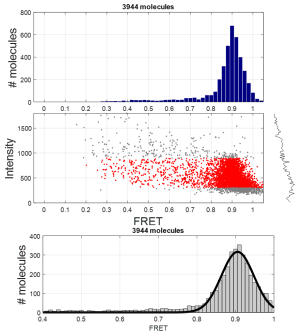
(j) GTGTGTGT



(k) TTGTTTGT

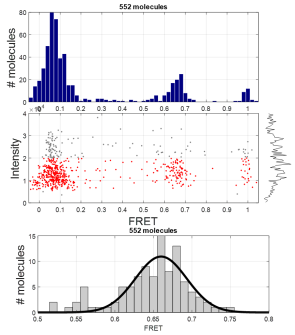


(l) ATATATAT

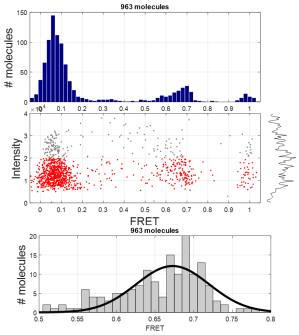


12-07-2021 High-throughput FRET measurements

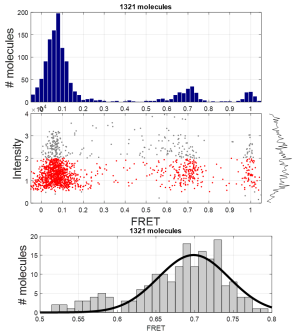
(a) AAAAAAAA



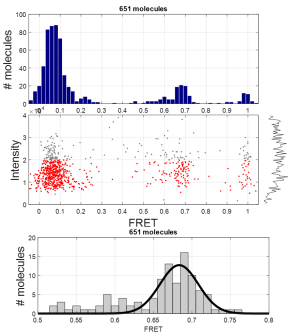
(b) AAAAGGGG



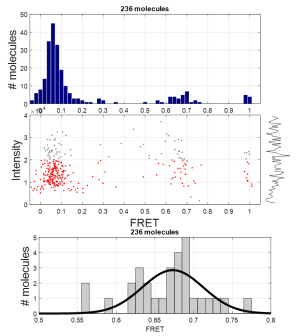
(c) TTTTGGGG



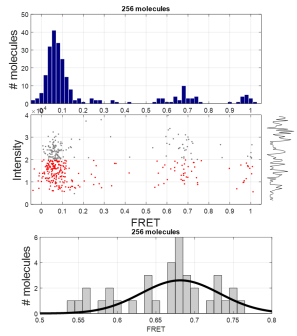
(d) CCCCTTTT



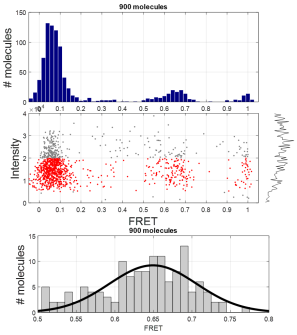
(e) TTTTTTTT



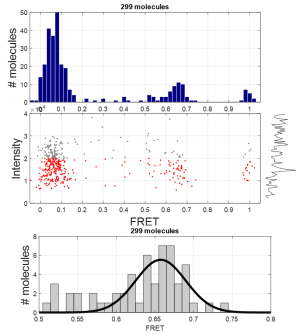
(f) CCCCCCC



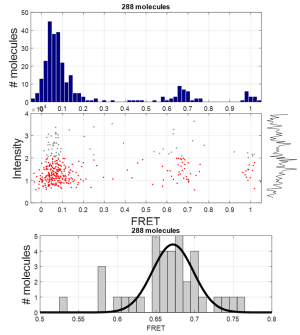
(g) ACACACAC



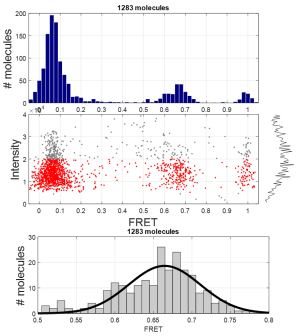
(h) AAAACCCC



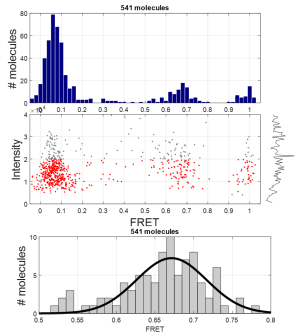
(i) GTGGGTGG



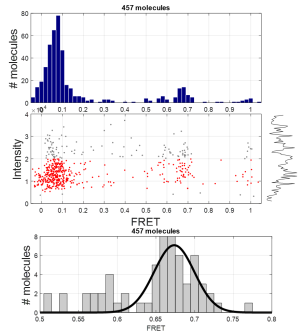
(j) GTGTGTGT



(k) TTGTTTGT



(l) ATATATAT



Bibliography

[1] Base stacking. 2021. URL https://www.open.edu/openlearn/science-maths-technology/science/biology/nucleic-acids-and-chromatin/content-section-2.2.2.

[2] Harmonic potentials, bond stretches. *Big Chemical Encyclopedia*, 2021.

[3] Energy terms in potential function. 2021.

[4] Geraldine Aubert and Peter M Lansdorp. Telomeres and aging. *Physiological reviews*, 88(2): 557–579, 2008.

[5] Lei Bao, Xi Zhang, Lei Jin, and Zhi-Jie Tan. Flexibility of nucleic acids: from dna to rna. *Chinese Physics B*, 25(1):018703, 2015.

[6] Alessandro Bosco, Joan Camunas-Soler, and Felix Ritort. Elastic properties and secondary structure formation of single-stranded dna at monovalent and divalent salt conditions. *Nucleic acids research*, 42(3):2064–2074, 2014.

[7] Véronique Calleja, Pierre Leboucher, and Banafshé Larijani. Protein activation dynamics in cells and tumor micro arrays assessed by time resolved förster resonance energy transfer. *Methods in enzymology*, 506:225–246, 2012.

[8] Junghuei Chen and Nadrian C Seeman. Synthesis from dna of a molecule with the connectivity of a cube. *Nature*, 350(6319):631–633, 1991.

[9] Bridget E Collins, F Ye Ling, Daniel Duzdevich, and Eric C Greene. Dna curtains: Novel tools for imaging protein–nucleic acid interactions at the single-molecule level. *Methods in cell biology*, 123:217–234, 2014.

[10] Franziska Doll, Jessica Hassenrück, Valentin Wittmann, and Andreas Zumbusch. Intracellular imaging of protein-specific glycosylation. *Methods in enzymology*, 598:283–319, 2018.

[11] Alissa Drees and Markus Fischer. High-throughput selection and characterisation of aptamers on optical next-generation sequencers. *International journal of molecular sciences*, 22(17): 9202, 2021.

[12] Fabian Drube, Karen Alim, Guillaume Witz, Giovanni Dietler, and Erwin Frey. Excluded volume effects on semiflexible ring polymers. *Nano letters*, 10(4):1445–1449, 2010.

[13] S Hartmann, D Weidlich, and D Klostermeier. Single-molecule confocal fret microscopy to dissect conformational changes in the catalytic cycle of dna topoisomerases. *Methods in enzymology*, 581:317–351, 2016.

[14] Theodore Johnson, Jian Zhu, and Roger M Wartell. Differences between dna base pair stacking energies are conserved over a wide range of ionic conditions. *Biochemistry*, 37(35):12343–12350, 1998.

[15] Anthony D Keefe, Supriya Pai, and Andrew Ellington. Aptamers as therapeutics. *Nature reviews Drug discovery*, 9(7):537–550, 2010.

[16] Av Leeuwenhoek. Observationes de anthonu lewenhoeck, de natis e semine genitali animalculis. *Phil Trans Roy Soc*, 12:1040–1043, 1753.

[17] VG Malathi and P Renuka Devi. ssdna viruses: key players in global virome. *Virusdisease*, 30 (1):3–12, 2019.

[18] Noam Mamet, Itai Rusinek, Gil Harari, Zvi Shapira, Yaniv Amir, Erez Lavi, Adva Zamir, Noam Borovsky, Noah Joseph, Maria Motin, et al. Ab-initio discovery of tumoricidal oligonucleotides in a dna sequencing machine. *bioRxiv*, page 630830, 2019.

[19] Dustin B McIntosh, Gina Duggan, Quentin Gouil, and Omar A Saleh. Sequence-dependent elasticity and electrostatics of single-stranded dna: signatures of base-stacking. *Biophysical journal*, 106(3):659–666, 2014.

[20] Takashi Norisuye and Hiroshi Fujita. Excluded-volume effects in dilute polymer solutions. xiii. effects of chain stiffness. *Polymer Journal*, 14(2):143–147, 1982.

[21] Ruman Rahman, Nicholas R Forsyth, and Wei Cui. Telomeric 3-overhang length is associated with the size of telomeres. *Experimental gerontology*, 43(4):258–265, 2008.

[22] Michael Rubinstein, Ralph H Colby, et al. *Polymer physics*, volume 23. Oxford university press New York, 2003.

[23] Nadrian C Seeman. Nucleic acid junctions and lattices. *Journal of theoretical biology*, 99(2): 237–247, 1982.

[24] Nadrian C Seeman. Dna in a material world. *Nature*, 421(6921):427–431, 2003.

[25] Kyung-Mi Song, Seonghwan Lee, and Changill Ban. Aptamers and their biological applications. *Sensors*, 12(1):612–631, 2012.

[26] Jeffrey A Speir and John E Johnson. Nucleic acid packaging in viruses. *Current opinion in structural biology*, 22(1):65–71, 2012.

[27] Siyang Sun, Venigalla B Rao, and Michael G Rossmann. Genome packaging in viruses. *Current opinion in structural biology*, 20(1):114–120, 2010.

[28] Xiao-Wei Wang and Stéphane Blanc. Insect transmission of plant single-stranded dna viruses. *Annual Review of Entomology*, 66:389–405, 2021.

[29] Diana Wu, Trevor Feagin, Peter Mage, Alexandra Rangel, Leighton Wan, Anping Li, John Coller, Michael Eisenstein, Sharon Pitteri, and H Tom Soh. Automated platform for high-throughput screening of base-modified aptamers for affinity and specificity. *bioRxiv*, 2020.