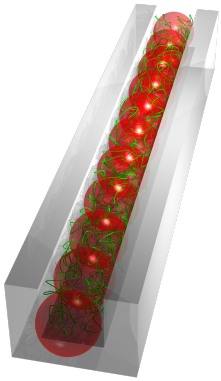
Masters's Thesis 2011: xxxx-xxxx okänt löpnummer

Building Software for an Ordered Study of Single Long Polymer Chains Confined in Nano-channels

Eric Sandlund



Examinator:
Prof. Bernhard Mehlig

Supervisors:
Dr. Jonas Tegenfeldt
Dr. Fredrik Persson

Department of Applied Physics
Complex Adaptive Systems
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2011

Building Software for an Ordered Study of Single Long Polymer Chains Confined in Nano-channels
Eric Sandlund

Masters's Thesis 2011: xxxx-xxxx okänt löpnummer
Department of Applied Physics
Complex Adaptive Systems
Chalmers University of Technology
SE-412 96 Göteborg
Sweden

Göteborg, Sweden 2011

Building Software for an Ordered Study of Single Long Polymer Chains Confined in Nano-channels

Eric Sandlund
Department of Applied Physics
Complex Adaptive Systems
Chalmers University of Technology

Abstract

Since the discovery of bio-molecules and their significance they have been the object of much study. In order to better understand the structure and dynamics of bio-molecules devices and methods to capture and manipulate features of the molecule has been developed, such as confining long polymers in nano-channels and epifluorescence microscopy.

In this report focus has been given to the study of the data properties of data collected from viewing confined long polymers homogeneously stained along the length of the molecule in order to reduce noise and trace significant features along the molecule in the data. An overview of the polymer physics involved will also be presented along with an overview of the techniques used to capture the data.

From the the study of the experimentally obtained data software tools for viewing large single molecules confined in nano-channels is developed. The software is used to view the data in an ordered way by reducing noise, using redundancy, and presenting subsets of the data in an interface. The software is also able to detect and trace simple features along the molecule.

Keywords:

Sammanfattning

Ända sedan upptäckten av bio-molekyler och deras betydelse har de varit ämne för mycket studier. För att bättre förstå uppbyggnaden och dynamiken i bio-molekylerna har olika apparater och metoder utvecklats för att kunna fånga in och manipulera olika aspekter av molekylerna. Exempel på metoder som utvecklats i detta syfte är nano-channel devices och epifluorens mikroskopi.

I denna rapport kommer fokus ligga på studier av dataegenskaperna hos den data som insamlats vid studier av fångade homogent infärgade långa polymer molekyler, i syfte att studera brus reduktion i data. Översikter över relevant polymerfysik och teknisk utrustning för att fånga molekylerna kommer också behandlas.

Från studierna av experimentell data och brusreduktion kommer mjukvara utvecklas för att kunna presentera data på ett strukturerat och ordnat sätt. Mjukvaran kommer kunna reducera brus och presentera olika delar av datan efter behov. Enkel feature tracing kommer också att finnas.

Contents

List of Figures

List of Abbreviations

AAM	Active Appearance Model, a statistical model matching method for matching models to new images.
ACM	Active Contour Model(s), a Dynamic programming framework.
ASM	Active Shape Model, a discrete representation of the Active Contour Model framework.
DP	Dynamic Programming, a computational method for solving complex systems.
DTW	Dynamic Time Warping, an algorithm for measuring similarities between sequences.
ETT	Main window for TT Extraction in the SMATool software.
FIR	Finite Impulse Response filter.
IRLS	Iteratively Re-weighted Least Squares estimator of a linear regression.
MAT	MATLAB specific file format.
NPV	Nearest Pixel Value.
OLS	Ordinary Least Squares estimator of a linear regression.
ROI	Rough region(s) of interest in the raw data.
SA	Sequence Alignment, a bioinformatics term used for techniques that align and arrange sequences, usually RNA, DNA or protein sequences.
SAW	Self-avoiding random walk.
SMAT	The software SMATool 1.0, developed by E. Sandlund
SNR	Signal to noise ratio.
TFT	Main window of Trace Molecule Toll in the SMAT software.
TIF	Tagged Image File version 6.0, image file format.
TME	Main window for molecule edge tracing in SMAT.
TT	Time trace, the graphical representation of the 1D spatial extension of the channel model over time.
TXT	Text file format.
WFA	Window Function Average.
WLC	Worm-like chain model, a polymer model describing a semi-flexible chain.

Chapter 1

Introduction

Bio-molecules and long polymer chains have long been a tempting area of research and have captured the imaginations of many researchers specifically because this is the matter that which builds up life.

Since DNA was first isolated and discovered by Swiss physician F. Miescher[10] in the winter of 1868/9 this fascinating molecule has become the center of much research. In 1928 F. Griffith called his observation of DNA the *transforming principle* in an experiment where he infected mice with virulent and non-virulent bacteria[14][21]. The transforming principle was later confirmed by O. Avery with his coworkers in 1943[3] making the DNA molecule central as the information carrier in living cells.

Trying to understand and predict the behavior of bio molecules has since their discovery been the goal of many researchers as the understanding and especially the ability to predict the behavior of bio molecules has the potential to have a great impact on our ability to cure and understand diseases and on our understanding of evolution.

The work presented in this thesis will introducing tools for the ordered study of DNA and other large single molecules. The tools introduced will be collected into a software interface for easy access by the user. An overview of the polymer physics known in the area, the experimental setup used to collect empirical data as well as the methods used to analyze the experimental data will also be considered.

1.1 Empirical studies of single long polymers

The problem of understanding the fluctuations and mechanical properties of single DNA molecules confined in nano-slits or channels and the variation in these properties when polymerization or degradation, e.g. in the form of melting, becomes relevant when studying experimentally the polymer in real time and accuracy in measurements and information extraction is attempted.

In the beginning of the history of polymer research the empirical study of the subject has in the roughest sense been an indirect art. Conclusion about the dynamics and function of bio molecules were drawn by experiments such as F. Griffith's famous experiment in 1928, where he discovered the ability of bacteria to transfer genetic information[21].

Resent research into the area of nano scale physics and the ability to create nano sized structures has enabled a flood of new research that before was unfeasible. Among new research areas enabled by nano structure science is the realtime optical study of dynamics of nano sized objects contained by nano scale structures, specifically macromolecules such as bio molecules.

In the area of polymer physics single molecule polymer chain dynamics, has become an interesting field of research as it has the potential to answer a great many questions until now only hinted at by early theoretical models.

An example of research combining nanoscale science with polymer physics is the work done by J. Tegenfeldt in B. Austin's group [34] showing in 2004 that DNA confined in nano-slits scales

linearly with the contour length of the DNA.

Among the problems to be solved when attempting to study single polymers is the experimental setup, regulating the flexibility and quality of the study, and the indication or labeling of the polymers. When the experimental data is analyzed noise and distortion handling needs to be considered as well.

1.1.1 Experimental setup

The purpose of the experimental design when studying long polymers is to create a way to see and study the polymer. In this thesis the design will be confined to studies of highly confined polymers labeled in some way. The confinement is usually a slit or a channel with length scales such that the polymer must conform to the geometrics of the channel or slit and appears stretched to the shape of the channel when studied. For DNA these length scales typically goes from below 50 nm for extremely confined to well above a few hundred nm for confined DNA with a relatively large amount of freedom.

For the purpose of this thesis the general design of the experimental setup is assumed to be, as indicated by fig. 1.1, made up of the *nano channel device*, the *Microscope* and the software. To these major parts comes the methods of labeling the object in the device and reading and storing the experimental data created by the capture device.

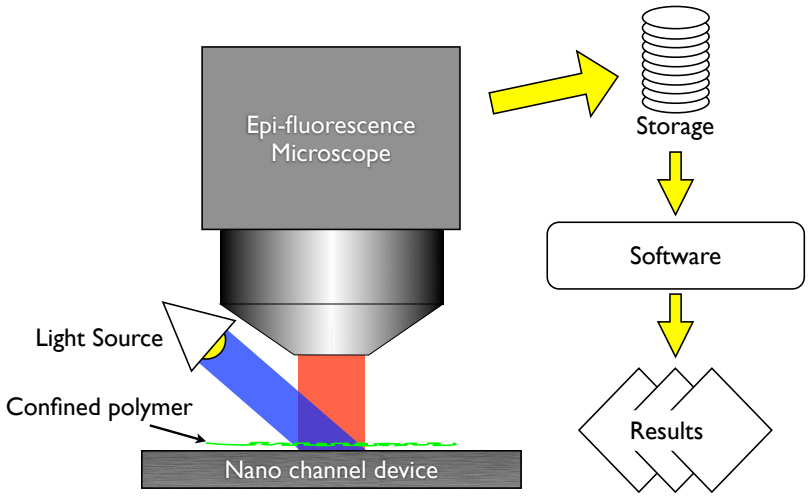


Figure 1.1: The figure illustrates the major parts of the basic experimental setup assumed in this thesis. The major components is the nano channel device, the labeling induced or emitted from the subject in the device with the light source, the microscope, the A/D converter necessary to create a digital image of the data, the storage device for the data, the software applied to the data to analyze and present the results.

The nano channel device is usually made up of a chip with a nano scaled structure designed to confine an object, e.g. DNA, reservoirs for the object and methods for feeding the object to the nano scaled structure, which includes methods for regulating the flow of the object through the device. There are two common ways of regulating the flow of the object. The first, electrophoresis, is a well studied method that can accurately manipulate the flow of charged particles and single molecules with the use of electric fields. The second is fluid flow which uses pressure to regulate fluidic flow through the nano structure. A good general device can perform both of these methods for regulating the flow of the object through the nano structure. In fig. 1.2 the schematic of a general purpose nano channel device is illustrated. The device is designed for both electrophoresis and fluid flow regulation of the object confinement.

The microscope is the device that translates the physical signal emitted from the confined

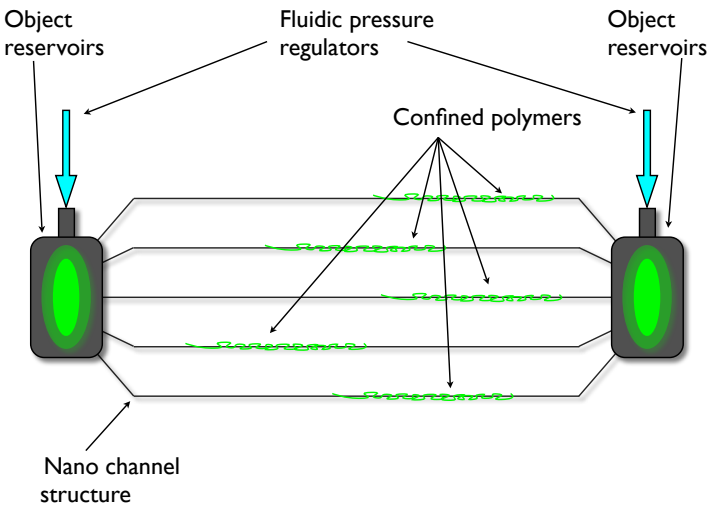


Figure 1.2: The figure illustrates the schematic of the setup of the nano channel device. In the device the object is confined spatially inside the nano channel structure and fed from the reservoirs. The fluidic pressure and electromagnetic field regulators are the tools for regulating the confinement of the subject inside the nano channel structure.

object in the nano channel device to an electrical signal and compiles the electrical signal into a digital image. There are several different techniques for capturing information about a confined object and all these needs to be considered when constructing the experimental setup. Different methods include force reading, electromagnetic field reading and photon reading. In this thesis light capturing devises will be considered, i.e. photon reading. Examples include Epifluorescence Microscopy and EMCCD. When using capture devices for reading light emission from the object the object needs to be dyed with fluorescent markers. This needs to be considered when choosing the type and amount of dye. The epifluorescence microscope is convenient to use as the device confining the subject can be designed separately and placed beneath.

The software, which is the last major part of the experimental setup and the major topic of this thesis, is designed to take the raw data from the capture device and analyze it in a structured way and present aspects of the experimental data. The software should be flexible enough to be able to present different aspects of the data and filter the result as the user wishes.

1.1.2 Polymer indicators and labeling

When studying real polymers, e.g. DNA or proteins, confined in nano-channels and attempting to extract comparable features and study spatial information along the molecule the polymer needs to be indicated or labeled in some way. The labeling method used determines the way the polymer or object can be studied.

A convenient way of labeling a object in order to read spatial information is to use fluorescent markers. One of the most common probes used is fluorescent dye molecules. The dye molecules are bounded to the object and excited by light of some specific wavelength and subsequently emits light of another specific wavelength. Using dye molecules is convenient because they can be used to either stain the object evenly or to stain only regions along the object with specific attributes.

Staining an object evenly can yield information about density fluctuations and the dynamics of the fluctuations within the object. Manipulating the environment around the subject and causing the dye molecules to detach from the subject can also yield information of the spatial composition of the subject. E.g. for DNA dyed with an intercalating dye molecule, increasing the temperature over time and partially melting the DNA creates a melting map of the AT and GC concentrations

along the molecule[27].

In general any way to label the polymer can be used as long as the labeling method can yield spatial information about the polymer. A rough method of reading spatial information is end-end labeling, where only the ends of the polymer is labeled. With end-end labeling the global length fluctuations of the polymer can be studied with accurate length calculations within the confinement.

1.1.3 Noise and distortions of the data

When studying very small objects, specifically single molecules, and analyzing experimental data noise and distortion handling needs to be considered. With confined polymers and reading labels evenly attached along the molecule there are two major noise and distortion factors to consider; the dynamics of the molecule and the dynamics of the experimental setup.

The dynamics of the molecule consists of the thermal fluctuations and the polymer physics dynamics of the confined and labeled polymer. The pure polymer physics dynamics of the object is a vast area of research and there exists several theoretical models that explains certain characteristics of the molecule in different types of confinement. The added dynamics of the labeling of the molecule creates a much more complex analytical model where the research is sparse. Empirical studies has been conducted and theoretical models as been attempted to be modified[17][34].

The noise and distortions created by the experimental setup is usually very complex and depends entirely on the types of equipment used. Normally the experimental setup is designed to improve the signal to noise ratio as much as possible making the added noise and distortions as small as possible. Usually the experimental setup noise is modeled as a stochastic added signal and removed using ordinary signal noise filtering. Creating an improved model of the noise from the experimental setup involves measurements on known objects and then adapting the measurements to the model. Usually these kinds of noise and distortions is of much higher frequency than the distortions and fluctuations created by the dynamics of the object.

In general when characterizing and identifying the observed fluctuations of the object the fluctuations can be either studied globally as length fluctuations or locally as correlated local stretching and squeezing of the molecule. Studying global fluctuations creates a much easier model where only net fluctuations along the whole object needs to be considered. Studying local fluctuations requires much more complex models and significantly larger computational power.

1.2 Approaches to decoding DNA mappings

When the raw data is taken from the microscope it is sent to a computer to be processed in order to allow the presentation of aspects of the data in an ordered way. Decoding experimental data is the inverse problem of creating theoretical models to explain the behavior of the experiments. Decoding the information from visualizations of DNA maps involves the realization and identification of the processes that interact on the data creating distortion on the map away from the ideal map of the DNA.

Much of the methods and research in the area of signal analysis and image processing concerns the problem of either removing noise or identifying known elements in a signal or image. Other applications in this area concerns the extraction of specific features or attributes from the signal or image. In this report the image processing and signal analysis involves the problem of identification of specific attributes and features and subsequent operations on the image based on these features.

Existing image processing techniques are able to find geometric forms, text and faces in images[33][35]. Other methods can recognize speech and reduce echoes and noise in signals[7].

1.2.1 Dynamic Programming

Dynamic Programming (DP) is the description of the process of solving complex problems by dividing larger decisions into smaller sub-decisions. The term first created by the american math-

ematician R. E. Bellman at the end of the 1940s and the beginning of the 1950s to describe the process of decision making in planning[13]. Later the IEEE recognized DP as a systems analysis and engineering topic.

DP is used both as a computer programming method and a mathematical optimization method. In order to make DP applicable and effective on a problem the problem exhibit both optimal substructure and overlapping subproblems that is only slightly smaller. Optimal substructure means that the optimal solution to the problem can be constructed efficiently from the optimal solutions to its subproblems. Overlapping subproblems means that the subproblems are reused over and over and that the number of subproblems are relatively small.

Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm designed to measure similarities between two sequences that can vary in time and speed. The first example of the method is in voice recognition[18] where the voice of the speaker can vary in both time and pitch, i.e. speed. Since then the method has been used in hand signature recognition [19], classification of killer whale vocalization[5] among other.

The DTW method is a method for matching two sequences together, with certain restrictions, by warping the sequences non-linearly in the time dimension[20]. See fig. 1.3

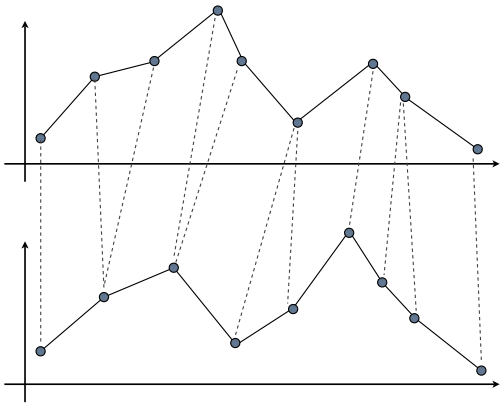


Figure 1.3: The figure shows an example of Dynamic Time Warping.

1.2.2 Sequence Alignment

Sequence Alignment (SA) is a term used in bioinformatics to encompass techniques used to arrange and align sequences of biomolecules, such as RNA, DNA or proteins, and to identify regions of similarity that could be implicit of functionality within the sequences.

Computational approaches to SA usually fall into two distinct categories; global and local sequence alignment. Global SA is a global optimization problem that requires the sequences to each span the whole preferred length. Local SA is a local optimization problem that tries to find similarities locally where the global sequences can diverge significantly.

Methods used for SA include DP and different stochastic optimization methods.

Needleman-Wunsch

The Needleman-Wynsch algorithm is a global sequence alignment algorithm specially developed for alignment of sequences in biomolecules, e.g. proteins or nucleotides. The algorithm was first presented by Saul B. Needleman and Christian D. Wunsch in 1970[22]. The algorithm was then elaborated to include insertion/delition constraints[30]. The new elaborated algorithm also

increased the computational speed from cubic to quadratic. The Needleman-Wynsch algorithm is the first example of a DP algorithm used on biological sequence matching.

Smith-Waterman

The Smith-Waterman algorithm is a local sequence alignment algorithm first presented by T. F. Smith and M. S. Waterman in 1981[32]. The Smith-Waterman algorithm is a development of the Needleman-Wynsch algorithm, and as such a DP method, but uses constraints that makes local sequence alignments visible. The algorithm matches subsequences of different lengths and optimizes a similarity measure between these local sequence alignments. Today the Smith-Waterman algorithm is seldom used because new developments have made algorithms with better scaling[16] and better accuracy[2].

1.2.3 Image processing

Image processing concerns the manipulation of images such that the result is a new image where a set of characteristics or attributes is promoted or extracted. As such the area of image processing is vast and can encompass almost any type of operation on images. In this report image processing concerns the identification of attributes in images that directly represents the underlying dynamics and attributes of the system that is imaged. Inferring the dynamics of the molecules studied.

Here examples of image processing using a statistical and a "flexible" model will be presented as two different approaches to decoding information and extracting specific features from data that can be represented as images.

Active Appearance Model

The Active Appearance Model (AAM) is a statistical model matching method introduced by T.F. Cotes, et al.[8]. The method uses a statistical model and a gray level appearance of the object of interest. During a learning phase the model is taught the relationship between the parameter displacement and the error introduced between the training image set and the created model image.

The method performs well when searching for deformable objects.

Active Contour Model

Active Contour Models (ACM), or *Snakes*, is a general theory or framework for finding contours in noisy 2D images. The general idea is to create a representation of a rubber band which is iteratively deformed along some contour in a data set represented as an image. The ACM framework attempts this by minimizing a energy function associated to the current contour as a sum of an *external* and an *internal* energy. The external energy is associated with contour in the data set, and minimizes as the ACM approaches. The internal energy is associated with the relevant shape sought, and should give low energy to shapes approaching the sought.

In the ACM the snake is defied as a set of points $\mathbf{v}_i = (x_i, y_i)$ where $i = 1...n-1$, an internal energy $E_{internal}$ and an external energy $E_{external}$. The external energy can further be represented as the sum of the constraint introduced by the user E_{user} and the constraint introduced from the data set E_{data}. The energy function associated with the model can thus be defined as

$$E^*_{snake} = \int_0^1 E_{snake}(\mathbf{v}(s))ds = \int_0^1 E_{internal}(\mathbf{v}(s)) + E_{external}(\mathbf{v}(s))ds \tag{1.1}$$

The Active Shape Model (ASM) is a discrete representation of the ACM approach in 2D and was developed by Tim Cootes and Chris Taylor in 1995[9]. The ASM is a statistical model that iteratively deforms to the manifestation of the model in a data set. The method has been widely used in image recognition an medical modeling.

1.3 Thesis outline

This thesis is organized as follows:

Chapter 2: *Polymer physics of confined DNA*

The basics of polymer physics will be presented in this chapter with relevant theories and models. The chapter will focus on simple models to explain the basic dynamics of polymers in free solution and then give an overview of the two largest theories concerning confined polymers in thin channels.

Chapter 3: *Experimental basics*

Details of the general basic experimental setup will be presented in this chapter.

Chapter 4: *Analysis of single molecules confined in nano channels*

Methods developed for the analysis of experimental data extracted from single molecules confined in nano channels will be presented in this chapter. Methods presented in this chapter will be aimed at enhancing the data extracted without knowing the specifics of the type of sample in the experiment.

Chapter 5: *The Software*

A detailed description of the structure and usage of the developed software created as a tool for the analysis of confined DNA.

Chapter 2

Polymer physics of confined DNA

In this chapter the basics of polymer physics relevant to this thesis will be presented. The chapter will give an overview of the polymer physics that has been developed to explain the behavior of biological macro molecules, which is the basics for the mechanics and dynamics of the DNA considered in this thesis. Focus will lie on the polymer physics dealing with simple polymer model to explain the basic dynamics and confined polymers in long thin channels.

The chapter will follow the structure

1. Simple polymer models
2. More complex polymer models
3. Polymers confined in relatively large pores
4. Highly confined polymers

The chapter will start out by giving an outline of the most basic polymer models used to explain the dynamics of polymer chains. The chapter continues to present two more complex models, where constraints is added to the simple polymer model and the chain is restricted in movement. Mapping between the models will also be considered.

When the polymer models has been explained the chapter will move on to confined polymers. Two approaches in *size* regimes will be considered when explaining the dynamics of confined polymers. The two approaches represent the major theories in the area. The first approach will be the de Gennes approach, presented by P. G. de Gennes in 1977[25]. In this regimes the channel or slit width is large and the polymer is highly entangled. The second approach will be the approach presented by T. Odijk in 1983[23] where the DNA is highly contained and the polymer is unentangled.

2.1 Polymer Models

In this section the linear polymer chain, which is the simplest theoretical polymer structure, will be considered. Two models of the linear chain will be described. The *Ideal Chain* and the *Worm-like Chain*. The first model, the ideal chain, is a discreet model and the second, the worm-like chain, a continuos model of the polymer chain. The section will consider both freely jointed and restricted chains and the mapping between them. Focus will be on the relation between contour length and magnitude of the polymer in solution.

2.1.1 Ideal chain

The ideal chain, which is a realization of a random walk in space, is the simplest model of a polymer where interactions between monomers are considered only if the monomers are adjacent.

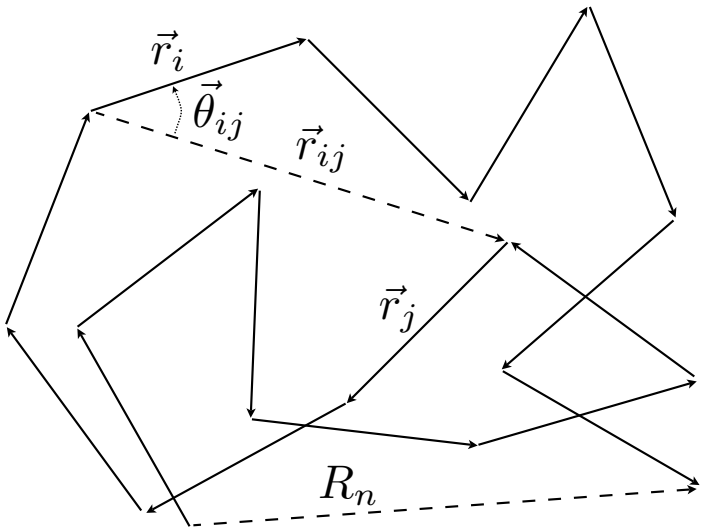


Figure 2.1: The ideal chain is an example of a random walk. The figure illustrates a possible conformation of an ideal chain with notations for the end-to-end vector as well as the the angle between two arbitrary monomer vectors.

The most simple representation of the model is the freely jointed chain. In the freely jointed chain the diameter is set to zero and the chain made up of n rigid links, describing the monomers. The bond vector r_i, joining the monomers, is of constant length l, i.e. $|\vec{r}_i| = l \quad \forall \quad i$ along the chain, and the vectors are freely jointed making every bond direction possible. See fig. 2.1. The end-to-end vector of the chain is given by the sum of all the bond vectors,

$$\vec{R}_n = \sum_{i=1}^{n} \vec{r}_i, \tag{2.1}$$

making the average end-to-end vector $\langle \vec{R} \rangle = 0$, in the freely jointed chain.

Instead of looking at the average end-to-end distance one can look at the mean square magnitude, $\langle R^2 \rangle$, which is more relevant. The mean square magnitude is calculated as

$$\langle R^2 \rangle \equiv \langle \vec{R}_n^2 \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \vec{r}_i \cdot \vec{r}_j \rangle = l^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \cos \theta_{ij} \rangle = nl^2 \tag{2.2}$$

which becomes obvious when noting that the bond length is assumed constant of length l and that eq. 2.1 gives $\langle \cos \theta_{ij} \rangle = 0 \quad \forall \quad i \neq j$, which is consistent with that any bond direction is equally possible..

2.1.2 Restricted chain

In real chains the bond vectors is restricted by bond angles and stearic hindrances to rotation and can never be totally uncorrelated. According to W. Kuhn, a Swiss physical chemist, this can be taken into account in the model by replacing the chain with longer segments, i.e. *Kuhn segments*, each consisting of several bond vectors. Introducing the Kuhn segments accommodates for the correlation introduced in real chains. Flory reasons that equivalently that n and l may be preserved by replacing the resulting expression in eg. 2.2 by

$$\langle R^2 \rangle = C_\infty n l^2 \tag{2.3}$$

where C_∞, called Flory's characteristic ratio, is determined by the monomeric units of the polymer[15].

The reason for using the ideal chain is that it can provide a scaffolding on which to map chains of higher correlation. The mapping is done by re-normalizing the chain into long enough segments such that each segment becomes uncorrelated. Kuhn segments is such a mapping.

Using Kuhn segments and replacing a chain consisting of n bonds of length l with N segments of length b, where b is called the Kuhn length, the maximum end-to-end distance R_{max} and $\langle R^2 \rangle$ is rewritten to

$$\begin{aligned} Nb &= R_{max} \\ Nb^2 &= \langle R^2 \rangle = C_\infty n l^2, \end{aligned} \tag{2.4}$$

which is solved by

$$\begin{cases} N = \dfrac{R_{max}^2}{C_\infty n l^2} \\ b = \dfrac{C_\infty n l^2}{R_{max}} \end{cases}. \tag{2.5}$$

Mapping real chains on the freely jointed chain R_{max} becomes the contour length L_C of the real chain, and the expression

$$\langle R^2 \rangle = Nb^2 \tag{2.6}$$

becomes an arbitrary description of a polymer chain. Compare to eq. 2.2 and 2.3.

2.1.3 Worm-Like Chain

The Worm-Like Chain model (WLC) , also sometimes called a Kratky-Porod Worm-Like Chain, is a convenient model created to describe semi-flexible chains. Semi-flexible chains are chain that are rigid over much larger length scales than the size of a monomer, as is the case for most bio-polymers such as double stranded DNA and unstructured RNA, etc..

In the WLC model the polymer chain is modeled as continuously flexible, compared to the ideal chain where each monomer is rigid where the chain can be called discretely flexible. Defining tangent vectors $\vec{t}_i$ and $\vec{t}_j$ of unity length separated by an angle θ_{ij} to describe the correlation along the chain between i and j, it can be shown that the expected value of the directional function along the polymer decays exponentially, i.e.

$$\langle \vec{t}_i \times \vec{t}_j \rangle = \langle \cos \theta_{ij} \rangle = e^{-\frac{|j-i|}{P}}, \tag{2.7}$$

where P denotes the formal definition of persistence length. The Kuhn length, of the previous section, denotes double the persistence length, $2P = b$. The persistence length defines the characteristic length scales over which the directional correlation of the tangent vector is lost.

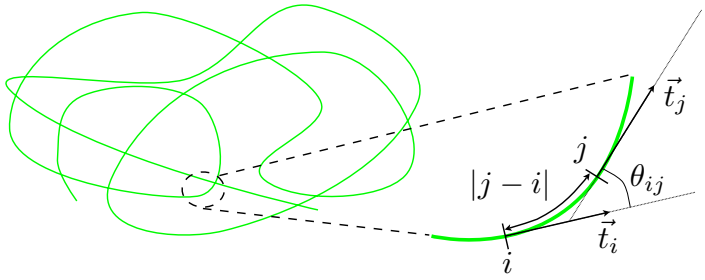


Figure 2.2: The figure illustrates a possible conformation of a continuously semi-flexible coiled chain, with notations for tangent unity vectors $\vec{t}_i$ and $\vec{t}_j$ along the chain at positions i and j. The notations are used to describe correlation along the chain.

In fig. 2.2 a possible conformation of the WLC is illustrated. Deriving an expression for $\langle R^2 \rangle$ along the contour length L_C for the continuos chain, using analogous reasoning as for eq. 2.2,

yields

$$\begin{aligned}\langle R^2\rangle &= \langle \int_0^{L_C} \vec{t}_i di \times \int_0^{L_C} \vec{t}_j dj\rangle \\ &= \int_0^{L_C} di \int_0^{L_C} \langle \vec{t}_i \times \vec{t}_j\rangle dj \\ &= \int_0^{L_C} di \int_0^{L_C} e^{-\frac{|j-i|}{P}} dj \\ &= 2PL_C\left[1-\frac{P}{L_C}\left(1-e^{-\frac{L_C}{P}}\right)\right].\end{aligned} \tag{2.8}$$

In the limit $L_C >> P$ the result in eq. 2.8 reduces to $\langle R^2\rangle \approx 2PL_C$, which is consistent with the result in eq. 2.6 of an arbitrary freely jointed polymer chain with a Kuhn length that is two times the persistence length.

2.2 Confined polymers

In this section the effects confining polymers by enclosure will be considered. Confinements will in this section always be small enough that the polymer is prevented from assuming its free solution coiled up state. Two approaches will be presented. In the less confined region where the persistence length is much smaller than the confinement the de Gennes blob theory will be considered, and in the more confined region where the persistence length is larger then the confinement Odijk's theory will be considered. According to Reisner et al. there are, when studying long polymer chains confined in thin channels, two relevant regimes. The first regime is when the confinement in a channel is much larger then the persistence length, and the second regime is where the confinement is much smaller than the persistence length[28].

2.2.1 de Gennes regime

One of the first great contributors to the area of empirical research of single chain polymer physics confined in porous media is P. G. de Gennes. de Gennes made great contributions in the regime $P \ll D \ll R_F$, where D denotes the channel diameter and R_F denotes the Flory radius which relates to the mean-square magnitude in eq. 2.6 but with added contributions from excluded volume effects that increases the size of the polymer coil. The excluded volume effects is introduced when the polymer chain is modeled as a self-avoiding random walk (SAW). In channel confinement of rectangular cross-section described by D_1 and D_2 the channel diameter D is replaced by a geometric average[28] $D_{av} = \sqrt{D_1 D_2}$.

In the theory presented by de Gennes[12] the DNA is modeled as a string of "blob"s of diameter D_{av}. Enclosed in the blob the DNA is assumed to behave like a discrete unconfined Flory coil, i.e. the DNA behaves as a 3D SAW. Inside the channel, outside the blob, the blobs behave as a 1D SAW of solid repelling bubbles. Inside the blobs the DNA segments interact hydrodynamically but between the blobs hydrodynamic interactions are negligible, see fig. 2.3.

With the de Gennes theory the extension r_x of the DNA enclosed in the blobs along the channel is described by

$$r_x = D_{av}\frac{L_C}{L_b}, \tag{2.9}$$

where L_b is the length of the DNA enclosed by the blob. The length of DNA inside the blob is described within the expression for the Flory radius for self-avoiding persistent polymers with[29]

$$R_F \sim (wP)^{\frac{1}{5}} L_C^{\frac{1}{5}}, \tag{2.10}$$

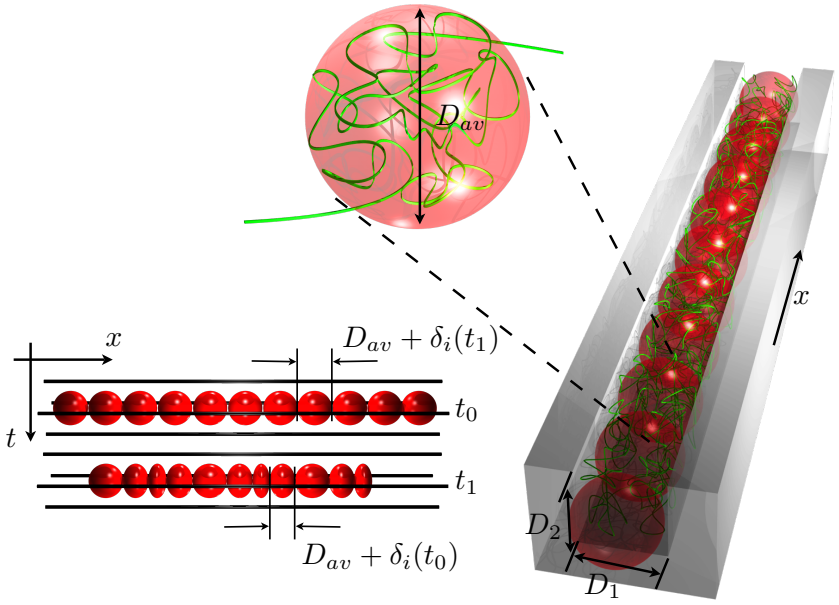


Figure 2.3: Illustration of the de Gennes blob model with notation for channel dimensions D_1 and D_2 and the geometric average diameter $D_{av} = \sqrt{D_1 D_2}$ of the blob. The illustration also shows the effect of thermal fluctuations on the DNA within the blob.

where w represents the effective width of the DNA, eq. 2.9 can then be reformulated to express the extension of the DNA along the channel as[4]

$$r_x \sim L_C \left(\frac{wP}{D_{av}^2} \right)^{\frac{1}{3}}. \tag{2.11}$$

The expression in eq. 2.11 clearly illustrates the linear relationship between the contour length L_C of the DNA and the extension r_x of the DNA along the channel in thermal equilibrium. When considering the effects of thermal fluctuations $\delta(t)$ on the DNA a free energy model is attached to each blob.[11] From the free energy, which is on the order of $k_B T$ where k_B is Boltzmann's constant, a spring constant can be calculated to determine the mean square average of the thermal fluctuations. The mean square average thermal fluctuation is described by [34]

$$\langle \delta r_x^2 \rangle \sim (PwD_{av})^{\frac{1}{3}} L_C, \tag{2.12}$$

which decreases with the channel diameter.

2.2.2 Odijk regime

In the regime where $D \ll P$ and excluded volume effects are insufficient to explain the behavior of the polymer T. Odijk have developed a theory for the dynamics of the polymer chain. Considering the the Odijk regime where the persistence length is much smaller than the width of the confining channel and the dynamics of semi-flexible chains, described in Section 2.1.3, it is intuitive that the chain no longer forms coils in the channel due to the much smaller space available, making excluded volume interactions negligible compare to the dynamics created through deflections against the channel wall.

Odijk showed that a new length scale λ appears when a polymer modeled as a WLC is confined in a channel with a diameter $D \ll P$[23][24]. The new length scale is given by

$$\lambda^3 \approx D^2 P \tag{2.13}$$

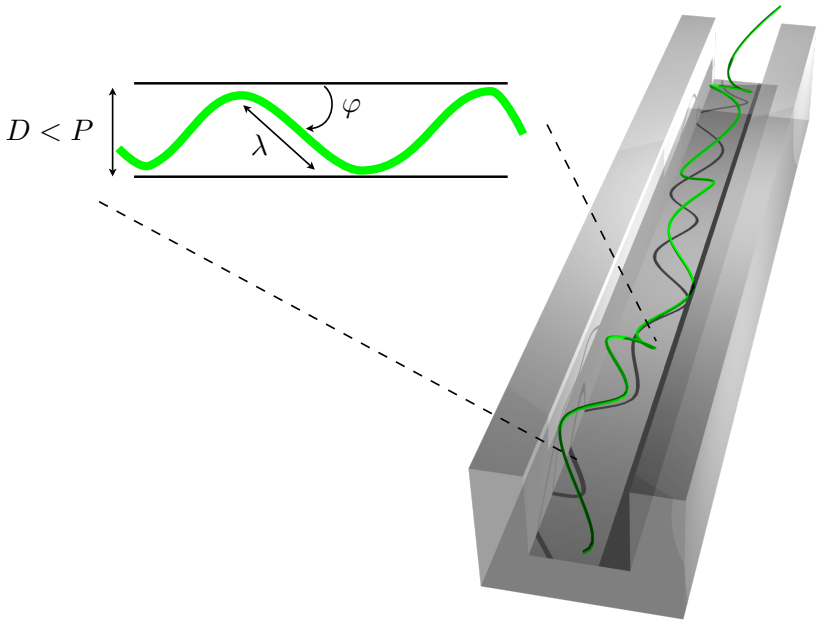


Figure 2.4: In the Odijk region the confinement diameter is smaller than the persistance length, $D < P$, and the polymer chain is no longer able to form coils. The chain stretches out by a series of deflection against the confinement wall, and elongation of the polymer is described by deflections lengths, λ and deflections angles, φ.

Within the length scale of λ the WLC confined in the channel is modeled as a completely stiff rod and the segment has a relative extension that is given by $\lambda \cos\varphi$, if φ is the deflection angle of the chain against the channel wall, illustrated in fig. 2.4. The mean square average of the deflection angle $\langle\varphi^2\rangle$ is given by the expression[25]

$$\langle\varphi^2\rangle \approx \beta \left(\frac{D}{P}\right)^{\frac{2}{3}}, \tag{2.14}$$

where β is a constant of proportionality.

Using a Taylor expansion approximation of the relative extension of the chain along the channel and the mean square average of the deflection angle the expression

$$\frac{r_x}{L_C} = \langle\cos\varphi\rangle \approx L_C \left[1 - \frac{\langle\varphi^2\rangle}{2!}\right] = L_C \left[1 - \beta \left(\frac{D}{P}\right)^{\frac{2}{3}}\right] \tag{2.15}$$

can be derived.

Considering the length scale of the Odijk regime and the dynamics of deflection the average defined in the previous section (section 2.2.1) cannot be used. Instead the relative extension of the chain in a channel of a rectangular cross section, height D_h and width D_w, is described by

$$\frac{r_x}{L_C} \approx \left[1 - \epsilon \left[\left(\frac{D_h}{P}\right)^{\frac{2}{3}} + \left(\frac{D_w}{P}\right)^{\frac{2}{3}}\right]\right], \tag{2.16}$$

which was derived by Burkhardt in 1997[6]. In eq. 2.16 the proportionality constant β has been replaced by a new constant ϵ.

Chapter 3

Experimental basics

In this chapter the basic experimental setup used for the project considered in this thesis will be outlined. The purpose of this thesis has not been to collect or produce experimental data, but to analyze and read experimental data in an ordered way, making the experimental setup secondary. The analysis has been designed to be as general as possible with regards to types of experiments generating the data analyzed. Therefor this chapter will only give a rough outline and suggestions of the experimental setup assumed for the experimental data. Different approaches to experimental setups and types of experiments will be presented with indications of which aspects of the experimental data will be relevant and essential for the analysis later in this thesis.

The experimental basics presented in this chapter will have the following general structure

1. General methods for reading and labeling polymers in an experimental setup
2. Notes optical reading of fluorescent markers
3. Specifics of the nano-channel devices
4. About the experimental data format and structure

3.1 Labeling

In this section labeling and reading DNA will be considered. When studying bio-molecules optically labeling of the molecules is needed in order to detect features along the molecule. For the experiments and experimental data considered in this report optical reading of labeled samples, with fluorescent probes are used.

A barcoding scheme first introduced by Reisner et. al. will also be considered to detect sequence specific features along a DNA molecule.

3.1.1 Fluorescent Labeling

Labeling using fluorescent probes is a convenient way to track the bio-molecule in a non-destructive way. Fluorescent labeling is a method for tracking specific features in bio-molecules by means of fluorescent emissions at specific frequencies. The emissions are generated from small fluorescent probes that are attached to the bio-molecule, which absorbs photons when exposed to light of specific wave lengths and briefly enters an excited state before either dispersing the excited energy or emitting photons of specific and lower frequency. the fluorescent labeling of DNA can be done using several different types of probes including quantum dots, fluorescent probes, fluorescent proteins and small fluorescent dye molecules. The most commonly used probe when labeling DNA is small dye molecules, which are chosen to have high binding affinity to the DNA and high fluorescent enhancement when binding to the DNA[26]. High binding affinity ensures a high

staining ratio and high fluorescent enhancement ensures high contrast of the bound dye molecules in the data.

Both the contour length and the electrostatic characteristics of the molecule is affected by the labeling and needs to be considered when studying the dynamics of confined DNA, discussed in Chapter 2. Another consideration when using fluorescent labeling is photobleaching, which is the gradual loss of fluorescence in the probes due to photo-induced damage. A third consideration introduced by F. Westerlund et. al.[36] is the Fluorescent enhancement induced by the channel width. The fluorescent enhancement needs to be considered when designing the chip confining the molecule.

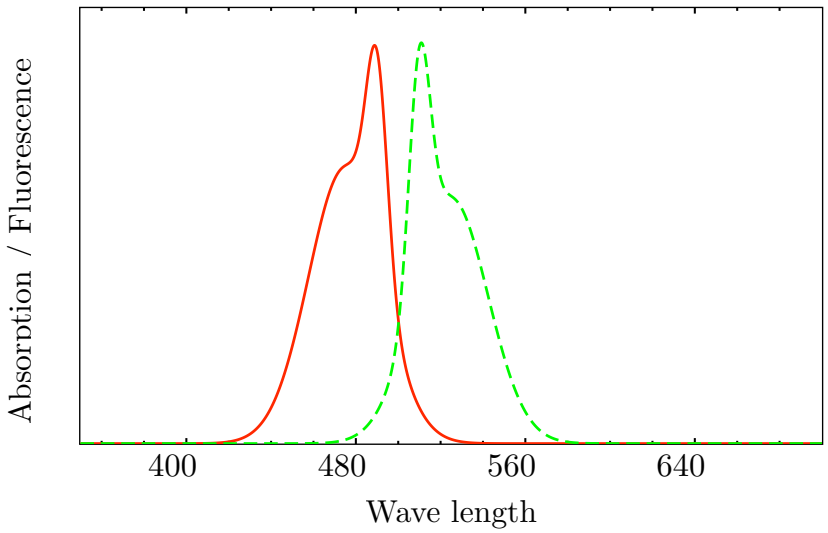


Figure 3.1: The graph shows a illustration of a generic absorption fluorescence spectra with different and slightly overlapping spectra.

The choice of which type of labeling to use when setting up an experiment to extract data from a confined polymer effects the data collected in many different ways, as indicated previously, and needs to be considered individually for each experiment depending on which features of the confined object is to be studied. In this report fluorescent dye molecules were used for labeling of the object molecule. No specific concentration of the dye molecule were considered as the added dynamics of the dye molecules were assumed to add only some constant to the models.

3.1.2 Barcoding scheme

When the object molecule is labeled it can be detected by fluorescent microscope. Usually the labeling of the molecule is even along the object molecule and not sequence specific, which limits the detection of the dynamics to concentration fluctuations due to entanglement.

In order to detect sequence specific features a scheme that considers the sequence is needed. W. Reisner et. al. introduced a scheme where DNA is dyed using an intercalating dye molecule, YOYO-1, and partial denaturation of the DNA molecule by using formamide and local heating is used[27]. The scheme detects sequence specific features along the DNA by utilizing the difference in bounding energies between the hydrogen bounds of the AT and the GC bases. The different binding energies allows the DNA to be denaturated, or melted, at lower temperatures in the AT rich regions, while the GC rich regions melt at higher temperatures. With the scheme the labeled molecule is detected by the microscope as a intensity map along the contour of the molecule. The intensity indicates attached dye molecules and low level of intensity can be due to thermal stretching, i.e. low entanglement, or low level of labeling, due to melting.

3.2 Optical reading

The method of reading labeled polymers confined in a nano channel needs to be considered when choosing type of labeling. In this report fluorescent labeling and fluorescence microscopy will be considered. When using fluorescent microscopy the object studied is illuminated with light of a specific wavelength which is absorbed by fluorophores attached to the object molecule. When the fluorophores are illuminated excitatory light of another and longer wave length is emitted by the fluorophores and captured by the microscopes CCD-chip.

In fig. 3.2 an illustration of an epifluorescence microscope setup is shown. Using a epifluorescence microscope instead of a fluorescence microscope is convenient as the excitatory light is passed from above instead of through the object. Using this setup allows the separation of the device holding the object from the microscope. The epifluorescence microscope improves the SNR as well as only emitted and reflected excitatory light is passed to the CCD chip.

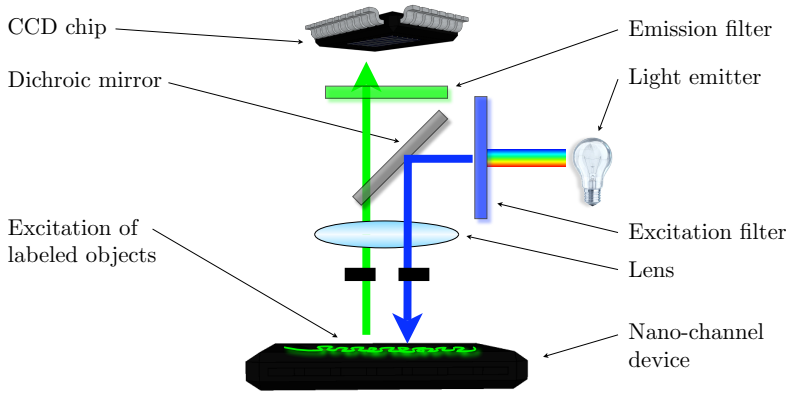


Figure 3.2: The figure illustrates the schematic of the epifluorescence microscope used as capture device for the labeled object. The different parts of the microscope are labeled. Using a epifluorescence microscope instead of a regular fluorescence microscope, where the excitatory light is passed through the object from below, improves the signal to noise ratio as only reflected excitatory light reaches the lens and is passed to the CCD chip. Adding the emission filter further improves the signal strength to the CCD chip.

3.3 Nano channel device

The nano-channel device is the device constructed to confine and manipulate the polymer in nano channel structure. The device is separated from the microscope when using the epifluorescence microscope and consists of the chip holder and the chip containing the nano-channel structure, see fig. 3.3. The chip holder is a device designed to allow manipulation of the confined objects. The chip holder in the figure is designed to allow both electrophoresis and fluidic flow control of the confined objects. The chip containing the nano-channel structure is held in place with screws and contains the nano-channel structure designed to confine the object molecules.

3.4 Experimental data

The experimental data collected from the CCD chip is stored as 2D intensity images where the intensity in the images indicate concentrations of active fluorophores. The images are stored in the file format TIF which allows the data to be stored as several images in one file.

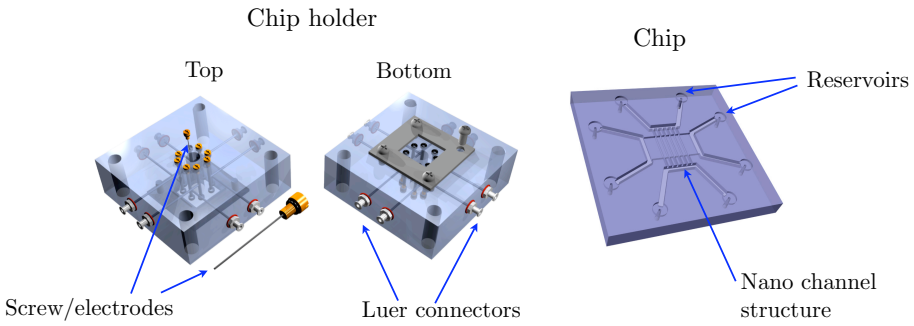


Figure 3.3: The figure is reproduced from F. Persson's PhD thesis[26] and is a schematic illustration of the nano channel device. To the left is the bottom and top view of the chip holder. Note the screw/electrodes and the luer connectors for controlling the electrophoresis and fluidic flow in the chip. The chip, which contains the nano channel structure confining the DNA, is shown to the right in the figure with annotations for the reservoirs and the nano channel structure. The design of the nano structure of the chip can be modified as needed, with different channel width and configurations.

Chapter 4

Analysis of single molecules confined in nano channels

In this chapter methods will be presented dealing with the analysis of data extracted from single molecules confined in nano channels. The methods presented in this chapter can be implemented on a broad range of data from observations of confined single molecules. In general the data is assumed to be a series of observations over time of one or more single molecules confined in nano channels and the analysis will deal with one molecule at a time.

The analysis methods presented in this chapter is aimed at enhancing the data without information of the nature of the experiment or the type of molecule in the observations. Only correlations in the existing data will be considered.

The chapter will follow the structure

1. Rough region if interest
2. Defining the channel(s) in the raw data.
3. Extracting a time trace from the raw data along the data region of the channel.
4. Finding edges of the molecule
5. Tracing local features along the molecule

The first step is defining one or more rough regions of interest (ROI). Defining ROI makes it possible to analyze and enhance observations of single molecules in raw data containing observations of samples with multiple single molecules. Choosing ROI instantly reduces the amount of data that needs to be considered for each analysis.

The second step, defining the channel in the raw data, is important as it is along the channel relevant information about the molecule in the sample observed can be found. This step introduces methods for noise handling in the raw data, algorithms for finding and selecting values in the raw data indicating the channel and fitting channel models to the selected values.

After the channel has been defined in the raw data, methods will be introduced that extract data along the defined channel containing observations of the molecule in the sample. The data collected is then compiled to show the dynamics of the observed molecule in a kymograph.

The edges of the molecule is then traced using the compiled data of the previous step. Steps 2-4 sort from the raw data and the ROI the values containing information of observations of the molecule and extracts these preparing for the final step.

When the location of the molecule in the raw data is determined for all observations the last step traces features along the molecule and through time. When features are traced along the molecule and through time both information about the dynamics and of the structure of the molecule can be illustrated, making the analysis methods presented here a powerful tool when studying single molecules.

4.1 Rough Region of Interest

The data considered for analysis is assumed to be a collection of observations, denoted frames, of a single sample taken consecutively over time. Each frame will be treated as a lattice with two spatial dimensions, length and width, where the dynamics of the observed sample is captured. The observations of molecules within the sample, denoted intensity maps, fluctuate and move on the lattice. Each collection of frames should represent a series of "snap-shots" of the sample in sequence with a constant or clearly defined relation in time, i.e. each frame in a sequence should have a high correlation in time. Normally the time correlation delay between frames is assumed to be constant and will not be considered. In fig. 4.1 the data structure is illustrated with the relevant features such as spatial dimensions and constant time dependance of the frames. Intensity maps is also illustrated in the figure.

The raw data can contain multiple molecules in addition to high intensity noise such as stuck molecules in the nano channel and clusters of coiled molecules outside the nano channel. To reduce the amount of data needed to be considered and remove general unwanted data one or more ROI is defined for a set of frames, by a hand, and numbered. The ROI is identified in the raw data by $[\Delta x, \Delta y]$, see fig. 4.2. When features along the molecule is considered later in this chapter restrictions in time can also be applied but initially the ROI is only defined spatially and data is extracted from the defined ROI in each frame.

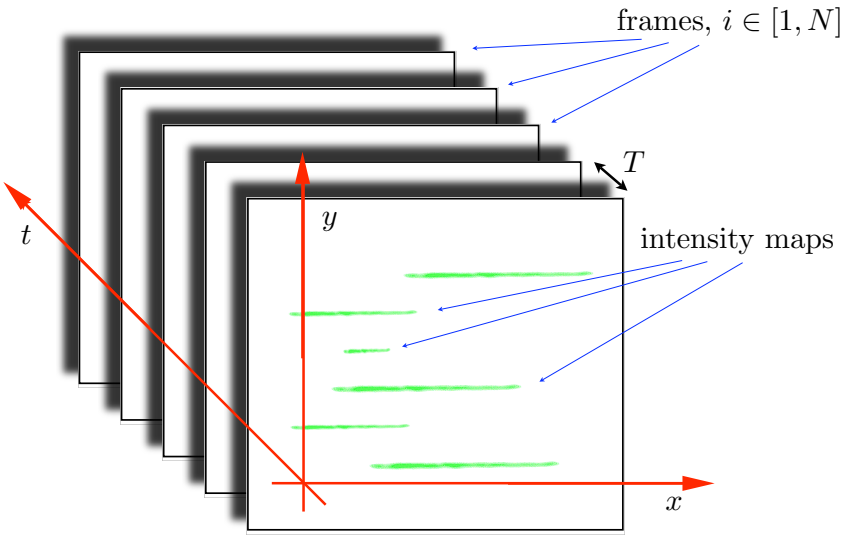


Figure 4.1: The data considered for analysis in this report is represented as a series of frames, $1...N$, in sequence with a strong time dependance. The frames are generally correlated in time with a constant delay T. The spatial dimensions of each frame $[x, y]$ defines the position of each intensity map, which fluctuates and moves in the frames.

4.2 Locating the channel

In this section methods for locating the position and direction of the data values indicating the channels in the raw data will be presented. Noise handling will be introduced first as a method to generally enhance the signal to noise ratio (SNR) in the raw data. After general noise handling the actual values indicating the channel will be considered. After the values indicating the channel is selected, a channel model will be fitted to the data. When the position of the data values along the defined channel is subsequently extracted any noise handling is removed before the extraction to ensure minimum loss of data. In this way noise handling is only a temporary tool to find or define

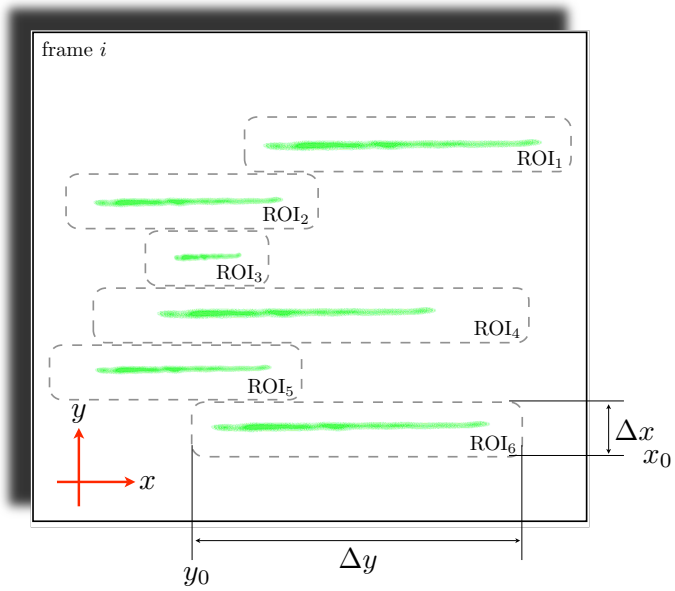


Figure 4.2: Each frame can contain multiple intensity maps in addition to regions of high contrast noise. In the figure 6 different ROI are defined and numbered. Each ROI is identified by $[\Delta x, \Delta y]$ and a number. Each ROI are chosen for each set of frames and do not move frame one frame to another.

positions of specific features and raw data is always passed on. Noise handling is only considered when illustrating the data.

4.2.1 Noise handling in the frames

To enable accurate selection of ROI and later select the values indicating the channel some noise handling is needed to enhance the SNR in the raw data.

Three different general methods were separately considered to handle the noise in the raw data; multiple frames, filtering and thresholds. These methods were implemented through a user interface as needed, see ch. 5 on the GUI's used, with different parameter settings available as needed. No automatic filtering was employed giving the user full control over the noise handling. Generally the multiple frames method and some type of mild low pass filtering is the most relevant noise handling when choosing ROI and threshold only when the values indicating the channel is selected. In fig. 4.3 a ROI has been selected in a low pass filtered frame.

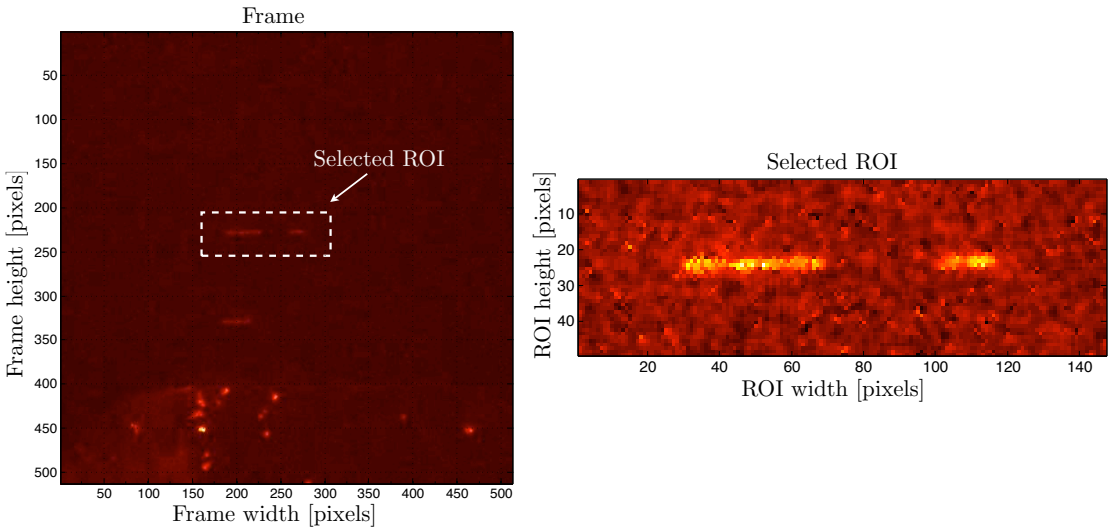
Multiple Frames

The multiple frames method handles noise by averaging away noise with low persistence in time and increasing the SNR of signals with stronger persistence trough time. The multiple frames noise handling increases the time correlation of the frames.

When employing the method to select ROI or values indicating the channel, where only rough features of the molecule is relevant, the method can be implemented with good results using rough averages, even averages of all frames were used when the SNR were very low. See fig. 4.4.

Filtering

Filtering was done by employing Finite Impulse Response (FIR) filters, both of 1D and 2D design. The filters of 1D design considers each row or column of the frame lattice separately. 1D filter



(a) (b)

Figure 4.3: Filtering the frame using a mild low pass filter, in the figure a Wiener filter of size 3x3 pixels was used, increases the SNR enough to easily select a ROI.

designs therefor only be used effectively on frames where the observations of the channel is roughly aligned with either side of the frame. When using data filtering the frame lattice is considered a signal in two dimensions. The filters can be spatial filters, applying masks to the frame, frequency filters adding masks to the the frequency response of the frame, or filters based on statistical calculations of the frame, such as the wiener filter that uses statistical estimates of a neighborhood of each pixel.

The Data filtering were usually employed with rather large strength, blurring local features but effectively enhancing global features. See fig. 4.4. Filtering generally decreases the over all signal strength, but the signal strength can easily be raised after filtering by re-scaling the intensity without significant loss in SNR.

Thresholds

Intensity thresholds were used to cut off low level intensity values in the data, isolating and identifying signal intensity regions with significantly higher SNR of signal from molecules than from the device.

Thresholds were effective when selecting data values indicating the channel as the allowed intensity intervals can be chosen very small and still yield enough data points from the channel to accurately determine algorithmically its position and direction. See fig. 4.5 and fig. 4.4.

4.2.2 Selecting intensity values indicating the channel

Selecting good intensity values indicating the observations of the channel in the frames is crucial to have accuracy in defining the channel position and direction using a channel model fitting.

After the noise handling, described in previous sections, the SNR is increased enough to select the values defined as indicating the channel. The values are selected algorithmically by extracting all local maxima from the filtered ROI selected in previous steps. See fig. 4.5 for example of algorithmically selected channel values at different threshold values with and without filtering.

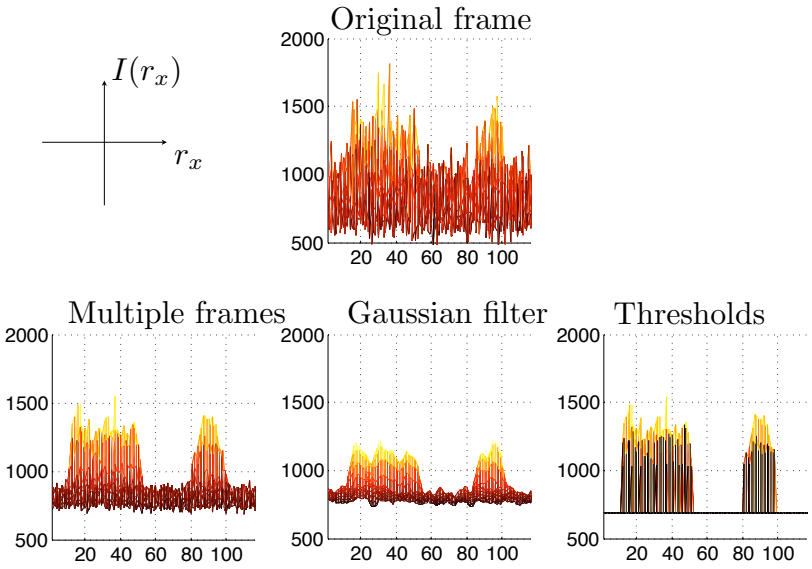


Figure 4.4: Illustration of noise handling methods used when selecting ROI and values indicating the channel in the frame. In the illustration multiple frames is an average of $N = 10$ frames, the Gaussian filter symmetrical of width $n = 10$ and standard deviation $\sigma = 10$, the threshold set to $I_{max} = 50$ % of the intensity interval of the ROI.

4.2.3 Fitting channel model

In this report the channel is always assumed to be linear and only linear fitting methods will be considered. In general the channel model can assume any pre-defined non-linear or piece-wise linear shape and the fitting can assume any non-linear regression of the general shape of the channel to the selected values. Fitting a channel model to the selected values indicating the channel is done by fitting a geometrical representation of the assumed channel shape to the selected values.

In this section two linear fitting methods will be considered. The simplest method is the Ordinary Least Squares (OLS) method and the more competent the Iteratively Re-weighted Least Squares (IRLS) method.

Ordinary Least Squares

The Ordinary Least Squares estimator is the simplest implementation of a linear regression considered in this report. Consider a set of n selected values $\{y_i, x_i\}_{i=1}^n$ and the linear model

$$\hat{y} = \alpha + \beta\hat{x}, \tag{4.1}$$

OLS fits the linear model in eq. 4.1 to the selected values by minimizing the sum of all squared residuals in the objective fungtion

$$\min_{\alpha,\beta} Q(\alpha,\beta) = \min_{\alpha,\beta} \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2, \tag{4.2}$$

where the residuals represent the vertical distance between each selected value and the linear model.

It can be shown that eq. 4.2 is satisfied by

$$\begin{cases} \hat{\beta} &= \dfrac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{j=1}^n y_j}{\sum_{i=1}^n (x_i^2) - \frac{1}{n}(\sum_{i=1}^n x_i)^2} = \dfrac{\mathrm{Cov}[x,y]}{\mathrm{Var}[x]} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\,\bar{x}. \end{cases} \tag{4.3}$$

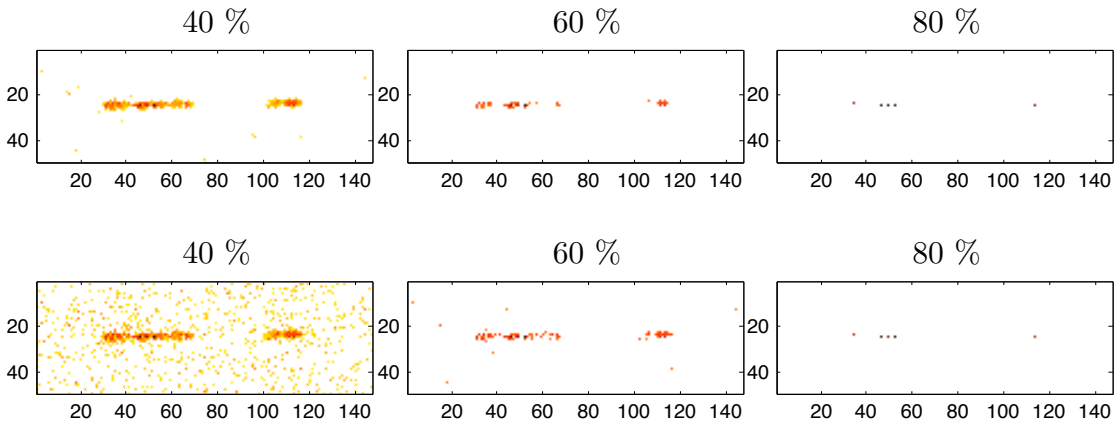


Figure 4.5: Algorithmically selected intensity values in ROI at different threshold values, with filtering (upper row) and without filtering (lower row). The filtering used is a Wiener filter of width and height $n = m = 5$. Intensity values are selected by locating local maxima regions. Axis values indicate pixel positions relative the selected ROI.

The OLS method becomes less reliable when stray intensity values from outside the channel is included in the fitting, as the OLS method weights all values equally and values outside the channel will pull the linear model away from the actual channel position if non-channel values are present.

Iteratively Re-weighted Least Squares

To decrease the effect of selected values from outside of the channel position, weights are attached to each selected value, giving selected values higher weight if the values cluster along the linear model fitted.

In this report the IRLS method will be considered where the squared residuals ε_i^2 are iteratively re-weighted. The objective function Q in eq. 4.2 becomes in IRLS

$$\min_{\alpha,\beta} Q(\alpha,\beta) = \min_{\alpha,\beta} \sum_{i=1}^{n} w_i(\beta)\hat{\varepsilon}_i^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} w_i(\beta)(y_i - \alpha - \beta x_i)^2, \tag{4.4}$$

where w is a bisquare function

$$\begin{aligned} w_i &= \begin{cases} (1-r_i^2)^2, & |r_i| < 1 \\ 0, & \text{otherwise} \end{cases} \\ r_i &= \frac{y_i - \alpha - \beta x_i}{\gamma \hat{\sigma}_i \sqrt{1-h_i}}. \end{aligned} \tag{4.5}$$

$\hat{\sigma}$ is a estimation of the standard deviation of the error term ε_i, h_i is the leverage term estimating the influence of the selected value y_i on the regression model, γ is a constant.

Solving the IRLS is done algorithmically with the update rule

$$\beta^{(t+1)} \leftarrow \min_{\beta} \sum_{i=1}^{n} w_i(\beta^{(t)})\hat{\varepsilon}_i^2(\beta^{(t)}). \tag{4.6}$$

When using channel model fitting in this report only the IRLS will be considered as it performs significantly better than OLS without being orders of magnitude more computational heavy.

4.3 Extracting Time Trace

When the channel position and direction is defined the values corresponding to the model fitted to the selected values can be extracted from all frames in the raw data and a kind of kymograph created. A kymograph is a graphical representation of spatial positions over time. In this report the 1D spatial extension of the channel model is plotted as a function of the frame number or time, which in this report will be denoted Time Trace (TT) of the channel.

When creating a TT from the raw data along the located channel model as much information from each frame as possible needs to be extracted. In this section two approaches will be considered when creating a TT. First the simplest most straightforward approach, Nearest Pixel Value (NPV), will be considered. The more complex, Window Function Averaging (WFA), will be considered next. After both approaches has been considered some notes about raw data compilation into a TT vill be presented.

4.3.1 Nearest Pixel Value

Extracting TT data values from the raw data frames using the NPV approach is the most straight forward way to create TT. The NPV uses the previously fitted channel model, see section 4.2.3, and selects the pixel values from the raw data frame that is closest to the model, corresponding to the minimum distance. If r_x is the position along the fitted channel, the NPV selects the TT values I_{TT} for all frames n as

$$\begin{aligned} &I_{\mathrm{TT}}(n, r_x) = I_{\mathrm{frame},n}(\hat{x}, \hat{y}) \\ &\hat{y} = \mathrm{round}(\alpha + \beta\hat{x}) \end{aligned}, \tag{4.7}$$

where $I_{\mathrm{frame},n}$ is the raw data frame n and $\hat{x}$ is the discrete position x of the selected pixel relative the frame. α, β is the fitted channel model parameters.

The NPV approach for TT value extraction works best if the channel is roughly aligned to the frame, vertically or horizontally, or the resolution of the raw data frame is such that the width in pixels of the channel is large. Extracting channel data for a time trace using the NPV setup without alignment or enough resolution in the raw data frame can result in loss of relevant data and sharp discontinuities in the TT, as illustrated in fig. 4.6.

4.3.2 Window Function Average

A more complex approach to extracting TT data values from the raw data frames is WFA. With the WFA a weighted window function is defined and applied to the raw data frame as the TT values are selected, generally the window function is an asymetric 2D pulse $P(\hat{x}, \hat{y})$ defined on a finite interval

$$I(\hat{x}, \hat{y}) = \frac{1}{MN} \sum_{i=-\frac{N}{2}}^{\frac{N}{2}} \sum_{j=-\frac{M}{2}}^{\frac{M}{2}} P(\hat{x} + i, \hat{y} + j), \tag{4.8}$$

where N, M define the window size.

With the WFA eq. 4.7 then becomes

$$\begin{aligned} &I_{\mathrm{TT}}(n, r_x) = \frac{1}{MN} \sum_{i=-\frac{N}{2}}^{\frac{N}{2}} \sum_{j=-\frac{M}{2}}^{\frac{M}{2}} P(\hat{x} + i, \hat{y} + j) I_{\mathrm{frame},n}(\hat{x} + i, \hat{y} + j) \\ &\hat{y} = \mathrm{round}(\alpha + \beta x) \end{aligned}, \tag{4.9}$$

which weights the surrounding of each pixel $I_{\mathrm{frame},n}(\hat{x}, \hat{y})$ along the defined channel model in the raw data frame into the selected TT data value $I_{\mathrm{TT}}(n, r_x)$.

An example of a window function is the symmetric Gaussian

$$P(\hat{x}, \hat{y}) = e^{-\left(\frac{(\hat{x}-\hat{x}_0)^2}{2\sigma_{\hat{x}}^2} + \frac{(\hat{y}-\hat{y}_0)^2}{2\sigma_{\hat{y}}^2}\right)}, \tag{4.10}$$

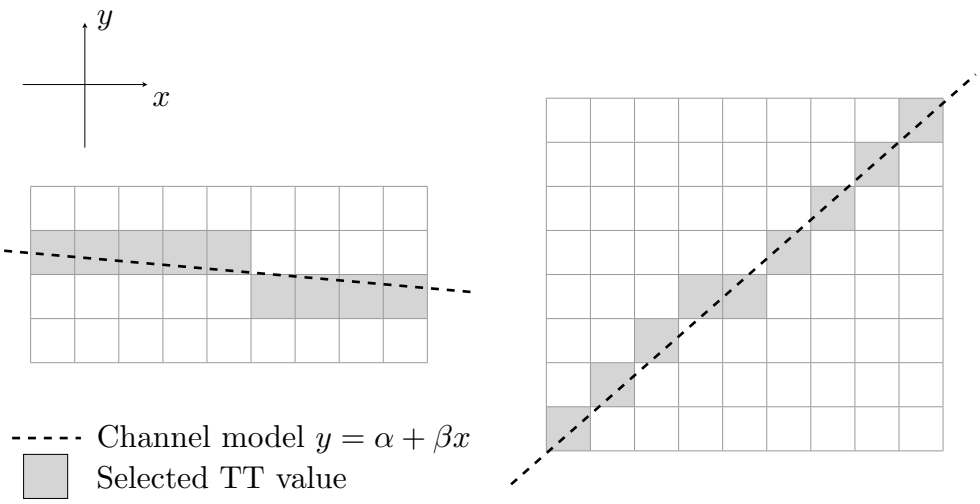


Figure 4.6: When extracting values for TT in the raw data frame using the NVP method the nearest pixel value to the fitted channel model is selected. To the left the channel model is roughly aligned to the frame and to the right the channel is un-aligned.

where $\sigma_{\hat{x}}, \sigma_{\hat{y}}$ determines the slope and width of the window function. In fig. 4.7 a channel has been defined and values selected and extracted both using NPV and WFA. The window function used is a symmetrical Gaussian with $\sigma_{\hat{x}} = \sigma_{\hat{y}} = 2$ defined on a square window of size $N = 5$.

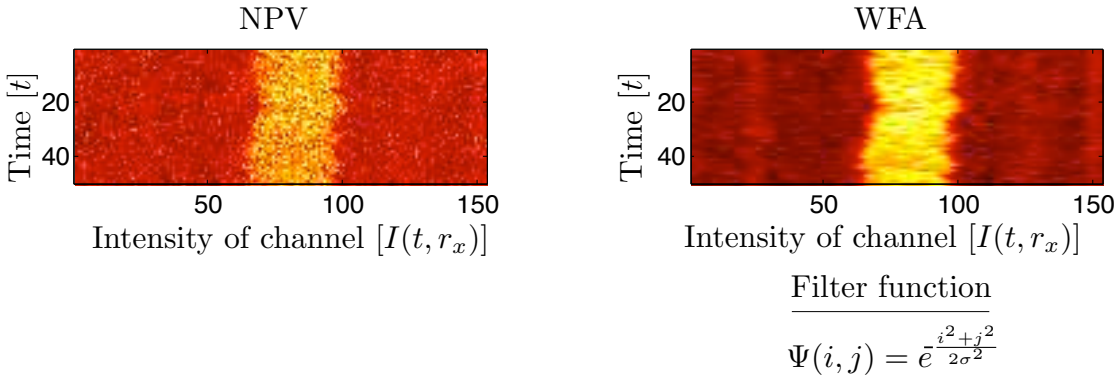


Figure 4.7: The figure shows a TT compiled from a set of frames containing an observation of λ-DNA in a noisy channel. On the left the TT data values is extracted using NPV and on the right the TT data values is extracted using WFA, with a symmetrical Gaussian window function with $\sigma_{\hat{x}} = \sigma_{\hat{y}} = 2$ defined on a square window of size $N = 5$.

4.3.3 Raw Data Compilation

When compiling raw data into TT in this report only the first approach, NPV, were used. NPV was found to extract enough information along the channel model extension for the purposes of this report without unnecessarily reducing the resolution of the raw data. Using WFA to create TT can be compared to first applying noise handling to the frame and then creating TT using NPV, assuming a strong alignment of the selected channel values to the frame. The WFA approach also

lowers the resolution of the raw data compiled in the TT as the WFA acts as a low-pass filtering on the TT. In fig. 4.8 a illustration of the difference between NPV and WFA is presented. The illustration shows a simulation of an intensity map i ROI. The intensity map were created with potential functions where

$$I(x,y) = \sum_{i=1}^{K} e^{-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma_i^2}}, \tag{4.11}$$

and K denotes the number of maximas in the map.

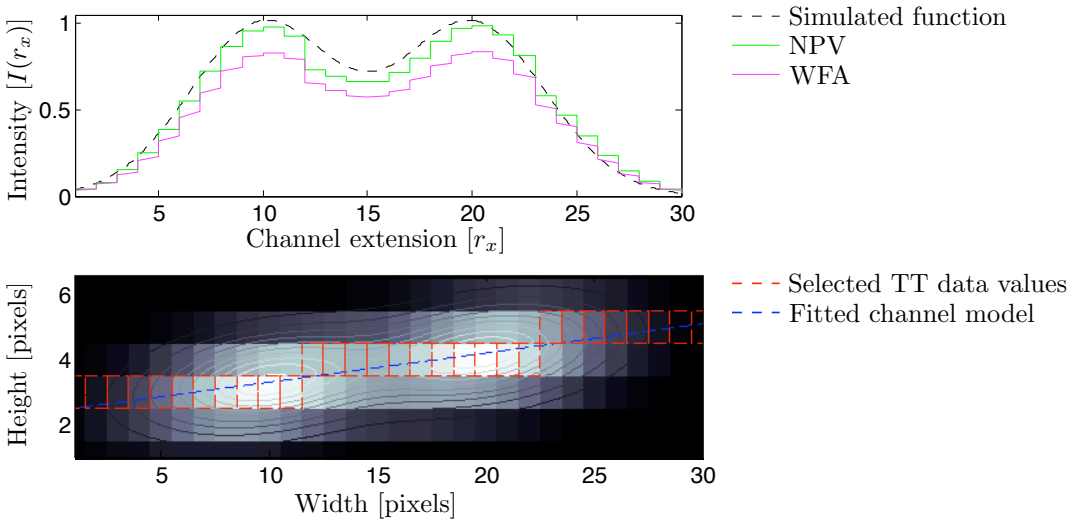


Figure 4.8: Simulating a ROI with a potential field simulating the intensity map, the difference between NPV and WAF using a Gaussian window function can be illustrated. In the upper plot of the channel intensity $I(r_x)$ as a function of the extension r_x it is clearly illustrated that the WFA dampens the noise of the simulated intensity map and that NPV is a better fit and sufficient. The lower plot show the simulated ROI and intensity map with annotations of the selected TT values and the fitted channel model.

4.4 Finding edges of the molecule in time trace

Ideally the edges of the molecule in the TT would be found by setting a threshold value above the signal intensity of the channel in the device and below the signal intensity of the molecule in the channel and then selecting the values chosen by that threshold. As the signal to SNR usually is very low in the frames of the raw data this approach is insufficient. A more sophisticated approach is to create a model pulse with enough rough features of the actual intensity map of the molecule and then fitting this model to the TT for each frame.

In this section edges of molecules were found by creating a smooth statistical model pulse with relatively sharp edges with a variable number of features between each with sharp but smooth transitions, and then fitting the model to the intensity map of each frame in the TT using a non-linear regression fitting.

4.4.1 Slope model and slope fitting

Model functions considered in this report is constructed using the error function as slopes for the pulses constructed. There are a number of different stepping functions with an exponentially

rising slopes bound to a finite function value interval that could be used, see fig. 4.9. The error function

$$y(x) = \frac{a}{2}\left(1 + \text{erf}\left(\frac{x - x_0}{\sigma\sqrt{2}}\right)\right) \tag{4.12}$$

was for this report chosen by convenience as it in a convenient way can be modified by position, amplitude and pitch with the parameters x_0, a and σ respectively.

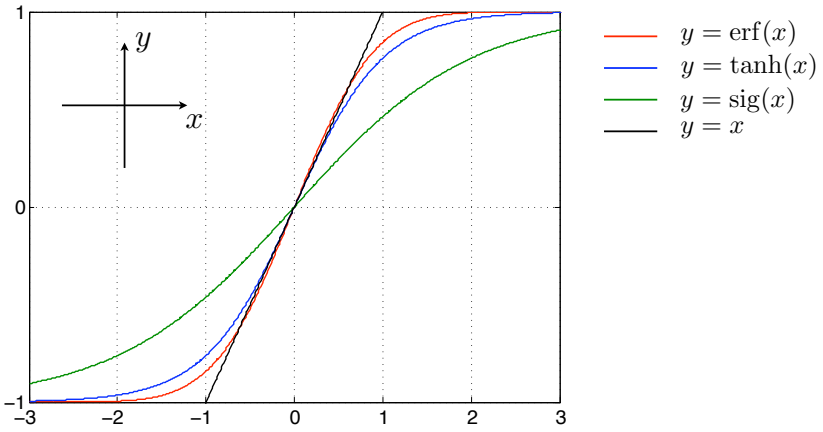


Figure 4.9: When creating a continuos model pulse with reasonable parameters a stepping function is needed. The figure shows four different stepping function with a straight line, $y = x$ as reference. As all functions rises exponentially in the interval $[-1, 1]$ the significant difference between the functions are the pitch of the slope and the parameter characteristics. In this report the error function $\text{erf}(x)$ were chosen by convenience as its position, amplitude and slope easily can be modified with the parameters x_0, a and σ respectively, see eq. 4.12.

Using exponential slopes to create a pulse model is convenient as the fitting of the exponential slopes to sample data gives good fitting values and the model pulses are readily created. In fig. 4.10 left and right sloping error function slopes has been fitted to the intensity map, $I(x)$, of a DNA molecule. The fitting method is a simple OLS and as can be seen the edges of the molecule has significantly better fitting values than the rest of the intensity map. Note also the effect of the slope direction of the model slope.

The residual equations used for the left, $L(X)$, and the right, $R(x)$, slope fitting respectively has the form:

$$\begin{aligned} L^2(x) &= \frac{1}{2\xi}\sum_{n=x-\xi}^{x+\xi}\left[I(n) - \frac{1}{2}\left(1 + \text{erf}\left(\frac{n-x}{\sigma\sqrt{2}}\right)\right)\right]^2 \\ R^2(x) &= \frac{1}{2\xi}\sum_{n=x-\xi}^{x+\xi}\left[I(n) - \frac{1}{2}\left(1 - \text{erf}\left(\frac{n-x}{\sigma\sqrt{2}}\right)\right)\right]^2 \end{aligned} \tag{4.13}$$

where ξ is half the slope width used.

4.4.2 Model function fitting

When the stepping function is chosen a model pulse is constructed using multiple slopes in sequence with different amplitudes and slope signs, i.e. directions, and steepness. The model pulse has the general form

$$\Psi_M(x) = I_0 + \sum_{i=1}^{M}\frac{a_i}{2}\left[1 + \text{erf}\left(\frac{x - \left(x_0 - \frac{L_0}{2}\right) - L_i}{\sigma_i\sqrt{2}}\right)\right] \tag{4.14}$$

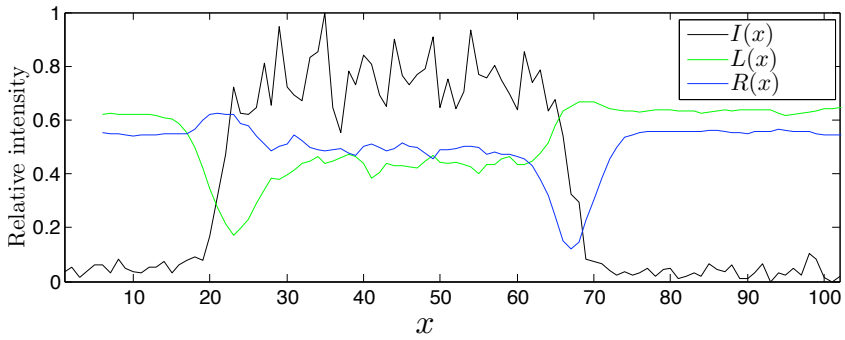


Figure 4.10: Using model slopes created with error functions and fitting the models to the data using ordinary least squares fitting, see eq. 4.13, first with a left slope and then a right slope will give significantly better fitting values in the regions of the intensity map where the amplitude changes as an indication of a molecule. In the figure both the data intensity map and the result from the fitting of the model slope in both directions is shown. The left and right slope in the figure are the square root of the equations in 4.13 respectively, i.e. $L(x)$ and $R(x)$. As can be seen the edges is clearly visible in the fitting values.

In eq. 4.14 the steepness of the slope is determined by σ_i, the number of slopes by M, the length between two slopes by L_i, a_i determines the relative amplitude of each slope. The global parameters x_0 and L_0 determines the global location of the composite pules and its total width respectively. In fig. 4.11 a model pulse with 4 slopes is shown together with notations indicating the different parameters.

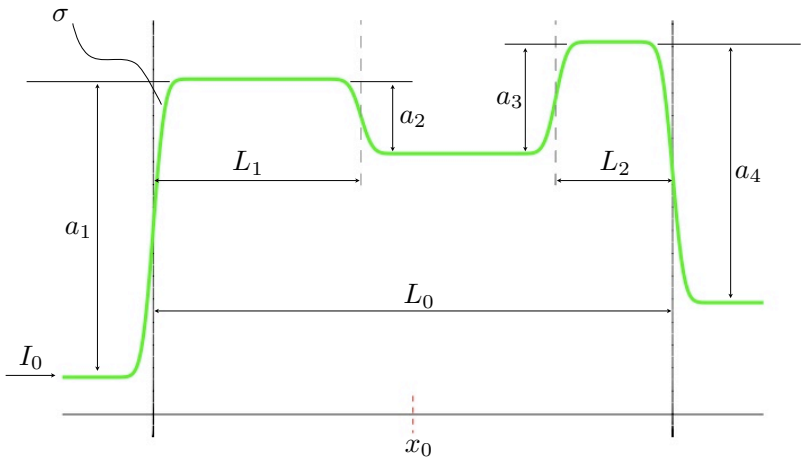


Figure 4.11: The figure illustrated a model pulse with slopes created using erf-functions. Annotations of the pulse parameters is added to the illustration.

In order to bind the model parameters to always be in the region that conserves the general form and position of the model, constraints needs to be put on all the parameters. Generally the restrictions set on the parameters is in the form of pre defined intervals that cannot be exceeded. The exact restrictions put on the parameters will not be defined in this section as they depend on the raw data. In chapter 5 on the GUI's used further information on exactly how the restrictions are treated is considered.

Finding the edges of the molecule in each frame of the time trace was done by using a non-linear regression fit on the slope model applied to the data. In fig. 4.12 two model pulses, one with only two outer slopes and one with two inner slopes as well, were fitted using a non-linear regression fit on an intensity map of a DNA molecule.

The non-linear regression algorithm used is an Levenberg-Marquardt algorithm[1] that uses non linear least squares to compute a iterative fit of the model function to the data. The algorithm requires relatively good starting values for the fit to be able to converge. In Chapter 5 about the GUI's used handles are presented for manual setting of the initial search values.

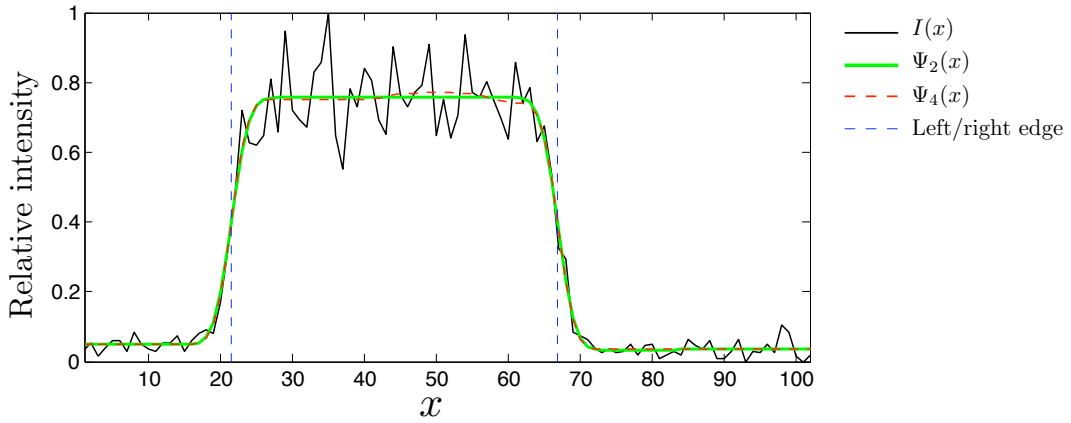


Figure 4.12: Finding the edges of the molecule is done by applying non-linear regression of a double humped error function pulse, $\Psi_4(x)$, with independent intensities between and outside the humps. The molecule is defined to exist between the outer slopes of the regression model. A single humped model pulse, $\Psi_2(x)$ is also shown as reference.

4.5 Tracing local features along the DNA

Tracing features along the molecule confined in the channel through the TT is the last part covered in this chapter.

In this section basic ideas will be presented on tracing features along the DNA trough time.

4.5.1 Average time trace

The simplest method of presenting the features along the DNA as a one dimensional array is a simple average of frames in the TT. This method removes noise effectively by using the inherent redundancy created by the length of time the molecule is observed.

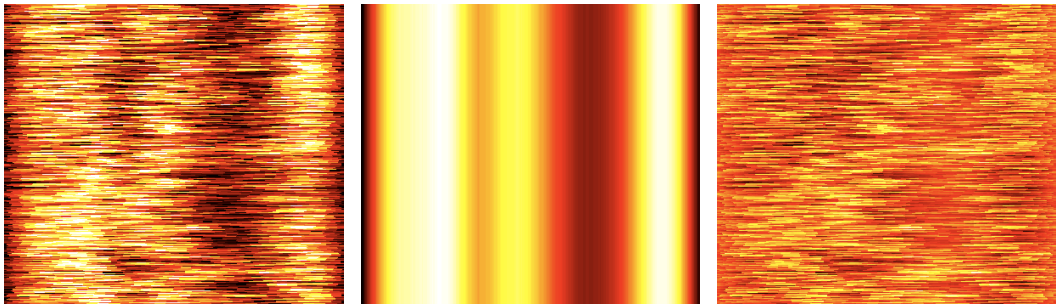
Problems arrises when observing and taking into account the fluctuations along the molecule present in both length fluctuations and warping of features along the length of the molecule.

4.5.2 Uniform stretching of molecule

One way to analyze the DNA molecule is to handle the fluctuations of the molecule by only considering the length of the molecule and ignore all internal local fluctuations. In the channel when the molecule is observed the molecule length is assumed to have had time to converge to a state with least entropy, i.e. the length of the molecule is fluctuating around some effective length. In Chapter 5, on the GUI's used, methods to study the distribution of length fluctuations is presented making it possible to study diffusion and effective length. Using the fluctuations to calculate the effective length of the molecule a value for the length of the molecule can be studied.

Considering all length fluctuations as global the molecule can be uniformly stretched along all of its length and an average of several frames can be made. In fig. 4.13 an example of a uniformly stretched and aligned molecule is illustrated. The figure also shows the average of all frames in the aligned and stretched time trace as well as the resulting removed data.

Using this method creates a severe smoothing of the features along the molecule. Assuming that each feature along the molecule confined in the channel fluctuates around the effective position of that particular feature and the fluctuations is roughly Gaussian in distribution the average becomes a gaussian filtering on the features with strength corresponding to the effective length of all local fluctuations.



(a) Stretched and aligned. (b) Average of all frames. (c) Subtracted data.

Figure 4.13: Uniform stretching with average is one of the simplest ways to obtain a one dimensional picture of the molecule from the redundant noisy data in the original frames. From left the figures presented is a aligned and stretched molecule, an average of all frames (rows) in the first figure and the resulting removed noise and fluctuations in the right picture.

4.5.3 Tracing local features through time

A more labour intensive approach is to define local features along the molecule and attempting to follow these features through time. When enough features are traced, traces of the features through time can be constructed and then an average can be obtained from this locally aligned and stretched time trace. This method corresponds to shortening the effective fluctuation length for each local feature along the molecule.

In fig. 4.14 local maximas along the intensity map is defined as the local features, and arrows indicate the direction and length of the fluctuations of the defined features trough four frames for in a small subsection of the channel.

In the scope of the project for this report no attempts to align traced features has been made due to limits in time and resources for this project. An attempt to trace features trough time has however been made and can be explored using the GUI presented in Chapter 5.

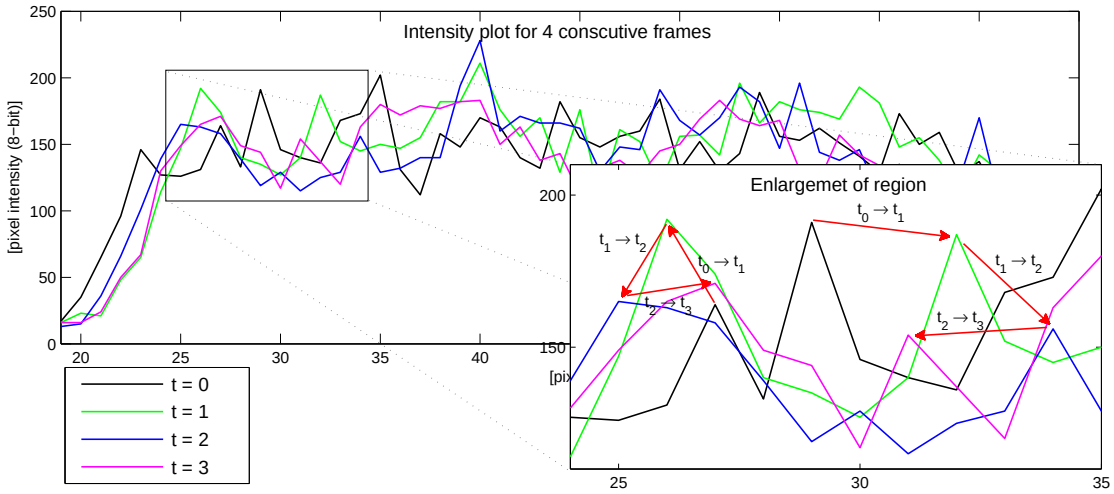


Figure 4.14: The figure show how the features, defined as local maximas, evolve through time. Four frames are shown together, with an enlargement of a small region, where the local maxima features are traced out with red arrows.

Chapter 5

The Software

In this chapter the software developed for the project considered in this report will be presented. The software is called *SMATool* 1.0, which is short for Single Molecule Analysis Tool, and was developed by E. Sandlund as a project for the degree of master of science. The purpose of the software is to act as a tool to display and analyze experimental data obtained from observations of single molecules confined in nano-channels. The software is intended to be general in its application and not assume any specific type of molecule in the observation.

The chapter will follow the structure

1. Some notes about the data input for the software
2. Using the software
3. Specifics on methods and functions within the software

5.1 Data considered for the software

The data considered for the software needs to be of a specific type to be readable by the software. The software assumes the data is collected from observations of single molecules confined on a sample in nano channels, described in chapter 4, and the observations is assumed to be stored in in a series of consecutive frames, where the observations of the single molecules moves on the frames as intensity maps.

5.1.1 Filetype for frame collections

The file format of the collections of frames, denoted stacks, is assumed to be uncompressed Tagged Image Format version 6.0 (TIF) with frames stored in a stack of the same file. The intensity resolution of each pixel of the TIF-file is assumed to be 16-bit.

For optimal performance the channels observed in the data should align roughly either horizontally or vertically.

5.1.2 Filetype for Time Trace

TT files read and produced by the software is assumed to be TIF-files with one frame in the stack, with an assumed bit depth of 16-bit. Each row in the image file should representlat one frame in the raw data of extracted channel values, and each column a specific position along the fitted channel.

5.1.3 Save filetypes

Saved files can be stored as either or all of TIF-, TXT- or MAT-files. TXT-files is basic ascii text files (TXT) containing meta data produced by the software. MAT-files (MAT) are double-precision, binary, files specific of MATLAB. In the MAT files variables used by the software can be stored.

Saved intensity data is usually stored as a TIF file, with or without stacked frames. Meta data about the intensity data is stored as a TXT with the same name as the stored intensity data file. Session data in the software is stored as MAT files.

5.2 Starting the software

Execution of the software *SMATool* (SMAT) is done from within the numerical computing environment MATLAB, developed by MathWorks. The software is then divided into two parts. One part extracts ROI and fits a channel model to the located values from observations of the nanochannel containing the single molecule, after fitting the TT can be compiled from the selected channel data values. In the other part the TT can be viewed and different features, such as length fluctuations and diffusion, can be viewed and analyzed.

Starting the GUI is done within the root directory of the GUI files, i.e. the same directory as the folder `@SMATool`, from the MATLAB command prompt with the command

```
>> SMATool
```

When the GUI is launched the window in fig. 5.1 appears and the user can choose weather to proceed with extracting time trace in a ROI from a new set of raw data (*Extract Time Trace*), trace edges of molecule in existing TT (*Trace Molecule Edges*) or trace features in an existing TT, where the molecule edges are traced (*Trace Features Tool*).

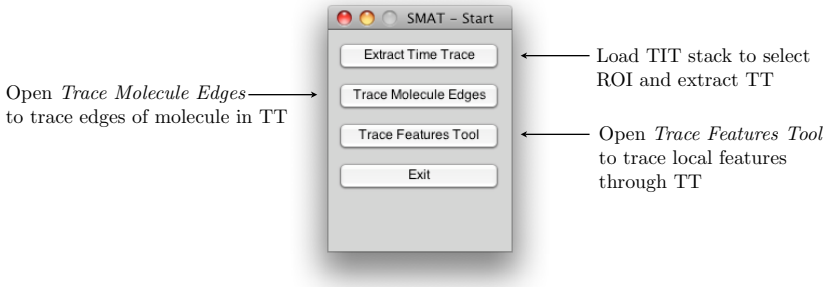


Figure 5.1: A startup window appears after launch of SMATool, where *Extract Time Trace*, *Trace Molecule Edges* or *Trace Features Tool* can be selected to either select a ROI and extract TT, trace molecule edges in an existing TT or trace features in an existing TT where the molecule edges are traced.

5.3 Extract Time Trace

When selecting and clicking *Extract Time Trace* in the startup window, see fig. 5.1, a load file window is opened, see fig. 5.2, where a TIF stack can be selected to load into the software for TT extraction.

After a TIF stack is loaded the new window Extract Time Trace (ETT) appears. In ETT a ROI is selected and TT extracted from a fitted channel model, see fig. 5.3.

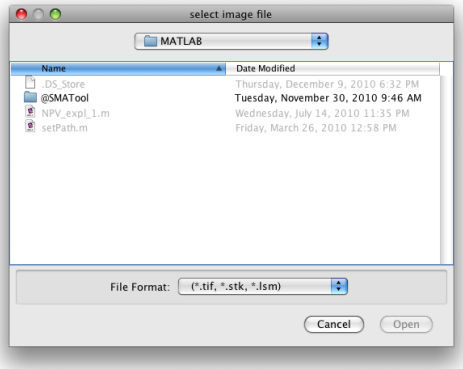


Figure 5.2: Load files window in SMATool are opened to selected to be loaded into the software.

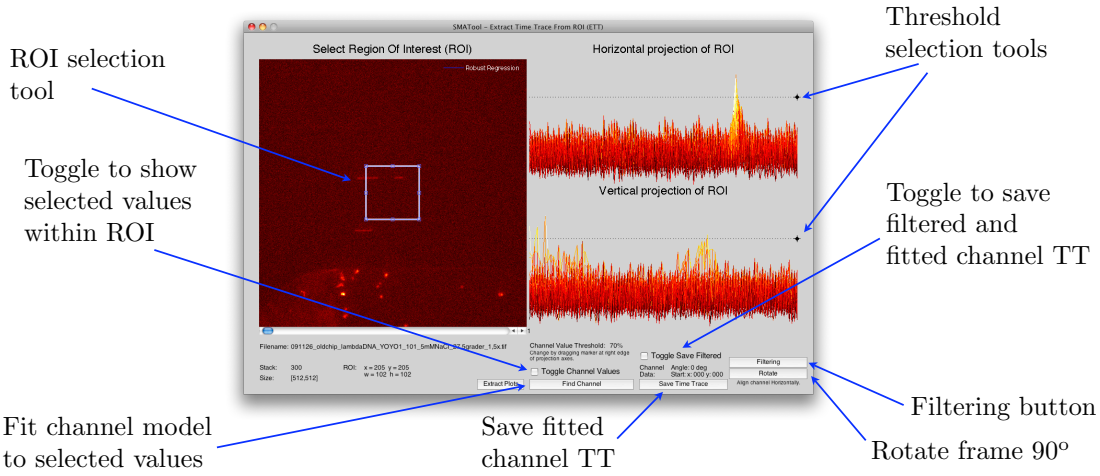


Figure 5.3: Start window for ETT. ROI selection appears after a TIF stack is loaded into the software as well as horizontal and vertical projections of the ROI. In the window ROI can be selected and viewed, channel model fitted to selected channel value and saved.

5.3.1 Select Region Of Interest (ROI)

In the ETT window the first frame of the TIF stack is displayed to the left with a scroll bar underneath, where the currently displayed frame can be changed. In the current frame view a resizable rectangle selection tool is displayed. The tool is used to define and change the ROI. Moving the selection is done by clicking and dragging it to the new position. Resizing the selection is done either by clicking and dragging the corners or the edges of the selection tool. When the ROI selection or the current frame number is changed the ETT window automatically updates all graphs and information labels to display the new current frame and settings. See fig. 5.3.

5.3.2 Projection axes of ROI

To the right in the ETT main window the horizontal and vertical projection of the selected ROI is displayed as intensity graphs with axes scaled to display the full spectra in the selected ROI. In order to extract a nice TT the channel needs to be roughly horizontally aligned in the frame. When the channel is aligned correctly and a good ROI is chosen the horizontal projection of ROI

should contain the smallest region of high intensity values an the vertical projection the longest region.

5.3.3 Setting and changing threshold values

On the right edge of the plots there is a selection tool for changing the threshold values for channel value selection, see fig. 5.3. The tool is used by dragging the marked dot to the left of the projection plot up or down. Both the vertical and the horizontal selection tool can be used interchangeably.

5.3.4 Filtering and rotating of frame

To add noise handling to the daw data frames the *Filtering*-button in the bottom right corner of the ETT window is pressed and the window in fig. 5.4 opens. In the filtering window different noise handling settings can be set and added to the current frame and adjusted as needed. To add a filter to the current frame the selected filter type checkbox is checked, by clicking it. Adjusting the parameters for the selected filter is done by entering new values into the filter parameter settings boxes of the filtering window. Notes on the filter types and its parameters can be read to the right of the filters in the window. Details of the software functions used by the filtering settings is described in Section 5.6. In fig. 5.5a a Gaussian filter has been added to the current frame as an illustration.

To rotate the whole frame, if the channels in the data are aligned vertically, the *Rotate*-button is pressed once.

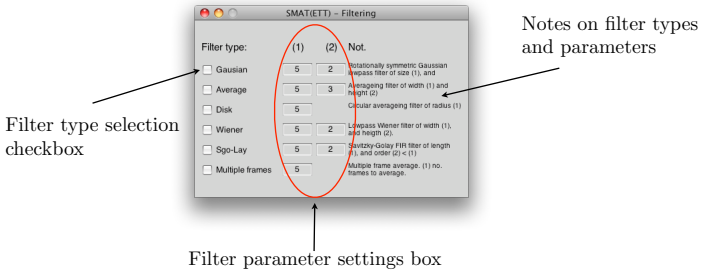


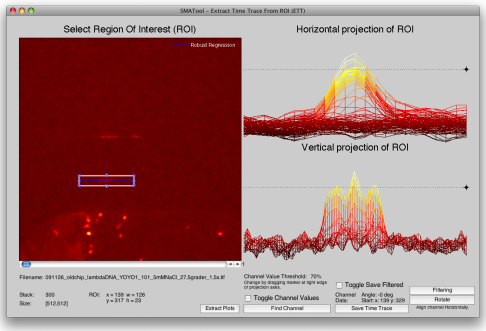
Figure 5.4: Filtering selection window for TT extraction. In the window different noise handling settings can be set and adjusted. Filters is added by checking the filter type selection checkbox and parameters adjusted by entering new values into the filter parameter settings box.

5.3.5 Find Channel and Toggle Channel Values

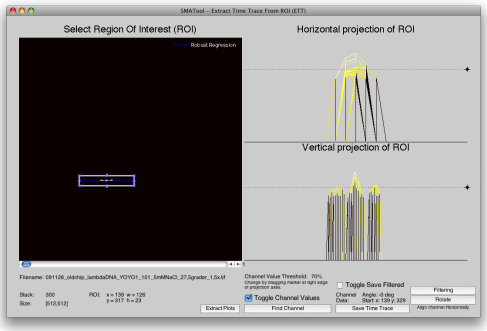
To find a channel in the selected ROI filtering is usually added and then the *Find Channel*-button is pressed. To check the which selected channel values were chosen by the software after noise handling and selecting intensity threshold the *Toggle Channel Values*-checkbox is checked. See fig. 5.3 for an illustration of filtering and fig. 5.5 for a fitted and filtered example and an illustration of the *Toggle Channel Values*-checkbox.

5.3.6 Extract Plots

In the ETT window all plots can be extracted from the window with the *Extract Plots*-button. The current frame and the two projection plots will be copied to new window with controls for editing and saving the individual plots.



(a) Gaussian filter added to the current frame.



(b) Selected channel values viewing are toggled on.

Figure 5.5: Illustration of the ETT window when noise handling (a) has been added and selected channel values in the ROI has been toggled (b). The blue line is the fitted channel model to the selected values.

5.3.7 Save Time Trace

With the *Save Time Trace*-button the data values along the computed channel is extracted and can saved. When the button is pressed a new window appears where the progress of the extraction is displayed. When the extraction is done the TT can be saved by pressing the *Save*-button in the new window. The extraction is done by default on the original data values, making the TT unfiltered. By checking the *Toggle Save Filtered*-checkbox the filtered TT is also extracted. In the extraction window when the *Save*-button is pressed only the unfiltered TT is save. To save the filtered TT the plots must be extracted from the window by pressing the *Plots*-button.

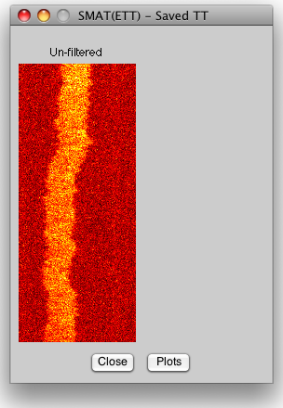


Figure 5.6: When the *Save Time Trace*-button is pressed the TT is extracted in a new window where the TT can be saved and the plots extracted.

5.4 Trace Molecule Edges

When selecting and clicking *Trace Molecule Edges* in the startup window, see fig. 5.7, the empty Trace Molecule Edges window (TME) appears. When the TME window is loaded there are two options for loading data into the software.

Load TIF-file opens a raw TT-file in the required TIF file format. With this option an new TT can be traced and the molecule edges in the TT detected.

Load MAT-file opens data from a previous session of the program by loading a MAT-file. With this option the data from the previous session is loaded into the software including data file paths. In order for the *Load MAT-file* option to function the original TT data file needs to be located in the same directory as in the previous session.

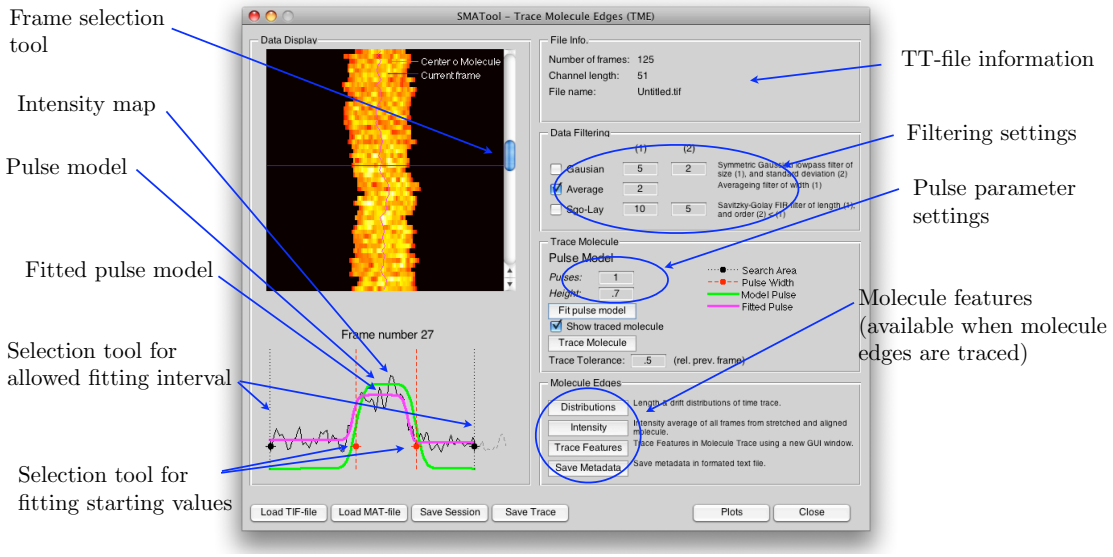


Figure 5.7: All functions available in the TME main window is set and adjusted in the main window. The functions are grouped in different panels according do type.

5.4.1 Data Display

In the *Data Display*-panel the TT is displayed above and the current frame as an intensity plot is displayed below, see fig. 5.7. To the right of the whole TT the current frame can be changed by dragging the scroll bar up or down. The current frame number is displayed below the TT.

Below the TT the current frame is displayed as an intensity plot. In the current frame plot draggable selection tools for defining the allowed fitting interval for the pulse model, black vertical dotted lines, and defining the initial pulse model width, red vertical dotted lines, is placed. To adjust these simply click and drag the solid dots on the respective dotted lines.

5.4.2 File Info

In the *File Info*-panel information about the loaded TT-file is displayed. The number of frames in the TT, the length of each frame in the TT and the file name can be read.

5.4.3 Data filtering

In the *Data Filtering*-panel filtering can be added and adjusted to the TT. By checking the filtering checkboxes and adjusting the values in the parameter settings boxes, (1) and (2), filtering is added to the TT, see Section 5.6.1 for details on methods used.

When the mouse pointer is moved over the different filters and the parameters tooltips appears with notes on the filters and their parameters. Notes on the filter types and parameters can also be read to the right in the panel.

5.4.4 Trace Molecule

In the *Trace Molecule*-panel the pulse model fitting is controlled. Parameters for the number of sub-pulses and the max height of the model pulse is set. The settings for the allowed search interval and starting values for the pulse edges is set by dragging the selection tools in the current frame intensity plot in the *Data filtering*-panel. The model pulse is fitted to either only the current frame, *Fit pulse model*-button, or to all the frames, *Trace Molecule*-button, see Section 5.6.4 for details on functions used.

When the *Trace Molecule*-button is pressed all frames are fitted to the model pulse beginning with the current frame. After all frames are traced the filtering can be removed or adjusted in the *Data Filtering*-panel and molecule features can be studied.

By checking the *Show traced molecule*-checkbox all pixels outside the traced molecule can be toggled on or off. When toggled off the pixels outside the traced molecule appear black as in fig. 5.7.

5.4.5 Molecule Features

When the edges of a molecule is traced, or a already trace molecule is loaded with the *Load MAT-file*-button, the *Molecule Features*-panel is activated and different features of the traced molecule can be studied. The functions available are:

Distributions

By pressing the *Distributions*-button the window in fig. 5.8 appears and the length and drift distributions of the traced molecule can be studied.

Intensity

By pressing the *Intensity*-button the average intensity of the traced molecule can be studied, see fig. 5.9. In the intensity window the mean and standard deviation of the traced molecule is displayed and can be studied.

Trace Features

By pressing the *Trace Features*-button the window in fig. 5.10 appears where features along the traced molecule can be attempted to be traced, see Section 5.5.

Save Metadata

By pressing *Save Metadata*-button information of the molecule edge trace in the current session can be saved as a TXT-file.

5.4.6 Save and export data

When a session needs to be saved the *Save Session*-button is pressed. After a file name has been chosen the software saves the current session in a MAT file. In the session data file information form the current session, containing file name and path to TT file etc., is stored.

When a molecule is successfully traced and needs to be saved the *Save Trace*-button is pressed and after new file name and path is selected the software stores the traced molecule data in two files. File one is a TIT file containing a uniformly stretched, see Section 4.5.2, TT on basis of the current trace. File two is a MAT file containing about the current trace, such as molecule center and width.

If the plots needs to be exported to a new MATLAB figure in order to be edited or modified the *Plots*-button is pressed. All current axes in the open window will then be exported into a new editable window where the individual axes and plots can be edited and saved.

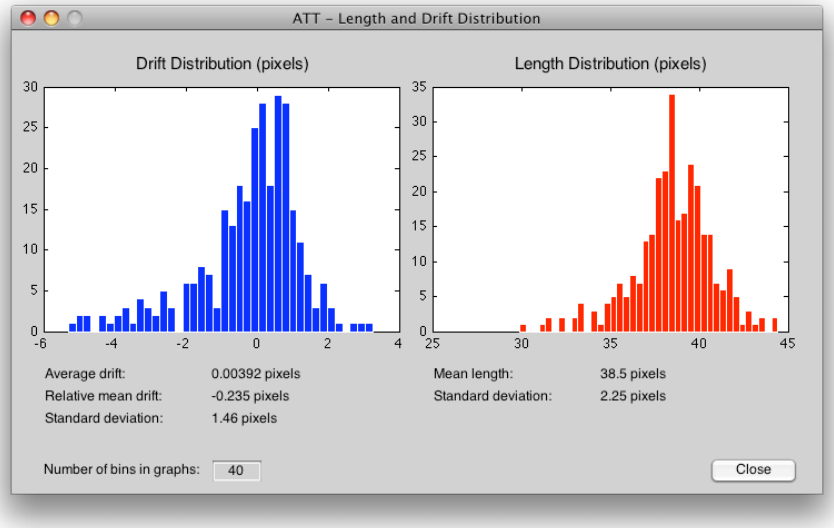


Figure 5.8: When the *Distributions*-button is pressed the length and drift distributions of the traced molecule is displayed in two bar graphs.

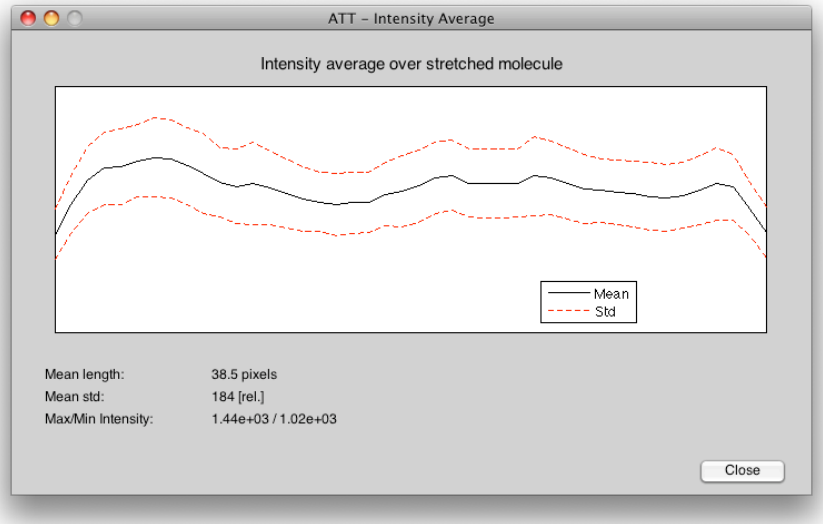


Figure 5.9: When the *Intensity*-button is pressed a window containing the averaged intensity of the traced molecule is displayed. Annotations for the mean and standard deviation of the intensity is also displayed.

5.5 Trace Molecule Features

By pressing the *Trace Features Tool*-button in the SMAT start window or pressing *Trace Features* in TME the main window of Trace Features Tool (TFT) appears.

In the window the TT is displayed uniformly stretched. In the uniformly stretched TT all frames are stretched to match the longest frame in the stack. Uniform stretching of the TT removes all global fluctuations on the molecule and allows the study of only local fluctuations along the molecule.

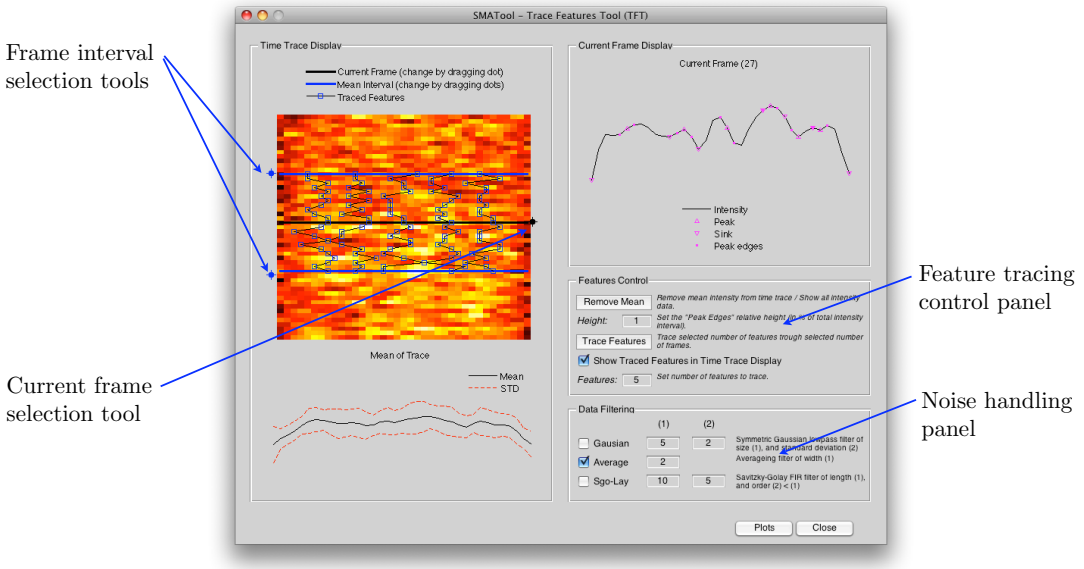


Figure 5.10: When the *Trace Features*-button in TME or *Trace Features Tool* in the start window is pressed a window appears where features along the molecule can be traced.

5.5.1 Time Trace Display

In the *Time Trace Display*-panel the stretched TT is displayed above with selection tools for selecting frame interval and the current frame.

When the *Trace Features*-button is pressed the traced features is also displayed on-top of the stretched TT, see fig. 5.10.

Below the stretched TT the mean of the selected frame interval is displayed with annotation of standard deviation.

5.5.2 Current Frame Display

In the *Current Frame Display*-panel the selected frame is displayed with local peak and sink features annotated. Markers for the edge of the local peak feature are also displayed. The local peak edge definition can be modified in the *Feature Control*-pane.

5.5.3 Data Filtering

In the *Data Filtering*-panel standard noise handling can be applied to the TT, see Section 5.6.1 for details on the noise handling functions used. All noise handling used in TFT are only performed frame-wise, i.e. all frames are handled separately, to avoid loss in time resolution.

5.5.4 Feature Control

In the *Feature Control*-panel control for locating and displaying local features along the molecule is located.

By pressing the *Remove mean*-button the mean of the selected frame interval is removed from the TT. The mean intensity in the *Time Trace Display*-panel is updated to shows the mean and standard deviation of all noise in the selected frame interval. In the *Current Frame*-panel the current frame noise is displayed.

The height parameter controls the height of a the defined local feature peak in the *Current Frame*-panel.

By pressing the *Trace Features*-button local peak features is traced in the selected frame interval through time, see Section 5.6.5 for details on the algorithm used. If the *Show Traced Features*-checkbox is toggled on the traced features are displayed on-top of the TT in the *Time Trace Display*-panel. The number of traced features is controlled by entering new values in the *Features*-parameter box.

5.6 Details on methods and functions used in SMAT

In this section the methods and functions used in the software will be considered.

5.6.1 Noise handling

Noise handling in the software can be divided into tree parts, multiple frame averages, filtering and thresholds. The parts can be used together or separately depending on the desired effect.

Multiple frames average

The multiple frame average function performs a mean calculation of several frames, see fig. 5.11, preserving the mean intensity and size of the original frames. The function averages frames as

$$I_{avr}(x,y) = \frac{1}{k}\sum_{i=1}^{k} I(i,x,y), \tag{5.1}$$

where x, y is the position of the intensity data values I in the frames. k denotes the number of frames to be averaged.

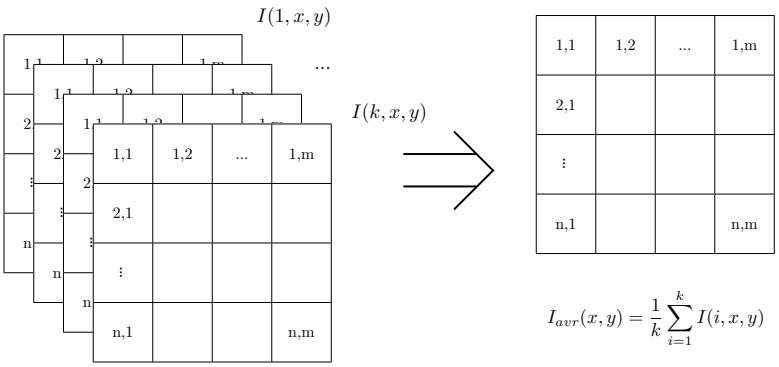


Figure 5.11: An illustration of the multiple frames average function used in SMATool.

Filtering

Filtering used by the software is of FIR type and is determined by a set of input parameters. In the software one or two filter parameters can be set and adjusted for each filter the parameters are named (1) and (2).

In the ETT part of the software all types of filters can be added and used as the user wants by clicking the filtering button in the mail window. The filters used in the ETT part are all, except for the Savitzky-Golay, filters of two dimensional type. The Savitzky-Golay filter considers each row of the current frame separately.

In ATT part of the software only the Gaussian, the averaging and the Savitzky-Golay filter types can be used, and the filters in the ATT part are all of one dimensional type. The Gaussian and the averaging filters creates a zero-phased filtering by filtering the data in both the forward and the reversed direction, using a direct form II transposed implementation of the filter algorithm. Using a zero-phased filter is convenient eliminates phase shifts in the data after filtering.

Simple Average

Simple average filtering is the most simple filter type used by the software. For the two dimensional filtering the average function filters the data by computing a full two dimensional convolution of the data with a averaging mask $\{\vec{c}: c_{ij} = \frac{1}{nm} \quad \forall \quad i,j \in [1..n, 1..m]\}$, where n is the height and m is the width of the averaging mask. n and m corresponds to the two adjustable parameters for the averaging filter.

For the one dimensional filtering the software computes one filtering in each direction using a direct II transposed implementation of the filter algorithm. The parameter set in the software for the oner dimensional filter is the length of the averaging window.

Disk

The disk filtering uses a circular averaging mask within a square matrix. The mask has the side length $n = 2r + 1$, where r is the radius of the circular mask and corresponds to the parameter value set in the software.

Symmetrical Gaussian

The general two dimensional Gaussian low-pass filter convolutes the data with the mask function

$$c_{ij} = c_0 e^{-\left(\frac{i-\frac{n}{2}}{2\sigma_i} + \frac{j-\frac{m}{2}}{2\sigma_j}\right)}, \tag{5.2}$$

where $i = 1..n$ is the height and $j = 1...m$ is the width of the mask, and c_0 is a constant term that sums the mask to zero. In the symmetrical Gaussian, which is implemented by the software in ETT, $\sigma_i = \sigma_j = \sigma$ and $n = m = k$, corresponding to the two adjustable parameters set in the software.

In the one dimensional Gaussian filter eq. 5.2 becomes

$$c_i = c_0 e^{-\left(\frac{i-\frac{n}{2}}{2\sigma}\right)}, \tag{5.3}$$

where $i = 1..n$ is the width of the filter window. The one dimensional filter filters the data in both directions using the direct II transposed form. The two parameters σ and n is set in the software to determine the filter strength and window length.

Wiener

The Wiener filter is a low-pass filter that uses a pixel-wise adaptive Wiener method on a neighborhood of size $n \times m$. The filter estimates the local mean μ and variance σ around each pixel in the neighborhood and computes the filter with the function

$$I_{est,ij} = \mu + \frac{\sigma^2 - v^2}{\sigma^2}(I_{ij} - \mu), \tag{5.4}$$

where I_{ij} is the data pixel computed and v is the noise variance given by the variance of all local estimated variances.

Savitzky-Golay

The Savitzky-Golay[31] filter computers a local polynomial regression, of degree k, on a local filtering window of size $n \geq k + 1$. k and n corresponds the pre parameters set in the software. The filter is good at preserving local features such as local min/max.

Thresholds

Thresholds cut intensity values at specific values eliminating noise outside the specified intensity value. The threshold function is effective when choosing values representing the channel, using thresholds instantly eliminates data values outside the relevant intensity region where the molecule can be observed in the data and reduces the amount of data considered for the channel model fitting.

5.6.2 Find Channel

When the ROI is set and the *Find Channel*-button is pressed the software selects the relevant channel values and fits the channel model to the selected channel values, see Section 4.2.2 and 4.2.3 for details on different approaches. The software selects the channel values by using the filtered frame displayed in the ETT window and selecting all values within the threshold, see section 5.3.3 for setting the threshold. The selected values is then passed to the fitting algorithm, which fits the selected values to a linear channel model using a iteratively re-weighted least squares model, see Section 4.2.3.

5.6.3 Extracting Time Trace

The TT is always selected by the software using the NPV method described in Section 4.3.1. For details on the method se previous section. With the NPV the values are extracted from edge to edge of the ROI, making the width of the TT equal to the width of the ROI.

5.6.4 Pulse fitting to detect molecule edges

When the model pulse is fitted the the intensity data in the TME window a pulse model of erf-functions are fitted to the intensity profile of the current frame using a IRLS algorithm, see Section 4.2.3 for details on IRLS and Section 4.4.2 for details on pulse model construction and fitting.

The parameter values are in the SMAT software always bound to within the unity interval, which corresponds to the total intensity interval in the TT current TT.

5.6.5 Trace Features

The algorithm for tracing features through the selected frames interval uses a set of features equally spaced along the frame length and calculates the minimum distance to all local peak features in the frame. The algorithm then stretches the features to the nearest peak feature, keeping the sequence of each feature, see fig. 5.12.

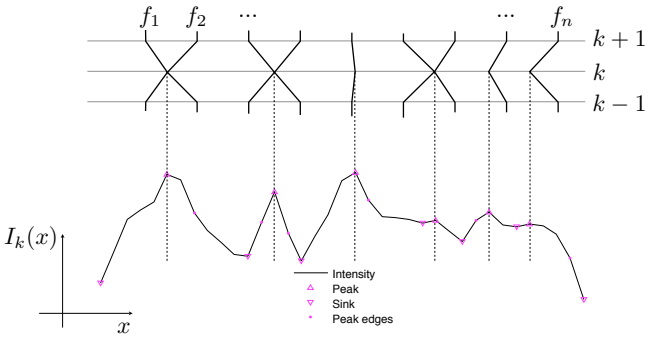


Figure 5.12: For each frame $k \in [1, K]$ in the selected frame interval all features $\{f_i\}_{n=1}^{N}$ are adjusted from equally spaced along the frame to the nearest peak feature in the frame. The algorithm keeps the sequence in the traced features.

Bibliography

[1] Inc. © 1984-2010-The MathWorks. Nonlinear regression - matlab.

[2] S. Altschul and B. Erickson. Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, 48:603–616, 1986. 10.1007/BF02462326.

[3] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 149(2):297–326, 1979.

[4] F. Brochard and P. G. de Gennes. Dynamics of confined polymer chains. *Journal of Chemical Physics*, 67(1):52–56, 1977.

[5] J. C. Brown, A. Hodgins-Davis, and P. J. O. Miller. Classification of vocalizations of killer whales using dynamic time warping. *The Journal of the Acoustical Society of America*, 119(3):EL34–EL40, 2006.

[6] T. W. Burkhardt. Free energy of a semiflexible polymer in a tube and statistics of a randomly-accelerated particle. *Journal of Physics A: Mathematical and General*, 30(7):L167, 1997.

[7] H.-D. Cheng and K.-S. Fu. VLSI architecture for dynamic time-warp recognition of handwritten symbols. In *Acoustics, Speech and Signal Processing, IEEE Transactions on*, volume 34, pages 603–613, 1986.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV (2)*, pages 484–498, 1998.

[9] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, (66):38–59, 1995.

[10] R. Dahm. Discovering DNA: Friedrich miescher and the early years of nucleic acid research. *Journal of Human Genetics*, 122(6):565–581, January 2008.

[11] M. Daoud and P. G. de Gennes. Statistics of macromolecular solutions trapped in small pores. *Le Lournal De Physique*, 38(1):85–93, 1977.

[12] P. G. de Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press, 1st ed. edition, November 1979.

[13] S. Dreyfus. Richard Bellman on the Birth of Dynamic Programming. *Operations Research*, 50(1):48–51, 2002.

[14] A. W. Dwonie. Pneumococcal transformation – a backward view. fourth griffith memorial lecture. *Journal of General Microbology*, 73:1–11, 1972.

[15] P. J. Flory. Spatial configuration of macromolecular chains. *Nobel Lecture, Chemistry 1971-1980*, pages 156–177, 1974.

[16] O. Gotoh. An improved algorithm for mathching biological sequences. *Journal of Molecular Biology*, 162(3):705–708, 1982.

[17] L. J. Guo, X. Cheng, and C.-F. Chou. Fabrication of size-controllable nanofluidic channels by nanoimprinting and its application for dna stretching. *Nano Letters*, 4(1):69–73, 2004.

[18] S. Chiba H. Saoke. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[19] R. Jayadevan, R. K. Satish, and M. P. Pradeep. Dynamic time warping based static hand printed signature verification. *Journal of Pattern Recognition Research*, 4(1):52–65, 2009.

[20] D. Lemire. Faster retrieval with a two-pass dynamic-time-warping lower bound. *ArXiv e-prints*, November 2008.

[21] M. G. Lorenz and W. Wackernagel. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58(3):563–602, 1994.

[22] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.

[23] T. Odijk. On the statistics and dynamics of confined or entangled stiff polymers. *Macromolecules*, 16(8):1340–1344, 08 1983.

[24] T. Odijk. Similarity applied to the statistics of confined stiff polymers. *Macromolecules*, 17(3):502–503, 1984.

[25] T. Odijk. Scaling theory of dna confined in nanochannels and nanoslits. *Physical Review E*, 77(6):060901, 2008.

[26] F. Persson. *Nanofluidics for Single Molecule Biophysics*. PhD thesis, DTU Nanotech, 2009.

[27] W. Reisner, N. B. Larsen, A. Silahtaroglu, A. Kristensen, N. Tommerup, J. O. Tegenfeldt, and H. Flyvbjerg. Single-molecule denaturation mapping of dna in nanofluidic channels. *Proceedings of the National Academy of Sciences*, 2010.

[28] W. Reisner, K. J Morton, R. Riehn, Y. M. Wang, Z. Yu, M. Rosen, J. C. Sturm, S. Y. Chou, E. Frey, and R. H. Austin. Statics and dynamics of single dna molecules confined in nanochannels. *Physical Review Letters*, 94(19):196101, May 2005.

[29] M. Rubinstein and R. H. Colby. *Polymer Physics*. Oxford University Press, USA, 2003.

[30] D. Sankoff. Matching Sequences under Deletion/Insertion Constraints. *Proceedings of the National Academy of Sciences of the United States of America*, 69(1):4–6, 1972.

[31] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 07 1964.

[32] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

[33] L. Wolf T. Serre and T. Poggio. Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition (CVPR 2005)*, 2005.

[34] J. O. Tegenfeldt, C. Prinz, H. Cao, S. Chou, W. W. Reisner, R. Riehn, Y. M. Wang, E. C. Cox, J. C. Sturm, P. Silberzan, and R. H. Austin. The dynamics of genomic-length dna molecules the dynamics of genomic-length dna molecules in 100-nm channels. *PNAS*, 101(30):10979–10983, 2004.

[35] J.-C. Terrillon, H. Fukamachi, S. Akamatsu, and M. N. Shirazi. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, FG '00, pages 54–63, Washington, DC, USA, 2000. IEEE Computer Society.

[36] F. Westerlund, F. Persson, A. Kristensen, and J. O. Tegenfeldt. Fluorescence enhancement of single dna molecules confined in si/sio2 nanochannels. *Lab on a Chip*, 10(16):2049–2051, 2010.