Optimal Estimation of Diffusion Coefficients from Noisy Time-Lapse-Recorded Single-Particle Trajectories

Vestergaard, Christian L.; Flyvbjerg, Henrik

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Optimal Estimation of Diffusion Coefficients from Noisy Time-Lapse-Recorded Single-Particle Trajectories

Christian L. Vestergaard

Kongens Lyngby 2012
DTU Nanotech

Department of Micro- and Nanotechnology
Technical University of Denmark
Building 345B, DK-2800 Kongens Lyngby, Denmark
Phone +45 4525 5700, Fax +45 4588 7762
info@nanotech.dtu.dk
www.nanotech.dtu.dk

DTU Nanotech-PhD-2012

Summary (English)

Optimal Estimation of Diffusion Coefficients from Noisy Time-Lapse-Measurements of Single-Particle Trajectories

Single-particle tracking techniques allow quantitative measurements of diffusion at the single-molecule level. Recorded time-series are mostly short and contain considerable measurement noise. The standard method for estimating diffusion coefficients from single-particle trajectories is based on least-squares fitting to the experimentally measured mean square displacements. This method is highly inefficient, since it ignores the high correlations inherent in these. We derive the exact maximum likelihood estimator for the diffusion coefficient, valid for short time-series, along with an exact benchmark for the maximum precision attainable with any unbiased estimator, the Cramér-Rao bound. We propose a simple analytical and unbiased covariance-based estimator based on the autocovariance function and derive an exact analytical expression of its moment generating function. We find that the maximum likelihood estimator exceeds the precision set by the Cramér-Rao bound, but at the cost of a small bias, while the covariance-based estimator, which is born unbiased, is almost optimal for all experimentally relevant parameter values. We extend the methods to particles diffusing on a fluctuating substrate, e.g., flexible or semiflexible polymers such as DNA, and show that fluctuations induce an important bias in the estimates of diffusion coefficients if they are not accounted for. We apply the methods to obtain precise estimates of diffusion coefficients of hOgg1 repair proteins diffusing on stretched fluctuating DNA from data previously analyzed using a suboptimal method. Our analysis shows that the proteins have different effective diffusion coefficients and that their diffusion coefficients are correlated with their residence time on DNA. These results imply a multi-state model for hOgg1's diffusion on DNA.

Summary (Danish)

Optimal estimering af diffusionskoefficienter fra tidsseriemålinger af enkeltpartikeltrajektorier med målestøj

Enkeltpartikelteknikker gør det muligt kvantitativt at følge enkelte molekylers diffusion. Tidsserier målt med disse teknikker er overvejende korte og indeholder meget målestøj. Standardmetoden for estimation af diffusionskoefficienter fra enkeltpartikeltrajektorier er baseret på mindste kvadraters fit til eksperimentelt målte *mean squared displacements*. Denne metode er højst ineffektiv, da den ignorerer de høje korrelationer mellem disse. Vi udleder den eksakte maksimum *likelihood* estimator for diffusionskoefficienten, sammen med et eksakt benchmark for den maksimale præcision nogen middelret estimator kan opnå, Cramér-Rao grænsen. Vi foreslår en simpel analytisk og middelret covariansbaseret estimator, baseret på autocovariansfunktionen, og udleder et eksakt analystisk udtryk for dens momentgenererende funktion. Vi finder at maksimum likelihood estimatorens præcision er højere end Cramér-Rao grænsen, på bekostning af en lille skævhed, mens den covariansbaserede estimator, som er født middelret, er optimal for alle eksperimentelt relevante parameterværdier. Vi udvider metoderne til partikler, som diffunderer på et fluktuerende substrat, såsom DNA, og viser at hvis der ikke tages højde for fluktuationerne, inducerer de en skævhed i de estimerede diffusionskoefficienter. Vi anvender metoderne til præcis estimering af diffusionskoefficienter for hOgg1 proteiner, som diffunderer på strukket fluktuerende DNA, udfra data, som tidligere er blevet analyseret med en suboptimal metode. Vores analyse viser at proteinerne har forskellige effektive diffusionskoefficienter og at deres diffusionskoefficienter er korrelerede med deres residenstider på DNA'et. Disse resultater peger på en multitilstandsmodel for hOgg1's diffusion på DNA.

Preface

Ph.D. thesis

Optimal Estimation of Diffusion Coefficients from Noisy Time-Lapse-Recorded Single-Particle Trajectories

This thesis has been written as a partial fulfillment of the requirements for obtaining a PhD degree at the Technical University of Denmark (DTU). The PhD project has been conducted in the period of December, 2008 to March, 2012 at

Department of Micro- and Nanotechnology, DTU Nanotech
Technical University of Denmark
DK-2800 Kongens Lyngby
Denmark

under supervision of

Henrik Flyvbjerg, associate professor, dr. scient.
Department of Micro- and Nanotechnology, DTU Nanotech
Technical University of Denmark
DK-2800 Kongens Lyngby
Denmark

This work was funded 1/3 by Rigshospitalet, and 2/3 jointly by the Danish Agency for Science, Technology and Innovation (Forsknings- og innovationsstyrelsen) and DTU Nanotech.

Christian L. Vestergaard

Acknowledgements

I would like to thank Paul Blainey, Broad Institute, MA, for providing experimental data, insightful discussions, and for pointing out important details related to collection of real data that a theoretician tends to forget. I wish to thank Christian Gradinaru for his collaboration in deriving statistical properties of the mean squared displacements and mean squared displacement-based estimators, and my collegues Kim Mortensen and Jonas N. Pedersen for numerous discussions, on- and off-topic, and for critical lecture of this manuscript. Last, but not least I wish to thank my supervisor Henrik Flyvbjerg for invaluable advice, guidance and countless late hours spent in front of the blackboard.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Diffusion is ubiquitous in biology and many cellular processes rely on diffusion as a passive means of transport. Quantitative knowledge of the diffusion coefficient is paramount for the precise understanding of these processes. Recent developments in fluorescent labels have made it possible to track diffusion of single molecules, e.g., proteins on biopolymers such as DNA [1, 2, 3] or microtubules [4, 5, 6], on surfaces [7], in lipid membranes [8, 9, 10] and inside cells [11, 12, 13], with time-lapse photography. Data mostly consist of relatively short time series with considerable experimental localization error. This makes it a challenge to determine diffusion coefficients. This challenge is even higher when individuality of diffusion coefficients is a concern, since one then cannot average over multiple trajectories of different molecules to reduce statistical error. The standard approach relies on Einstein's classic result for the mean squared displacement (MSD) of a particle undergoing free diffusion. It estimates the diffusion coefficient and measurement noise by fitting a straight line to experimental values of the MSD [14]. However, even though the MSD method gives the right value on average, it does not mean that it is a good way to estimate the diffusion coefficient. Its precision depends on the number of points used in the fit [15] and for good signal-to-noise ratio (SNR) the precision actually *decreases* the more points we use in the fit [8, 16]. One can improve the MSD estimator by trying to choose the optimal number of points to include in the fit [15]. A more rigorous method, generalized least squares (GLS), takes into account the correlations between the experimental values of the MSDs in the fitting procedure. The GLS estimator's non-linear dependence on the parameters of interest, the diffusion coefficient D and the variance σ^2 of the noise on position determination, means that results about the optimality of the GLS estimator for linear dependence on parameters [17] do not hold. The complicated dependence of the MSDs on data makes it more

than difficult to derive a maximum likelihood estimator based on the MSDs. We can do better than that, however, with a simpler approach: We calculate the probability distribution of the individual displacements of the diffusing particle during a time-lapse [18] and use spectral decomposition to decorrelate the measurements and construct a maximum likelihood estimator (MLE) for D and σ^2. This estimator is known to be asymptotically optimal in the long time-series limit, i.e., it is unbiased and reaches the lower limit on the variance of any unbiased estimator, set by information theory and known as the Cramér-Rao bound [17]. In this limit, the spectral decomposition reduces to the Fourier transform and the MLE to a maximum likelihood fit to the power spectrum [18].

However, for some systems it is difficult or impossible to obtain long time-series. Most experiments with individual biological molecules are limited by fluorophore lifetimes, proteins diffuse out of the field-of-view in confocal microscopy, and proteins, which diffuse on biopolymers, detach. Data predominantly consist of short time-series, where optimality of the MLE is not guaranteed. In this range we find that a simple covariance-based estimator (CVE) is generally to be preferred, since it practically reaches the Cramér-Rao bound *and* is unbiased, whereas the more complicated MLE and GLS estimator are biased. When the amplitude of the positional noise is known a priori, this information can be used to considerably increase the precision of diffusion coefficient estimates. In this case, the MLE and GLS estimator are both unbiased, and the MLE, CVE, and GLS estimator all practically reach the Cramér-Rao bound.

For diffusion on many cellular structures, the recorded movement also contain a fluctuating term due to thermal motion of the substrate, e.g., for diffusion on DNA or in lipid membranes. If the time-scale of the fluctuations is much shorter than the time-lapse, these fluctuations will only contribute to the movement as a constant-amplitude noise term, which can be absorbed in the positional noise amplitude σ. If the fluctuation time-scale is comparable to the time-lapse or longer, however, these fluctuations need to be taken into account. We extend our methods to diffusion on fluctuating substrates and derive a MLE, which explicitly accounts for substrate fluctuations and is optimal for long time-series. We derive an expression for the bias of the CVE for diffusion on a fluctuating substrate and show how this can be used to obtain unbiased estimates of diffusion coefficients for short time-series, where the MLE fails.

We estimate diffusion coefficients of fluorescently marked hOgg1 repair proteins on DNA from time-lapse measurements. The data has previously been analyzed using MSD-based methods [19]. We measure diffusion coefficients in the range 0.1 μm^2/s-0.5 μm^2/s. We show that the DNA fluctuations induce a bias in the estimates of diffusion coefficients of up to 0.2 μm^2/s, i.e, we overestimate diffusion coefficients by up to 200%, if the fluctuations are not taken into account. The increased resolution our methods offer, allows us to inves-

tigate diffusion coefficients at the single molecule level. Analysis of the data shows that the identical hOgg1 proteins have different diffusion coefficients and that protein residence times on DNA follow a non-trivial distribution. Furthermore, the individual proteins' diffusion coefficients are correlated with the proteins' residence times on DNA. These results suggest a two-state model for diffusion on DNA as proposed in [20, 21].

The diffusion coefficient is the parameter of interest, since it characterizes the physical system, while the measurement noise only describes the experiment. Hence, the focus in this thesis is on the performance of the different methods in estimating the diffusion coefficient. The methods presented here estimate the noise as well, however, and their performance at that is briefly addressed in the results section (Chap. 2).

Chapter 2

Results

In this chapter the main results of the thesis are presented. The chapter is divided into three sections: (i) Estimation of diffusion coefficients from single-particle tracking (SPT) measurements of a particle undergoing free diffusion, i.e., Brownian motion. Results in this section are also applicable to measurements of a particle diffusing on a stiff substrate such as a microtubule or an actin filament. (ii) Estimation of diffusion coefficients of a particle diffusing on a fluctuating substrate, e.g., a stretched flexible or semi-flexible polymer, or a supported or tethered lipid bilayer. (iii) Application of the methods developed in the preceding chapters to experimental SPT measurements of hOgg1 repair proteins diffusing on flow-stretched DNA.

2.1 Diffusion coefficient estimation for free diffusion

In this section we review existing methods for estimating diffusion coefficients from SPT measurements of freely diffusing particles. We derive exact expressions for the likelihood function of measured displacements and the Cramér-Rao lower bound, along with a simple and unbiased covariance-based estimator (CVE).

In Sec. 2.1.1 we present some basic properties of the mean squared displacement (MSD) along with the generalized least-squares (GLS) estimator and compare it to existing MSD-based estimators. In Sec. 2.1.2 we derive simple expressions for the exact likelihood function and Cramér-Rao bound, give a computationally efficient algorithm for the maximum likelihood estimator (MLE), and compare this to the approximate MLE given in [18]. We show

that motion blur decreases estimator precision. We propose the CVE and give exact analytical expressions for its variance and characteristic function. Section 2.1.3 compares the precisions of the MLE and CVE on Monte Carlo generated data.

2.1.1 Mean squared displacement-based methods

We review the statistical properties of the MSD and existing MSD-based estimators of the diffusion coefficient. We also introduce the GLS estimator, which properly accounts for the correlations in the MSD.

MSD-based estimation methods are all based on Einstein's classic result for Brownian motion in an isotropic medium, derived in his seminal 1905 paper on the molecular theory of heat [22]. Adding the effects of positional noise, due to, e.g., the limited number of photons emitted by the fluorophore, and motion blur, due to the finite camera shutter-time, we find that the expected squared displacement of a Brownian particle with diffusion coefficient D diffusing in d dimensions during a time-interval t is

$$\left\langle (\mathbf{x}(t) - \mathbf{x}(0))^2 \right\rangle = 2dDt + 2(\sigma^2 - 2RdD\Delta t) \ , \tag{2.1.1}$$

where σ is the amplitude of the positional noise and $2R\Delta t$ is the motion blur [18] (App. A.1). The motion blur coefficient $R \in]0, 1/4[$ is determined by the particular shutter mechanism of the camera and is defined in App. A.1.2. For instantaneous camera shutter $R = 0$, while $R = 1/6$ for full time-integration, where the camera shutter is open for the full time-lapse. Since diffusion is a scale-free process, the noise amplitude σ sets the scale, while the performance of estimators (for given time-series length N and motion blur $2RD\Delta t$) is determined by the ratio between diffusion and noise, which we define as the signal-to-noise ratio (SNR),

$$\mathrm{SNR} \equiv \frac{\sqrt{D\Delta t}}{\sigma} \ , \tag{2.1.2}$$

where Δt is the time-lapse between successive measurements. In practical applications we typically want to keep a signal-to-noise ratio larger than one to be sure that we actually record what we expect and avoid spurious behavior, especially when recorded time-series are short.

For diffusion in an isotropic medium, a particle diffusing in d dimensions is equivalent to d independent particles diffusing in one dimension. So we assume from now on that $d = 1$ and note that all results can be generalized to higher dimensions. The MSD method consists of least-squares fitting a straight line to the experimentally measured MSDs, identifying its slope with $2Dt$ and its offset with $2\sigma^2 - 4RD\Delta t$. The experimental MSDs are usually estimated by

$$\rho_n = \frac{1}{N - n + 1} \sum_{i=0}^{N-n} (x_{i+n} - x_i)^2 \ . \tag{2.1.3}$$

The averaging is done over all possible time-separations $n\Delta t$ to obtain maximal information content in the individual estimate ρ_n. Since all ρ_ns are calculated from the same time-series $\{x_0, x_1, \ldots, x_N\}$, they are highly correlated (Fig. 2.1a). Since all ρ_ns are calculated from the same time-series $\{x_0, x_1, \ldots, x_N\}$, they are highly correlated (Fig. 2.1a).

Ignoring the correlations between the ρ_ns leads to an estimator that performs poorly, i.e., an estimator with a precision several times lower than the Cramér-Rao bound. For a high SNR (SNR > 2) ordinary and weighted least squares fits (OLS/WLS) actually perform *worse* the *more* points are fitted! (Fig. 2.1b.) This result is counterintuitive if one thinks of more points as more information. The points (n, ρ_n) are *not* independent data points, however, as they all are based on the same information, the time-series $\{x_i\}_{i=0,1,\ldots,N}$. One should think of (n, ρ_n), $n = 1, 2, \ldots, n_{\max}$ as a committee of $n_{\max}$ members, who all have access to the same information, but treat it differently, and less reliably for larger n. A fit to $n_{\max}$ data points corresponds to a vote by the committee. Unweighted fitting (OLS) is most democratic and the vote clearly gives a worse result the larger the number $n_{\max}$ of members in the committee is. Weighted fitting (WLS) gives more votes to members, which are known to be more reliable, but even here the result of the vote gets worse the larger $n_{\max}$ is, since the members influence each other.

For small SNR, some intermediate number of points $n_{\max}$ can be used to obtain a reasonable fit (Fig. 2.1c). This problem has traditionally been addressed by fitting to a given number of points $\{\rho_1, \rho_2, \ldots, \rho_{n_{\max}}\}$, where $n_{\max}$ is chosen ad hoc (values for $n_{\max}$ ranging from 2 to $N/2$ have been reported [7, 19, 23, 24, 25, 26, 27, 28, 10]). This approach is clearly far from optimal and the question of how to choose $n_{\max}$ has recently been addressed [15]. Also, a least-squares method that deals correctly with correlations already exists in the statistical literature. It is known as the generalized least squares (GLS) estimator and is defined in App. A.2. If the covariance matrix of $\boldsymbol{\rho}$, Σ_ρ, is independent of D and σ^2, the GLS estimator is the best linear unbiased estimator (BLUE) [17] and is thus guaranteed to outperform all other linear estimators based on $\boldsymbol{\rho}$, such as OLS and WLS, no matter how many points are included in the fit. The linear GLS estimator has lower variance than the OLS and WLS estimators, and it practically reaches the Cramér-Rao bound (Fig. 2.1b,c). For diffusion with positional noise, no Σ_ρ independent of D and σ^2 exists, and the GLS estimator should be found using an iterative relaxation algorithm as described in [29]. This iterative GLS is equal to the MLE based on $\boldsymbol{\rho}$ under the assumption that $\boldsymbol{\rho}$ is Gaussian distributed [30]. This assumption is not correct, however, which means that the iterative GLS is biased and its variance is significantly higher than predicted theoretically, see Fig. 2.1d.

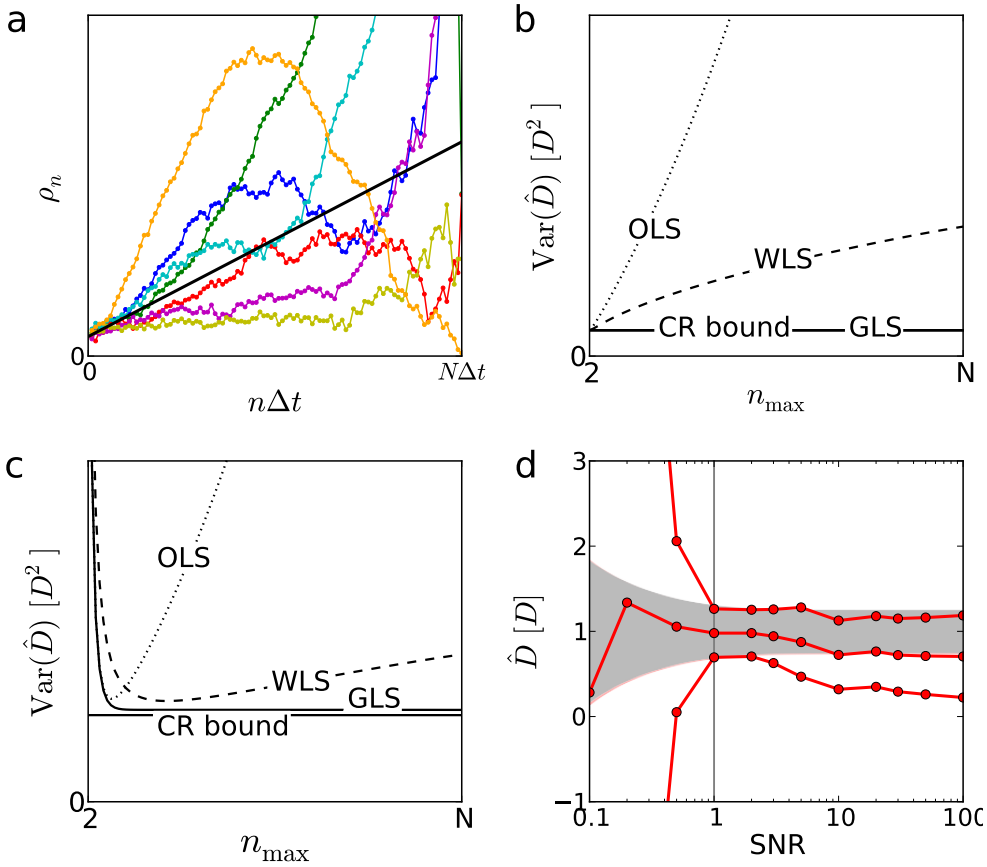

Figure 2.1: a) Experimental MSDs, $\{\rho_n\}_{n=1,\ldots,N}$, calculated from simulated Brownian motion trajectories, compared to their expected value. (Theoretical MSD: full line, black; experimental MSDs: colored points.) The experimental MSDs are highly correlated and their variance increases with time-displacement $n\Delta t$. b) Variance of the MSD estimate of the diffusion coefficient D as a function of the number of MSD points used in the fit for SNR = 10. (OLS fit to the experimental MSDs: dotted line; WLS fit: dashed line; GLS estimator, which attains the Cramér-Rao (CR) bound: full line.) c) Variance of the MSD estimate of D as a function of the number of MSD points used in the fit for SNR = 1/3. d) Mean plus/minus standard error of the GLS estimator. (Numerical results for the iterative GLS: circles, red; Cramér-Rao bound: grey area.) The GLS estimator is biased and does not reach the Cramér-Rao bound in practice, except for SNR ~ 1.

2.1.2 Methods based on individual displacements

We derive methods based on the set of measured displacements of a freely diffusing particle, $\mathbf{\Delta x} = (\Delta x_1, \Delta x_2, \ldots, \Delta x_N)^T$, where $\Delta x_n = x_n - x_{n-1}$ is the displacement during one time-lapse Δt. Since diffusion is translationally invariant, the displacements are, contrary to $\boldsymbol{\rho}$, a sufficient statistic, which means that they contain all the relevant information available in the experimental measurements $\mathbf{x}$. The displacements are Gaussian distributed with mean zero and covariance matrix $\Sigma_{\Delta x}$ [18] (App. A.1),

$$(\Sigma_{\Delta x})_{ij} = [2D\Delta t(1-2R) + 2\sigma^2]\delta_{i,j} + [2RD\Delta t - \sigma^2]\delta_{i,j\pm 1} \ . \tag{2.1.4}$$

2.1.2.1 The exact likelihood function and maximum likelihood estimator

The maximum likelihood estimator (MLE) of the parameters D and σ^2 is the set of values which maximizes the likelihood function $\mathcal{L}(D, \sigma^2|\mathbf{\Delta x}) = p(\mathbf{\Delta x}|D, \sigma^2)$ given the measured displacements $\mathbf{\Delta x}$, where $p(\mathbf{\Delta x}|D, \sigma^2)$ is the probability density of $\mathbf{\Delta x}$ when D and σ^2 are given.

We use that $\mathbf{\Delta x}$ can be transformed to a set of independent variables $\widetilde{\mathbf{\Delta x}} = \mathbf{P}^{-1}\mathbf{\Delta x}$ using an orthogonal transformation $\mathbf{P}$ (App. A.3.1). The transformation matrix $\mathbf{P}$ is given in Sec. 3.1. The logarithm of the likelihood function is then reduced to a sum over independent entries. This allows for computationally efficient maximum likelihood estimation and calculation of the Cramér-Rao lower bound, which bounds the variance of any unbiased estimator and approximately gives the variance of the MLE. The Cramér-Rao bound is defined as the inverse of the Fisher information matrix $\mathcal{I}$, given in Sec. 3.1.

If one ignores boundary terms in $\Sigma_{\Delta x}$, which are of order $1/N$, the transformation $\mathbf{P}^{-1}$ reduces to the discrete Fourier transform, and we recover the log-likelihood, MLE, and information matrix given in [18]. This approach is valid for long time-series, i.e, $N \gg 1$. When the boundary terms cannot safely be ignored the estimators differ. A comparison of the performance of the two estimators on Monte Carlo generated data for different SNR and time-series lengths N is shown in Fig. A.1. The numerical analysis gives two important results: (i) The MLE is biased (Figs. 2.2 and A.1). This bias arises because we forbid the estimates of D and σ^2 to take physically meaningless negative values in order to avoid numerical problems in the optimization algorithm (App. A.3.1). (ii) For large signal-to-noise ratio (SNR > 1) the terms neglected in the approximate MLE based on the Fourier transform are unimportant and the estimator practically attains the precision of the exact MLE, even for small N. On the other hand, the approximate MLE performs much worse than the exact MLE for low SNR, even for relatively large N, i.e., the approximate MLE only converges slowly to the exact MLE in this regime (Fig. A.1). For

all parameter values, the exact MLE is more precise than its approximation. So in the remainder of this section we only consider the exact MLE.

2.1.2.2 Covariance-based estimator

A simple alternative to the MLE and MSD estimators can be derived directly from the covariance matrix given by Eq. (2.1.4). By combining unbiased estimators for the two first values of the autocovariance of Δx_n (the only, which are non-zero) we find unbiased estimators of D and σ^2 (App. A.3.2). The covariance-based estimator (CVE) is given in Sec. 3.1 along with its variance and characteristic function. The characteristic function gives all higher moments of the estimator and its exact distribution by numerical Fourier transformation and can thus be used to calculate exact confidence intervals for the covariance-based estimates of D. Examples of the distribution of the CVE are shown in Fig. A.3b. The CVE has been proposed previously in literature [31, 32], but its properties have not previously been derived. In [31] it was proposed as a maximum likelihood estimator based on the faulty assumption that measured displacements are uncorrelated; and in [32] it was used as part of a bootstrap-based estimator. The CVE is practically optimal for experimentally relevant values of the parameters, i.e., when the SNR is larger than one (Fig. A.3a). The question of when the CVE is optimal and situations where the MLE should be preferred is addressed in detail in Sec. 2.1.3.

The CVE has several advantages over the MLE and MSD estimators: (i) It is given by a simple analytical expression and is thus orders of magnitude faster than the MLE and MSD estimators, which are only given implicitly and must be found by a numerical optimization algorithm; (ii) it is unbiased and its variance can be calculated exactly, this only holds asymptotically for the MLE and MSD estimators; (iii) exact confidence intervals for the estimates and their distributions can be found from the characteristic function of the CVE derived in App. A.3.2.2, while no such results can be found for the MLE and MSD estimators, except asymptotically, at $N \to \infty$.

2.1.2.3 Motion blur's effect on estimator precision

It has recently been shown that by increasing motion blur while keeping the SNR constant, one lowers the Cramér-Rao bound in the large N-limit [18]. This result is confirmed by our exact results for all values of N. Thus, one can theoretically make estimation more efficient by engineering the experiment to obtain maximum motion blur (Fig. A.3a). Motion blur does decrease the variance of the MLE and makes it less biased; the MLE becomes effectively unbiased for $N \gtrsim 20$ for maximal motion blur ($R = 1/4$). It is a rather surprising result that more information can be extracted from the measurements by increasing the noise, and it turns out that it is too good to be true. The result

relies on the implicit assumption that the motion blur can be increased independently of the positional noise. This assumption is not true, since motion blur increases the width of the measured point-spread function (PSF) emitted by the recorded particle and thus increases the positional error (App. A.3.1.4). This means that the Cramér-Rao bound cannot be lowered by increasing the motion blur (Fig. A.2b). On the contrary, motion blur tends to *increase* the Cramér-Rao bound and consequently the variance of estimates of the diffusion coefficient, though only by a negligible amount for most experimentally relevant signal-to-noise ratios.

2.1.3 Numerical results

We compare the precision of the estimators by using them on Monte Carlo generated data. In the first part of this section we treat the case when both D and σ^2 must be estimated from the time-series. In the second part we compare the precision of the estimators when σ^2 is known a priori, and only D is estimated from the time-series. Only data for full time-integration ($R = 1/6$) are shown, but results for other values of the motion blur coefficient are similar and do not change the conclusions.

2.1.3.1 Unknown noise amplitude

Estimation of the diffusion coefficient As shown in Sec. 2.1.1, the MSD-based estimators are sub-optimal, while the MLE and the CVE are close to optimal. The MLE and the CVE are compared in Fig. 2.2. The variance of the MLE is smaller than the variance of the CVE for all parameter values and actually violates the Cramér-Rao bound for high SNR. This is possible because the MLE is biased, which means that the total error of the MLE is smaller than that of any unbiased estimator, but it comes at a cost of a systematic error in the estimate. This complicates statistical analysis of estimates from multiple time-series, since averages and other statistics do not converge to their true values. The CVE is constructed to be unbiased, and, as Fig. 2.2 shows, it practically reaches the Cramér-Rao bound as long as the SNR is larger than one. In experiments the SNR typically lies in the interval from 2 to 20 [4, 5, 19, 25, 33] where the CVE is the optimal estimator of the diffusion coefficient.

Estimation of the noise amplitude Even though the positional noise is not a parameter of main interest, it reports experimental conditions, so one may want to estimate its amplitude, even if the noise can be estimated from the measured PSF of the fluorophore (Sec. 3.1.1), since it then provides an independent check of this estimate. Both the MLE and CVE provide approximately optimal estimates of σ^2. However, the Cramér-Rao bound grows as

a quadratic function of the SNR, and hence as $1/\sigma^2$, for a SNR larger than one. This means that even though the estimators are optimal, they are very imprecise when noise is low, and cannot in general be used to obtain a reliable estimate of σ^2. (The standard deviation of the estimator for σ^2 is larger than the value of σ^2 for a SNR larger than $\sqrt{N/10}$.) Note that when the noise is so low that it is difficult to estimate its amplitude precisely, it is also irrelevant to know the precise value of its amplitude.

2.1.3.2 Known noise amplitude

When the particle's position is estimated by fitting to the PSF, Eq. (3.1.17) gives an estimate of the amplitude of the positional noise σ, which can be inserted in the estimation routines for D instead of estimating both D and σ^2. This allows us to use all the information available in the time-series to estimate the diffusion coefficient D. As shown in Fig. 2.3, a more precise estimate of D can be obtained by using a priori knowledge of σ^2. When the error on the a priori estimate of σ^2 is negligible, the standard error of $\widehat{D}$ is reduced by a factor ≈ 1.5 for high SNR (SNR > 5) and for full time-integration ($R = 1/6$), and a factor ≈ 1.8 in the absence of motion blur ($R = 0$). We furthermore see that the MLE is unbiased when σ^2 is known a priori and that both the MLE and the CVE reach the Cramér-Rao lower bound for SNR > 1. For SNR < 1 the CVE is suboptimal, while MLE almost reaches the Cramér-Rao bound. When σ^2 is known a priori the GLS estimator performs much better than when σ^2 is unknown and almost reaches the precision of the MLE. It is considerably more complicated than the MLE and much more computationally demanding, due to the multiple matrix inversions and products performed in each time-step of the optimization algorithm. So the MLE or the CVE are always to be preferred.

In most experimental situations (where the SNR > 1) the MLE and the CVE are both optimal and the moment-based estimator should be preferred due to its simplicity and analytical tractability.

2.2 Diffusion coefficient estimation on a fluctuating substrate

In this section we develop methods for estimating diffusion coefficients from trajectories of single particles diffusing on a fluctuating substrate, i.e., a DNA molecule or another medium for which we can model its motion, rigorously or phenomenologically.

In Sec. 2.2.1 we present the theoretical framework, which allows us to model substrate fluctuations and diffusion on a fluctuating substrate. In Sec. 2.2.2

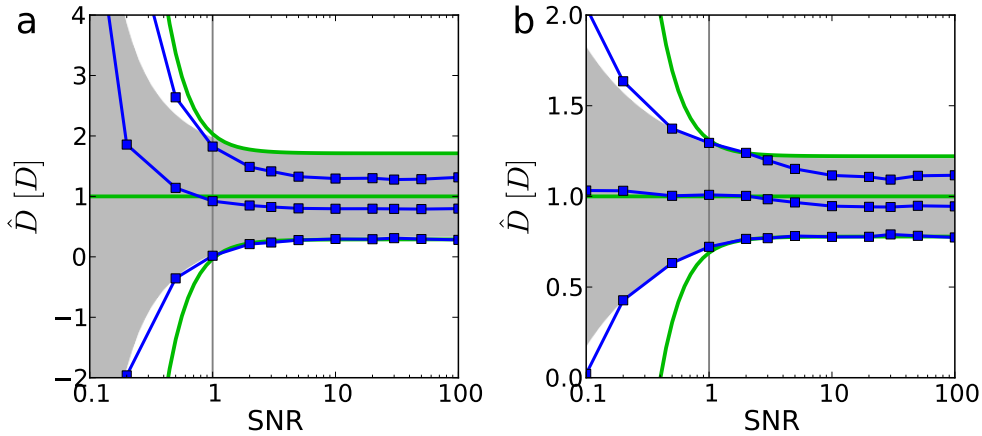

Figure 2.2: Mean plus/minus standard error of the MLE and CVE applied to an ensemble of 1,000 Monte Carlo generated time-series for unknown σ^2. (MLE: connected squares, blue; CVE: full lines, green; Cramér-Rao bound: grey area.) $R = 1/6$ for both simulations, which corresponds to full time-integration. a) Time-series length $N = 10$ and b) $N = 100$. The MLE reaches and even surpasses the Cramér-Rao bound for high SNR, where it is biased, and rapidly converges to the Cramér-Rao bound for SNR < 1. The CVE is unbiased and attains the Cramér-Rao bound for SNR > 1, while it is suboptimal for SNR < 1.

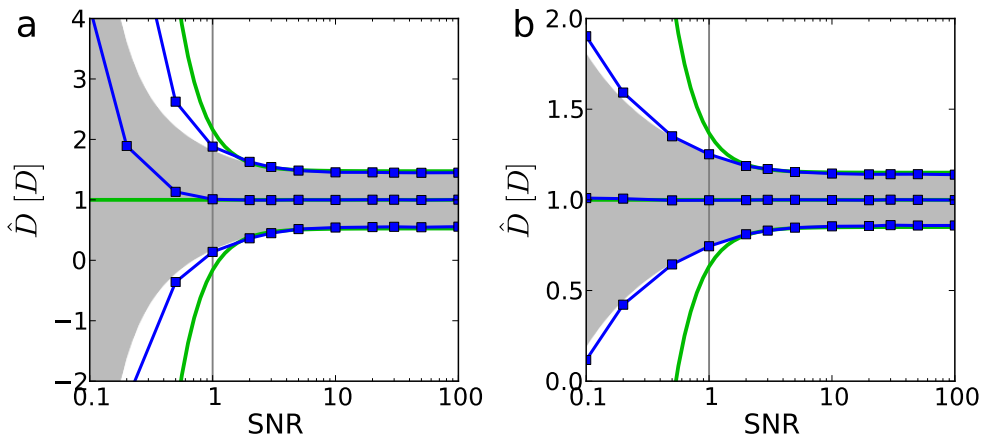

Figure 2.3: Mean plus/minus standard error of the MLE and CVE applied to an ensemble of 10,000 Monte Carlo generated time-series for a priori determined σ^2. (MLE: connected squares, blue; CVE: full lines, green; Cramér-Rao bound: grey area.) $R = 1/6$ for both simulations, which corresponds to full time-frame integration. a) Time-series length $N = 10$ and b) $N = 100$. For known σ^2 both the MLE and the CVE are unbiased and attain the Cramér-Rao bound for SNR > 1. For SNR < 1 the MLE rapidly converges to the Cramér-Rao bound, while the CVE is suboptimal.

we derive a MLE based on the power spectrum of the measured displacement and present an extension to the CVE derived in Sec. 2.1.2.2, which allows us to obtain unbiased estimates of diffusion coefficients for short time-series. Section 2.2.3 investigates the performance of the MLE and CVE on Monte Carlo generated data.

For experimental measurements of diffusion on cellular structures, one wants to use fluorophores that are as small as possible to avoid altering the diffusing particle's movement [34]. Since smaller fluorophores are typically less bright and bleach faster, photon economy is paramount and measurements are recorded at full time-integration, i.e., the camera shutter-time τ is equal to the time-lapse Δt. We thus only present results for full time-integration. Results are easily generalized to faster camera shutter (App. B.2).

2.2.1 Statistics of diffusion on a fluctuating substrate

A particle diffusing on a fluctuating substrate is not an isotropic system, contrary to a freely diffusing particle. The observed position of the particle will in general depend both on time t and the particle's position on the substrate, $s(t)$,

$$x(s(t), t) \quad . \tag{2.2.1}$$

For a flat substrate, e.g., a supported or tethered lipid bilayer, the observed motion is just the sum of the particle's diffusion and the substrate's movement. For a substrate that is not flat, e.g., a DNA molecule, the substrate's tendency to curl up means that the measured displacements of the particle are smaller than its actual displacements along the substrate.

We assume that the substrate can be described by a linear theory. The assumption of linearity is well substantiated both theoretically and experimentally for taut DNA pulled by the ends [35, 36, 37], while for DNA stretched by a laminar flow we expect the linear approximation to be correct on the upstream part of the DNA, where it is taut [38, 39]. This means that the substrate movement can be modeled as a sum of independent normal modes, of which only those corresponding to the lowest frequencies (longest wavelengths) contribute significantly to the measured movement (App. B.1).

Since in general the total diffusion length of the particle is small compared to the length-scale of the substrate, the substrate movement can be assumed uncorrelated with the diffusive motion of the particle, and the covariance of the measured displacements is (App. B.2)

$$\begin{aligned}(\Sigma_{\Delta x})_{i,j} \quad = \quad & \left[\frac{4}{3}\zeta(\overline{s})^2 D\Delta t + 2\sigma^2\right]\delta_{i,j} + \left[2\zeta(\overline{s})^2 D\Delta t - \sigma^2\right]\delta_{i,j\pm 1} \\ & +\overline{\rho}_{\Delta x}(\overline{s}, |i-j|\Delta t) \quad ,\end{aligned} \tag{2.2.2}$$

where $\overline{s}$ is the mean position of the particle in substrate coordinates, $\zeta = \sqrt{\langle(\partial x/\partial s)^2\rangle}$ is the local degree of stretching of the substrate and $\overline{\rho}_x$ is the autocovariance function between the substrate's displacements (App. B.2). The form of ζ and $\overline{\rho}_{\Delta x}$ depends on the specific substrate, and is given in Sec. 3.2 for a stretched flexible or semi-flexible polymer, such as DNA.

For diffusion on a fluctuating substrate, we cannot find an exact spectral decomposition, which completely decorrelates the measured displacements as we could for free diffusion, but we can use the Fourier transform, which decorrelates the data to a very good approximation (to order $1/N$). The fluctuations of the substrate contribute to the measured power spectrum with a sum of Lorentzian terms, one for each normal mode, where only the lowest one or two modes contribute in practice (Sec. 2.3.3 and App. B.1). The measured power spectrum is thus the sum of the diffusive movement, the positional noise, and the substrate movement (Figs. 2.4e,f).

2.2.2 Estimation

Using knowledge of the statistics derived in the previous section, we derive methods for estimating diffusion coefficients from the measured displacements of single particles diffusing on a fluctuating substrate. We present a MLE, which optimally estimates diffusion coefficients and parameters describing substrate fluctuations for long time-series (Sec. 2.2.2.1). If substrate fluctuations have been determined a priori, an unbiased CVE can be used, which is optimal for short time-series (Sec. 2.2.2.2).

2.2.2.1 Maximum likelihood estimation

We can estimate the diffusion coefficient, noise and parameters describing the substrate motion by maximum likelihood fitting to the power spectrum of the measured displacements of a diffusing particle (Fig. 2.4e,f). For long time-series this MLE is optimal (Sec. 2.2.3). For diffusion on a polymer, such as a DNA strand, we also record the transversal (y-direction) movement of the strand, which only consists of positional noise and substrate fluctuations (Figs. 2.4a-d). This movement is coupled to the longitudinal (x-direction) movement and thus provide additional information about the substrate movement and positional noise (App. B.1). Thus, both the y- and x-direction power spectra should be included in the MLE procedure (App. B.3.1). For time-series shorter than around $N = 35$ (depending on parameter values), the MLE algorithm may fail to converge. In this case an unbiased CVE should be used as described below.

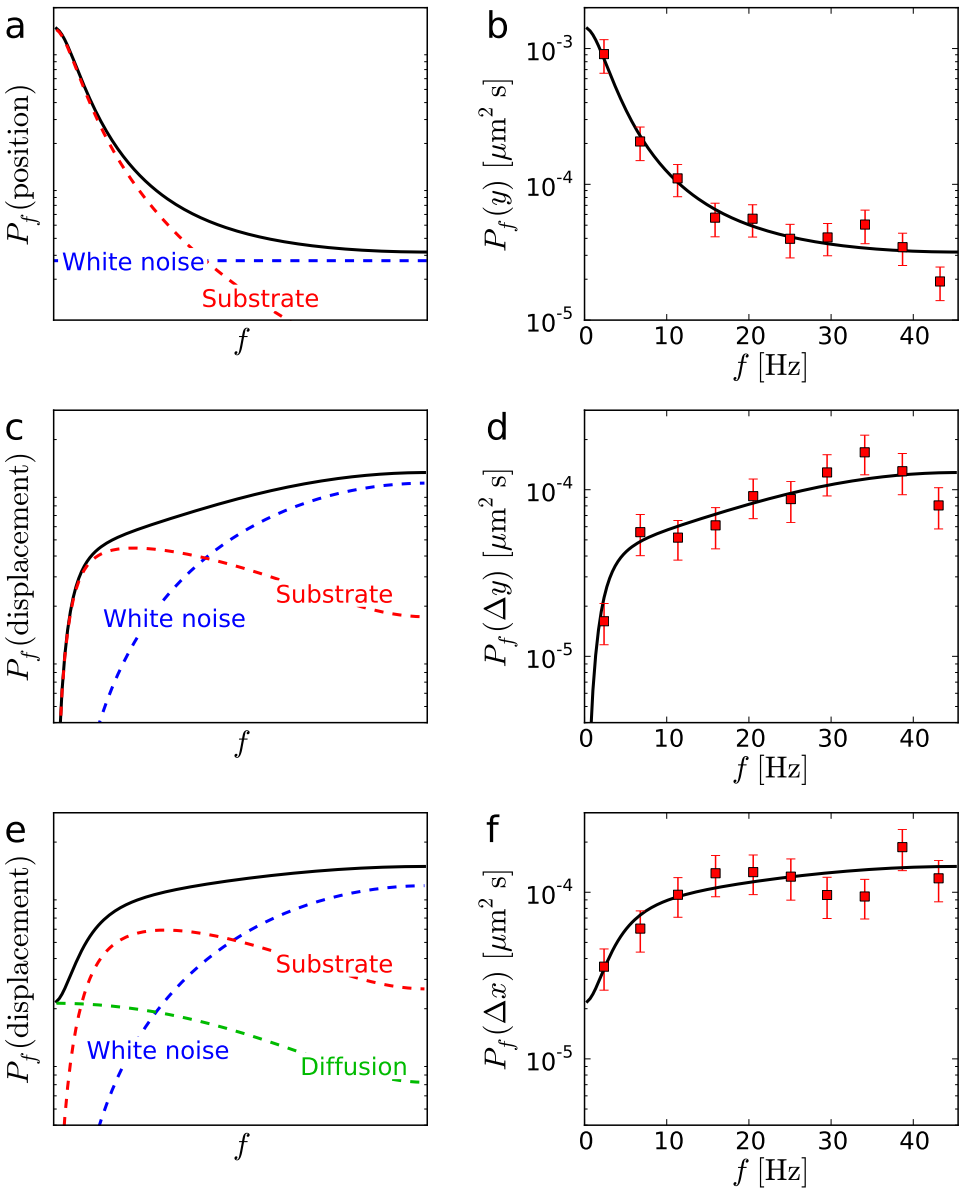

Figure 2.4: Power spectra of measured displacements of different types of stochastic movement. (Sum of individual contributions to the power spectrum and fit to measured data: black line; experimentally measured values: red squares; individual contributions to power spectrum: dashed lines.) a) Composition of position spectrum of substrate fluctuations measured in presence of positional noise. b) Experimentally measured power spectrum of transversal (y-direction) positions of a protein diffusing on DNA. c) Composition of displacement spectrum of substrate fluctuations measured in presence of positional noise. d) Experimental measurements of transversal (y-direction) displacements of a protein diffusing on DNA. e) Composition of displacement spectrum of diffusion on a fluctuating substrate measured in the presence of positional noise. f) Experimental measurements of longitudinal (x-direction) displacements of a protein diffusing on DNA.

2.2.2.2 Unbiased covariance-based estimation

If parameters describing the substrate fluctuations can be determined a priori they should be used in estimating diffusion coefficient as described in Sec. 3.2.2, since this increases precision considerably (Sec. 2.2.3). In this case the parameters describing substrate motion can be used to calculate the bias caused by substrate movement for the CVE defined in Sec. 2.1.2. The bias can be subtracted from the CVE to obtain unbiased estimates of diffusion coefficients for short time-series (Apps. B.3.2 and C.4.1). This unbiased CVE is by construction unbiased and reaches the Cramér-Rao lower bound for most parameter values.

For short time-series of particles diffusion on a substrate, where the fluctuations have not been determined a priori and where the MLE described above fails, the DNA movement parameters estimated from longer time-series can be used to calculate the bias of the CVE and thus obtain unbiased estimates of diffusion coefficients even for short time-series. The procedure is described in detail in Sec. 3.2.1.2.

2.2.3 Numerical results

We test the estimators on simulated data of a particle diffusing on a fluctuating DNA strand. In general, only the lowest mode of the DNA movement contributes significantly to the particle's movement (Figs. B.1-B.6 and D.2). So we neglect higher modes of the DNA motion and only simulate the lowest mode. Since the mathematical description of diffusion on a fluctuating substrate contains two or three additional parameters compared to free diffusion, we cannot define a simple signal-to-noise ratio, which defines estimator performance, as we could in the previous section (Sec. 2.1). The signal-to-noise ratio is a function of the frequency (Figs. 2.4a,c,e) and a complicated interplay of parameters define estimator precision. Thus, we will in this section not investigate the whole parameter space, but a subset. A subset which, however, fully comprises parameter values observed in experimental measurements of hOgg1 proteins diffusing on flow stretched DNA (Sec. 2.3 and App. C.4). The lowest mode of the DNA fluctuations is described by three parameters: the correlation constant c_1, the longitudinal amplitude $\mathcal{X}_{1,1}$, and the transversal amplitude $\mathcal{Y}_1$. For the experimental data analyzed in Sec. 2.3, $c_1\mathcal{X}_{1,1}$ and $c_1\mathcal{Y}_12$ do not change along the DNA, while c_1 changes approximately twofold (Figs. C.12 and C.11). The estimators' precisions are less sensitive to the values of D and σ^2. So we let the parameters $c_1\mathcal{X}_{1,1}$, $c_1\mathcal{Y}_k^2$, D and σ^2 be constant and equal to the mean of their experimentally determined values for hOgg1 proteins diffusing on DNA (Sec. 2.3).

Measurements of diffusion on stretched DNA are simulated as described in App. B.4. We set $D = 0.3\ \mu\text{m}^2/\text{s}$, $\sigma^2 = 1500\ \text{nm}^2$, $c_1\mathcal{X}_{1,1} = 2.1\ \mu\text{m/s}$, and

$c_1\mathcal{Y}_k^2 = 0.20\ \mu\text{m}^2/\text{s}$, while the correlation constant c_1 is varied tenfold, between 10 and 100 Hz. (Simulation results are shown for other values of D and σ^2 in Table. B.1.)

DNA fluctuations induce a bias in the estimates of the diffusion coefficient, which can be multiple times larger than the diffusion coefficient itself, if these are not accounted for (Fig. 2.5). The MLE appropriately accounts for the DNA fluctuations and is practically optimal for long time-series (Fig. 2.5b). For short time-series both the CVE and the MLE fail to give reliable estimates of the diffusion coefficient. The MLE performs even worse than the CVE.

When the parameters describing the DNA fluctuations have been determined a priori both the CVE and the MLE are practically optimal, for both short and long time-series (Fig. 2.6).

A priori determination of both the DNA movement parameters and the positional noise gives a slight increase in the precision of the CVE of the diffusion coefficient, while it induces a bias in the MLE for slow DNA fluctuations (Fig. 2.7). In this case the CVE should be used. There is, however, almost no noticeable increase in precision when σ^2 is known a priori over the case where both D and σ^2 are unknown for the parameter values examined.

2.3 hOgg1 proteins diffusing on flow-stretched DNA

In this section we use the methods we have developed for diffusion on a fluctuating substrate to reanalyze a data set of hOGG1 repair proteins diffusing on λ-DNA previously analyzed using MSD-based methods [19]. The human 8-Oxyguanine DNA glycosylase (hOgg1) protein is crucial in the repair of oxidative damage of guanine bases in DNA [40, 41, 42]. Our analysis of the data shows that DNA fluctuations induce a significant bias in the estimated diffusion coefficients if they are not accounted for. The increased precision over earlier methods and precise knowledge of the estimation uncertainty allows us to conclude that proteins show different diffusion coefficients and that these diffusion coefficients are correlated with the proteins' residence time on DNA. We show that these results implies a multi-state model for hOgg1 diffusion on DNA.

In Sec. 2.3.1 we review the experimental setup and preliminary data analysis. In Sec. 2.3.2 we analyze the distribution of protein residence times on the DNA, which shows that unbinding from the DNA is not a simple Poisson process. In Sec. 2.3.3 we employ our methods to estimate the proteins' diffusion coefficients. Finally, in Sec. 2.3.4 we show using numerical simulations that a single-state model for diffusion on DNA cannot account for the observed distribution of diffusion coefficients and residence times.

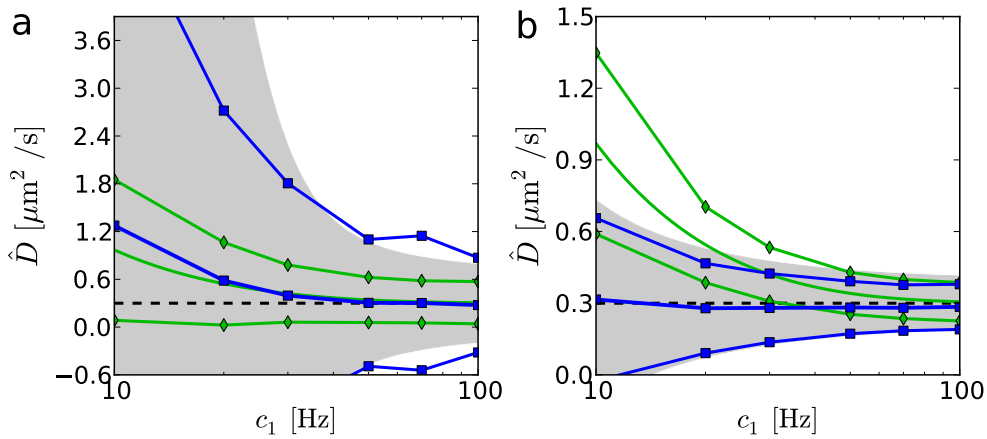

Figure 2.5: Mean plus/minus standard error of the MLE, which accounts for DNA fluctuations, and the CVE, which does not, applied to an ensemble of 1,000 Monte Carlo generated time-series of diffusion on a fluctuating substrate, where both positional noise amplitude σ^2 and parameters determining the substrate motion, (c_1,$\mathcal{X}_{1,1}$,$\mathcal{Y}_1$), are unknown. (MLE: connected squares, blue; moment estimator: full lines and diamonds, green; Cramér-Rao bound: grey area). $\tau = \Delta t$ for both simulations, which corresponds to full time-frame integration. True parameter values are $D = 0.3$ μm^2/s, $\sigma^2 = 1500$ nm^2, $c_1\mathcal{X}_{1,1} = 2.1$ μm/s, and $c_1\mathcal{Y}_k^2 = 0.20$ μm^2/s. a) Time-series length $N = 10$ and b) $N = 100$.

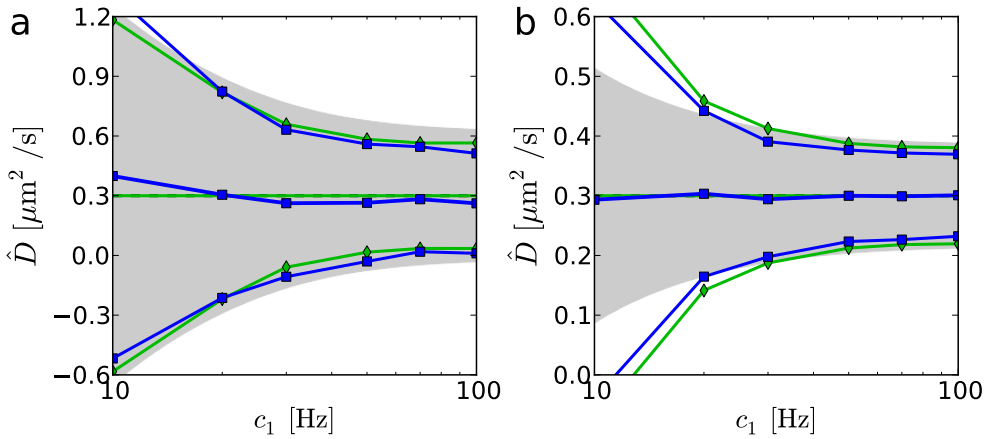

Figure 2.6: Mean plus/minus standard error of the MLE and CVE, which account for DNA fluctuations, applied to an ensemble of 1,000 Monte Carlo generated time-series of diffusion on a fluctuating substrate, where positional noise amplitude σ^2 is unknown, while parameters determining the substrate motion, (c_1,$\mathcal{X}_{1,1}$,$\mathcal{Y}_1$), are known a priori. (MLE: connected squares, blue; moment estimator: full lines and diamonds, green; Cramér-Rao bound: grey area). $\tau = \Delta t$ for both simulations, which corresponds to full time-frame integration. True parameter values are $D = 0.3$ μm^2/s, $\sigma^2 = 1500$ nm^2, $c_1\mathcal{X}_{1,1} = 2.1$ μm/s, and $c_1\mathcal{Y}_k^2 = 0.20$ μm^2/s. a) Time-series length $N = 10$ and b) $N = 100$.

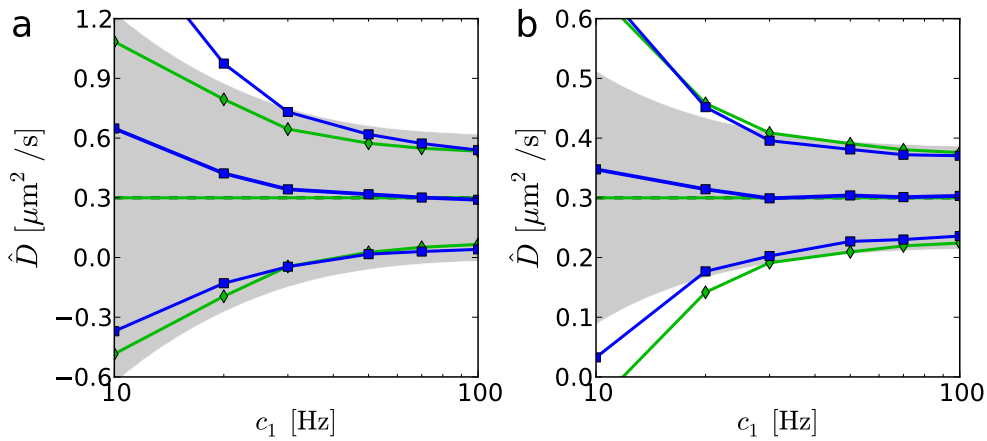

Figure 2.7: Mean plus/minus standard error of the MLE and CVE, which account for DNA fluctuations, applied to an ensemble of 1,000 Monte Carlo generated time-series of diffusion on a fluctuating substrate, where both positional noise amplitude σ^2 and parameters determining substrate motion, (c_1,$\mathcal{X}_{1,1}$,$\mathcal{Y}_1$), are known a priori. (MLE: connected squares, blue; moment estimator: full lines and diamonds, green; Cramér-Rao bound: grey area). $\tau = \Delta t$ for both simulations, which corresponds to full time-frame integration. True parameter values are $D = 0.3$ μm^2/s, $\sigma^2 = 1500$ nm^2, $K_x = 2.1$ μm/s, and $K_y = 0.20$ μm^2/s. a) Time-series length $N = 10$ and b) $N = 100$.

2.3.1 Experimental setup and preliminary data analysis

The experimental setup and a snapshot of raw experimental data are shown in Figure 2.8. The data consist of hundreds of individual time-series of the longitudinal x- and transversal y-positions of fluorescently labeled hOgg1 proteins diffusing on λ-phage DNA stretched by a strong shear flow. Experiments are performed at a pH of 7.5 and the DNA is approximately 90% stretched [19]. The positions of the proteins are obtained by fitting a 2D-Gaussian to the measured point-spread function emitted by the fluorescent protein.

Time-series, whose transversal (y-direction) motion are not consistent with DNA motion with positional noise, such as proteins getting stuck on the coverslip, have been discarded (see App. C.1 for details). Furthermore, time-series shorter than $N = 12$ ($t = 0.132$ s) have been discarded, since we have no way of reliably checking whether the measured movement is consistent with diffusion on DNA for very short trajectories. Also, to ensure homogeneous conditions for the time-series analyzed, proteins close to the tethering point and the free end of the DNA are not included in the analysis. Proteins close to the tethering point tend to get stuck on the coverslip, effectively filtering out long time-series, while proteins close to the free end also tend to have shorter residence times on DNA, probably due to increased DNA motion (App. C.1).

2.3.2 Protein residence time on DNA

If the interactions between proteins and DNA are constant during the time the proteins stay bound to the DNA and are the same for all proteins, the distribution of residence times will be an exponential distribution (Ap. C.3). As Figs. 2.9a-b show, the exponential distribution does not account for the spread observed in the experimentally measured distribution. A goodness of fit test gives $p = 4 \cdot 10^{-5}$, i.e., the data do not support the hypothesis that unbinding from DNA is a Poisson process (see App. C.2 for design of the test). A natural extension is to consider a constant-rate two-state model for the protein-DNA interactions, i.e., proteins can switch between two states with different binding energies between protein and DNA, and the rates for switching between states are constant. This model predicts that the distribution of the residence times is a sum of two exponentials (App. C.3). Another simple extension to the one-state model assumes that proteins bind with different but constant binding energies to the DNA and that these energies are normal distributed. This results in log-normal distributed characteristic residence times (App. C.3). A third extension of the simple one-state hypothesis is to assume that the protein experiences a rugged landscape of binding energies dependent on base-pair sequence while diffusing on the DNA as described in [20] (Sec. 2.3.4).

Simulations of the rugged energy landscape model show that this cannot alone produce the residence time distribution observed experimentally (Fig. 2.12a),

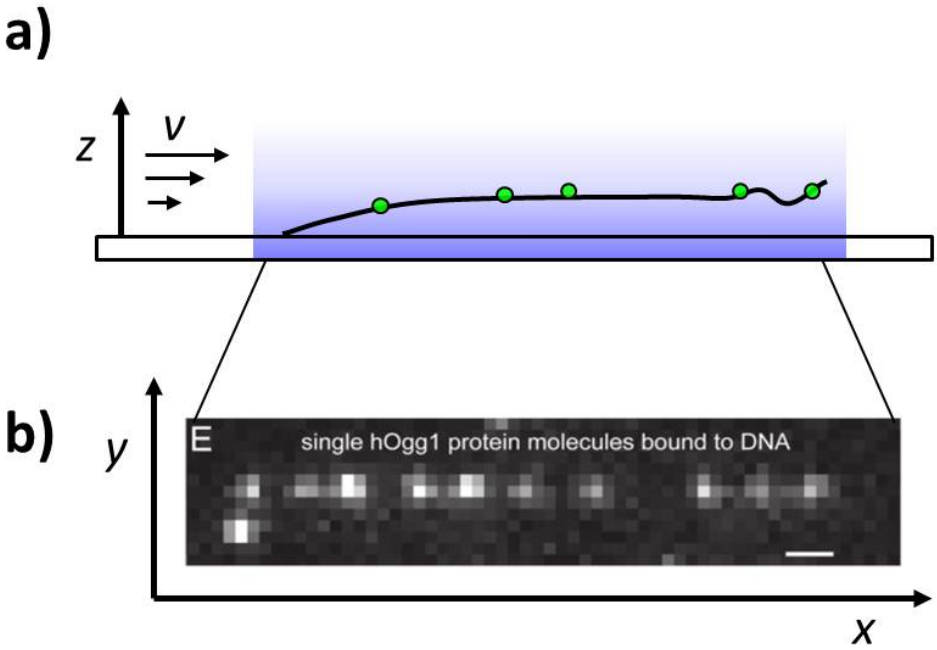

Figure 2.8: a) Experimental setup. A λ-phage DNA molecule is end-biotinylated and fused to the coverslip. The DNA is stretched by a shear flow to approximately 90% of its contour length ($L = 48,502$ base-pairs, equal to 17.9 μm) [19]. Fluorescent hOGG1 proteins diffusing on the DNA molecule are recorded using total internal reflection microscopy (a hOgg1 protein is approximately 5 nm in size). b) Image of fluorescent hOGG1 molecules diffusing on the stretched λ-DNA molecule (image from [19], the scale-bar is 1 μm).

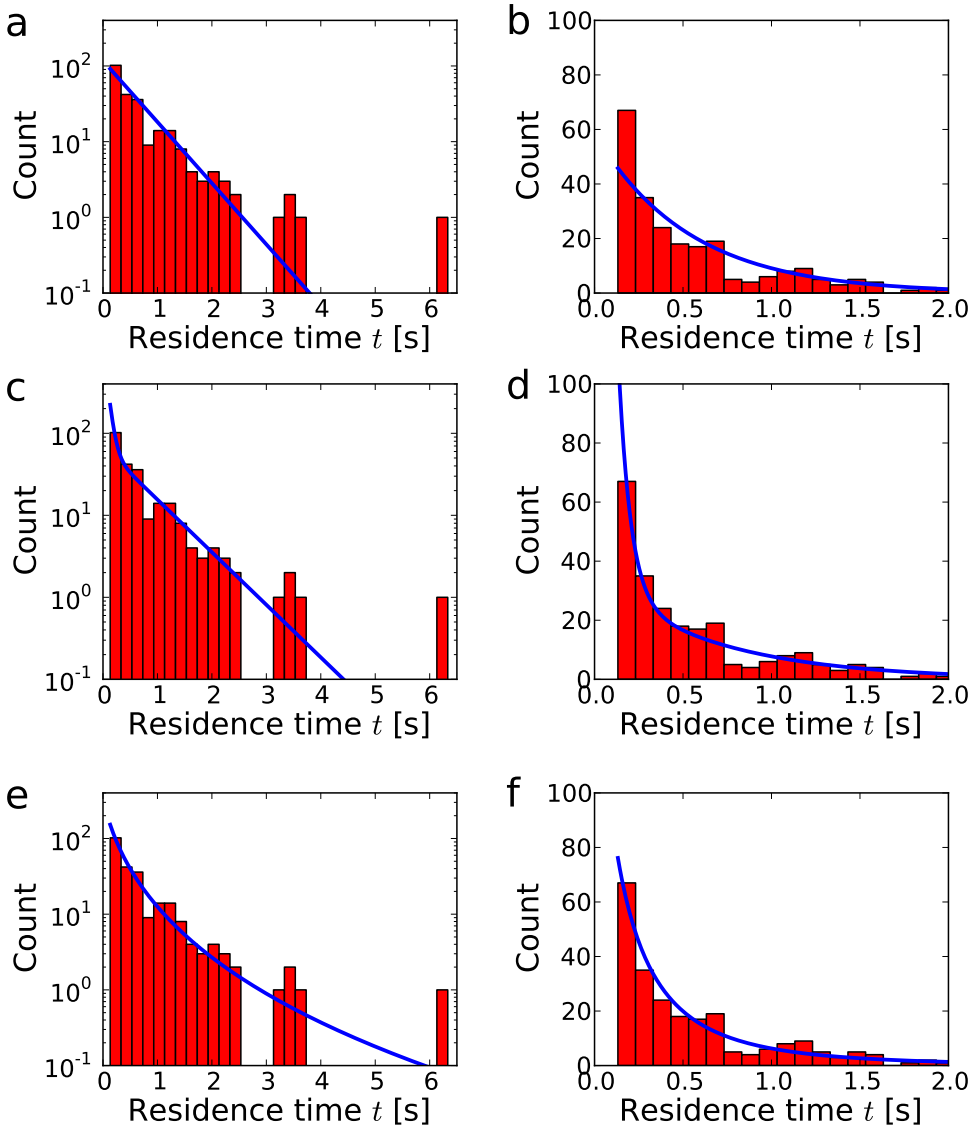

Figure 2.9: Distribution of residence times on DNA. Only time-series longer than $N \geqslant 12$, i.e., $t \geqslant 0.132$ s are included. (Measured distribution: red bars; theoretical distribution: blue line.) a) Single exponential fit to distribution of residence times shown on semilog scale; b) zoom on the first two seconds of the time-axis of a) shown on linear scale. Pearson's chi-squared test (goodness of fit) gives $p = 4{\cdot}10^{-6}$. c) Fit to two exponentials shown on semilog scale; d) zoom on the first two seconds of the time-axis of c) shown on linear scale. Pearson's chi-squared test gives $p = 0.04$. e) Fit to an exponential distribution with log-normal distributed characteristic residence times shown on semilog scale; f) zoom on the first two seconds of the time-axis of e) shown on linear scale. Pearson's chi-squared test gives $p = 0.02$. Akaike-weights for the two models, which are consistent with data are: $w_2 = 0.82$ for two exponentials and $w_{\mathrm{logN}} = 0.18$ for log-normal distributed characteristic residence times, i.e., the two-state model agrees better with data.

while both the two-state model (Figs. 2.9c-d) and the normal distributed energies model (Figs. 2.9e,f) or a combination of the three models may explain the observed residence time distribution.

Figure 2.9 illustrates another point: It is important to be able to properly analyze short time-series, since the majority of the recorded time-series are short, and thus, much of the experimental information is contained in the short time-series.

2.3.3 Parameter estimation

Measurements of YOYO-labeled DNA and Brownian dynamics simulations show that the time-scale of longitudinal fluctuations of the free end of the DNA is of the order of 20 ms [19], i.e., two times the time-lapse. This result agrees with fits to the measured power spectra (Fig. C.11). This means that we need to take DNA fluctuations into account when estimating diffusion coefficients, since the DNA fluctuations do not resemble white noise (Fig. 2.4b). Ignoring the DNA fluctuations leads to a bias, which can be larger than the value of the diffusion coefficient itself (Figs. 2.11 and C.15). The mean bias is $\overline{b(D)} = 0.10 \pm 0.01\ \mu\text{m}^2/\text{s}$, while it may be as large as $0.2\ \mu\text{m}^2/\text{s}$ (Fig. C.15).

We estimate diffusion coefficients from the measured time-series as described in Sec. 3.2. (Three sample trajectories are shown in Fig. 2.10.) For time-series longer than $N = 35$, diffusion coefficients D, positional noise amplitudes σ^2, and parameters describing DNA motion (longitudinal amplitude $\mathcal{X}_{1,1}$, transversal amplitude $\mathcal{Y}_1$, and correlation constant c_1) are estimated (App. C.4). The substrate fluctuations are modeled using a single mode. Including more modes does not affect diffusion coefficient estimates and is only possible for time-series longer than $N = 145$ (Fig. D.2). For time-series of $N = 35$ or shorter, the MLE algorithm may not converge. For these short time-series we use the CVE and subtract the bias caused by DNA fluctuations as described in Sec. 3.2. We thus obtain unbiased diffusion coefficient estimates even for short time-series (Fig. 2.11 and App. C.4).

As Fig. 2.8 shows, multiple proteins diffuse on the DNA at a time. This leads to infrequent collisions between the proteins. This crowding does not induce correlations between estimated diffusion coefficients and residence times, nor does it alter the mean and dispersion of diffusion coefficient estimates considerably (App. C.4.2). This insensitivity to molecular crowding is a specificity of the displacement-based methods, since these estimators rely on the shortest displacements, which are least affected by the crowding. This contrasts with MSD-based estimators, which rely on longer displacements and are thus highly affected by crowding.

The increased resolution offered by the methods developed in this thesis re-

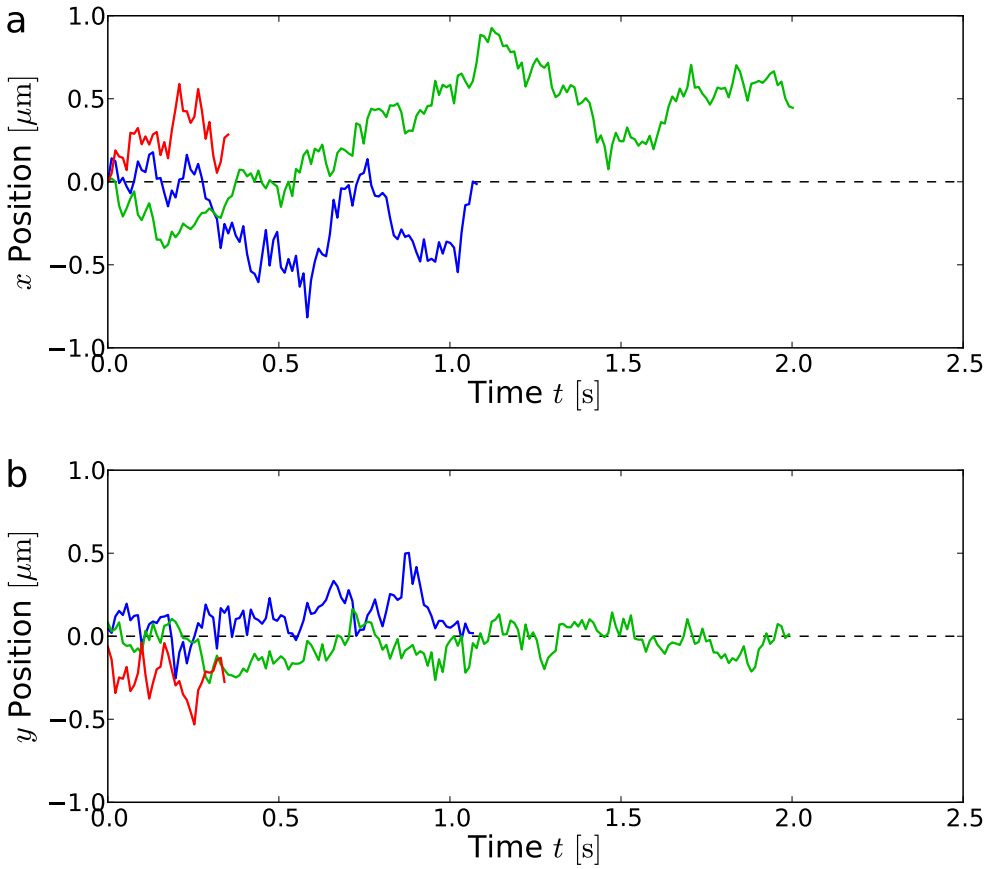

Figure 2.10: Trajectories of hOgg1 proteins diffusing on flow-stretched DNA. Trajectories are set to start at zero. a) Position of the proteins in the longitudinal (x) direction. b) Position of the proteins in the transversal (y) direction. The proteins are measured until they either unbind from the DNA or bleach. Whichever of these events happens first determines the length of the time-series, but the time-scale of bleaching is in the experiments much longer than the proteins' mean residence time on DNA [19].

veals a clear correlation between the proteins' residence times and diffusion coefficients (Fig. 2.11), and a dispersion in the proteins diffusion coefficients (App. C.4).

2.3.4 Random walker in rugged energy landscape

A model for protein diffusion on DNA developed in [20] treats the protein-DNA interactions as a random energy landscape, with uncorrelated normal distributed binding energies, and the protein as a random walker in this quenched landscape. This induces correlations between protein residence times on DNA and effective diffusion coefficients and provides a possible simple explanation of the observed trend in diffusion coefficients and the residence time distribution observed. However, for experimentally relevant physical parameters, the random walker's movement effectively self-averages during a time-lapse. So neither a trend between diffusion coefficients and residence times, nor a non-exponential residence time distribution are seen (Fig. 2.12 and App. C.5).

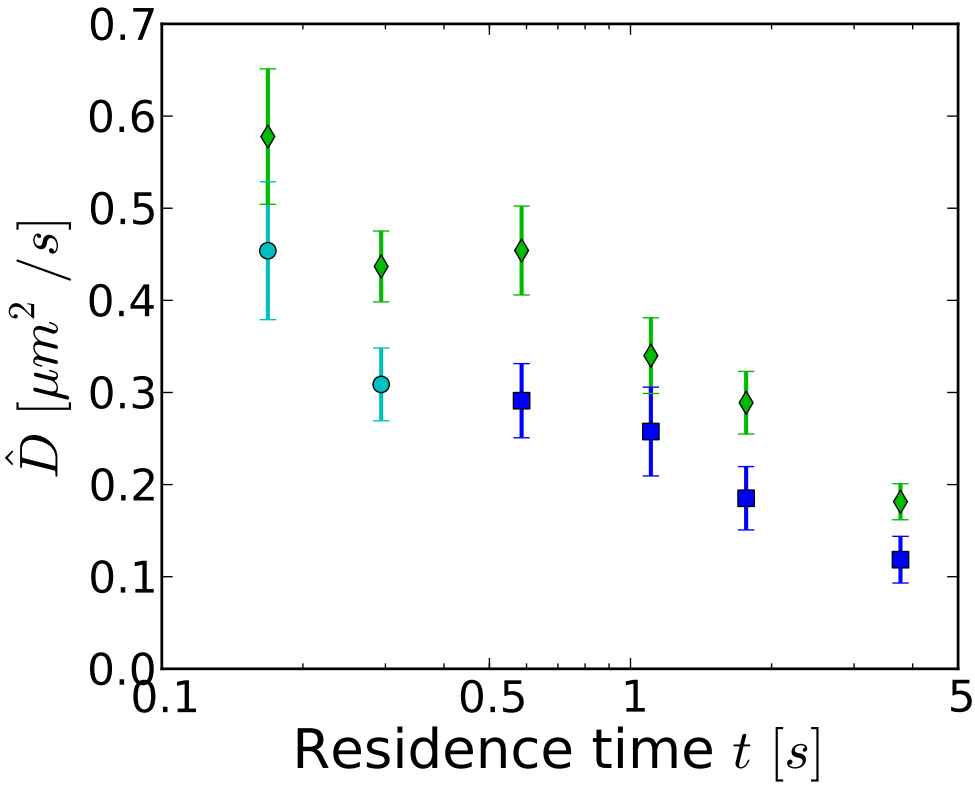

Figure 2.11: Diffusion coefficient estimates $\widehat{D}$ versus protein residence time on DNA t. (CVE, which does not account for DNA fluctuations: green diamonds; MLE, which explicitly takes DNA fluctuations into account: blue squares; CVE where bias due to DNA movement is subtracted: cyan circles.) The MLE only works for longer time-series ($N > 35$), The CVE is biased, and its bias is uncorrelated with time-series length. Using the CVE corrected for bias allows estimation for short time-series. A clear correlation between time-series lengths and diffusion coefficients is seen.

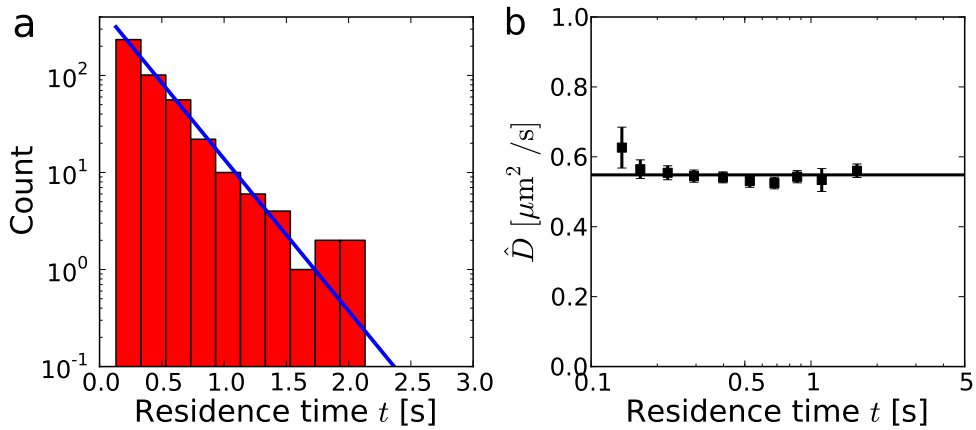

Figure 2.12: a) Distribution of residence times for Monte Carlo simulated data of a random walker in a rugged energy landscape. A Pearson's chi-squared test for exponential distribution gives a p-value of $p = 0.38$. b) Diffusion coefficient versus residence time. A chi-squared test for variance to test the hypothesis that D is constant gives a p-value of $p = 0.74$.

Chapter 3

Methods

This chapter reviews the methods developed in the thesis, and explains which methods should be applied, and how to apply them. In Sec. 3.1 methods for diffusion coefficient estimation from single-particle tracking (SPT) measurements of a particle undergoing free diffusion are presented. Section 3.2 presents methods for diffusion coefficient estimation from SPT measurements of a particle diffusing on a fluctuating substrate.

In all cases we assume that data consist of a time-series of $N+1$ measured positions $\{x_0, x_1, \ldots, x_N\}$, possibly in multiple dimensions, of a diffusing particle recorded at equidistant times $\{0, \Delta t, \ldots, N\Delta t\}$. We denote by N the length of such a time-series, measured in units of the time-lapse Δt.

3.1 Free diffusion

Methods presented in this section should be used for diffusion coefficient estimation from SPT measurements of a particle undergoing free diffusion or diffusion on a stiff substrate, e.g., a stiff polymer such as a microtubule or an actin filament. The methods presented in this section are divided in two subsections: estimation when the positional noise amplitude is unknown (Sec. 3.1.1); and estimation when the positional noise amplitude has been determined a priori (Sec. 3.1.2).

3.1.1 Unknown noise amplitude

When the amplitude of the positional noise is not known a priori the CVE described below should generally be used, since it is unbiased and practically reaches the Cramér-Rao bound for a large range of signal-to-noise levels (SNR > 1). If long time-series are recorded and the signal-to-noise ratio is very low (SNR < 1) the MLE should be used, since the CVE is not efficient in this range.

3.1.1.1 Covariance-based estimation

Covariance-based estimator (CVE) The CVE of the diffusion coefficient D is

$$\widehat{D} = \frac{\overline{\Delta x_n^2}}{2\Delta t} + \frac{\overline{\Delta x_n \Delta x_{n+1}}}{\Delta t} \ , \tag{3.1.1}$$

while the CVE of the positional noise variance σ^2 is

$$\widehat{\sigma}^2 = R\,\overline{\Delta x_n^2} + (2R-1)\,\overline{\Delta x_n \Delta x_{n+1}} \ , \tag{3.1.2}$$

where

$$\overline{\Delta x_n^2} = \frac{1}{N}\sum_{n=1}^{N} \Delta x_n^2 \ , \quad \overline{\Delta x_n \Delta x_{n+1}} = \frac{1}{N-1}\sum_{n=1}^{N-1} \Delta x_n \Delta x_{n+1} \ , \tag{3.1.3}$$

and the motion blur coefficient R is equal to $R = 1/6$ for full time-integration and $R = 0$ for instantaneous camera shutter. (For values of R for other camera shutters, see App. A.1.)

Variance The variance of the covariance-based estimator (CVE) is to second order in $1/N$

$$\mathrm{Var}\left(\widehat{D}\right) = \frac{6\alpha^2 + 4\alpha\beta + 2\beta^2}{N(\Delta t)^2} + \frac{4(\alpha+\beta)^2}{N^2(\Delta t)^2} \ , \tag{3.1.4}$$

with $\alpha \equiv D\Delta t$ and $\beta \equiv \sigma^2 - 2RD\Delta t$ (App. A.3.2.1).

The variance of $\widehat{\sigma}^2$ and the covariance between $\widehat{D}$ and $\widehat{\sigma}^2$ can be found in App. A.3.2.1.

For importance weighting, e.g., when calculating the weighted mean, the time-series length N should be used as weight, since it is known exactly. This avoids complications such as bias due to correlations between the estimated parameters and the estimated variances.

Characteristic function The characteristic function of the moment estimator, $\widetilde{p}$, derived in App. A.3.2.2, is defined by

$$\ln \widetilde{p}(\omega) = \frac{N+1}{2} \ln \left[\frac{4}{A}\right] + \frac{1}{4} \ln \left[(A+C)^2 - B^2\right] - \frac{1}{2} \left(S_{1,+} + S_{1,-} + S_{2,+} + S_{2,-}\right) \ , \tag{3.1.5}$$

where

$$S_{1,\pm} = \frac{N+1}{2} \ln \left[1 + \sqrt{1 - \left(\frac{B \pm \sqrt{B^2 - 4AC}}{2A}\right)^2}\right] \ , \tag{3.1.6}$$

$$S_{2,\pm} = \frac{1}{2} \ln \left[1 - \left(\frac{1 - \sqrt{1 - \left(\frac{B \pm \sqrt{B^2-4AC}}{2A}\right)^2}}{1 + \sqrt{1 - \left(\frac{B \pm \sqrt{B^2-4AC}}{2A}\right)^2}}\right)^{N+1}\right] \ , \tag{3.1.7}$$

and A, B, and C are functions of ω,

$$A = 1 + \frac{2\omega}{N}(D+Q) \ , \quad B = \frac{2\omega}{N-1}\left[2D + \left(1 + \frac{1}{N}\right) Q\right] \ , \quad C = -\frac{4\omega}{N} Q \ , \tag{3.1.8}$$

with

$$Q = \frac{s^2 - 2RD\Delta t}{\Delta t} \ .$$

Higher moments of the CVE are found by differentiating $\widetilde{p}$ w.r.t ω and letting ω tend to zero,

$$\left\langle \widehat{D}^k \right\rangle = i^k \left.\frac{\partial^k \widetilde{p}}{\partial \omega^k}\right|_{\omega=0} \ . \tag{3.1.9}$$

The probability density of the CVE of D is equal to the inverse Fourier transform of the characteristic function,

$$p\left(\widehat{D} \middle| D, \sigma^2\right) = \mathcal{F}^{-1}[\widetilde{p}(\alpha)]\left(\widehat{D}\right) \ . \tag{3.1.10}$$

Thus, the probability density of the CVE and confidence intervals for covariance-based estimates can be found by numerical Fourier transformation of $\widetilde{p}$. This can be done effectively using a fast-Fourier transform (FFT) algorithm with corrections for end contributions as described in [43].

3.1.1.2 Maximum likelihood estimation

Maximum likelihood estimator (MLE) The MLE uses the transformed displacements $\breve{\mathbf{d}}$, given by

$$\breve{\Delta x}_k = (\mathbf{P}^{-1}\mathbf{\Delta x})_k = \sqrt{\frac{2}{N+1}} \sin\left[\frac{\pi k n}{N+1}\right] \Delta x_n \ . \tag{3.1.11}$$

These are independent and Gaussian distributed with variances equal to (Sec. A.3.1)

$$\psi_k = 2D\Delta t + \left(2 - 2\cos\frac{\pi k}{N+1}\right)(\sigma^2 - 2RD\Delta t) \ . \tag{3.1.12}$$

The exact log-likelihood function for the parameters D and σ^2, given the data $\mathbf{\Delta x}$, is equal to the sum

$$\ln \mathcal{L}(D, \sigma^2 | \mathbf{\Delta x}) = -\frac{1}{2}\sum_{k=1}^{N}\left\{\ln\psi_k + \frac{\widetilde{\Delta x}_k^2}{\psi_k}\right\} \ . \tag{3.1.13}$$

The MLE is obtained by maximizing the log-likelihood (3.1.13) numerically. An algorithm for the MLE, which is approximately two times faster, is given in Sec. A.3.1.2.

Variance The variance of the MLE is to order $1/N$

$$\mathrm{var}\left(\widehat{D}, \widehat{\sigma}^2\right) = \mathcal{I}^{-1} \ , \tag{3.1.14}$$

where

$$\mathcal{I} = \frac{1}{2}\sum_{k=1}^{N}\frac{1}{\psi_k^2}\begin{pmatrix} \left(\frac{\partial\psi_k}{\partial D}\right)^2 & \frac{\partial\psi_k}{\partial D}\frac{\partial\psi_k}{\partial\sigma^2} \\ \frac{\partial\psi_k}{\partial D}\frac{\partial\psi_k}{\partial\sigma^2} & \left(\frac{\partial\psi_k}{\partial\sigma^2}\right)^2 \end{pmatrix} \ , \tag{3.1.15}$$

with

$$\frac{\partial\psi_k}{\partial D} = 2\Delta t - 2R\Delta t\left(2 - 2\cos\frac{\pi k}{N+1}\right) \ , \quad \frac{\partial\psi_k}{\partial\sigma^2} = 2 - 2\cos\frac{\pi k}{N+1} \ . \tag{3.1.16}$$

3.1.2 Known noise amplitude

The positional noise amplitude can usually be estimated directly when the diffusing particles' positions are determined as described below. In this case the CVE should be used for SNR > 1, while the MLE should be used for SNR < 1.

3.1.2.1 Estimating the noise amplitude a priori

An isolated diffusing and freely rotating fluorescent molecule emits photons in a stochastic manner. If the diffusion length $\sqrt{2D\Delta t}$ of the observed particle is smaller than or comparable to the width of the PSF (around 100 nm for green light observed through a 100X oil-immersion objective), the position of the particle can be estimated by fitting a circular 2D Gaussian plus a constant

background term to its measured point spread function [44]. This results in a positional noise with variance

$$\sigma^2 = \frac{s_a^2}{P}\left(\frac{16}{9} + \frac{8\pi s_a^2 b^2}{Pa^2}\right) \quad , \tag{3.1.17}$$

for camera pixel width a, background b, measured number of photons P and $s_a^2 = s^2 + a^2/12$, where s is the width of the measured PSF (Sec. A.1.3).

Since the positional noise is Gaussian distributed, the error on the estimate of σ^2 is $\mathrm{Var}\left(\widehat{\sigma}^2\right) = 2\sigma^4/P$, where P is the number of photons recorded per frame. An even more precise estimate of σ^2 is obtained by averaging over the positional noise of all $N+1$ positions in the time-series and, if the different time-series have the same noise amplitude, averaging over all M time-series, to obtain an error on $\widehat{\sigma}^2$ of $\mathrm{Var}\left(\widehat{\sigma}^2\right) = 2\sigma^4/(N+1)P$ and $\mathrm{Var}\left(\widehat{\sigma}^2\right) = 2\sigma^4/M(N+1)P$, respectively. Since P is typically on the order of 1000 or more, and M typically is on the order of hundreds to thousands, the contribution to the variance of $\widehat{D}$ from the error on $\widehat{\sigma}^2$ is usually negligible.

3.1.2.2 Covariance-based estimation

Covariance-based estimator (CVE) When an estimate of the noise amplitude is obtained a priori, the CVE of the diffusion coefficient reduces to

$$\widehat{D} = \frac{\overline{\Delta x_n^2} - 2\widehat{\sigma}^2}{2(1-2R)\Delta t} \quad , \tag{3.1.18}$$

where $\widehat{\sigma}^2$ is the estimate of σ^2 obtained as described in Sec. 3.1.2.1.

Variance The variance of the CVE is

$$\mathrm{Var}\left(\widehat{D}\right) = \frac{2\alpha^2 + 3\alpha\beta + 4\beta^2}{N(1-2R)^2(\Delta t)^2} + \frac{\mathrm{Var}\left(\widehat{\sigma}^2\right)}{(\Delta t)^2(1-2R)^2} \quad . \tag{3.1.19}$$

3.1.2.3 Maximum likelihood estimation

Maximum likelihood estimator (MLE) When an estimate of the noise amplitude is obtained a priori, the MLE is obtained by maximizing $\ln \mathcal{L}(D, \widehat{\sigma}^2 | \mathbf{\Delta x})$ given by Eq. (3.1.13) with respect to D only.

Variance The Cramér-Rao bound, which approximately gives the variance of the MLE, is given by

$$\mathcal{I}^{-1} = \mathcal{I}_D^{-1} + \mathcal{I}_D^{-2}\mathcal{I}_{D,\sigma^2}^2 \mathrm{Var}\left(\widehat{\sigma}^2\right) \quad , \tag{3.1.20}$$

where

$$\mathcal{I}_D = \frac{1}{2}\sum_{k=1}^{N}\frac{1}{\psi_k^2}\left(\frac{\partial\psi_k}{\partial D}\right)^2 \tag{3.1.21}$$

and

$$\mathcal{I}_{D,\sigma^2} = \frac{1}{2}\sum_{k=1}^{N}\frac{1}{\psi_k^2}\frac{\partial\psi_k}{\partial D}\frac{\partial\psi_k}{\partial\sigma^2} \quad . \tag{3.1.22}$$

3.2 Diffusion on a fluctuating substrate

Methods presented in this section should be used for measurements of diffusion on a fluctuating substrate, e.g., a supported or grafted lipid bilayer or a flexible or semiflexible polymer such as DNA. If the substrate's fluctuations are considerably faster than the measurement frequency $1/\Delta t$, they only contribute to the noise term σ^2 and the methods described in Sec. 3.1 should be used. The time-scale of the fluctuations should be checked, e.g., by measuring the position of a fixed fluorophore on the substrate and comparing the power spectrum of the positions to a straight line (see Fig. 2.4a,b). For diffusion on DNA or other polymers, the measured transversal positions of the diffusing particles can be used. The substrate fluctuations are described by a set of parameters, which we denote $\boldsymbol{\phi}$. The number of parameters and their form depend on the nature of the substrate fluctuations. Methods presented in this section are derived for full time-integration, i.e., shutter time τ equal to the time-lapse Δt. (For other values of τ see App. B.2.) They apply to diffusion on DNA or another polymer. For diffusion on a flat substrate, the methods differ slightly and should be altered as described in Sec. 3.2.3.1.

This section is divided into three subsections: estimation when parameters describing substrate motion are unknown (Sec. 3.2.1); estimation when parameters describing DNA motion have been determined beforehand (Sec. 3.2.2); and an overview of analytical expressions for the form of the statistics used in estimation for a selection of experimental setups (Sec. 3.2.3).

Reviews of methods for stretching DNA and visualizing protein-DNA interactions are found in [45, 46].

3.2.1 Unknown substrate fluctuations

When substrate fluctuations cannot be determined a priori they must be estimated directly from the measured time-series along with the diffusion coefficient. This is done using the MLE. For short time-series ($N <$ 30-40), the MLE algorithm may fail to converge and the CVE should be used.

3.2.1.1 Long time-series, maximum likelihood estimation

The MLE procedure uses the power spectra of the measured displacements, where we for diffusion on DNA measure both the longitudinal (x) and transversal (y) displacements, $\mathbf{\Delta x}$ and $\mathbf{\Delta y}$. The power spectra are calculated as

$$\widehat{P}_f(\Delta\cdot) = \frac{1}{N\Delta t}\left|\mathcal{DFT}[\Delta\cdot]_f\right|^2 \quad , \tag{3.2.1}$$

where $\mathcal{DFT}$ is the discrete Fourier transform, which is calculated effectively using a fast-Fourier transform (FFT) algorithm, and $f = 1/N\Delta t, 2/N\Delta t, \ldots, f_{\mathrm{Nyq}}$. The Nyquist frequency is $f_{\mathrm{Nyq}} = 1/2\Delta t$ if N is even, and $f_{\mathrm{Nyq}} = (N-1)/2N\Delta t$ if N is odd[1].

We denote the expected values of the power spectra by $P_f^{(\Delta\cdot)}$.

Maximum likelihood estimator (MLE) The MLE is the set of parameter values $(\widehat{D}, \widehat{\sigma}^2, \widehat{\boldsymbol{\phi}})$, which maximizes the log-likelihood function given the measured power spectra

$$\left\{\widehat{P}\right\}_f \equiv \left\{\widehat{P}_f(\Delta x), \widehat{P}_f(\Delta y)\right\}_{f=1/N\Delta t,\ldots,f_{\mathrm{Nyq}}} \quad .$$

The log-likelihood function is given by (App. B.3.1)

$$\ln\mathcal{L}\left(D,\sigma^2,\boldsymbol{\phi}|\left\{\widehat{P}\right\}_f\right) = \ln\mathcal{L}\left(D,\sigma^2,\boldsymbol{\phi}|\left\{\widehat{P}(\Delta x)\right\}_f\right) + \ln\mathcal{L}\left(D,\sigma^2,\boldsymbol{\phi}|\left\{\widehat{P}(\Delta y)\right\}_f\right) \quad , \tag{3.2.2}$$

where

$$\ln\mathcal{L}\left(D,\sigma^2,\boldsymbol{\phi}|\left\{\widehat{P}(\Delta\cdot)\right\}_f\right) = \sum_{f=1/N\Delta t}^{f_{\mathrm{Nyq}}}\left\{\ln P_f^{(\Delta\cdot)} + \frac{\widehat{P}_f(\Delta\cdot)}{P_f^{(\Delta\cdot)}}\right\} \quad . \tag{3.2.3}$$

The expected values of the power spectra are

$$P_f^{(\Delta y)} = \sigma^2\Delta t + \sum_{m=-M}^{M-1}\left(\frac{1-\cos(2\pi f\Delta t)}{\pi(f\Delta t+m)}\right)^2\left\langle P_{f+\frac{m}{\Delta t}}\left[y^{\mathrm{DNA}}(\overline{s})\right]\right\rangle \quad , \tag{3.2.4}$$

and

$$\begin{aligned} P_f^{(\Delta x)} \simeq\ & 2\zeta(\overline{s})^2 D(\Delta t)^2 + [2-2\cos(2\pi f\Delta t)]\left(\sigma^2\Delta t - \frac{\zeta(\overline{s})D(\Delta t)^2}{3}\right) \\ & + \sum_{m=-M}^{M-1}\left(\frac{1-\cos(2\pi f\Delta t)}{\pi(f\Delta t+m)}\right)^2\left\langle P_{f+\frac{m}{\Delta t}}\left[x^{\mathrm{DNA}}(\overline{s})\right]\right\rangle \quad . \end{aligned} \tag{3.2.5}$$

[1]Since the measurements are real numbers, the spectrum is symmetric and we only need to analyze half of it, i.e., for frequencies in the range $f = 1/N\Delta t, \ldots, f_{\mathrm{Nyq}}$.

The number of aliases M to include in Eqs. (3.2.4) and (3.2.5) should be large enough to include all aliases, which contribute significantly to the power spectrum. In general $M = 2$ is sufficient. The expected power spectral values of the DNA motion are

$$\left\langle P_f\left[y^{\mathrm{DNA}}(\overline{s})\right]\right\rangle \simeq \sum_{k=1}^{K} \frac{2c_k \mathcal{Y}_k(\overline{s})^2}{c_k^2 + (2\pi f)^2} \ , \tag{3.2.6}$$

$$\left\langle P_f\left[x^{\mathrm{DNA}}(\overline{s})\right]\right\rangle \simeq \sum_{k,l=1}^{K} \frac{2(c_k + c_l)\mathcal{X}_{k,l}(\overline{s})^2}{(c_k + c_l)^2 + (2\pi f)^2} \ . \tag{3.2.7}$$

The eigenfunctions $\mathcal{Y}_k$ and $\mathcal{X}_{k,l}$, and the local degree of stretching of the substrate, ζ, depend on the type of the substrate motion and are given in Sec. 3.2.3. Only the lowest K modes of the substrate movement are included in the estimation procedure. Since we are not interested in modeling the substrate movement, but to obtain unbiased estimates of diffusion coefficients, only the first one or two modes need to be included. If no theoretical framework exists, which accurately describes the substrate motion, the modes must be fitted individually from the time-series data, and only one mode should be included.

Variance We define $\boldsymbol{\theta} = (D, \sigma^2, \boldsymbol{\phi})^T$. The variance of the MLE is to order $1/N$ given by $\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) = \mathcal{I}^{-1}$, where the Fisher information matrix $\mathcal{I}$ is

$$\mathcal{I}_{ij} = \sum_{f=1/N\Delta t}^{f_{\mathrm{Nyq}}} \left\{ \left(P_f^{(\Delta x)}\right)^{-2} \frac{\partial P_f^{(\Delta x)}}{\partial \theta_i} \frac{\partial P_f^{(\Delta x)}}{\partial \theta_j} + \left(P_f^{(\Delta y)}\right)^{-2} \frac{\partial P_f^{(\Delta y)}}{\partial \theta_i} \frac{\partial P_f^{(\Delta y)}}{\partial \theta_j} \right\} \ . \tag{3.2.8}$$

The derivatives $\partial P_f^{(\Delta\cdot)}/\partial\theta_i$ are found by differentiating Eqs. (3.2.4) and (3.2.5) w.r.t. to the parameters θ_i with the appropriate eigenfunctions inserted as given in Sec. 3.2.3.

3.2.1.2 Short time-series, covariance-based estimation

For short time-series the MLE fails. For a diffusing particle with mean position $\overline{s}$ on the substrate, estimates of the parameters $\boldsymbol{\phi} = (\{c_k\}, \{x_{k,l}(\overline{s})\})^T$ describing the substrate movement can be obtained by averaging over ML estimates obtained from long time-series of particles with a mean position close to $\overline{s}$. In the averages individual estimates should be weighted with time-series length N_m [47]. For the parameter ϕ_i,

$$\overline{\phi}_i = \frac{\sum_{m=1}^{M} N_m \widehat{\phi}_{i,m}}{\sum_{m=1}^{M} N_m} \ . \tag{3.2.9}$$

The variance of the weighted average is estimated by

$$\mathrm{Var}\left(\overline{\phi_i}\right) = \frac{\sum_{m=1}^{M} N_m \left(\widehat{\phi}_{i,m} - \overline{\phi_i}\right)^2}{(M-1)\sum_{m=1}^{M} N_m} \quad . \tag{3.2.10}$$

The CVE is then calculated as described in Sec. 3.2.2.1 using the averaged DNA parameter values.

3.2.2 Known substrate fluctuations

When the parameters describing the substrate motion can be determined a priori, the precision of diffusion coefficient estimates may be significantly increased by using this information. In this case, the CVE corrected for bias should be used to estimate diffusion coefficients, since it is unbiased for short time-series, where the MLE may be biased.

3.2.2.1 Covariance-based estimation

When the parameters of the substrate movement have been determined a priori, an unbiased CVE of the diffusion coefficient can be obtained by calculating and subtracting the bias caused by the substrate movement (Apps. B.3.2 and C.4.1)

Covariance-based estimator (CVE) An unbiased CVE of the diffusion coefficient is given by

$$\widehat{D} = \frac{\overline{\Delta x_n^2}}{2\Delta t} + \frac{\overline{\Delta x_n \Delta x_{n+1}}}{\Delta t} - b\left(D \middle| \widehat{\boldsymbol{\phi}}(\overline{\boldsymbol{s}})\right) \quad , \tag{3.2.11}$$

where the bias $b(D)$ is given by

$$b\left(D \middle| \widehat{\boldsymbol{\phi}}(\overline{\boldsymbol{s}})\right) = \sum_{kl=1}^{K} \frac{\left(1 - e^{-(\widehat{c}_k + \widehat{c}_l)\Delta t}\right)^3 \widehat{\mathcal{X}}_{k,l}(\overline{s})^2}{\left(\widehat{c}_k + \widehat{c}_l\right)^2 \Delta t^3} \quad . \tag{3.2.12}$$

The form of c_k and $\mathcal{X}_{k,l}(\overline{s})$ depend on the substrate motion and are given in Sec. 3.2.3 for different types of substrate motion.

Variance The variance of the unbiased CVE is to order $1/N^2$ (App. B.3.2)

$$\mathrm{Var}\left(\widehat{D}\right) = \mathrm{Var}_0(D, \sigma^2) + \mathrm{Var}_1(D, \sigma^2, \boldsymbol{\phi}) + \mathrm{Var}_2(\boldsymbol{\phi}) + \mathrm{Var}_{b(D)}(\boldsymbol{\phi}) \quad , \tag{3.2.13}$$

where Var_0 is given by Eq. (3.1.4), and

$$\mathrm{Var}_1(D,\sigma^2,\boldsymbol{\phi}) = \frac{(6\alpha+2\beta)\rho_0+8\alpha\rho_1+4\alpha\rho_2-2\beta\rho_3}{N(\Delta t)^2} + \frac{4(\alpha+\beta)\rho_0+4\beta\rho_2+2\beta\rho_3}{N^2(\Delta t)^2} \ , \tag{3.2.14}$$

$$\mathrm{Var}_2(\boldsymbol{\phi}) = \frac{1}{(\Delta t)^2}\sum_{j=1}^{N-1}\left\{\frac{N-j}{N^2}\rho_j^2+\frac{4(N-j)}{N(N-1)}\rho_{j-1}\rho_j+\frac{2(N-j-1)}{(N-1)^2}\left(\rho_{j-1}\rho_{j+1}+\rho_j^2\right)\right\} + \frac{3\rho_0^2/2+\rho_1^2}{N(\Delta t)^2}+\frac{\rho_0^2+\rho_1^2}{N^2(\Delta t)^2} \ , \tag{3.2.15}$$

and

$$\mathrm{Var}_{b(D)}(\boldsymbol{\phi}) = \sum_{ij}\frac{\partial b(D)}{\partial\phi_i}\frac{\partial b(D)}{\partial\phi_j}\mathrm{Cov}\left(\widehat{\phi}_i,\widehat{\phi}_j\right) \ . \tag{3.2.16}$$

We have defined

$$\rho_j \equiv \overline{\rho}_{\Delta x}(\overline{s},j\Delta t) = 2\overline{\rho}_x(\overline{s},j\Delta t)-\overline{\rho}_x(\overline{s},|j-1|\Delta t)-\overline{\rho}_x(\overline{s},(j+1)\Delta t) \ , \tag{3.2.17}$$

where

$$\overline{\rho}_x(\overline{s},0) = \sum_{kl=1}^{K}\frac{2\left(c_k+c_l\right)\Delta t-2\left(1-e^{-(c_k+c_l)\Delta t}\right)}{\Delta t^2(c_k+c_l)^2}\,\mathcal{X}_{k,l}(\overline{s})^2 \ , \tag{3.2.18}$$

and

$$\overline{\rho}_x(\overline{s},j\Delta t) = \sum_{kl=1}^{K}\frac{e^{-(c_k+c_l)(j-1)\Delta t}\left(1-e^{-(c_k+c_l)\Delta t}\right)^2}{\Delta t^2(c_k+c_l)^2}\,\mathcal{X}_{k,l}(\overline{s})^2 \ . \tag{3.2.19}$$

Known noise amplitude If the noise amplitude has been determined a priori higher precision can be obtained by using the following unbiased CVE to estimate D:

$$\widehat{D} = \frac{\overline{\Delta x_n^2}-2\widehat{\sigma}^2}{2(1-2R)\Delta t}-b\left(D|\widehat{\sigma}^2,\widehat{\boldsymbol{\phi}}\right) \ , \tag{3.2.20}$$

where

$$b\left(D|\widehat{\sigma}^2,\widehat{\boldsymbol{\phi}}\right) = \sum_{kl=1}^{K}\frac{2(c_k+c_l)\Delta t-3+4e^{-(c_k+c_l)\Delta t}-e^{-2(c_k+c_l)\Delta t}}{(1-2R)\Delta t^3(c_k+c_l)^2} \ . \tag{3.2.21}$$

The variance of this unbiased CVE is

$$\mathrm{Var}\left(\widehat{D}\right) = \frac{\mathrm{Var}\left(\overline{\Delta x_n^2}\right)}{4(1-2R)^2(\Delta t)^2}+\frac{\mathrm{Var}\left(\widehat{\sigma}^2\right)}{(1-2R)^2(\Delta t)^2}+\mathrm{Var}_{b(D)}(\sigma^2,\boldsymbol{\phi}) \ , \tag{3.2.22}$$

where

$$\frac{\mathrm{Var}\left(\overline{\Delta x_n^2}\right)}{4(\Delta t)^2} = \frac{2(\alpha+\beta)^2+(1-1/N)\beta^2}{N(\Delta t)^2}+\frac{2(\alpha+\beta)\rho_0-2(1-1/N)\beta\rho_1}{N(\Delta t)^2} + \frac{\rho_0^2}{2N(\Delta t)^2}+\sum_{j=1}^{N-1}\frac{(N-j)\rho_j^2}{N^2(\Delta t)^2} \ . \tag{3.2.23}$$

3.2.2.2 Maximum likelihood estimation

Maximum likelihood estimation (MLE) When the parameters describing substrate movement $\boldsymbol{\phi}$ have been determined a priori the MLE is found by maximizing the log-likelihood $\ln \mathcal{L}(D, \sigma^2 | \widehat{\boldsymbol{\phi}}, \{\widehat{P}_f\})$ w.r.t. only D and σ^2, and treating $\widehat{\boldsymbol{\phi}}$ as fixed parameters, where $\ln \mathcal{L}$ is given as in Sec. 3.2.1.1.

Variance We denote by $\boldsymbol{\theta} = (D, \sigma^2)^T$ the parameter estimate in the maximum likelihood fit. Uncertainties in the estimates of the substrate movement parameters induce additional uncertainties in the ML estimates of $\boldsymbol{\theta}$. This is taken into account by the following formula for the variance of the MLE (App. B.3.1.1)

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \simeq \mathcal{I}_\theta^{-1} + \mathcal{I}_\theta^{-1} \mathcal{I}_\phi^\theta \mathrm{Var}\left(\widehat{\boldsymbol{\phi}}\right) \mathcal{I}_\theta^\phi \mathcal{I}_\theta^{-1} \ , \tag{3.2.24}$$

where $\mathcal{I}_\theta$ is the information matrix of $\widehat{\boldsymbol{\theta}} = (\widehat{D}, \widehat{\sigma}^2)^T$,

$$(\mathcal{I}_\theta)_{ij} = \sum_{f=1/N\Delta t}^{f_{\mathrm{Nyq}}} \left\{ \left(P_f^{(\Delta x)}\right)^{-2} \frac{\partial P_f^{(\Delta x)}}{\partial \theta_i} \frac{\partial P_f^{(\Delta x)}}{\partial \theta_j} + \left(P_f^{(\Delta y)}\right)^{-2} \frac{\partial P_f^{(\Delta y)}}{\partial \theta_i} \frac{\partial P_f^{(\Delta y)}}{\partial \theta_j} \right\} \ , \tag{3.2.25}$$

and $\mathcal{I}_\phi^\theta$ is the information matrix describing the covariance between ϕ and θ,

$$\left(\mathcal{I}_\phi^\theta\right)_{ij} = \sum_{f=1/N\Delta t}^{f_{\mathrm{Nyq}}} \left\{ \left(P_f^{(\Delta x)}\right)^{-2} \frac{\partial P_f^{(\Delta x)}}{\partial \theta_i} \frac{\partial P_f^{(\Delta x)}}{\partial \phi_j} + \left(P_f^{(\Delta y)}\right)^{-2} \frac{\partial P_f^{(\Delta y)}}{\partial \theta_i} \frac{\partial P_f^{(\Delta y)}}{\partial \phi_j} \right\} \ , \tag{3.2.26}$$

where $\mathcal{I}_\theta^\phi = (\mathcal{I}_\phi^\theta)^T$.

3.2.3 Statistics of fluctuating substrates

3.2.3.1 Flat substrate

For diffusion on a fluctuating flat isotropic substrate, the movement in the x- and y-directions are equivalent and the power spectra of the measured displacements in both directions are given by

$$\begin{aligned} P_f^{(\Delta\cdot)} \simeq\ & 2D(\Delta t)^2 + [2 - 2\cos(2\pi f \Delta t)] \left(\sigma^2 \Delta t - \frac{D(\Delta t)^2}{3}\right) \\ & + \sum_{m=-M}^{M-1} \frac{2\left(\frac{1-\cos(2\pi f \Delta t)}{\pi(f\Delta t + m)}\right)^2 c_1 \mathcal{Y}_1^2}{c_1^2 + [2\pi(f + m/\Delta t)]^2} \ , \end{aligned} \tag{3.2.27}$$

while the bias of the CVE is given by

$$b\left(D\middle|\phi\right) = \frac{\left(1 - e^{-c_1\Delta t}\right)^3 \mathcal{Y}_1^2}{c_1^2 \Delta t^3} \quad . \tag{3.2.28}$$

The local degree of stretching is equal to one,

$$\zeta(\overline{s}) = 1 \quad .$$

The free parameters determining the substrate fluctuations c_1 and $\mathcal{Y}_1^2$ can be determined by recording a fluorophore fixed to the substrate, and diffusion coefficients can be estimated as described in Secs. 3.2.1 or 3.2.2.

3.2.3.2 DNA pulled by the ends

For a DNA molecule stretched by pulling at its ends, e.g., using optical tweezers [48, 49] or by fusing its ends to the coverslip [21, 50, 51, 52], the eigenfunctions of the DNA fluctuations are (App B.1.1): For $\omega_k = \pi k/L$,

$$\mathcal{Y}_k(\overline{s}) = \frac{\sqrt{2}\ \sin(\omega_k \overline{s})}{\sqrt{L\omega_k^2(\omega_k^2 l_\mathrm{p} + f_0)}} \quad , \tag{3.2.29}$$

and

$$\mathcal{X}_{k,l}(\overline{s}) = \frac{\frac{\sin[(\omega_k+\omega_l)\overline{s}]}{w_k+w_l} + \frac{\sin[(\omega_k-\omega_l)\overline{s}]}{\omega_k-\omega_l}}{L\sqrt{(\omega_k^2 l_\mathrm{p} + f_0)(\omega_l^2 l_\mathrm{p} + f_0)}} \quad . \tag{3.2.30}$$

For a DNA strand pulled by the ends, the tension is constant along the DNA. This means that the local degree of stretching is constant and equal to

$$\zeta = \frac{x(L)}{L} \quad , \tag{3.2.31}$$

where $x(L)/L$ is the overall degree of stretching of the DNA.

The free parameters determining the fluctuations are the DNA's persistence length l_p, the normalized external force $f_0 \equiv F/k_B T$, and the diffusing particle's mean position on the DNA, $\overline{s}$. The position on the DNA is determined by

$$\overline{s} = x(\overline{s})\frac{x(L)}{L} \quad .$$

Optical tweezers directly measure the DNA fluctuations with high precision and allows very precise determination of the parameters $l_\mathrm{p} \approx 50$ nm and f_0. This can be used in fits to obtain more precise diffusion coefficient estimates. For DNA fused to the coverslip at the ends, l_p and f_0 should be determined from the transverse fluctuations of particles bound to the DNA.

3.2.3.3 DNA in plug flow

For DNA stretched in a plug flow, the eigenfunctions are given by (App. B.1.2)

$$\mathcal{Y}_k(\overline{s}) = \mathcal{Z}_k(\overline{s}) = \sqrt{\frac{4k_BT}{\gamma_\parallel v}} \frac{J_0\left(\alpha_k\sqrt{1-\overline{s}/L}\right)}{\alpha_k J_1(\alpha_k)} \ , \tag{3.2.32}$$

and

$$\mathcal{X}_k(\overline{s}) = \frac{k_BT}{\gamma_\parallel v} \int_0^{\overline{s}} \frac{J_1\left(\alpha_k\sqrt{1-s/L}\right) J_1\left(\alpha_l\sqrt{1-s/L}\right)}{L(L-s)J_1(\alpha_k)J_1(\alpha_l)} \, ds \ , \tag{3.2.33}$$

The parameters determining the fluctuations are $\kappa \equiv \sqrt{k_BT/\gamma_\perp v}$ and $\overline{s}$. DNA fluctuations can be probed by attaching a fixed fluorescent marker to the DNA molecule and observing its motion, thus determining κ. The position $\overline{s}$, however, needs to be fitted individually, since $x(s)$ is not well defined mathematically (App. B.1.2).

For a DNA molecule in a strong plug flow, $\zeta(\overline{s}) \simeq 1$, except near the free end of the DNA, where we cannot estimate diffusion coefficients reliably.

3.2.3.4 DNA in shear flow

Perhaps the most simple way of performing TIRF microscopy measurements of diffusion on DNA is by attaching one end of a DNA molecule to the coverslip and stretching it by applying a shear flow [19, 25, 32, 53, 54].

We do not have a linear theory for DNA in shear flow and cannot describe the DNA movement globally (App. B.1.3). We can however fit to an effective one-mode theory locally, where the DNA motion parameters necessarily must be estimated as well. We let

$$\boldsymbol{\phi} = (c_1, \mathcal{X}_{1,1}, \mathcal{Y}_1)^T$$

be free parameters and estimate $(D, \sigma^2, \boldsymbol{\phi})$ directly as explained in Sec. 3.2.1.

For DNA in a strong shear flow $\zeta(\overline{s}) \simeq 1$, except near the free end, where measurements should be discarded.

Chapter 4

Conclusions

(i) We have derived a computationally simple expression for the exact likelihood function for single-particle tracking measurements of Brownian motion valid for short time-series. This allowed us to derive an efficient algorithm for the exact maximum likelihood estimator (MLE) along with an expression for the Cramér-Rao bound, which bounds the precision of any unbiased estimator. (ii) We propose an unbiased covariance-based estimator (CVE) and show that it is optimal in the sense that it is unbiased and reaches the Cramér-Rao bound for experimentally reasonable values of the signal-to-noise ratio (SNR), i.e., SNR > 1. (iii) We have shown that the standard MSD-based estimators are suboptimal and that their precision may be many times lower than the precision the CVE and MLE. (iv) Using Monte-Carlo simulations, we showed that the statistical error of the MLE is actually smaller than the Cramér-Rao bound, but at the cost of a small bias, while the ultimate least-squares estimator based on the MSDs, the GLS estimator, is suboptimal. (v) We found that motion blur increases the positional noise and (vi) that the variance of diffusion coefficient estimates can be reduced by a factor of up to three by estimating the noise amplitude independently. (vii) We have derived the statistics of single particle tracking of a particle diffusing on a fluctuating substrate, such as a stretched DNA molecule. (viii) From these we have derived the Cramér Rao bound and a MLE for diffusion on a fluctuating substrate, optimal for long time-series. (ix) For short time-series we propose an unbiased covariance-based estimator, which gives precise estimates when the MLE fails. (x) We showed from both simulated and experimental data that substrate fluctuations can induce an important bias in diffusion coefficient estimates if they are not taken into account. (xi) We have applied our methods to experimental data of hOgg1 proteins diffusing on DNA and shown that we are able to see molecular individuality. (xii) The superior resolution that our

methods offer and the ability to analyze short time-series has allowed us to show that hOgg1 proteins have multiple diffusion modes on DNA.

4.1 Outlook

We apply the methods we have developed to data of proteins diffusing on DNA stretched by a shear flow, but the methods work with alternative methods for stretching as well. In particular, stretching the DNA strand using optical tweezers allows for direct and precise measurements of the DNA fluctuations, which in turn increases the precision of diffusion coefficient estimates considerably. Additionally, the degree of extension can be controlled precisely with optical tweezers. This makes it possible to determine the optimal degree of extension of the DNA, by varying the force applied by the tweezers. In practice, we want the DNA to be stretched as much as possible to give an optimal signal-to-noise ratio when measuring, while we do not want the stretching to alter the protein-DNA interactions.

We expect that the methods presented in this thesis can readily be extended to diffusion coefficient estimation of molecules diffusing in lipid membranes, where substrate fluctuations certainly are present [55]. We expect that our methods work well for diffusion coefficient estimation of a particle in a crowded or confined environments, since they only rely on the shortest displacements and thus are less affected by crowding effects. An important generalization of this work would be to consider anomalous diffusion, which occurs in many biological processes, and where data analysis is traditionally performed using MSD-based methods [56].

Analysis of additional experimental data combined with Monte Carlo simulations of candidate theories can shed light on the specific mechanism hOgg1 employs when diffusing on DNA. Possible models are, e.g., a simple two-state constant rate model for diffusion as suggested in [21] with an added coupling between diffusion coefficient and characteristic residence time, or a two-state protein, which interacts specifically with DNA base-pairs as suggested by Slutsky and Mirny [20].

Appendix A

Diffusion coefficient estimation for free diffusion

A.1 Experimental tracking of a diffusing particle

In this section we review the mathematical properties of free diffusion (Brownian motion). We derive expressions for the statistics of of experimental time-series measurements of Brownian motion, where we account for positional noise due to diffraction and motion blur due to finite camera shutter time.

A.1.1 Brownian motion and the Wiener process

Brownian motion is described mathematically by the Wiener process. For a particle undergoing Brownian motion with diffusion coefficient D, the position of the particle is given by $x(t) = \sqrt{2D}W_t + x_0$, where W_t is the Wiener process. The Wiener process W_t has the following defining properties [57]:

i) W_t has independent increments

ii) The increments of W_t are normal-distributed with $W_t - W_s \sim \mathcal{N}(0, t-s)$.

iii) W_t is almost surely continuous.

iv) $W_0 = 0$.

We shall extensively use i) and ii), while iii) and iv) serve to uniquely define W_t.

A.1.2 Motion blur and positional noise

Pictures taken by a camera are not instantaneous snapshots. The camera shutter stays open for a finite time-interval to allow collection of enough photons for a decent image. This creates motion blur as we know it from photos taken with a hand held camera in low light conditions. If the shutter time is small compared to the time-lapse the relative effect of motion blur is small and can safely be ignored. In fluorescence microscopy, however, the number of photons that a fluorescent molecule can emit is limited. So one wants to maximize the number of photons collected. This is usually done by leaving the shutter open during the whole time-lapse. In this case the camera integrates the particle's position over the full time-lapse and we need to take motion blur into account. Furthermore, diffraction in microscope optics means that we do not measure the position of the diffusing particle, but a diffuse point spread function (PSF), approximately 200 nm wide. The average position can be estimated by fitting a 2D Gaussian function plus a constant background term to the PSF. This leads to an error σ on position determination, which is much smaller than the width of the PSF [44].

Combining these effects we find that the position x_n of a fluorescent molecule recorded with time-lapse photography at time t_n is

$$x_n = \int_0^{\Delta t} \varsigma(t) x_{\text{true}}(t_n - t')dt' + \sigma\xi_n \ , \tag{A.1.1}$$

where ξ_n is standard white noise and $\varsigma(t)$ is a generic shutter function, with $\int_0^{\Delta t} \varsigma(t)dt = 1$. Since x_{true} has normal distributed increments (property ii)) and ξ_n is normal distributed, the measured displacements $\Delta x_n = x_n - x_{n-1}$ are normal distributed and their probability distribution is completely characterized by the first two moments, $\langle \Delta x_n \rangle = 0$, and $(\Sigma_{\Delta x})_{ij} = \langle \Delta x_i \Delta x_j \rangle$. Using properties i) and ii) of Brownian motion, we find [18],

$$(\Sigma_{\Delta x})_{ij} = [2D\Delta t(1 - 2R) + 2\sigma^2]\delta_{i,j} + [2RD\Delta t - \sigma^2]\delta_{i,j\pm 1} \ , \tag{2.1.4}$$

where the motion blur coefficient is

$$R = \frac{1}{\Delta t} \int_0^{\Delta t} S(t)[1 - S(t)]dt \ , \tag{A.1.2}$$

and $S(t) = \int_0^t \varsigma(t')dt'$ [18]. For instantaneous (delta-function) camera shutter $R = 0$, and for full time-frame averaging $R = 1/6$. The maximal motion blur coefficient is $R = 1/4$, which corresponds to a double pulse shutter, i.e., two instantaneous snapshots of the particle, one at the start and end of each frame.

A.1.3 Width of the point spread function and positional noise amplitude

The diffraction-limited point spread function (PSF) emitted by a freely rotating fluorescent molecule or a fluorescent bead recorded by an EMCCD camera is well approximated by a two-dimensional (2D) Gaussian function plus a constant background term. For an isolated fluorophore of this kind with fixed position, fitting a 2D Gaussian plus a constant to the PSF allows us to estimate the position of the molecule much more precisely than the width of the PSF. This results in a positional noise with variance [44]

$$\sigma_0^2 = \frac{s_a^2}{P}\left(\frac{16}{9} + \frac{8\pi s_a^2 b^2}{Pa^2}\right) \quad , \tag{A.1.3}$$

for camera pixel width a, background b, measured number of photons P and $s_a^2 = s_0^2 + a^2/12$, where s_0 is the width of the PSF ($s_0 \approx 100$ nm).

For a diffusing fluorophore of this kind, the motion blur due to the finite camera shutter-time makes the measured PSF wider [15, 58]. For a particle diffusing only in the image plane, the width of the distribution of the particle's positions is $s_D = \sqrt{2RD\Delta t}$ (Sec. A.1.3.1). Since the PSF of a fixed fluorophore and the position distribution of a diffusing particle are both Gaussian and are independent the effective width s of the PSF is given by

$$s^2 = s_0^2 + 2RD\Delta t \ . \tag{A.1.4}$$

It should be noted that since diffusion is not a stationary process, the contribution to the PSF from the diffusive movement is only symmetrical on average, not for a single experiment. However, as long as the typical diffusion length $\sqrt{2D\Delta t}$ is smaller than or comparable to the width of the PSF of a fixed fluorophore, s_0, treating the measured PSF as circular is a good approximation [59].

A.1.3.1 Derivation of the PSF width

A freely diffusing fluorescent molecule emits photons as a Possion process, i.e., with a fixed rate. For the moment we forget about the dispersion of the photons due to diffraction in the microscope, which we assume is independent of the particle's position during the time-lapse Δt and can thus be added later by convoluting the PSF with the position distribution of the diffusing particle[1]. When a photon is emitted we can thus record the particle's position x_i. Since the photon emission process is independent of the particle's position and we only are interested in average behavior, we can assume that the x_is are

[1]This assumption is good when the diffusion length $\sqrt{2D\Delta t}$ is much smaller than the microscope's field-of-view.

equidistant in time. The width of the recorded position distribution is thus given by the sample variance

$$\mathrm{var}(x) = \frac{1}{P}\sum_i \varsigma_i (x_i - \overline{x})^2 \ ,$$

where P is the total number of photons recorded, s_i is the discrete shutter function, which determines whether the photon emitted at time t_i is recorded, and $\overline{x}$ is the average position. Since P is large[2] the sum is well approximated by an integral and the average sample variance is

$$s_D^2 = \int_0^{\Delta t} \varsigma(t) \left\langle (x(t) - \overline{x})^2 \right\rangle dt \ , \tag{A.1.5}$$

with $\overline{x} = \int_0^{\Delta t} \varsigma(t) x(t) dt$. We insert the expectation value $\langle x(t)x(t')\rangle = 2D\min(t,t')$ into Eq. (A.1.5) and perform partial integration to get

$$\begin{aligned} s_D^2 &= 2D\left[\int_0^{\Delta t} \varsigma(t)dt - \int_0^{\Delta t} \varsigma(t) \int_0^{\Delta t} \varsigma(t')\min(t,t')dt'dt\right] \\ &= 2D\left[-\int_0^{\Delta t} \varsigma(t) \int_0^{t} \varsigma(t')t'dt'dt + \int_0^{\Delta t} \varsigma(t)S(t)tdt\right] \\ &= 2D\int_0^{\Delta t} S(t)[1 - S(t)]dt \\ &= 2RD\Delta t \ , \end{aligned} \tag{A.1.6}$$

where we have used the definition of the motion blur coefficient R, from Eq. (A.1.2).

A.2 Mean square displacement based methods

This section presents some results about the MSDs and the GLS estimator based on the MSDs. We derive the covariance matrix of the MSDs. The GLS is defined and its theoretical variance is derived.

A.2.1 Covariance matrix of the MSD

The covariance matrix Σ_ρ is derived in [15] for full time-integration ($R = 1/6$). For other values of the motion blur coefficient, Σ_ρ is derived in the same manner as described in [15]. This gives: For $m + n \leqslant N + 1$,

$$\begin{aligned} (\Sigma_\rho)_{m,n} &= \left(\frac{8mn^2}{K_n} - \frac{8n(n^2-1)}{3K_n} + \frac{4n^2(n^2-1)}{3K_mK_n}\right)\alpha^2 + \frac{16n}{K_n}\,\alpha\beta \\ &\quad + \left(\frac{8}{K_n} - \frac{4n}{K_mK_n} + \frac{4\delta_{m,n}}{K_n}\right)\beta^2 \ ; \end{aligned}$$

[2] P is typically in the range 1000-10000.

and for $m + n \geqslant N + 1$,

$$(\Sigma_\rho)_{m,n} = \left(\frac{4K_m^3}{3K_n} - \frac{16K_m^2 n}{3K_n} + \frac{(8n - 4/3)K_m}{K_n} + \frac{8mn^2}{K_n} - \frac{8n^3}{K_n} + \frac{16n}{3K_n}\right)\alpha^2 + \frac{16n}{K_n}\,\alpha\beta + \frac{4(1+\delta_{m,n})}{K_n}\,\beta^2 \;;$$

where $\alpha \equiv D\Delta t$ and $\beta \equiv \sigma^2 - 2RD\Delta t$.

A.2.2 The generalized least squares estimator

The generalized least squares (GLS) estimator is defined as

$$\begin{pmatrix}\widehat{D}\\ \widehat{\sigma}^2\end{pmatrix} = \left(\mathbf{A}^T\Sigma_\rho^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T\Sigma_\rho^{-1}\boldsymbol{\rho}\ , \tag{A.2.1}$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_N)^T$, $\mathbf{A}$ is a $2 \times N$ matrix with $A_{1,n} = 2\Delta t(n - 2R)$ and $A_{2,n} = 2$, and Σ_ρ is a weight matrix proportional to the covariance matrix of $\boldsymbol{\rho}$.

The variance of the linear GLS estimator is found by taking the expectation value of the outer product of Eq. (A.2.1),

$$\begin{aligned}\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) &= (\mathbf{A}^T\Sigma_\rho^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathrm{Var}\,(\boldsymbol{\rho})\,\mathbf{A}(\mathbf{A}^T\Sigma_\rho^{-1}\mathbf{A})^{-1}\\ &= (\mathbf{A}^T\Sigma_\rho^{-1}\mathbf{A})^{-1}\ ,\end{aligned} \tag{A.2.2}$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{D}, \widehat{\sigma}^2)^T$. The iterative GLS estimator uses and estimates $\widehat{\Sigma}_\rho$ instead of the true covariance matrix Σ_ρ. The estimated covariance matrix $\widehat{\Sigma}_\rho$ is correlated with $\boldsymbol{\rho}$, which means that variance of the iterative GLS does not reduce to $(\mathbf{A}^T\Sigma_\rho^{-1}\mathbf{A})^{-1}$ as in Eq. (A.2.2), and that the variance of the iterative GLS estimator is higher than the variance of the linear GLS estimator. The difference is not of order $\mathcal{O}_p(1/N^2)$, since $\boldsymbol{\rho}$ is not a stationary statistic, i.e., the variance of the GLS estimator does not necessarily approach Eq. (A.2.2) in the large N-limit.

A.2.2.1 Known positional noise amplitude

When σ^2 is known a priori, the GLS estimator is reduced to

$$\widehat{D} = (\mathbf{a}^T\Sigma_{\rho-2\sigma^2}^{-1}\mathbf{a})^{-1}\mathbf{a}^T\Sigma_{\rho-2\sigma^2}^{-1}(\boldsymbol{\rho} - 2\widehat{\sigma}^2)\ , \tag{A.2.3}$$

where $a_n = 2\Delta t(n - 2R)$ and

$$\Sigma_{\rho-2\sigma^2} = \mathrm{Var}\left(\boldsymbol{\rho} - 2\widehat{\sigma}^2\right) = \Sigma_\rho + 4\mathrm{Var}\left(\widehat{\sigma}^2\right)\ . \tag{A.2.4}$$

Even when the noise amplitude σ is known a priori, Σ_ρ depends non-linearly on the unknown parameter D. So one still needs to use an iterative procedure to find the GLS estimate.

A.3 Methods based on individual displacements

In this section we derive results for the displacement-based estimators: the MLE (Sec. A.3.1) and the CVE (Sec. A.3.2).

A.3.1 Likelihood-based approach

We derive a simple analytical expression of the exact likelihood function. We derive an effective algorithm for maximum likelihood estimation and investigate the bias of the MLE. We finally investigate the effect on motion blur on the Cramér-Rao bound and thus the precision of unbiased estimators, and show that, contrary to the surprising result reported in [18], the estimation precision is lowered, not increased, by increasing the motion blur.

A.3.1.1 The exact likelihood function

The covariance matrix $\Sigma_{\Delta x}$ (Eq. (2.1.4)) of the displacements is a tridiagonal Toeplitz matrix and can be transformed into a diagonal matrix $\psi = \mathbf{P}^{-1}\Sigma_{\Delta x}\mathbf{P}$ by the orthogonal matrix $\mathbf{P}$, which is given by

$$P_{ij} = \sqrt{\frac{2}{N+1}}\ \sin\left[\frac{\pi i j}{N+1}\right] \quad . \tag{A.3.1}$$

Thus, the transformed displacements $\widetilde{\mathbf{\Delta x}}$, given by Eq. (3.1.11) are independent and Gaussian distributed with variances given by Eq. (3.1.12), and the exact log-likelihood function for the parameters D and σ^2, given the data Δx, reduces to Eq. (3.1.13).

A.3.1.2 Effective algorithm for the MLE

Finding the MLE of D and σ^2 is a two-dimensional optimization problem, but it can be reduced to a one-dimensional problem by using the scale-invariance of diffusion. We can rewrite the log-likelihood by defining the following new parameters:

$$\lambda\cos^2\phi = 2D\Delta t\ , \quad \lambda\sin^2\phi = 2\sigma^2 \quad . \tag{A.3.2}$$

Then $\psi_k = \lambda(\phi)F_n(\phi)$, where $F_n(\phi) = a_n\cos^2\phi + b_n\sin^2\phi$, and $a_n = 1 - 2R\left(1-\cos\frac{\pi n}{N+1}\right)$ and $b_n = 1-\cos\frac{\pi n}{N+1}$. The log-likelihood can then be rewritten as

$$\ln\mathcal{L}(\phi|\mathbf{\Delta x}) = -\frac{1}{2}\sum_{n=1}^{N}\left(\frac{\widetilde{\Delta x}_n^2}{\lambda(\phi)F_n(\phi)} + \log[\lambda(\phi)F_n(\phi)]\right) \quad , \tag{A.3.3}$$

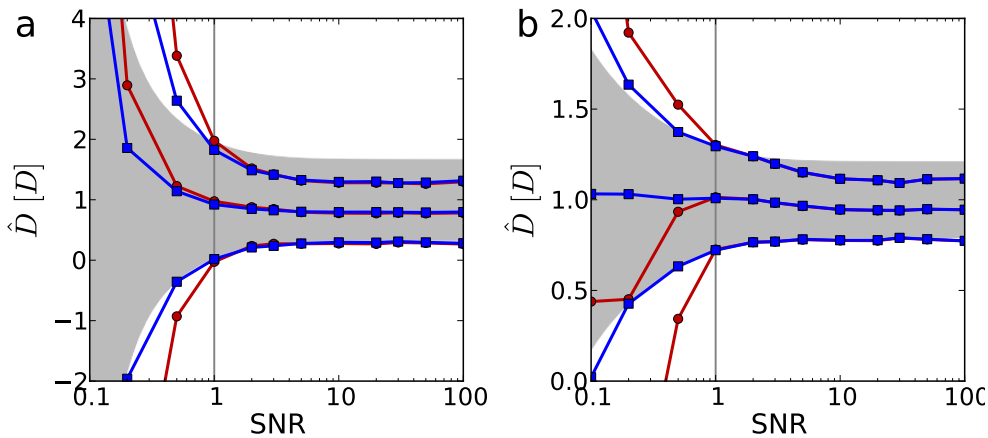

Figure A.1: Mean plus/minus standard error of the approximate and exact maximum likelihood estimates of the diffusion coefficient D (approximate MLE based on the Fourier-transformed displacements: connected circles, red; exact MLE based on the transformed displacements given by Eq. (3.1.11): connected squares, blue; Cramér-Rao bound: grey area). Results are obtained from an ensemble of 1,000 individual Monte Carlo generated time-series of lengths a) $N = 10$ and b) $N = 100$. $R = 1/6$ for both simulations, which corresponds to full time-integration. Both the exact and the approximate MLEs are biased for high SNR and approximately unbiased for low SNR. The estimators are approximately equal for SNR > 1, while the exact MLE converges much faster to the Cramér-Rao bound than the approximate MLE for SNR < 1.

where λ is given by the stationarity condition

$$\lambda(\phi) = \frac{1}{N}\sum_{n=1}^{N}\frac{\widetilde{\Delta x}_n^{\,2}}{F_n(\phi)} \quad . \tag{A.3.4}$$

The one-dimensional optimization problem is considerably easier to solve than the two-dimensional problem. Each calculation of the likelihood function takes longer, however, since it now involves a double sum and calls to transcendental functions. In practice this means that reducing the dimension of the optimization problem speeds the MLE-algorithm up by a modest factor of approximately two.

When σ^2 is known a priori it is faster to minimize Eq. (3.1.13) directly as described in Sec. 3.1.2.3, since this does not involve calls to transcendental functions.

A.3.1.3 Bias of the MLE

The numerical simulations show that the MLE is biased (Figs. A.1 and 2.2). We here investigate the possible sources of this bias. The MLE may be biased due to skewness of the likelihood function. By Taylor expansion of the stationarity condition $\partial_{\theta_i}\ln[\mathcal{L}(\widehat{\boldsymbol{\theta}})] = \mathbf{0}$, where ∂_{θ_i} is the partial derivative w.r.t. θ_i, we can derive an approximation of the bias of the MLE. To order $1/N$ this bias is

$$b(\widehat{\boldsymbol{\theta}})_i = -\sum_{jkm}\mathcal{I}^{ij}\mathcal{I}^{km}\left(\left\langle\partial_{\theta_j}\partial_{\theta_k}\ln[\mathcal{L}(\boldsymbol{\theta})]\partial_{\theta_m}\ln[\mathcal{L}(\boldsymbol{\theta})]\right\rangle + \frac{1}{2}\left\langle\partial_{\theta_j}\partial_{\theta_k}\partial_{\theta_m}\ln[\mathcal{L}(\boldsymbol{\theta})]\right\rangle\right) \quad , \tag{A.3.5}$$

where $\mathcal{I}^{ij} \equiv (\mathcal{I}^{-1})_{ij}$. Since $\mathcal{L}$ is Gaussian and ψ_k is a first-degree polynomial in D and σ^2,

$$\left\langle\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\ln[\mathcal{L}(\boldsymbol{\theta})]\right\rangle = -2\sum_{n=1}^{N}\frac{\partial_{\theta_i}\psi_n\partial_{\theta_j}\psi_n\partial_{\theta_k}\psi_n}{\psi_n^3}$$

and

$$\left\langle\partial_{\theta_i}\partial_{\theta_j}\ln[\mathcal{L}(\boldsymbol{\theta})]\partial_{\theta_k}\ln[\mathcal{L}(\boldsymbol{\theta})]\right\rangle = \sum_{n=1}^{N}\frac{\partial_{\theta_i}\psi_n\partial_{\theta_j}\psi_n\partial_{\theta_k}\psi_n}{\psi_n^3} \quad .$$

Thus, the bias due to skewness is zero (to order $1/N$), but we introduce a bias when we require $\widehat{D}$ and $\widehat{\sigma}^2$ to be positive. Equations (A.3.3) and (A.3.4) can help in understanding this. For a given experimental realization of the measurements $\mathbf{\Delta x}$, there is a finite probability that the maximum of $\ln\mathcal{L}$ lies outside of $\phi \in [0, \pi/2]$. This probability is not symmetric in ϕ, the maximum is more likely to be in the region $\phi < 0$ for high SNR and more likely to be in the region $\phi > \pi/2$ for low SNR. Since we require $0 < \phi < \pi/2$ we introduce positive bias in $\widehat{\sigma}^2$ and a negative bias in $\widehat{D}$ for high SNR, and vice-versa for

low SNR. Since this bias stems from the dispersion of the measured likelihood function around its true maximum, we expect this bias to be of the same order as the standard error of the estimates, i.e., it decreases as $N^{-1/2}$. This is confirmed by numerical results (Figs. A.1 and 2.2).

A.3.1.4 Motion blur's effect on estimator precision

From Eqs. (A.1.3) and (A.1.4) we have that the variance on the position determination is

$$\sigma^2 = \sigma_0^2 \left(1 + 2R\frac{\mathrm{SNR}_0^2}{P}\right) + \frac{8\pi s_a^4 b^2}{P^2 a^2}\left[2R\frac{\mathrm{SNR}_0^2}{P} + \left(2R\frac{\mathrm{SNR}_0^2}{P}\right)^2\right] \quad , \qquad \text{(A.3.6)}$$

where $\mathrm{SNR}_0 \equiv \sqrt{D\Delta t}/\sigma_0$. When the motion blur coefficient or the diffusion length $\sqrt{D\Delta t}$ is small we recover Eq. (A.1.3) as expected. When the diffusion length is high compared to the width of the fixed fluorophore PSF, s_a, Eq. (A.3.6) shows that the actual SNR may decrease with diffusion length $\sqrt{D\Delta t}$ in the presence of motion blur (see Fig. A.2a). This means that the Cramér-Rao bound is *increased*, not be lowered, by increasing the motion blur (Fig. A.2b).

The results derived in this section rely on the assumption that the same average number of photons per time frame P is recorded with different camera shutters, and accordingly, different motion blur coefficients. Experiments are usually performed with full time-integration ($R = 1/6$) in order to maximize the photon count, since it ultimately defines the precision of position determination. With modern stroboscopic techniques and efficient fluorophores, it is possible to record with almost instantaneous shutter, which is required for minimal motion blur, but, given the minimal possible gains in estimator precision obtained by reducing the shutter time (Fig. A.2b), the experimental setup should be optimized to maximize the number of photons recorded, not to reduce the motion blur.

A.3.2 Covariance-based estimator

In this section we derive the variance of the CVE and we derive an exact analytical expression for the characteristic function of the CVE.

A.3.2.1 Variance of the covariance-based estimator

The variance of the CVE of D is computed directly from Eq. (3.1.1),

$$\mathrm{Var}\left(\widehat{D}\right) = \frac{\mathrm{Var}\left(\overline{\Delta x_n^2}\right)}{4(\Delta t)^2} + \frac{\mathrm{Var}\left(\overline{\Delta x_n \Delta x_{n+1}}\right)}{(\Delta t)^2} + \frac{\left\langle \overline{\Delta x_n^2}, \overline{\Delta x_n \Delta x_{n+1}} \right\rangle_{\mathcal{C}}}{(\Delta t)^2} \quad , \quad \text{(A.3.7)}$$

where $\langle x, y \rangle_{\mathcal{C}}$ designates the covariance of x and y. The individual terms of Eq. (A.3.7) are:

$$\begin{aligned} \mathrm{Var}\left(\overline{\Delta x_n^2}\right) &= \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{Var}\left(\Delta x_n^2\right) + \frac{2}{N^2} \sum_{n=1}^{N-1} \left\langle \Delta x_n^2, \Delta x_{n+1}^2 \right\rangle_{\mathcal{C}} \\ &= \frac{8\alpha^2 + 16\alpha\beta + 12\beta^2}{N} - \frac{4\beta^2}{N^2} \quad , \end{aligned} \quad \text{(A.3.8)}$$

where we have used that $\mathrm{Var}\left(\Delta x_n^2\right) = 8(\alpha+\beta)^2$, $\left\langle \Delta x_n^2, \Delta x_{n+1}^2 \right\rangle_{\mathcal{C}} = 2\beta^2$, and $\left\langle \Delta x_n^2, \Delta x_m^2 \right\rangle_{\mathcal{C}} = 0$ for $m > n+1$, with $\alpha = D\Delta t$ and $\beta = \sigma^2 - 2RD\Delta t$ as defined in Sec. 3.1.1.1;

$$\begin{aligned} \mathrm{Var}\left(\overline{\Delta x_n \Delta x_{n+1}}\right) &= \frac{1}{(N-1)^2} \sum_{n=1}^{N-1} \mathrm{Var}\left(\Delta x_n \Delta x_{n+1}\right) \\ &\quad + \frac{2}{(N-1)^2} \sum_{n=1}^{N-2} \left\langle \Delta x_n \Delta x_{n+1}, \Delta x_{n+1} \Delta x_{n+2} \right\rangle_{\mathcal{C}} \\ &= \frac{4\alpha^2 + 8\alpha\beta + 7\beta^2}{N-1} - \frac{2\beta^2}{(N-1)^2} \quad , \end{aligned} \quad \text{(A.3.9)}$$

since $\mathrm{Var}\left(\Delta x_n \Delta x_{n+1}\right) = 4(\alpha+\beta)^2 + \beta^2$, $\left\langle \Delta x_n \Delta x_{n+1}, \Delta x_{n+1} \Delta x_{n+2} \right\rangle_{\mathcal{C}} = \beta^2$, and $\left\langle \Delta x_n \Delta x_{n+1}, \Delta x_m \Delta x_{m+1} \right\rangle_{\mathcal{C}} = 0$ for $m > n+1$;

$$\begin{aligned} \left\langle \overline{\Delta x_n^2}, \overline{\Delta x_n \Delta x_{n+1}} \right\rangle_{\mathcal{C}} &= \frac{2}{N(N-1)} \sum_{n=1}^{N-1} \left\langle \Delta x_n^2, \Delta x_n \Delta x_{n+1} \right\rangle_{\mathcal{C}} \\ &= -\frac{8\alpha\beta + 8\beta^2}{N} \quad , \end{aligned} \quad \text{(A.3.10)}$$

since $\left\langle \Delta x_n^2, \Delta x_n \Delta x_{n\pm1} \right\rangle_{\mathcal{C}} = 4(\alpha\beta - \beta^2)$ and $\left\langle \Delta x_n^2, \Delta x_m \Delta x_{m+1} \right\rangle_{\mathcal{C}} = 0$ for $m > n$.

Inserting Eqs. (A.3.8)-(A.3.10) in Eq. (A.3.7) gives Eq. (3.1.4). The variance of $\widehat{\sigma}^2$, Eq. (A.3.11), and the covariance, Eq. (A.3.12), of $\widehat{D}$ are derived in the same manner and are, to second order in $1/N$,

$$\begin{aligned} \mathrm{Var}\left(\widehat{\sigma}^2\right) &= \frac{4(1-4R+6R^2)\alpha^2 + 8(1-2R+2R^2)\alpha\beta + (7-12R+8R^2)\beta^2}{N} \\ &\quad + \frac{4(1-2R)^2\alpha^2 + 8(1-2R)^2\alpha\beta + (5-20R+16R^2)\beta^2}{N^2} \end{aligned} \quad \text{(A.3.11)}$$

and

$$\begin{aligned}\mathrm{Cov}\left(\widehat{D},\widehat{\sigma}^2\right) &= -\frac{(1-2R)(4\alpha^2+4\alpha\beta+3\beta^2)+2\alpha^2-\beta^2}{N\Delta t} \\ &\quad -\frac{4(1-2R)(\alpha+\beta)^2+\beta^2}{N^2\Delta t} \ . \end{aligned} \tag{A.3.12}$$

The standard error of the CVE is compared to the Cramér-Rao bound in Fig. A.3a.

Known positional noise amplitude For known positional noise amplitude, the variance of $\widehat{D}$ is

$$\mathrm{Var}\left(\widehat{D}\right) = \frac{\mathrm{Var}\left(\overline{\Delta x_n^2}\right)}{4(\Delta t)^2(1-2R)^2} + \frac{\mathrm{Var}\left(\widehat{\sigma}^2\right)}{(\Delta t)^2(1-2R)^2}$$

and Eq. (3.1.19) is found by inserting Eq. (A.3.8) above.

A.3.2.2 Characteristic function of the CVE

We derive the characteristic function of the covariance-based estimator of D. By the definition of the CVE of D and the defining property of the Dirac-delta function we have the trivial equality

$$p(\widehat{D}|D,\sigma^2) = \int p(\mathbf{\Delta x}|D,\sigma^2)\ \delta\left(\widehat{D} - \frac{\overline{d_i^2}+2\overline{d_i d_{i+1}}}{2\Delta t}\right)\mathcal{D}\mathbf{\Delta x}\ .$$

We use that the Dirac-delta function can be represented as

$$\delta(x-x_0) = \int_{-\infty}^{\infty}\frac{e^{i\omega(x-x_0)}}{2\pi}\ d\omega\ ,$$

and that

$$\frac{1}{2}\overline{d_i^2} = \frac{1}{N}\mathbf{\Delta x}^T\mathbb{I}\mathbf{\Delta x}\ , \quad \overline{d_i d_{i+1}} = \frac{1}{N-1}\mathbf{\Delta x}^T\mathbb{C}\mathbf{\Delta x}\ ,$$

where $\mathbb{C} = \delta_{i,j\pm1}$. This gives

$$\begin{aligned} p(\widehat{D}|D,\sigma^2) &= \int_{-\infty}^{\infty}\frac{e^{i\omega\widehat{D}}}{2\pi}\int p(\mathbf{\Delta x}|D,\sigma^2)e^{-\frac{i\omega}{2\Delta t}\mathbf{\Delta x}^T\left(\frac{\mathbb{I}}{N}+\frac{\mathbb{C}}{N-1}\right)\mathbf{\Delta x}}\mathcal{D}\mathbf{\Delta x}\ d\omega \\ &= \int_{-\infty}^{\infty}\frac{e^{i\omega\widehat{D}}}{2\pi}\int\frac{e^{-\frac{1}{2}\mathbf{\Delta x}^T\left\{\Sigma_d^{-1}+\frac{i\omega}{\Delta t}\left(\frac{\mathbb{I}}{N}+\frac{\mathbb{C}}{N-1}\right)\right\}\mathbf{\Delta x}}}{(2\pi)^{N/2}\sqrt{\det\Sigma_d}}\ \mathcal{D}\mathbf{\Delta x}\ d\omega\ . \end{aligned}$$

We use that both Σ_d and $\mathbb{C}$ can be diagonalized using the orthogonal transformation matrix $\mathbf{P}$. We define

$$\lambda_{\mathbb{C}}(n) = [\mathbf{P}^{-1}\mathbb{C}\mathbf{P}]_{nn} = 2\cos\theta_n$$

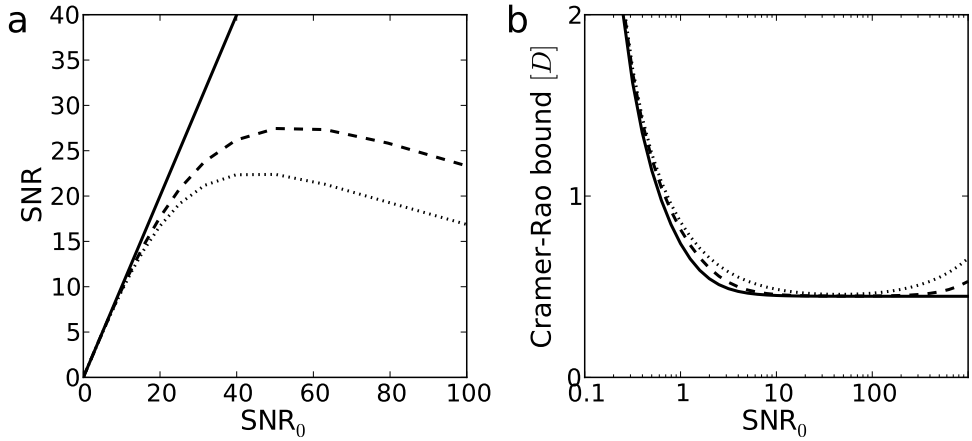

Figure A.2: a) Actual SNR as a function of $\mathrm{SNR}_0 = \sqrt{D\Delta t}/\sigma_0$ for typical parameter values of the camera setup, $a = 20$ nm, $P = 1000$, $b = 10$ and $s_a = 100$ nm ($R = 0$: full line; $R = 1/6$: dashed line; $R = 1/4$: dotted line). The SNR is not an increasing function of diffusion length per time-frame $\sqrt{D\Delta t}$ in the presence of motion blur. b) Cramér-Rao bound as a function of SNR_0 for $N = 10$. The contribution of motion blur to the positional noise means that the Cramér-Rao bound increases with SNR_0 for high SNR_0. Changing the parameters may change specificities of the curves but not their characteristic form nor their interrelationship.

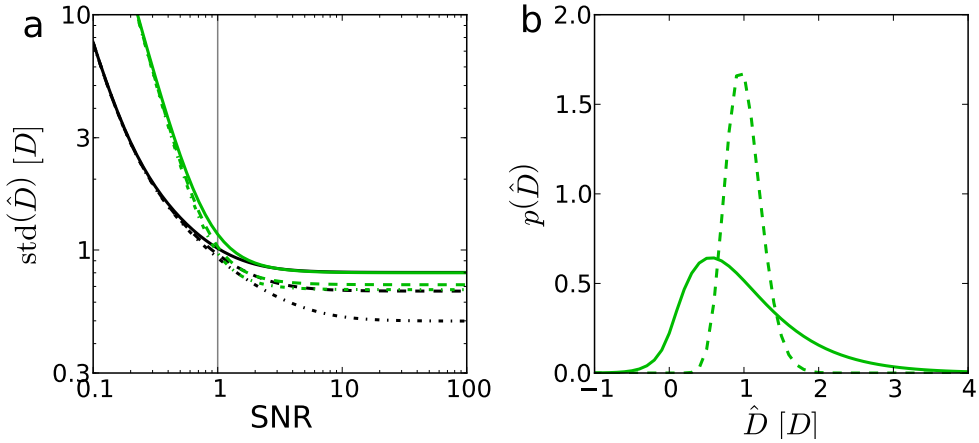

Figure A.3: (a) Standard deviation of the CVE (green) compared to the Cramér-Rao bound (black) for different motion blur coefficients ($R = 0$: full lines; $R = 1/6$: dashed lines; $R = 1/4$: dotted lines). Time-series length is equal to $N = 10$. The CVE practically reaches the Cramér-Rao bound for SNR > 1 except for very high motion blur $R \approx 1/4$, where its standard deviation is approximately 30% higher than the Cramér-Rao bound. (b) Distribution of the CVE for SNR $= 2$ and $R = 1/6$ ($N = 10$: full line; $N = 100$: dashed line). The distribution of the CVE approaches a Gaussian distribution as N increases.

and

$$\lambda_\Sigma(n) = [\mathbf{P}^{-1}\Sigma_d\mathbf{P}]_{nn} = 2D\Delta t + 2(\sigma^2 - 2RD\Delta t)\,(1-\cos\theta_n) \quad ,$$

where $\theta_n = \frac{\pi n}{N+1}$. Then,

$$\begin{aligned} p(\widehat{D}|D,\sigma^2) &= \int_{-\infty}^{\infty} \frac{e^{i\omega\widehat{D}}}{2\pi} \prod_{n=1}^{N} \int \frac{e^{-\frac{1}{2}\left\{\lambda_\Sigma(n)^{-1}+\frac{i\omega}{\Delta t}\left(\frac{1}{N}+\frac{\lambda_{\mathbb{C}}(n)}{N-1}\right)\right\}\widetilde{\Delta x}_n^2}}{\sqrt{2\pi\lambda_\Sigma(n)}}\, \mathcal{D}\widetilde{\mathbf{\Delta x}} d\omega \\ &= \int_{-\infty}^{\infty} \frac{e^{i\omega\widehat{D}}\; d\omega}{2\pi\sqrt{\prod_{n=1}^{N}\lambda_n(\omega)}} \quad , \end{aligned}$$

where

$$\lambda_n(\omega) = 1 + \frac{i\omega}{\Delta t}\lambda_\Sigma(n)\left(\frac{1}{N} + \frac{\lambda_{\mathbb{C}}(n)}{N-1}\right) \quad . \tag{A.3.13}$$

The characteristic function of the CVE, $\widetilde{p}$, is defined as the Fourier transform of the probability density. So

$$\widetilde{p}(\omega) = \prod_{n=1}^{N} \{\lambda_n(\omega)\}^{-\frac{1}{2}} \quad . \tag{A.3.14}$$

From Eq. (A.3.13) we see that λ_n is a second-order polynomial in $\cos\theta_n$. So it can be written on the form

$$\lambda_n = A + B\cos\theta_n + C\cos^2\theta_n \quad ,$$

with $A(\omega) = 1 + 2\alpha(D+Q)/N$, $B(\omega) = 2\alpha[2D + (1+1/N)Q]/(N-1)$, and $C(\omega) = -4\alpha Q/N$, where $Q = (s^2 - 2RD\Delta t)/\Delta t$. The logarithm of the characteristic function is thus given by

$$-2\ln[\widetilde{p}] = \sum_{n=1}^{N} \ln[A + B\cos\theta_n + C\cos^2\theta_n] \quad . \tag{A.3.15}$$

We use that

$$\sum_{n=1}^{N} f(\cos\theta_n) = -\frac{1}{2}(f(1)+f(-1)) + \frac{1}{2}\sum_{n=1}^{2N+2} f(\cos\theta_n)$$

to rewrite Eq. (A.3.15),

$$-2\ln[\widetilde{p}] = -\frac{1}{2}\ln[A+B+C] - \frac{1}{2}\ln[A-B+C] + \frac{1}{2}\sum_{n=1}^{2N+2}\ln[\lambda_n] \quad . \tag{A.3.16}$$

The last term can be further simplified,

$$\begin{aligned}\frac{1}{2}\sum_{n=1}^{2N+2}\ln[\lambda_n] &= \frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[C\left(\cos\theta_n+\frac{a_+}{2C}\right)\left(\cos\theta_n+\frac{a_-}{2C}\right)\right] \\ &= -(N+1)\ln[4C]+\frac{1}{2}\sum_{n=1}^{2N+2}\{\ln[a_+ + 2C\cos\theta_n]+\ln[a_- + 2C\cos\theta_n]\} \quad ,\end{aligned} \tag{A.3.17}$$

where we have defined

$$a_\pm = B \pm \sqrt{B^2-4AC} \quad .$$

To simplify the last two terms of Eq. (A.3.17) we define the sum

$$\begin{aligned}S(a,b) &= \frac{1}{2}\sum_{n=1}^{2N+2}\ln[a+b\cos\theta_n] \\ &= \frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[a+\frac{b}{2}\left(e^{i\theta_n}+e^{-i\theta_n}\right)\right] \\ &= \frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[c\left(\beta+e^{i\theta_n}\right)\left(\beta+e^{-i\theta_n}\right)\right] \quad ,\end{aligned}$$

with $c\beta = \frac{b}{2}$ and $c(1+\beta^2) = a$. This means that

$$c = \frac{b}{2\beta} \tag{A.3.18}$$

and

$$\beta_\pm = \frac{a}{b} \pm \sqrt{\left(\frac{a}{b}\right)^2 - 1} \quad , \tag{A.3.19}$$

Using Eqs. (A.3.18) and (A.3.19), we rewrite $S(a,b)$,

$$\begin{aligned}S(a,b) &= (N+1)\ln\left[\frac{b}{2\beta_\pm}\right]+\frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[\beta_\pm+e^{i\theta_n}\right]+\frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[\beta_\pm+e^{-i\theta_n}\right] \\ &= (N+1)\ln\left[\frac{b\beta_\pm}{2}\right]+\frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[1+\frac{e^{i\theta_n}}{\beta_\pm}\right]+\frac{1}{2}\sum_{n=1}^{2N+2}\ln\left[1+\frac{e^{-i\theta_n}}{\beta_\pm}\right] \quad .\end{aligned} \tag{A.3.20}$$

The two last terms in Eq. (A.3.20) can be rewritten as

$$\begin{aligned}\sum_{n=1}^{2N+2}\ln\left[1+\frac{e^{\pm i\theta_n}}{\beta_\pm}\right] &= -\sum_{n=1}^{2N+2}\sum_{k=1}^{\infty}\frac{\left(-\frac{e^{\pm i\theta_n}}{\beta_\pm}\right)^k}{k} \\ &= -\sum_{k=1}^{\infty}\frac{(-\beta_\pm)^{-k}}{k}\sum_{n=1}^{2N+2}e^{\pm ik\theta_n} \quad .\end{aligned} \tag{A.3.21}$$

The sum $\sum_{n=1}^{2N+2} e^{\pm ik\theta_n}$ is equal to zero because its argument makes a full circle in the complex plane, except when k is equal to an integer times $2N+2$, i.e.,

$$\sum_{n=1}^{2N+2} e^{\pm ik\theta_n} = \begin{cases} 2N+2 & \text{for} \quad k = 2(N+1)l \ , \quad l \in \mathbb{N} \\ 0 & \text{else} \end{cases} \ .$$

Thus,

$$\begin{aligned} \sum_{n=1}^{2N+2} \ln\left[1 + \frac{e^{\pm i\theta_n}}{\beta_\pm}\right] &= -2(N+1)\sum_{l=1}^{\infty} \frac{\left((-\beta_\pm)^{-2(N+1)}\right)}{2(N+1)l} \\ &= \ln\left[1 - \beta_\pm^{-2(N+1)}\right] \ . \end{aligned} \tag{A.3.22}$$

Inserting Eq. (A.3.22) in Eq. (A.3.20) gives

$$\begin{aligned} S(a,b) &= (N+1)\ln\left[\frac{b\beta_\pm}{2}\right] + \ln\left[1 - \beta_\pm^{-2(N+1)}\right] \\ &= -(N+1)\ln 2 + (N+1)\ln\left[a \pm \sqrt{a^2-b^2}\right] \\ &\quad + \ln\left[1 - \left(\frac{a \mp \sqrt{a^2-b^2}}{a \pm \sqrt{a^2-b^2}}\right)^{N+1}\right] \\ &= -(N+1)\ln 2 + \ln\left[\left(a \pm \sqrt{a^2-b^2}\right)^{N+1} - \left(a \mp \sqrt{a^2-b^2}\right)^{N+1}\right] \ , \end{aligned}$$

where we have used that $\beta_+\beta_- = 1$.

We insert this result into Eq. (A.3.17), using Eq. (A.3.16), to get

$$\begin{aligned} \ln[\tilde{p}] &= \frac{1}{2}(N+1)\ln[4C] + \frac{1}{4}\ln\left[(A+C)^2 - B^2\right] - \frac{1}{2}S(a_+, 2C) - \frac{1}{2}S(a_-, 2C) \\ &= \frac{1}{2}(N+1)\ln[16C] + \frac{1}{4}\ln\left[(A+C)^2 - B^2\right] \\ &\quad -\frac{1}{2}\ln\left[\left(a_+ + \sqrt{a_+^2 - 4C^2}\right)^{N+1} - \left(a_+ - \sqrt{a_+^2 - 4C^2}\right)^{N+1}\right] \\ &\quad -\frac{1}{2}\ln\left[\left(a_- + \sqrt{a_-^2 - 4C^2}\right)^{N+1} - \left(a_- - \sqrt{a_-^2 - 4C^2}\right)^{N+1}\right] \end{aligned} \tag{A.3.23}$$

with $a_\pm = B \pm \sqrt{B^2 - 4AC}$.

A.3.2.3 Computer-friendly expression

While Eq. (A.3.23) is relatively simple and good to work with analytically, it is not suitable for numerical analysis. In this subsection we look into the cases that give analytical problems and change the expression to cope with these.

Zero "measurement" noise When $\sigma^2 = 2RD\Delta t$, $Q = 0$ and the expressions for A, B, and C simplify to

$$A = 1 + \frac{2\omega D}{N} , \tag{A.3.24}$$

$$B = \frac{4\omega D}{N-1} , \tag{A.3.25}$$

$$C = 0 . \tag{A.3.26}$$

When C approaches zero it gives rise to numerical problems in Eq. (A.3.23), since $\ln C$ tends to infinity. This problem can be solved by using that $4C = a_+ a_-/A$ and thus also $2C/a_\pm = a_\mp/2A$. So we can rewrite $\ln \widetilde{p}$ as

$$\begin{aligned}
\ln \widetilde{p} &= \frac{N+1}{2} \ln\left[\frac{4}{A}\right] + \frac{1}{4} \ln\left[(A+C)^2 - B^2\right] \\
&\quad -\frac{1}{2} \ln\left[\left(1+\sqrt{1-\left(\frac{2C}{a_+}\right)^2}\right)^{N+1} - \left(1-\sqrt{1-\left(\frac{2C}{a_+}\right)^2}\right)^{N+1}\right] \\
&\quad -\frac{1}{2} \ln\left[\left(1+\sqrt{1-\left(\frac{2C}{a_-}\right)^2}\right)^{N+1} - \left(1-\sqrt{1-\left(\frac{2C}{a_-}\right)^2}\right)^{N+1}\right] \\
&= \frac{N+1}{2} \ln\left[\frac{4}{A}\right] + \frac{1}{4} \ln\left[(A+C)^2 - B^2\right] \qquad \text{(A.3.27)} \\
&\quad -\frac{1}{2} \ln\left[\left(1+\sqrt{1-\left(\frac{a_-}{2A}\right)^2}\right)^{N+1} - \left(1-\sqrt{1-\left(\frac{a_-}{2A}\right)^2}\right)^{N+1}\right] \\
&\quad -\frac{1}{2} \ln\left[\left(1+\sqrt{1-\left(\frac{a_+}{2A}\right)^2}\right)^{N+1} - \left(1-\sqrt{1-\left(\frac{a_+}{2A}\right)^2}\right)^{N+1}\right] .
\end{aligned}$$

Zero frequency When ω approaches zero, so does B and C, i.e., $A \to 1$, $B \to 0$, $C \to 0$ and consequently $a_\pm \to 0$, but the rearrangement made in the previous paragraph to take care of numerical problems when $Q = 0$ also takes care of this problem, since

$$\begin{aligned}
\ln \widetilde{p} &= \frac{N+1}{2} \ln 4 + \frac{1}{4} \ln 1 - \frac{1}{2} \ln\left[2^{N+1}\right] - \frac{1}{2} \ln\left[2^{N+1}\right] \\
&= 0 , \qquad \text{(A.3.28)}
\end{aligned}$$

and thus

$$\widetilde{p}(0) = \int_{-\infty}^{\infty} p\left(\widehat{D}|D, \sigma^2, t\right) d\widehat{D} = 1$$

as it should be.

Handling complex numbers numerically and the square root. The value of a non-algebraic function of a complex variable is calculated numerically by rewriting the complex number $z = x + iy$ as $z = |z|e^{i\theta}$, e.g., $\sqrt{z} = \sqrt{|z|}e^{i\frac{1}{2}\theta}$, $z^q = |z|^q e^{iq\theta}$ or $\ln z = \ln|z| + i\theta$. The angle θ is not uniquely defined, every angle $\theta + 2\pi n$ is equivalent. The computer handles this ambiguity by requiring the angle θ to lie in the interval $]-\pi, \pi]$. This means that if we want to take the square root of a variable that changes angle fast, e.g., $z = x^{N+1}$, which will for N big rotate a lot even if x only changes angle a little, we will end up with a numerical calculation that shows $\sqrt{z}$ suddenly changing angle between $-\pi/2$ and $\pi/2$, due to the fact that the angle of z is constrained to the interval $]-\pi, \pi]$.

We should calculate $\widetilde{p}$ as the exponential of a sum of logarithmic terms to avoid this effect. Furthermore, to avoid other possible numerical complications due to the angle of some terms changing very fast, we rewrite Eq. (A.3.27) as

$$\begin{aligned}
\ln\widetilde{p} &= \frac{N+1}{2}\ln\left[\frac{4}{A}\right] + \frac{1}{4}\ln\left[(A+C)^2 - B^2\right] \\
&\quad -\frac{N+1}{2}\ln\left[1+\sqrt{1-\left(\frac{a_-}{2A}\right)^2}\right] - \frac{N+1}{2}\ln\left[1+\sqrt{1-\left(\frac{a_+}{2A}\right)^2}\right] \\
&\quad -\frac{1}{2}\ln\left[1-\left(\frac{1-\sqrt{1-\left(\frac{a_-}{2A}\right)^2}}{1+\sqrt{1-\left(\frac{a_-}{2A}\right)^2}}\right)^{N+1}\right] - \frac{1}{2}\ln\left[1-\left(\frac{1-\sqrt{1-\left(\frac{a_+}{2A}\right)^2}}{1+\sqrt{1-\left(\frac{a_+}{2A}\right)^2}}\right)^{N+1}\right] \\
&\equiv \frac{N+1}{2}\ln\left[\frac{4}{A}\right] + \frac{1}{4}\ln\left[(A+C)^2 - B^2\right] - \frac{1}{2}\left(S_{1,+} + S_{1,-} + S_{2,+} + S_{2,-}\right)
\end{aligned}$$

. (A.3.29)

Here $S_{2,\pm}$ is very small, so a fast change in angle of this term does not give any numerical problems (Fig. A.4), and $S_{1,\pm}$ changes angle slowly.

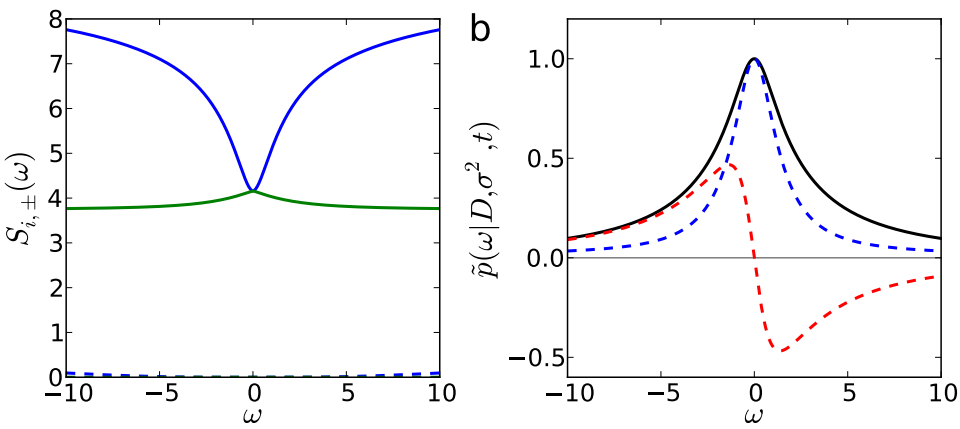

Figure A.4: a) Absolute values of $S_{i,\pm}$ as a function of ω. ($S_{1,+}$: full line, blue; $S_{1,-}$: full line, green; $S_{2,+}$: dashed line, blue; $S_{2,-}$: dashed line, green.) $S_{2,\pm}$ are much smaller than $S_{1,\pm}$. b) Characteristic function $\widetilde{p}$ of the CVE as a function of ω. (Absolute value of $\widetilde{p}$: full line, black; real part of $\widetilde{p}$: dashed line, blue; imaginary part of $\widetilde{p}$: dashed line, red.) Parameter values for both a) and b) are $N = 5$ and $Q = 0$.

Appendix B

Diffusion coefficient estimation on a fluctuating substrate

B.1 Spectral decomposition of DNA fluctuations

In this section we derive analytical expressions of the thermal movement of a flexible or semiflexible polymer, such as DNA, using spectral (eigenfunction) decomposition. We derive general expressions of the statistics of the fluctuations of a stretched DNA molecule based on any linear theory. We apply the general framework to three different cases of DNA stretched by external forces: DNA pulled by the ends (Sec. B.1.1); DNA in plug flow (Sec. B.1.2); and DNA in shear flow (Sec. B.1.3). The results derived in general hold for any semiflexible or flexible polymer unless otherwise specified.

The DNA's movement is strongly overdamped. So by the fluctuation-dissipation theorem [60] the equation of motion for the transversal motion of the DNA is [61]

$$\frac{\partial y}{\partial t}(s,t) = \mathcal{L}_s y(s,t) + \sqrt{\frac{2k_BT}{\gamma_\perp}}\eta_y(s,t)\,, \quad s \in [0,L]\,,\ t \in [0,\infty[\ , \tag{B.1.1}$$

where $\mathcal{L}_s$ is a linear operator in s and η_y is standard continuous-time white noise. We assume that the movement in the other transversal direction, z, is described by a similar linear model. Equation (B.1.1) has dimensions of a speed, so $\mathcal{L}_s$ has dimensions of frequency and $\gamma_\perp$ of force/(speed $\cdot$ length). The parallel friction coefficient is of the order of $\gamma_\parallel \sim 1$ fN s/μm [62], which means that $\gamma_\perp \sim 2$ fN s/μm^2 and thus that $k_BT/\gamma_\perp \sim 2$ μm^3/s.

In experimental measurements of diffusion on DNA the DNA is extended in its longitudinal direction to a large fraction of its contour length, but the stretching forces are not so large that enthalpic contributions to the DNA movement need to be taken into account [63], i.e., the contour length can be considered constant. This means that $|\mathbf{r}'| \equiv |\frac{\partial \mathbf{r}}{\partial s}| = 1$, with $\mathbf{r} = (x, y, z)^T$, and

$$x' = \sqrt{1 - ((y')^2 + (z')^2)} \simeq 1 - \frac{1}{2}\left((y')^2 + (z')^2\right) \quad . \tag{B.1.2}$$

The general solution to (B.1.1) is

$$y(s,t) = \sum_{k=1}^{\infty} A_k(t) y_k(s) \quad ,$$

where y_k is the solution to the eigenvalue problem

$$[\mathcal{L}_s + c_k] y_k(s) = 0 \quad , \tag{B.1.3}$$

with given boundary conditions, and

$$\begin{aligned} A_k(t) &= e^{-c_k t}\left(A_k(0) + \sqrt{\frac{2k_B T}{\gamma_\perp}} \int_0^t e^{c_k t'} \eta_k(t') dt'\right) \\ &= \sqrt{\frac{2k_B T}{\gamma_\perp}} \int_{-\infty}^t e^{-c_k(t-t')} \eta_k(t') dt' \quad , \end{aligned} \tag{B.1.4}$$

with

$$\eta_k(t) = \int_0^L y_k(s) \xi_y(s,t) dt \quad .$$

The time-dependent parts of the eigenmodes, A_k have dimensions of $(\text{length})^{3/2}$ and the space-dependent parts y_k have dimensions of $1/\sqrt{\text{length}}$. Thus, y has dimensions of length.

The moments of x can be expressed in terms of moments of y and z. Since y is a Gaussian process it is completely described by its two first moments, $\langle y(s,t) \rangle = 0$ and

$$\begin{aligned} \rho_y(s_2, s_1, t_2 - t_1) &\equiv \langle y(s_2, t_2) y(s_1, t_1) \rangle \\ &= \frac{k_B T}{\gamma_\perp} \sum_{k=1}^{\infty} \frac{y_k(s_2) y_k(s_1)}{c_k} e^{-c_k |t_2 - t_2|} \quad . \end{aligned} \tag{B.1.5}$$

From Eq. (B.1.2) we then have

$$\langle x(s) \rangle = s - \frac{1}{2} \int_0^s \left[\rho_{y'}(s) + \rho_{z'}(s)\right] ds \quad , \tag{B.1.6}$$

where we in the notation have omitted repeated variables and variables that are zero, $\rho_{y'}(s) = \rho_{y'}(s, s, 0)$. Since y' and z' are uncorrelated, Eq. (B.1.2) gives

$$\rho_{x'}(s_2, s_1, t) = \frac{1}{4}\left[\left\langle y'(s_1,t)^2, y'(s_2,0)^2 \right\rangle_{\mathcal{C}} + \left\langle z'(s_1,t)^2, z'(s_2,0)^2 \right\rangle_{\mathcal{C}}\right] \quad .$$

Using that y' and z' are Gaussian we find

$$\rho_{x'}(s_2, s_1, t) = \frac{1}{2}\left[\rho_{y'}(s_2, s_1, t)^2 + \rho_{z'}(s_2, s_1, t)^2\right] \quad ,$$

and

$$\rho_x(s_2, s_1, t) \quad = \quad \frac{1}{2}\int_0^{s_2}\int_0^{s_1}\left[\rho_{y'}(s, s', t)^2 + \rho_{z'}(s, s', t)^2\right] ds ds' \quad , \quad \text{(B.1.7)}$$

Finally, the local degree of stretching of the DNA, defined as $\zeta \equiv \sqrt{\langle(\partial x/\partial s)^2\rangle}$, is

$$\zeta(\overline{s}) = 1 - \frac{1}{2}\left[\rho_{y'}(s) + \rho_{z'}(s)\right] \quad ,$$

We define the the weighted eigenfunctions $\mathcal{Y}_k = \sqrt{k_B T/\gamma_\perp c_k}\, y_k$, $\mathcal{Z}_k = \sqrt{k_B T/\gamma_\perp c_k}\, z_k$, and

$$\mathcal{X}_{k,l}(s) = \sqrt{\frac{1}{2}\left(\int_0^s \mathcal{Y}_k'(s')\mathcal{Y}_l'(s')ds'\right)^2 + \frac{1}{2}\left(\int_0^s \mathcal{Z}_k'(s')\mathcal{Z}_l'(s')ds'\right)^2} \quad .$$

The autocovariance functions of a point $\overline{s}$ on the DNA are then

$$\rho_y(\overline{s}, t) = \sum_{k=1}^{\infty} \mathcal{Y}_k(\overline{s})^2 e^{-c_k t} \quad , \quad \rho_z(\overline{s}, t) = \sum_{k=1}^{\infty} \mathcal{Z}_k(\overline{s})^2 e^{-c_k t} \quad , \qquad \text{(B.1.8)}$$

$$\rho_x(\overline{s}, t) = \sum_{kl=1}^{\infty} \mathcal{X}_k(\overline{s})^2 e^{-(c_k + c_l)t} \quad , \qquad \text{(B.1.9)}$$

and the local degree of stretching is

$$\zeta(\overline{s}) = 1 - \frac{1}{2}\sum_{k=1}^{\infty}\left\{\mathcal{Y}_k'(\overline{s})^2 + \mathcal{Z}_k'(\overline{s})^2\right\} \quad . \qquad \text{(B.1.10)}$$

The weighted eigenfunctions have units of length, while ζ is unitless.

The theory can be compared to the an effective one-mode theory employed by Hatfield and Quake [64], which derives a spring constant from the force-strain relation of Marko and Siggia [35],

$$f_0(\chi) = \frac{1}{l_\mathrm{p}}\left(\frac{1}{4}(1-\chi)^{-2} - \frac{1}{4} + \chi\right) \quad ,$$

where $\chi \equiv x(L)/L$ is the degree of stretching of the worm-like chain polymer. The y-direction spring constant is

$$k^{(y)} = \frac{F(\chi)}{x} = \frac{k_B T}{l\mathrm{p}L}\left(\frac{1}{4}\chi^{-1}(1-\chi)^{-2} - \frac{1}{4}\chi^{-1} + 1\right)$$

and the x-direction spring constant is

$$k^{(x)} = \frac{\partial f(\chi)}{\partial x} = \frac{k_B T}{l\mathrm{p}L}\left(\frac{1}{2}(1-\chi)^{-3} + 1\right)$$

According to this theory the x-direction spring constant $k^{(x)}$ is much higher than $k^{(y)}$ for highly stretched DNA, while they are equal according to the theory developed in the section, since $c^{(x)} = k^{(x)}/\gamma_{\|}$ and $c^{(y)} = k^{(y)}/\gamma_{\perp}$, and $\gamma_{\perp} = 2\gamma_{\|}$. Hatfield and Quake [64] do not present any experimental verification of their result, while the experimental data analyzed in Sec. C.4 show that a factor 2 between the lowest modes as predicted by the theory developed in this section is reasonable (Fig. C.11).

B.1.1 DNA pulled by the ends

DNA can be stretched by applying a constant force to the ends of the DNA. This can be done in various ways [45, 46], i.e., using two optical tweezers [36], using one optical tweezer and a micropipette, binding the DNA ends to biotin-labeled micro-beads fused to the coverslip, or directly binding the DNA ends to the coverslip[1] [37]. We derive expressions for the DNA's movement and statistics based on the worm-like chain (WLC) model.

The DNA experiences a constant stretching force F along its contour and thus a potential energy in the transversal direction y, of

$$E[y] = \int_0^L \left\{ A \left(\frac{\partial^2 y}{\partial s^2} \right)^2 + F \left(\frac{\partial y}{\partial s} \right)^2 \right\} ds \ , \tag{B.1.11}$$

where $A = k_B T l_{\mathrm{p}} \ll F$ is the bending energy and l_{p} is the persistence length of DNA. The same potential energy acts on the DNA in the other transversal direction, z. The linear operator $\mathcal{L}_s$ is given as minus the functional derivative of the energy functional Eq. (B.1.11),

$$\mathcal{L}_s = -\frac{\delta E}{\delta y}[y] \ ,$$

where the functional derivative is defined as

$$\frac{dE}{dz(s)}[z] = \frac{dE}{d\epsilon}[z + \epsilon\delta_s]\bigg|_{\epsilon=0} \ ,$$

with $\delta_s(s') \equiv \delta(s' - s)$. Thus,

$$\mathcal{L}_s = -\frac{A}{\gamma_{\perp}} \frac{\partial^4}{\partial s^4} + \frac{F}{\gamma_{\perp}} \frac{\partial^2}{\partial s^2} \ , \tag{B.1.12}$$

which gives the following eigenvector equation

$$-z_k^{(4)}(s) + q z_k''(s) = \lambda_k z_k(s) \ ,$$

[1] The proximity of the coverslip to the DNA does not alter the eigenmodes considerably [37].

where $q \equiv F/A$ and $\lambda_k \equiv \gamma_\perp c_k/A$, and with the boundary conditions

$$z(0)_k = z_k(L) = z_k''(0) = z_k''(L) = 0 \ ,$$

i.e., the ends are fixed and there is no bending force at the ends.

Equation (B.1.1) with $\mathcal{L}_s$ given by Eq. (B.1.12) is solved by

$$z_k(s) = \sqrt{\frac{2}{L}} \sin(\omega_k s) \ ,$$

where $\omega_k = k\pi/L$ and $\lambda_k = \omega_k^2(\omega_k^2 + q)$, which means that

$$\begin{aligned} c_k &= \frac{\omega_k^2}{\gamma_\perp}(A\omega_k^2 + F) \\ &= \frac{k_B T}{\gamma_\perp}\omega_k^2(\omega_k^2 l_\mathrm{p} + f_0) \ , \end{aligned} \tag{B.1.13}$$

where we have defined $f_0 \equiv F/k_B T$. Thus, since the movement in the two transversal directions are equivalent, the weighted transversal eigenfunctions $\mathcal{Y}_k$ and $\mathcal{Z}_k$ are equal and are given by Eq. (3.2.29), while the longitudinal eigenfunction $\mathcal{X}_k$ is given by (3.2.30).

The first values of c_k are plotted in Fig. B.1a, while the relative contributions of the lowest couple of weighted eigenmodes to the transversal variance is shown in Fig. B.1b for slow DNA dynamics, or equivalently, high frequency measurements. The relative contribution of the eigenmodes to the power spectrum are shown in Fig. B.2. Only the three lowest modes contribute significantly to the measured movement. The third mode resembles white noise. So it just contributes to the positional noise and we do not need to take into account explicitly.

A freely linked chain (FLC) polymer pulled by the ends is described by the same eigenfunctions as the WLC, while the correlation constants c_k are given by $c_k = F\omega_k^2/\gamma_\perp$.

B.1.2 DNA in plug flow

The simplest flow we can think of is the plug flow, which has a constant velocity profile. The plug flow is a good approximation to the center of a Poiseuille flow, i.e., a flow with a parabolic velocity profile.

Since the DNA is highly stretched, it is a good approximation to neglect hydrodynamic interactions, such as screening [65]. This is the free-draining approximation. We furthermore assume that the DNA is so stretched that

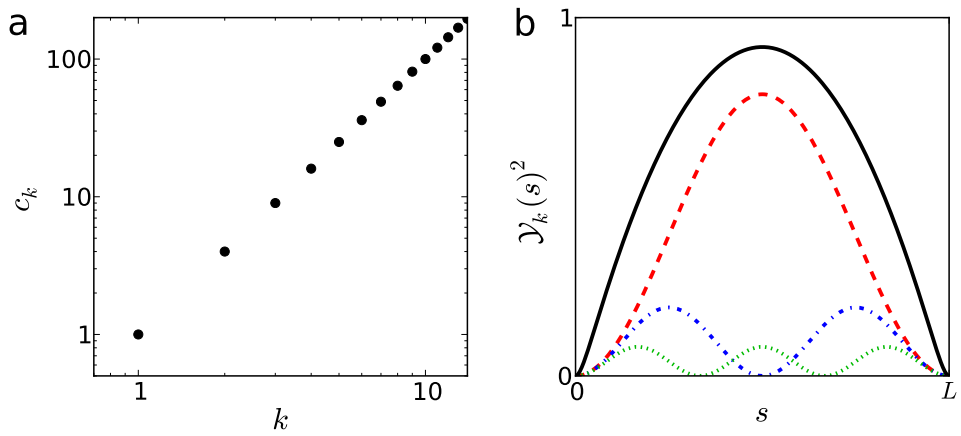

Figure B.1: a) Correlation coefficient c_k as a function of mode number k for DNA pulled by the ends. The correlation coefficients are normalized such that $c_1 = 1$. b) Relative contributions to the variance of the transversal motion of the three lowest modes, $\mathcal{Y}_k(s)^2$. (1st mode. dashed line, red; 2nd mode: dash-dotted line, blue; 3rd mode: dotted line, green; total variance: full line, black.)

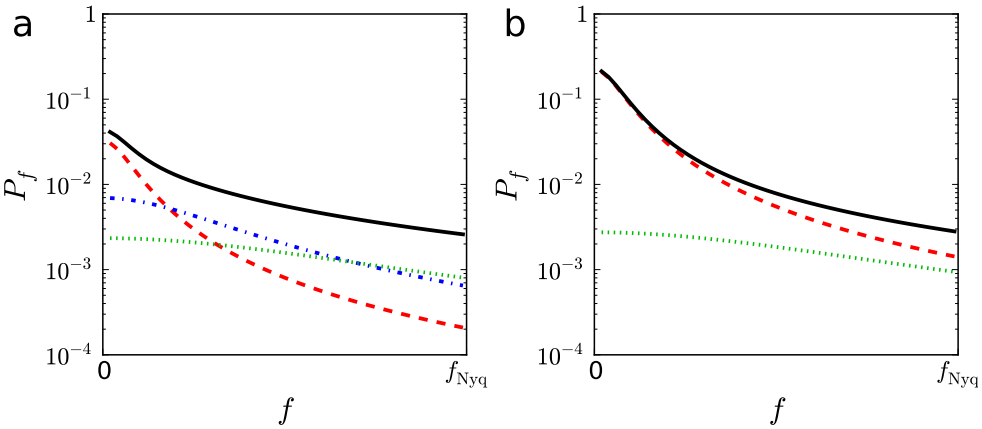

Figure B.2: Relative contributions to of the three lowest modes to the power spectrum of the measured transversal motion of a DNA strand pulled by the ends. The time-lapse is $\Delta t = 4/c_1$. (1st mode. dashed line, red; 2nd mode: dash-dotted line, blue; 3rd mode: dotted line, green: total power: full line, black.) a) Spectrum near the end of the DNA, $s = L/8$. b) Spectrum at the middle of the DNA, $s = L/2$.

all parts of the DNA are approximately aligned with the flow. The potential energy for the DNA is then [66]

$$E[\mathbf{r}] = \frac{1}{2}\int_0^L \left\{ A\left(\frac{\partial^2 \mathbf{r}_\perp}{\partial s^2}\right)^2 + \gamma_\parallel v(L-s)\frac{\partial x}{\partial s}\right\} ds \ ,$$

where $\mathbf{r}_\perp = (y,z)^T$ is the transverse components of $\mathbf{r}$. We use that the DNA is highly stretched such that $x' \simeq 1 - [(y')^2 + (z')^2]/2$. Thus the y-direction part of the energy is

$$E[y] \simeq \frac{1}{2}\int_0^L \left\{ A\left(\frac{\partial^2 y}{\partial s^2}\right)^2 + \gamma_\parallel v(L-s)\left(\frac{\partial y}{\partial s}\right)^2\right\} ds \ . \tag{B.1.14}$$

The model is not correct close to the free end, where the force acting on the DNA locally due to the hydrodynamic drag downstream inevitably is small. Here the DNA will curl up and neither the approximation that the DNA is parallel to the flow, nor the free draining approximation, hold. The change in orientation increases the drag on the DNA, since $\gamma_\perp = 2\gamma_\parallel$, while hydrodynamic screening decreases the drag. So these two effects counter each other to some extent. Furthermore, for strong flows, the part of the DNA, which is curled up is small compared to the total DNA length [38], which means that even if we make an error on the local scale near the free end, the total error is small on the global scale.

Static properties of this model have been derived by considering the energy E as a function of the y-direction part of the tangent vector r_y, which reduces E from fourth to second order [67]. However, we are interested in dynamic properties of the system. So we need to consider the full fourth-order problem defined by Eq. (B.1.14), since we cannot define a Langevin equation for r_y. From Eq. (B.1.14) the linear operator $\mathcal{L}_s$ is equal to

$$\mathcal{L}_s = -\frac{A}{\gamma_\perp}\frac{\partial^4}{\partial s^4} + \frac{\gamma_\parallel v}{\gamma_\perp}\frac{\partial}{\partial s}(L-s)\frac{\partial}{\partial s} \ , \tag{B.1.15}$$

which is self-adjoint with the appropriate boundary conditions:

$$y(0,t) = y''(0,t) = y''(L,t) = y^{(3)}(L,t) = 0 \ ,$$

i.e., the end is tethered at $s = 0$, there is no bending force at the ends and no torque at the free end. The movement in the two transversal directions are equivalent, so we just treat the y-direction here. We need to solve the fourth-order linear differential equation

$$[\mathcal{L}_s + c_k]y_k(s) = 0 \tag{B.1.3}$$

to find the space eigenfunctions y_k. There does not seem to be a solution of this differential equation in terms of known special functions [68, 69]. However, since we are only able to observe relatively slow behavior in SPT experiments[2],

[2]The time-lapse is on the order of the slowest relaxation time of the DNA.

only the lowest modes contribute significantly to y. These govern behavior at long length scales and depend very little on the DNA's persistence. If we ignore the persistence term in Eq. (B.1.15), Eq. (B.1.3) for plug flow reduces to

$$\left(\frac{\partial}{\partial s}(L-s)\frac{\partial}{\partial s}+\frac{Q_k^2}{4}\right)y_k(s)=0 \ , \tag{B.1.16}$$

with the boundary condition $y_k(0)=0$, where

$$Q_k^2 \equiv \frac{4\gamma_\perp c_k}{\gamma_\parallel v} \ .$$

Introducing the variable $q=Q_k\sqrt{L-s}$ gives

$$q^2 y_k''(q)+q y_k'(q)+q^2 y_k(q)=0 \ , \tag{B.1.17}$$

which is a zeroth-order Bessel differential equation. The solutions to Eq. (B.1.17) are of the form

$$y_k(q)=\mathcal{N}_k J_0(q)+\mathcal{B}_k Y_0(q) \ ,$$

where Y_0 diverges at $q=0$ and thus cannot be a solution. Furthermore, the boundary condition at $s=0$, implies that $Q_k=\alpha_k/\sqrt{L}$, where α_k is the k'th zero of the zeroth order Bessel function of the first kind, J_0. This gives

$$c_k=\frac{\gamma_\parallel v}{4\gamma_\perp L}\,\alpha_k^2=\frac{v\,\alpha_k^2}{8L} \ .$$

In order for $\{z_k\}$ to be an orthonormal basis, the $\mathcal{N}_k$ must satisfy

$$\mathcal{N}_k^{-2}=\int_0^L J_0\left(\alpha_k\sqrt{1-s/L}\right)^2 ds = L J_1(\alpha_k)^2 \ .$$

So the set of functions

$$y_k(s)=\frac{J_0\left(\alpha_k\sqrt{1-s/L}\right)}{\sqrt{L}J_1(\alpha_k)}$$

forms an orthonormal basis for the boundary value problem (B.1.1) with $\mathcal{L}_s$ given as in Eq. (B.1.16). This gives the weighted eigenfunctions for DNA fluctuations, Eqs. (3.2.32) and (3.2.33), and

$$\mathcal{Y}_k'(s)=\mathcal{Z}_k'(s)=\sqrt{\frac{k_BT}{\gamma_\parallel v}}\frac{J_1\left(\alpha_k\sqrt{1-s/L}\right)}{\sqrt{L(L-s)}J_1(\alpha_k)} \ . \tag{B.1.18}$$

One may notice that since we have neglected the persistence, the variance of y' is not well defined, i.e, the value of $\rho_{y'}(s_1,s_2,t)$ goes to infinity as t goes to zero. This means that according to this model a DNA molecule with a given

extension x must be infinitely long! The DNA is a fractal[3]. This also means that the local degree of stretching of the DNA is not well defined. However, as Fig. B.3 shows the largest part of the DNA is almost completely stretched and the DNA only curls up at the end, in agreement with the stem-and-flower model of Brochard-Wyart for polymers in strong flows [38]. This means that for a highly stretched DNA strand, $\zeta(\overline{s}) \simeq 1$, except near the free end of the DNA.

The first values of c_k are plotted in Fig. B.3a, while the relative contribution of the lowest eigenmodes to the transversal variance is shown in Fig. B.3b. The relative contribution of the eigenmodes to the power spectrum are shown in Fig. B.4. As for DNA pulled by the ends (App. B.1.1), only the two lowest modes contribute significantly to the measured movement.

B.1.3 DNA tethered to a surface in shear flow

Many experiments measuring diffusion of fluorescent proteins on DNA using TIRF microscopy are performed by tethering one of the DNA strand to a coverslip and stretching the DNA by imposing a flow over the coverslip. This is practical since it assures that the DNA strand is close to the coverslip, where the evanescent field can excite the fluorescent molecules (Fig. 2.8a). Close to the coverslip, where the DNA is located, the hydrodynamic flow is to a good approximation a shear flow. While this setup is practical for experimental reasons the motion is difficult to treat analytically [67]. The flow is non-conservative and tends to induce cyclic motion of the DNA [71, 72, 73, 74], while the presence of the coverslip will couple DNA eigenmodes. However, for strong shear, i.e, taut DNA, cyclic motion is not observed [73]. This means that the overdamped linear theory (Eq. (B.1.1)), at least locally, should describe the DNA motion reasonably well.

The z- and y-directions are not equivalent for DNA in plug flow. The z-direction motion is asymmetric due to the presence of the coverslip and the z-dependence of the flow speed. This also means that the y-motion depends on the z-position of the DNA. We thus start with investigating the z-direction movement, where we will for the moment forget about the wall, the influence of which we treat in the following section (Sec. B.1.3.1). The system is sketched in Fig. 2.8a. As in Sec. B.1.2, we neglect DNA persistence. We assume that for

[3]This mathematical particularity may be mended in various ways: by approximating the eigenfunctions of the full problem given by Eq. (B.1.15) by truncated polynomial series. (This has the disadvantage of being computationally very expensive.); by treating the persistence as a perturbation and approximating c_k by the first order perturbative solution using classical perturbation theory [70]. (Note that for DNA stretched by the ends, this is actually equal to the *exact* solution.); or simply by truncating the infinite sum over modes in Eq. (B.1.18). (The difficulty here is to properly choose *where* to truncate.) These strategies are either impractical or unreliable, but the good news is that we do not need to take this into account to model the DNA's fluctuations.

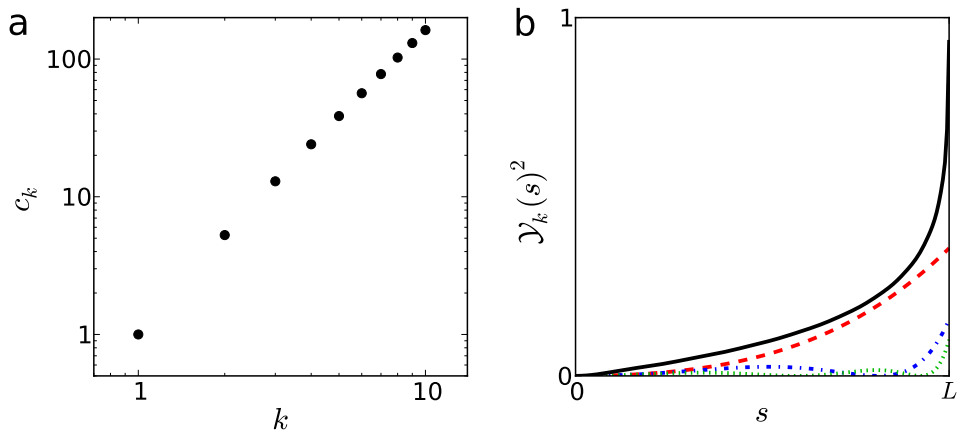

Figure B.3: a) Correlation coefficient c_k as a function of mode number k for DNA in a plug flow. The correlation coefficients are normalized such that $c_1 = 1$. b) Relative contributions to the variance of the transversal motion of the three lowest modes, $\mathcal{Y}_k(s)^2$. (1st mode. dashed line, red; 2nd mode: dash-dotted line, blue; 3rd mode: dotted line, green; total variance: full line, black.)

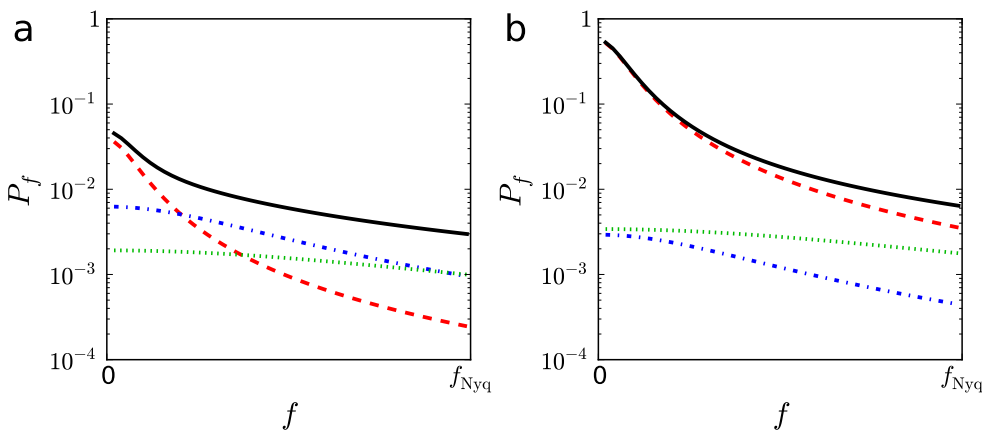

Figure B.4: Relative contributions to of the three lowest modes to the power spectrum of the measured transversal motion of a DNA strand stretched by a plug flow. The time-lapse is $\Delta t = 4/c_1$. (1st mode. dashed line, red; 2nd mode: dash-dotted line, blue; 3rd mode: dotted line, green: total power: full line, black.) a) Spectrum upstream on the DNA (near the tethered end), $s = L/4$. b) Spectrum downstream on the DNA (near the free end), $s = 3L/4$.

a given point s on the DNA, the upstream part of the DNA is approximately horizontal. This assumption cannot, of course, be true everywhere on the DNA, since this would mean that the DNA stuck to the coverslip. However, simulations show that the assumption is true for the largest part of the DNA and that the mean DNA contour looks as as sketched in Fig. 2.8a [65, 75, 76]. Under this assumption the mean force acting on the the DNA at the point s is

$$f(s) = \gamma v' \int_s^L \langle z(s') \rangle \, ds' \ ,$$

where γ is an effective drag coefficient and v' is the shear rate. Since f depends on z, the associated Langevin equation is non-linear. However, the non-linear Langevin equation

$$\frac{\partial z}{\partial t}(s,t) = \left(\frac{\partial}{\partial s} \frac{f(s)}{\gamma_\perp} \frac{\partial}{\partial s} + \sqrt{\frac{2k_B T}{\gamma_\perp}} \right) z(s,t) \tag{B.1.19}$$

is solved by $z_1(s,t) = A(t)s$. Furthermore, z_1 is the only function $A(t)z_i(s)$, which is a polynomial in s, that solves (B.1.19), since

$$\frac{\partial}{\partial s} \frac{f(s)}{\gamma_\perp} \frac{\partial}{\partial s} A(t) z_i(s) \propto \int_s^L z(s') ds' \frac{\partial^2 z_i(s)}{\partial s^2} - z_i(s) \frac{\partial z_i(s)}{\partial s} \ ,$$

which is only proportional to $z_i(s)$ itself if $z_i(s) \propto s$. Encouraged by this result[4], we assume that the force acting on the downstream DNA is

$$f(s) = \gamma v' (L^2 - s^2) ds' \ . \tag{B.1.20}$$

We have reduced the nonlinear boundary value problem to a linear problem. With the appropriate boundary conditions, i.e., z must be finite on $[0, L]$ and $z(0) = 0$, this boundary value problem is solved by the uneven Legendre polynomials normalized on $[0, L]$,

$$\sqrt{\frac{4k-1}{L}} P_{2k-1}(s/L) \ , \quad k = 1, 2, \ldots \ .$$

Finally, this gives us that

$$c_k = 2k(2k-1) \frac{\gamma v'}{\gamma_\perp} \ .$$

From the assumption (B.1.20) we have that the y-eigenfunctions are the same as the z-eigenfunctions. So the normalized eigenfunctions for the transversal motion are

$$\mathcal{Y}_k(s) = \mathcal{Z}_k(s) = \sqrt{\frac{k_B T (4k-1)}{2\gamma v' L k (2k-1)}} \ P_{2k-1}(s/L)^2 \ .$$

[4]The lowest and dominating mode is approximately proportional to s, thus we assume that z is, too. Even though it is not consistent with or initial assumption of approximately horizontal alignment of the DNA. The fact that the coverslip couples the modes will make the DNA more horizontal.

The first values of c_k are plotted in Fig. B.5a, while the relative contribution of the lowest couple of eigenmodes to the transversal variance is shown in Fig. B.5b. The relative contribution of the eigenmodes to the power spectrum are shown in Fig. B.6. Both the variance of transversal fluctuations and the power spectrum of this model for a DNA molecule in a shear flow resemble those of a DNA strand in a plug flow. The only qualitative difference is that plug flow tends to stretch the DNA more, in agreement with scaling-based theories [77].

B.1.3.1 Brownian motion in a truncated harmonic potential

We have above neglected the influence of the coverslip on the DNA movement. We here investigate the effect of the coverslip on the lowest mode of the DNA's z-direction motion, and show that due to symmetry the presence of coverslip does not change the statistics of the lowest mode of the longitudinal (x-direction) motion. We only see the z-direction motion indirectly through its contribution to the x-direction motion, where it only enters as z^2. We are thus only able to observe even moments of z The coverslip restricts the DNA to the upper half of the z-axis. Thus forcing the DNA out into the flow. If we can only resolve the lowest DNA mode, the observed system is equivalent to a particle undergoing Brownian motion in a harmonic potential centered around and truncated at z equal to zero. The position $\mathcal{Z}_1$ of a Brownian particle trapped in a truncated harmonic potential is distributed according to a truncated Gaussian distribution

$$p(\mathcal{Z}_1) = \sqrt{\frac{2\gamma_\perp c_1}{\pi k_B T}} e^{-\frac{\gamma_\perp c_1}{2k_B T}\mathcal{Z}_1^2}.$$

Even moments of a truncated Gaussian distribution are the same as for a regular Gaussian distribution. Thus, the presence of the coverslip does not affect the lowest longitudinal (x-direction) mode.

B.1.3.2 Higher modes

While the coverslip does not change the properties of the lowest mode of the longitudinal motion, when multiple modes are considered, the presence of the coverslip couples the modes. This kills our hopes of deriving a global linear description of the DNA motion. For DNA stretched by shear flow we must thus content ourselves with fitting the modes individually and locally. However, since our objective is not to model the DNA movement, but only obtain unbiased diffusion coefficient estimates, a local description suffices. Since we have no theory for the DNA movement, we cannot describe the local degree of stretching of the DNA, ζ. However, as for DNA stretched in plug flow, the the largest part of the DNA is almost completely stretched and the DNA only

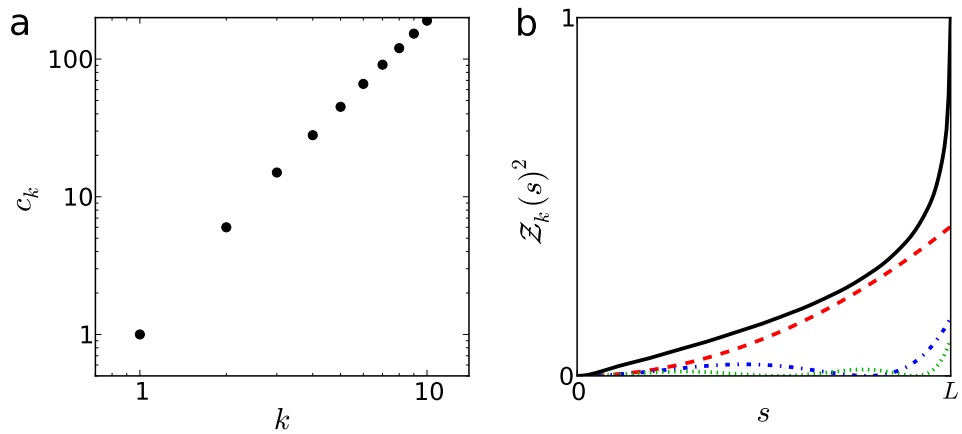

Figure B.5: a) Correlation coefficient c_k as a function of mode number k for DNA in a shear flow. The correlation coefficients are normalized such that $c_1 = 1$. b) Relative contributions to the variance of the transversal motion of the three lowest modes, $\mathcal{Y}_k(s)^2$. (1st mode. dashed line, red; 2nd mode: dash-dotted line, blue; 3rd mode: dotted line, green; total variance: full line, black.)

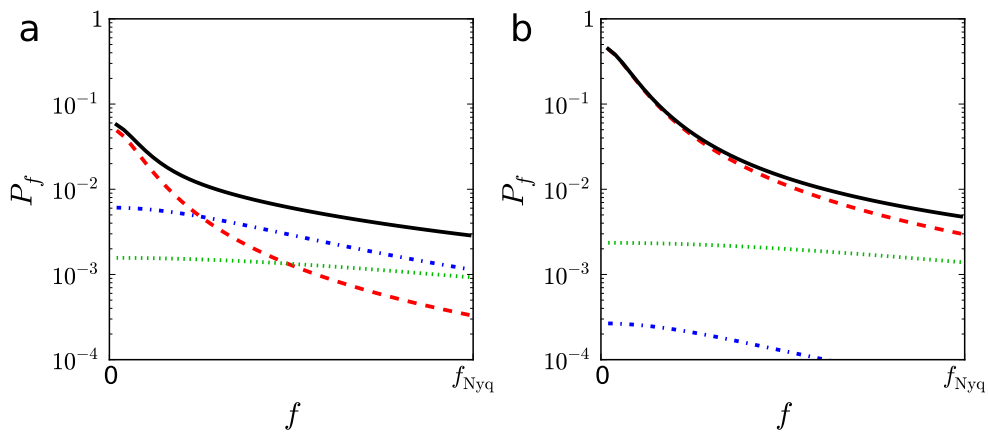

Figure B.6: Relative contributions to of the three lowest modes to the power spectrum of the measured transversal motion of a DNA strand stretched by a shear flow. The time-lapse is $\Delta t = 4/c_1$. (1st mode. dashed line, red; 2nd mode: dash-dotted line, blue; 3rd mode: dotted line, green: total power: full line, black.) a) Spectrum upstream on the DNA (near the tethered end), $s = L/4$. b) Spectrum downstream on the DNA (near the free end), $s = 3L/4$.

curls up near the end [38, 39] (Fig. B.5). Thus $\zeta(\overline{s}) \simeq 1$, except near the free end.

B.1.3.3 Other models of DNA in hydrodynamic flow

We here discuss other models for DNA in shear flow, which may serve to explain experimental results or as a starting point for building a simple theory for the DNA fluctuations. Based on scaling arguments, Brochard-Wyart has proposed a "horn"-model for DNA in shear flow [77]. The name stems from the observed form of the DNA, which looks like a horn. In this model the transverse fluctuations of the DNA scale as $(L-s)^{-1/2}$. Recent experimental measurements of DNA in strong shear flow by Laube et al. [39] suggest that the DNA motion is well described the "Stem-and-flower" model. Another scaling-based model initially derived for DNA in strong plug flow [38]. That DNA in shear flow may be adequately described by a plug-flow model is backed up by recent Brownian dynamics simulations. These suggest an approximate linear relationship between tensile force and position on the DNA [75] as for DNA in plug flow.

B.2 Statistics of diffusion on a fluctuating substrate

In this section we derive the statistical properties of a particle diffusing on a fluctuating substrate. In particular, diffusion on a fluctuating polymer such as a DNA strand. In Sec. B.2.1 we review general properties of Fourier transforms of data measured with finite time-resolution, and of stationary and non-stationary processes. In Sec. B.2.2 we derive the power spectra of the transversal and longitudinal movement of a particle diffusing on fluctuating DNA. In Sec. B.2.3 we derive the autocovariance function for diffusion on DNA.

As discussed in Sec. 2.2.1, the observed position of a particle moving on a substrate, i.e., the position observed in the lab reference-frame, is the position in space of the specific point on the substrate contour where the protein is situated at time t, $s(t)$,

$$x(s(t), t) \ . \tag{2.2.1}$$

The movement on the substrate, which determines s, is a Brownian motion with diffusion coefficient D, while $x(s,t)$ given $s(t)$ is a stochastic process described by the equations of motion governing the substrate movement.

B.2.1 Power spectra

We review some general properties of the Fourier transform of discretely measured data. In Sec. B.2.1.1 we derive an exact expression for the power spectrum of a continuous process measured discretely with finite camera shutter time in terms of the power spectrum of the underlying continuous process. In Sec. B.2.1.2 we investigate the power spectrum of a non-stationary process, namely Brownian motion. In Sec. B.2.1.3 we show a derivation of the distribution of the measured power spectrum.

B.2.1.1 Aliasing and finite shutter time

For experimental measurements the finite shutter time of the camera results in a low-pass filtering of the data measured. For a "infinite" Fourier transform $\mathcal{F}$, i.e.,

$$\mathcal{F}[x]_f = \int_{-\infty}^{\infty} e^{-i2\pi f t} x(t) dt \ , \tag{B.2.1}$$

Savin and Doyle [78] have shown that the averaging caused by the finite shutter time just alters the fourier transform by a multiplicative term of the form

$$\mathcal{F}\left[\overline{x}\right](f) = \frac{1 - e^{-i2\pi f \tau}}{i2\pi f \tau} \mathcal{F}[x](f) \ . \tag{B.2.2}$$

We also see an aliasing effect, since our measurements are discrete. Due to the finite sampling frequency, we are not able to distinguish time-modes with higher frequency than the Nyquist frequency f_{Nyq}. These thus contribute to the measured (aliased) power spectrum, and for a stationary process we have

$$P_f^{(a)} = \sum_{m=-\infty}^{\infty} P_{f+\frac{m}{\Delta t}} \ , \tag{B.2.3}$$

where $P_F^{(a)}$ is the aliased power spectrum. We expect to see a combination of these two effects in our data as pointed out by Wong and Halvorsen [79]. Since the experimentally measured time-series necessarily are of finite length, we should not compare to the "infinite" idealized Fourier transform, Eq. (B.2.1), but to the finite Fourier transform[5] over the interval $[0, t_N[$. Note, however, that the difference between the "infinite" and finite Fourier transforms is of order $1/t_N$.

From experimental data, we can calculate the Discrete Fourier Transform (DFT) of the measured motion blurred positions,

$$X_k \equiv \mathcal{DFT}[\overline{x}]_k = \Delta t \sum_{n=0}^{N-1} e^{-i2\pi \frac{kn}{N}} \frac{1}{\tau} \int_0^{\tau} x(t_n - t) dt \ . \tag{B.2.4}$$

[5]The finite Fourier transforms are equivalent to the Fourier coefficients, except for a normalization factor of $1/t_N$.

We derive the form of the experimental power spectrum from Eq. (B.2.4).

$$\begin{aligned}
X_k &= \frac{\Delta t}{\tau}\sum_{n=0}^{N-1} e^{-i2\pi kn/N}\int_0^{\Delta t} x(t_n - t)dt \\
&= \frac{\Delta t}{\tau}\sum_{n=0}^{N-1} e^{-i2\pi kn/N}\int_0^{\tau}\frac{1}{t_N}\sum_{k'=-\infty}^{\infty} e^{i2\pi k'(t_n-t)/t_N}\,\widetilde{x}_{k'}dt \\
&= \frac{1}{N\tau}\sum_{k'=-\infty}^{\infty}\widetilde{x}_{k'}\int_0^{\tau} e^{-i2\pi k't/t_N}dt\sum_{n=1}^{N-1} e^{-i2\pi(k-k')n/N} \\
&= \frac{\Delta t}{\tau}\sum_{m=-\infty}^{\infty}\frac{1-e^{-i2\pi k\tau/t_N}}{i2\pi\left(k/N+m\right)}\,\widetilde{x}_{k+mN} \quad , \qquad \text{(B.2.5)}
\end{aligned}$$

where we have used the definition of the Fourier series,

$$x(t) = \frac{1}{t_N}\sum_{k=-\infty}^{\infty} e^{i2\pi\frac{kt}{N}}\widetilde{x}_k \quad ,$$

where the finite Fourier transforms are given by

$$\widetilde{x}_k = \int_0^{t_N} x(t)e^{-i2\pi kt/t_N}dt \quad ,$$

and that $\sum_{n=0}^{N-1} e^{i2\pi\frac{n}{N}k} = 0$ except for $k = mN$ and $m \in \mathbb{Z}$, where $\sum_{n=0}^{N-1} e^0 = N$. For full time-integration, i.e., for $\tau = \Delta t$, (B.2.5) reduces to

$$X_k = \sum_{m=-\infty}^{\infty}\frac{1-e^{-i2\pi k/N}}{i2\pi\left(k/N+m\right)}\,\widetilde{x}_{k+mN} \quad .$$

Finally, the discrete power spectrum of the positional noise $\sigma\xi$ is

$$\begin{aligned}
\widehat{P}[\xi]_k &= \frac{\Delta t}{N}\sum_{mn=0}^{N-1} e^{-i2\pi\frac{k}{N}(n-m)}\left\langle\xi_n\xi_m\right\rangle \\
&= \sigma^2\frac{\Delta t}{N}\sum_{n=0}^{N_1} e^{-i2\pi\frac{kn}{N}} \\
&= \sigma^2\Delta t \quad .
\end{aligned}$$

Note that there is no aliasing or time-averaging effects, since the process ξ is inherently discrete.

We can then express the expectation value of the measured power spectrum $\widehat{P}$ in terms of the theoretical power spectrum P. For a stationary process measured with full time-integration, the power spectrum is

$$\left\langle\widehat{P}_k\right\rangle = \sigma^2\Delta t + \sum_{m=-\infty}^{\infty}\frac{2-2\cos\left(2\pi\frac{k}{N}\right)}{\left(2\pi\left(\frac{k}{N}+m\right)\right)^2}\left\langle P_{f_k+\frac{m}{\Delta t}}\right\rangle \quad , \quad \text{for } k \neq 0 \quad , \qquad \text{(B.2.6)}$$

where we have used that expectation value of the Fourier transform of a stationary process is zero except at zero frequency and the measured power spectrum is calculated from data as

$$\widehat{P}_k = \frac{1}{t_N} \left|\mathcal{DFT}[\{x_n\}_n]_k\right|^2 \quad .$$

We see that (B.2.6) is just the combination of (B.2.2) and (B.2.3), i.e., the result of Wong and Halvorsen [79] is exact.

For diffusion on DNA the y-direction movement is stationary. However the movement in the x-direction is not, due to the diffusive term.

Non-stationary processes If the process of which we calculate the power spectrum is not stationary the expectation value of the fourier transform is not zero and the expectation value of the power spectrum is

$$\begin{aligned}
\left\langle \widehat{P}_k \right\rangle &= \sigma^2 \Delta t + \sum_{mm'=-\infty}^{\infty} \frac{\left[2 - 2\cos\left(2\pi \frac{k}{N}\right)\right] \left\langle \widetilde{x}_{k+mN} \widetilde{x}^*_{k+m'N} \right\rangle}{4\pi^2 t_N \left(\frac{k}{N} + m\right)\left(\frac{k}{N} + m'\right)} \\
&= \sigma^2 \Delta t + \frac{2 - 2\cos\left(2\pi \frac{k}{N}\right)}{4\pi^2} \left(\sum_{m=-\infty}^{\infty} \frac{P_{f_k + \frac{m}{\Delta t}}}{\left(\frac{k}{N} + m\right)^2} + \sum_{m \neq m'} \frac{\left\langle \widetilde{x}_{x+mN} \right\rangle \left\langle \widetilde{x}^*_{x+m'N} \right\rangle}{t_N \left(\frac{k}{N} + m\right)\left(\frac{k}{N} + m'\right)} \right) \quad ,
\end{aligned}$$

This is only one of several problems one needs to be aware of before applying fourier analysis to non-stationary processes. We will explore additional problems in the following section.

B.2.1.2 Power spectrum of a non-stationary process

A lot of the neat properties that Fourier transforms have suppose that the stochastic process, which is Fourier transformed, is stationary. If the process is not stationary, many of these general results for Fourier transforms do not hold. In this section we investigate the power spectrum of the simplest non-stationary stochastic process, Brownian motion. Brownian motion is governed by the differential equation

$$\dot{x}_t = \eta_t \quad , \tag{B.2.7}$$

where η is a standard continuous-time white noise. We derive the properties of the Fourier transform of Brownian motion directly from the differential equation, Eq. (B.2.7).

We Fourier transform Eq. (B.2.7),

$$\begin{aligned}
\widetilde{\dot{x}}_\omega &= \int_0^{t_N} e^{-i\omega t} x_t dt \\
&= x_{t_N} - x_0 + i\omega \widetilde{x}_\omega \\
&= \widetilde{\eta}_\omega \quad ,
\end{aligned} \tag{B.2.8}$$

where $\langle \widetilde{\eta}_\omega \rangle = 0$ and $\langle \widetilde{\eta}_\omega \widetilde{\eta}_{\omega'} \rangle = t_N \delta_{\omega,\omega'}$. Equation (B.2.8) is rearranged to get

$$\frac{\left\langle \widetilde{x}_\omega \widetilde{x}^*_{\omega'} \right\rangle}{t_N} = \frac{\langle \widetilde{\eta}_\omega \widetilde{\eta}_{\omega'} \rangle + \left\langle (x_{t_N} - x_0)^2 \right\rangle - \langle (\widetilde{\eta}_\omega + \widetilde{\eta}_{\omega'})(x_{t_N} - x_0) \rangle}{\omega\omega' t_N} \ .$$

Since the process is non-stationary, the power spectrum is not a one dimensional, but a two-dimensional quantity, i.e., $\langle (x_{t_N} - x_0)^2 \rangle$ is not zero. For $\omega, \omega' \neq 0$,

$$\begin{aligned} \langle \widetilde{\eta}_\omega (X_{t_N} - x_0) \rangle &= \int_0^{t_N} e^{-i\omega t} \left\langle \eta_t \int_0^{t_N} \eta_{t'} dt' \right\rangle dt \\ &= \int_0^{t_N} = 0 \end{aligned}$$

and

$$\left\langle (x_{t_N} - x_0)^2 \right\rangle = t_N \ .$$

So

$$\frac{\left\langle \widetilde{x}_\omega \widetilde{x}^*_{\omega'} \right\rangle}{t_N} = \frac{\delta_{\omega,\omega'} + 1}{\omega\omega'} \ .$$

The delta function term is what we would expect to see from our experience with stationary processes and the second term comes from the no-longer-negligible boundary terms.

This simple example shows several problems with the power spectrum: (i) the Fourier modes are not independent, which complicates estimation; (ii) we can no longer expect the power spectrum to be exponentially distributed (see App. B.2.1.3); (iii) Boundary terms are not negligible in the large N-limit; and (iv) the (one-dimensional) traditional power spectrum is not a sufficient statistic, i.e., we throw away information about the data when we neglect the correlations between modes.

B.2.1.3 Distribution of the power spectrum

We show in this section that for a stationary Gaussian process, the measured power spectral components of the process is exponentially distributed to order $1/N$.

We use that for a Gaussian process y we can write

$$y_{n+1} = f(y_n) + \eta_n \ ,$$

where f is some function and $\eta_n = \int_{t_n}^{t_{n-1}} g(t)\eta_t dt$ is Gaussian distributed and has independent increments. If y is stationary then

$$e^{-i\omega\Delta t} \widetilde{y}_\omega = \widetilde{f(y)}_\omega + \widetilde{\eta}_\omega \ ,$$

to order $1/N$. The real and the imaginary parts of $\widetilde{\eta}$ are the cosine and sine transforms of η and are thus orthogonal on the interval $[0, t_N]$, i.e.,

$$\begin{aligned}\langle \Re(\widetilde{\eta}_\omega)\Im(\widetilde{\eta}_\omega)\rangle &= (\Delta t)^2 \sum_{mn=0}^{N-1} \cos(\omega\Delta t m)\sin(\omega\Delta t' n)\langle \eta_m \eta_n\rangle \\ &= \left\langle (\eta_n \Delta t)^2 \right\rangle \sum_{n=0}^{N-1} \cos(2\pi kn/N)\sin(2\pi kn/N) = 0 \ .\end{aligned}$$

Thus, $X \equiv \Re(\widetilde{\eta}_\omega)$ and $Y \equiv \Im(\widetilde{\eta}_\omega)$ are independent and identically distributed, with variance $\sigma^2 t_N$, and their distribution is

$$p(X,Y)dXdY = \frac{dXdY}{2\pi\sigma^2 t_N}\, e^{-\frac{X^2+Y^2}{2\sigma^2 t_N}}$$

and the marginal distribution of $P^{(y)} \equiv |\widetilde{\eta}_\omega|^2/t_N = (\Re(\widetilde{\eta}_\omega)^2 + \Im(\widetilde{\eta}_\omega)^2)/t_N$ is

$$p\left(P^{(y)}\right) dP^{(y)} = \frac{dP^{(y)}}{2\sigma^2}\, e^{-\frac{P^{(y)}}{2\sigma^2}} \ ,$$

i.e., an exponential distribution.

B.2.2 Power spectrum of diffusion on DNA

In this section we derive expressions for the power spectra of a particle diffusing on stretched DNA. We first derive the power spectrum of the transversal positions in Sec. (B.2.2.1) and then the more complicated expression for the power spectrum of the longitudinal displacements (Sec. B.2.2.2).

B.2.2.1 Transversal movement

The transversal position of a diffusing particle at time t is

$$y_t = y(s(t), t) \ .$$

Its expectation value $\langle y_t \rangle$ is trivially equal to zero, where the expectation value is taken first over the DNA's movement conditioned on the protein's position on the DNA and then over the proteins movement on the DNA[6] , i.e.,

$$\langle y(s(t),t)\rangle = \int_{-s_0}^{L-s_0} \frac{e^{-\frac{(s(t)-s_0)^2}{4Dt}}}{\sqrt{4\pi Dt}} \langle y(s(t),t)\rangle_{\mathrm{DNA}}\, d(s(t)-s_0) \ .$$

[6]There are two ways of calculating the following statistics, we have chosen to use conditional expectations. Another way, using a differential approach, runs into mathematical complications, i.e., expectation values, which are not well defined, since the DNA movement is defined by a first-order stochastic differential equation in time.

The autocovariance function of y_t is

$$\langle y_t y_{t'} \rangle = \frac{k_B T}{\gamma_\perp} \sum_{k=1}^{\infty} \frac{y_k[s(t)] y_k[s(t')]}{c_k} \; e^{-c_k|t-t'|} \; .$$

Since the distance traveled by the protein during the measurement is small compared to the length of the DNA we can expand the eigenfunctions around the particle's mean position on the DNA[7], $\overline{s}$,

$$y_k[s(t)] \simeq y_k(\overline{s}),$$

where higher order terms order can safely be ignored, since they are of order $2Dt/L^2 \ll 1$. Thus,

$$\langle y_t y_{t'} \rangle \simeq \sum_{k=1}^{\infty} \mathcal{Y}_k(\overline{s})^2 e^{-c_k|t-t'|} \; . \tag{B.2.9}$$

The power spectrum of y_t is

$$\begin{aligned}
\langle P_f[y(\overline{s})] \rangle &= \frac{1}{t_N} \int_0^{t_N} \int_0^{t_N} e^{-i2\pi f(t-t')} \langle\langle y_t y_{t'} \rangle\rangle \, dt' dt \\
&\simeq \frac{1}{t_N} \sum_{k=1}^{\infty} \mathcal{Y}_k(\overline{s})^2 \int_0^{t_N} \left(\int_0^t e^{-(i2\pi f + c_k)(t-t')} dt' + \int_t^{t_N} e^{-(i2\pi f - c_k)(t-t')} dt' \right) dt \\
&= \sum_{k=1}^{\infty} 2\mathcal{Y}_k(\overline{s})^2 \left(\frac{c_k}{c_k^2 + (2\pi f)^2} + \frac{\left(1 - e^{-c_k t_N}\right)\left[c_k^2 - (2\pi f)^2\right]}{t_N \left[c_k^2 + (2\pi f)^2\right]^2} \right) \; ,
\end{aligned} \tag{B.2.10}$$

for $f \neq 0$. When $c_k t_N \gg 1$, which is usually the case for experimental measurements, the second term in Eq. (B.2.10) can be ignored.

For a stationary process, such as y,

$$\widetilde{\Delta y} \equiv \mathcal{F}\left[y(t) - y(t - \Delta t)\right]_f \simeq \left(1 - e^{-i2\pi f \Delta t}\right) \widetilde{y}_f \; ,$$

when N is large. So

$$P_f(\Delta y) = [2 - 2\cos(2\pi f \Delta t)] P_f(y) \; . \tag{B.2.11}$$

Equations (B.2.11) and (B.2.10), along with Eq. (B.2.6) give Eqs. (3.2.4) and (3.2.6).

[7]Differentiation of the spatial eigenfunction gives a factor of α_k/L, where α_k is the wavenumber and c_k is a $(q-2)$th order polynomial in α_k times α_k^2, where q is the order of the differential operator $\mathcal{L}_s$. Thus each following term in the Taylor-expansion is approximately a factor $\alpha_k^2 Dt/L^2$ smaller than the preceding. This factor is small, since the typical diffusion length of the protein is very small compared to the length of the DNA ($\sqrt{2Dt} \sim$ 0.1-1 μm compared to $L \sim 18\mu$m for λ-phage DNA).

Distribution of the power spectrum The measured transversal position of the particle on the DNA is a sum of two Gaussian terms: the position of the DNA $y(\overline{s}, t_n)$ and the positional noise $\sigma\xi_n$. The power spectrum of the transversal position is thus exponentially distributed.

B.2.2.2 Longitudinal movement

Since the position of the freely diffusion particle is not a stationary process the power spectrum of the positions on the longitudinal axis, the x-axis, is not meaningful (Sec. B.2.1.2). Instead we look at the protein's displacement during one time-lapse, $\Delta x(t) = x(t) - x(t - \Delta t)$, which is equal to

$$\Delta x(t) = \int_0^{s(t)} \frac{\partial x}{\partial s}(s,t)ds - \int_0^{s(t-\Delta t)} \frac{\partial x}{\partial s}(s, t - \Delta t)ds \ .$$

Using that the length of the DNA is large compared to the diffusion length of the particle[8] we can Taylor expand Δx around $\overline{s}$ to get

$$\begin{aligned} \Delta x(t) &\simeq \frac{\partial x}{\partial s}(s,t)\Delta s(t) + \int_0^{\overline{s}} \left\{ \frac{\partial x}{\partial s}(s,t) - \frac{\partial x}{\partial s}(s, t - \Delta t) \right\} ds \\ &= \frac{\partial x}{\partial s}(\overline{s}, t)\Delta s(t) + \Delta x^{\mathrm{DNA}}(\overline{s}, t) \ , \end{aligned} \tag{B.2.12}$$

where $\Delta s(t) \equiv s(t) - s(t - \Delta t)$, and $\Delta x^{\mathrm{DNA}}(\overline{s}, t) \equiv x^{\mathrm{DNA}}(\overline{s}, t) - x^{\mathrm{DNA}}(\overline{s}, t - \Delta t)$ is the displacement of the point $\overline{s}$ on the DNA contour during one time-lapse[9]. We will adopt different strategies for calculating the power spectra of the two terms in Eq. (B.2.12). Since the DNA movement is assumed to be uncorrelated with the Brownian motion of the particle, the power spectrum of $\Delta x(t)$ is just the sum of the power spectrum of the two terms. The power spectrum of the longitudinal displacements of the DNA is

$$P_f\left(\Delta x^{\mathrm{DNA}}\right) = [2 - 2\cos(2\pi f \Delta t)] P_f\left(x^{\mathrm{DNA}}\right) \ . \tag{B.2.13}$$

For the diffusive part of the motion, we use that $\partial x/\partial s$ and s are uncorrelated. So from Eq. (2.1.4) the autocovariance of the diffusive part of the observed displacements is

$$\langle \Delta s_m \Delta s_n \rangle = \left(2\zeta(\overline{s})^2 D\Delta t - \frac{2\zeta(\overline{s})^2 D\tau}{3} \right) \delta_{m,n} + \frac{\zeta(\overline{s})^2 D\tau}{3} (\delta_{n,m+1} + \delta_{n,m-1}) \ ,$$

[8] During both one time-lapse and the whole measurement.

[9] Equation (B.2.12) is obtained by first order Taylor expansion. If we go to second order in the Taylor expansion we get the interesting result: $\langle\langle \Delta x(t) \rangle\rangle \simeq D\Delta t \langle\langle \partial^2 x/\partial s^2 \rangle\rangle$, i.e., the diffusion looks biased on a substrate with asymmetric tortuosity. Only ever so slightly, though.

where $\Delta s_n = \int_0^{\Delta t} \Delta s(t_n - t)dt/\Delta t$ and ζ is the local degree of stretching of the DNA, given by Eq. (B.1.10), since

$$\left\langle \frac{\partial x}{\partial s}(s,t)\frac{\partial x}{\partial s}(s,t') \right\rangle = \left(1 - \frac{1}{2}\left[\rho_{y'}(s) + \rho_{z'}(s)\right]\right)^2 + \frac{1}{2}\left[\rho_{y'}(s,t-t')^2 + \rho_{z'}(s,t-t')^2\right] \ ,$$

where $[\rho_{y'}(s,t-t')^2 + \rho_{z'}(s,t-t')^2]/2$ is negligible, since $\rho_{y'}(s) \ll 1$. So the expected values of the aliased power spectrum of the full time-integrated diffusive displacements are

$$\left\langle \widehat{P}_f(\Delta s) \right\rangle = 2\zeta(\overline{s})^2 D(\Delta t)^2 - \left(2 - \frac{2(N-1)}{N}\cos(2\pi f \Delta t)\right)\frac{\zeta(\overline{s})D\Delta t\tau}{3} \ . \tag{B.2.14}$$

Finally, by inserting Eqs. (B.2.13) and (B.2.14) in Eq. (B.2.6) we find the power spectrum of the measured displacements, Eqs. (3.2.5) and (3.2.7).

Distribution of the power spectrum Using Eqs. (B.2.12) and (B.1.10), the observed longitudinal displacement of a protein diffusing on DNA can be decomposed into

$$\Delta x_n = \sqrt{\zeta(\overline{s})}\Delta s_n + \Delta x_n^{\mathrm{DNA}} + \sigma \Delta \xi_n \ , \tag{B.2.15}$$

where Δs_n and $\Delta \xi_n = \xi_n - \xi_{n-1}$ are Gaussian. The x-direction DNA movement is not Gaussian, since it is a non-linear function of the transversal movement[10]. This means that $\Delta x_n^{\mathrm{DNA}}$ is not Gaussian. Since its Fourier transform is a weighted sum over all the measured $\Delta x_n^{\mathrm{DNA}}$ it is almost Gaussian, however. Furthermore, the measured Fourier transform is a sum of the Fourier transforms of the three independent terms given in Eq. (B.2.15), of which two are Gaussian and one almost Gaussian. The assumption that the measured Fourier transform is Gaussian distributed is thus a safe bet and we conclude that the power spectrum of the measured longitudinal displacements is exponentially distributed.

B.2.3 Autocovariance function for diffusion on DNA

In this section we derive the autocovariance function of the measured longitudinal displacements of a particle diffusing on DNA.

As in the previous section we assume that the distance traveled by the protein on the DNA during the time-lapse Δt is negligible. From Eq. (B.2.15) we then have

$$\langle d_n d_{n+j} \rangle \simeq \langle \Delta s_n \Delta s_{n+j} \rangle + \langle \Delta \xi_n \Delta \xi_{n+j} \rangle + \overline{\rho}_{\Delta x}(\overline{s}, j\Delta t, \tau) \ , \tag{B.2.16}$$

[10] In the general univariate case: we let $x(t) \equiv f(z_t)$, then $dx_t = f'(z_t)dz_t + \frac{1}{2}f''(z_t)(dz_t)^2$, using Itô calculus.

where $\overline{\rho}_{\Delta x}(\overline{s}, j\Delta t, \tau) = 2\overline{\rho}_x(\overline{s}, j\Delta t, \tau) - \overline{\rho}_x(\overline{s}, (j-1)\Delta t, \tau) - \overline{\rho}_x(\overline{s}, (j+1)\Delta t, \tau)$. The autocovariance of the motion blurred positions of the DNA, $\overline{\rho}_x$ is given by:

$$\begin{aligned}
\overline{\rho}_x(\overline{s}, 0; \tau) &= \frac{1}{\tau^2}\int_0^\tau\int_0^\tau \rho_x(\overline{s}, t'-t'')dt''dt' \\
&= \sum_{kl=1}^{\infty} \frac{\mathcal{X}_{k,l}(\overline{s})^2}{\tau^2}\int_0^\tau\int_0^\tau e^{-(c_k+c_l)|t'-t''|}dt''dt' \\
&= \sum_{kl=1}^{\infty} 2\frac{(c_k+c_l)\,\tau - \left(1-e^{-(c_k+c_l)\tau}\right)}{\tau^2(c_k+c_l)^2}\,\mathcal{X}_{k,l}(\overline{s})^2 \;,
\end{aligned}$$

for zero time-separation; and

$$\begin{aligned}
\overline{\rho}_x(\overline{s}, j\Delta t; \tau) &= \frac{1}{\tau^2}\int_0^\tau\int_0^\tau \rho_x(\overline{s}, j\Delta t - t' + t'')dt''dt' \\
&= \sum_{kl=1}^{\infty} \frac{\mathcal{X}_{k,l}(\overline{s})^2}{\tau^2}\, e^{-(c_k+c_l)j\Delta t}\int_0^\tau e^{-(c_k+c_l)t'}dt' \int_0^\tau e^{(c_k+c_l)t''}dt'' \\
&= \sum_{kl=1}^{\infty} \frac{2e^{-(c_k+c_l)j\Delta t}\{\cosh[(c_k+c_l)\tau]-1\}}{\tau^2(c_k+c_l)^2}\,\mathcal{X}_{k,l}(\overline{s})^2 \;, \quad \text{(B.2.17)}
\end{aligned}$$

for time-separation $j\Delta t$.

For full time-integration Eq. (B.2.16) gives Eq. (2.2.2), and Eqs. (B.2.17) and (B.2.17) reduce to Eqs (3.2.18) and (3.2.19).

B.3 Estimation

In this section we derive properties of the MLE and the CVE for diffusion on a fluctuating substrate. Maximum likelihood estimation is treated in Sec. B.3.1 and covariance-based estimation is treated in Sec. B.3.2

B.3.1 Maximum likelihood estimation

To extract as much information as possible from experimental data, we want to fit to both the y-direction and x-direction power spectra for diffusion on DNA. To do this, we assume that the measured y- and x-displacements are uncorrelated. This assumption is not true, since the x-direction DNA fluctuations are fully determined by the y- and z-direction fluctuations[11] However, as simulations show (Fig. B.7), the correlations are small and neglecting them does not influence the estimation significantly.

[11] Each transversal directions contribute to approximately half of the x-direction fluctuations.

B.3.1.1 Variance of the MLE for known DNA fluctuations

It is a well known result that the variance of the MLE is given by the inverse of the Fisher Information matrix asymptotically, i.e., to order $1/N$ [17]. In this section we derive an asymptotic result for the variance of the MLE where some parameters are estimated a priori with given uncertainties on these estimates.

Let $\boldsymbol{\phi}$ define the parameters estimated a priori and $\boldsymbol{\theta}$ define the parameters, which we want to estimate using maximum likelihood estimation. When we use the estimates $\widehat{\boldsymbol{\phi}}$ as fixed parameters in the estimation of $\boldsymbol{\theta}$, the errors on $\widehat{\boldsymbol{\phi}}$, i.e, the difference between the estimated and true values $\Delta\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*$, propagate to $\widehat{\boldsymbol{\theta}}$. Since $\widehat{\boldsymbol{\theta}}$ is only defined as an implicit function of $\boldsymbol{\phi}$, i.e., by the stationarity condition

$$l_{\theta_i}\left(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\phi}},\mathbf{x}\right) \equiv \frac{\partial \ln \mathcal{L}\left(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\phi}},\mathbf{x}\right)}{\partial\theta_i} = 0 \ , \tag{B.3.1}$$

we cannot use classical propagation of errors to calculate the variance of $\widehat{\boldsymbol{\theta}}$. We can, however, following a derivation similar to the one that shows the approximate equality between the inverse Fisher information and the variance of the MLE, derive a first order approximation (in $1/N$) to the variance of $\widehat{\boldsymbol{\theta}}$.

From the stationarity condition, Eq. (B.3.1),

$$\begin{aligned} 0 &= l_{\theta_i}\left(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\phi}},\mathbf{x}\right) \\ &= l_{\theta_i}\left(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,\mathbf{x}\right) + l_{\theta_i\theta_j}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X)\Delta\theta_j + l_{\theta_i\phi_k}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X)\Delta\phi_k + \mathcal{O}_p(1) \\ &= l_{\theta_i}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X) + (\mathcal{I}_\theta)_{ij}\Delta\theta_j + (\mathcal{I}_\phi^\theta)_{ik}\Delta\phi_k + \mathcal{O}_p(1) \ , \end{aligned}$$

since the central limit theorem dictates that

$$l_{\theta_i\theta_j}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X) = (I_\theta)_{ij} + \mathcal{O}_p(1) \ ,$$

where $\mathcal{I}_\theta$ is the information matrix corresponding to the variance of $\widehat{\boldsymbol{\theta}}$, $\mathcal{I}_\phi^\theta$ is the information matrix corresponding to the covariance between $\widehat{\boldsymbol{\theta}}$, and $\widehat{\boldsymbol{\phi}}$, and we sum over repeated indices. Thus,

$$\Delta\theta_m = -\left(\mathcal{I}_\theta\right)^{mi}\left[l_{\theta_i}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X) + \left(\mathcal{I}_\phi^\theta\right)_{ik}\Delta\phi_k\right] + \mathcal{O}_p\left(N^{-1}\right)$$

where $\mathcal{I}^{mi} = (\mathcal{I}^{-1})_{mi}$, and

$$\begin{aligned} \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)_{mn} &= \langle\Delta\theta_m\Delta\theta_n\rangle \\ &\simeq (\mathcal{I}_\theta)^{mi}\Big[\left\langle l_{\theta_i}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X)l_{\theta_j}(\theta^*|\phi^*,X)\right\rangle + \left(\mathcal{I}_\phi^\theta\right)_{ik}\langle\Delta\phi_k\Delta\phi_l\rangle\left(\mathcal{I}_\phi^\theta\right)_{lj} \\ &\quad + \left(\mathcal{I}_\phi^\theta\right)_{il}\langle l_{\theta_i}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X)\Delta\phi_l\rangle + \left(\mathcal{I}_\phi^\theta\right)_{jk}\left\langle l_{\theta_j}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X)\Delta\phi_k\right\rangle\Big](\mathcal{I}_\theta)^{jn} \\ &= (\mathcal{I}_\theta)^{mn} + (\mathcal{I}_\theta)^{mi}\left(\mathcal{I}_\phi^\theta\right)_{ik}\mathrm{Var}\left(\widehat{\boldsymbol{\phi}}\right)_{kl}\left(\mathcal{I}_\phi^\theta\right)_{lj}(\mathcal{I}_\theta)^{jn} \ , \end{aligned} \tag{B.3.2}$$

since $l_{\theta_j}(\boldsymbol{\theta}^*|\boldsymbol{\phi}^*,X)$ and $\widehat{\phi}_k$ are uncorrelated.

B.3.2 Covariance-based estimation

B.3.2.1 Bias of the covariance-based estimator

The simple covariance-based estimator, which optimally estimates D when there is no substrate fluctuations, is biased in the presence of DNA fluctuations. For full frame averaging, the expectation value of the CVE is

$$\left\langle \widehat{D} \right\rangle = \zeta(\overline{s})^2 D + \frac{\overline{\rho}_x(\Delta t) - \overline{\rho}_x(2\Delta t)}{\Delta t} \ ,$$

Thus, the bias of the CVE is[12]

$$b(D) = \sum_{kl=1}^{\infty} \frac{\left(1 - e^{-(c_k+c_l)\Delta t}\right)^3}{(c_k + c_l)^2 \Delta t^3} \mathcal{X}_{k,l}(\overline{s})^2 \ . \tag{3.2.11}$$

For fast DNA dynamics or long time-lapse, $(c_k + c_l)\Delta t$ tends to zero and so does the bias of the CVE, since DNA fluctuations are interpreted as white noise. One may reduce the bias of the moment-based estimator by adding higher autocorrelations, i.e.,

$$\widehat{D}_J = \frac{\overline{d_n^2}}{2\Delta t} + \frac{1}{\Delta t} \sum_{j=1}^{J} \overline{d_n d_{n+j}} \ .$$

This estimator has a bias of

$$\frac{\rho_x(j\Delta t) - \rho_x((j+1)\Delta t)}{\Delta t} = \sum_{kl=1}^{\infty} \frac{e^{-(j-1)(c_k+c_l)\Delta t}\left(1 - e^{-(c_k+c_l)\Delta t}\right)^3}{\Delta t^3 (c_k + c_l)^2} \mathcal{X}_{k,l}(s_0)^2 \ .$$

However, the variance of the CVE increases considerably by including the higher autocorrelations, similarly to MSD-based methods. Another option, which is much more precise, is to determine parameters of DNA motion a priori and use these to calculate and subtract the bias (Sec. 3.2.1). Provided that the estimates of the DNA parameters $\mathcal{X}_{k,l}$ and c_k are reasonably close to their true values this provides unbiased estimates of the diffusion coefficients, even for short time-series.

Known positional noise variance When the positional noise variance is known a priori the CVE is given by Eq. (3.2.20). For this estimator the bias due to substrate fluctuations is

$$\begin{aligned} b(D) &= \frac{\overline{\rho}_x(0) - \overline{\rho}_x(\Delta t)}{\Delta t(1 - 2R)} \\ &= \sum_{kl=1}^{\infty} \frac{2(c_k + c_l)\Delta t - 3 + 4e^{-(c_k+c_l)\Delta t} - e^{-2(c_k+c_l)\Delta t}}{(1 - 2R)\Delta t^3 (c_k + c_l)^2} \ . \end{aligned} \tag{B.3.3}$$

[12] Neglecting the part of the bias, which is due to incomplete stretching of the substrate. This part is taken into account as described in Sec. 3.2.

B.3.2.2 Variance of the covariance-based estimator

From Eq. (3.2.12) the variance of the unbiased covariance estimator is

$$\begin{aligned}\mathrm{Var}\left(\widehat{D}\right) &= \frac{\mathrm{Var}\left(\overline{\Delta x_n^2}\right)}{4(\Delta t)^2} + \frac{\mathrm{Var}\left(\overline{\Delta x_n \Delta x_{n+1}}\right)}{(\Delta t)^2} + \frac{\left\langle \overline{\Delta x_n^2}, \overline{\Delta x_n \Delta x_{n+1}} \right\rangle_{\mathcal{C}}}{(\Delta t)^2} \\ &\quad + \mathrm{Var}\left(\widehat{b}(D)\right) \quad . \end{aligned} \tag{B.3.4}$$

The variance of the bias estimate is found by standard propagation of errors and is given in Eq. (3.2.16). The other three terms are calculated as in Sec. A.3.2, where we must take contributions from DNA fluctuations into account as well. We define $\rho_j \equiv \overline{\rho}_{\Delta x}(\overline{s}, j\Delta t)$, then:

$$\begin{aligned}\mathrm{Var}\left(\overline{\Delta x_n^2}\right) &= \frac{\mathrm{Var}\left(\Delta x_n^2\right)}{N} + \frac{2}{N^2}\sum_{j=1}^{N-1}(N-j)\left\langle \Delta x_n^2, \Delta x_{n+j}^2 \right\rangle_{\mathcal{C}} \\ &= \frac{8(\alpha+\beta)^2 + 4(1-1/N)\beta^2}{N} + \frac{8(\alpha+\beta)\rho_0 - 8(1-1/N)\beta\rho_1}{N} \\ &\quad + \frac{2\rho_0^2}{N} + \sum_{j=1}^{N-1}\frac{4(N-j)\rho_j^2}{N^2} \quad , \end{aligned} \tag{B.3.5}$$

since

$$\begin{aligned}\mathrm{Var}\left(\Delta x_n^2\right) &= 8(\alpha+\beta)^2 + 8(\alpha+\beta)\rho_0 + 2\rho_0^2 \ , \\ \left\langle \Delta x_n^2, \Delta x_{n+1}^2 \right\rangle_{\mathcal{C}} &= 2\beta^2 - 4\beta\rho_1 + 2\rho_1^2 \ , \\ \left\langle \Delta x_n^2, \Delta x_{n+j}^2 \right\rangle_{\mathcal{C}} &= 2\rho_j^2 \ , \quad j > 1 \ , \end{aligned}$$

where we have defined $\alpha = \zeta(\overline{s})^2 D\Delta t$ and $\beta = \sigma^2 - \zeta(\overline{s})^2 D\Delta t/3$;

$$\begin{aligned}\mathrm{Var}\left(\overline{\Delta x_n \Delta x_{n+1}}\right) &= \frac{\mathrm{Var}\left(\Delta x_n \Delta x_{n+1}\right)}{N-1} \\ &\quad + \sum_{j=1}^{N-2}\frac{2(N-j-1)}{(N-1)^2}\left\langle \Delta x_n \Delta x_{n+1}, \Delta x_{n+j}\Delta x_{n+j+1} \right\rangle_{\mathcal{C}} \\ &= \frac{4(\alpha+\beta)^2 + 3\beta^2}{N-1} - \frac{2\beta^2}{(N-1)^2} \\ &\quad + \frac{4(\alpha+\beta)\rho_0 - 6\beta\rho_1 + 4(\alpha+\beta)\rho_2 - 2\beta\rho_3}{N-1} \\ &\quad + \frac{4\beta\rho_1 - 4(\alpha+\beta)\rho_2 + 4\beta\rho_3}{(N-1)^2} \\ &\quad + \frac{\rho_0^2 + \rho_1^2}{N-1} + \sum_{j=1}^{N-2}\frac{2(N-j-1)}{(N-1)^2}\left(\rho_{j-1}\rho_{j+1} + \rho_j^2\right) \quad , \end{aligned} \tag{B.3.6}$$

since

$$\begin{aligned}
\mathrm{Var}\left(\Delta x_n \Delta x_{n+1}\right) &= 4(\alpha+\beta)^2 + \beta^2 + 4(\alpha+\beta)\rho_0 - 2\beta\rho_1 + \rho_0^2 + \rho_1^2 \ , \\
\langle \Delta x_n \Delta x_{n+1}, \Delta x_{n+2}\Delta x_{n+3}\rangle_{\mathcal{C}} &= \beta^2 - 2\beta\rho_1 + 2(\alpha+\beta)\rho_2 + \rho_0\rho_2 + \rho_1^2 \ , \\
\langle \Delta x_n \Delta x_{n+1}, \Delta x_{n+2}\Delta x_{n+3}\rangle_{\mathcal{C}} &= -\beta\rho_3 + \rho_1\rho_3 + \rho_2^2 \ , \\
\langle \Delta x_n \Delta x_{n+1}, \Delta x_{n+j}\Delta x_{n+j+1}\rangle_{\mathcal{C}} &= \rho_{j-1}\rho_{j+1} + \rho_j^2 \ , \quad j > 2 \ ;
\end{aligned}$$

$$\begin{aligned}
\left\langle \overline{\Delta x_n^2}, \overline{\Delta x_n \Delta x_{n+1}} \right\rangle_{\mathcal{C}} &= \sum_{j=0}^{N-2} \frac{2(N-j-1)}{N(N-1)} \left\langle \Delta x_n^2, \Delta x_{n+j}\Delta x_{n+j+1}\right\rangle_{\mathcal{C}} \\
&= -\frac{8(\alpha+\beta)\beta}{N} + \frac{-4\beta\rho_0 + 8(\alpha+\beta)\rho_1 - 4[1-1/(N-1)]\beta\rho_2}{N} \\
&\quad + \sum_{j=0}^{N-2} \frac{4(N-j-1)\rho_j\rho_{j+1}}{N(N-1)} \ , \qquad \text{(B.3.7)}
\end{aligned}$$

since

$$\begin{aligned}
\left\langle \overline{\Delta x_n^2}, \overline{\Delta x_n \Delta x_{n+1}} \right\rangle_{\mathbb{C}} &= -4(\alpha+\beta)\beta - 2\beta\rho_0 + 4(\alpha+\beta)\rho_1 + 2\rho_0\rho_1 \ , \\
\left\langle \overline{\Delta x_n^2}, \overline{\Delta x_{n+1} \Delta x_{n+2}} \right\rangle_{\mathcal{C}} &= -2\beta\rho_2 + 2\rho_1\rho_2 \ , \\
\left\langle \overline{\Delta x_n^2}, \overline{\Delta x_{n+j} \Delta x_{n+j+1}} \right\rangle_{\mathcal{C}} &= 2\rho_j\rho_{j+1} \ , \quad j > 1 \ .
\end{aligned}$$

Combining Eqs. (B.3.5)-(B.3.7) gives Eq. (3.2.13).

B.4 Numerical results

B.4.1 Simulating diffusion on DNA

Since we only simulate the lowest mode of the DNA motion we can simulate the coupling between the two transversal modes and the longitudinal mode exactly without having to simulate the entire DNA strand. We can calculate the DNA's x-direction motion from its local y- and z-direction motion using that

$$\begin{aligned}
x(s,t) &= s - \frac{1}{2}\int_0^s \left(y'(s',t)^2 + z'(s',t)^2\right) ds' \\
&= s - \frac{1}{2}\int_0^s \left(A_1^{(y)}(t)^2 + A_1^{(z)}(t)^2\right) z_1'(s')^2 ds' \\
&= s - \frac{y(s',t)^2 + z(s',t)^2}{2} \int_0^s \frac{y_1'(s')^2}{y_1(s)} ds' \\
&= s - \frac{\mathcal{X}_{1,1}(s)}{2\mathcal{Y}_1(s)^2}\left[y(s',t)^2 + z(s',t)^2\right] \ , \qquad \text{(B.4.1)}
\end{aligned}$$

since the two transversal directions are equivalent[13].

Experimental data are usually measured with full time-integration, i.e., with camera shutter time τ equal to the time-lapse Δt. So to simulate positions we need to integrate over the full time-lapse,

$$\overline{y}_{n+1}^{\mathrm{DNA}} = \frac{1}{\Delta t}\int_{t_n}^{t_n+\Delta t} y^{\mathrm{DNA}}(t_n+t)dt \ , \tag{B.4.2}$$

We do this by approximating the integral with a sum,

$$\overline{y}_{n+1}^{\mathrm{DNA}} \simeq h\sum_{q=1}^{1/h} y^{\mathrm{DNA}}(t_n+qh\Delta t) \ , \tag{B.4.3}$$

where (B.4.3) approaches (B.4.2) as $h \to 0$. The motion of a transversal mode is equivalent to the motion of a Brownian particle trapped in an optical trap. We can thus use the theory of optical tweezers [80] and simulate y^{DNA} according to

$$y^{\mathrm{DNA}}(t_{m+1}) = e^{-c_1 h\Delta t} y^{\mathrm{DNA}}(t_m) + \Delta y_h \eta_m \ , \tag{B.4.4}$$

where $m = 1, 2, \ldots, (N+1)/h$, η_n is standard white noise, and

$$\Delta y_h \equiv \sqrt{\frac{\mathcal{Y}_1\left(1-e^{-c_1 h\Delta t}\right)}{2c_1}} \ .$$

The y-position of the DNA measured at time t_n is thus

$$\overline{y}_n^{\mathrm{DNA}} = h\sum_{q=1}^{h} y^{\mathrm{DNA}}(t_{n/h+q}) \ . \tag{B.4.5}$$

The diffusive movement of the protein is simulated in a similar fashion,

$$x^{\mathrm{Diff}}(t_{m+1}) = x^{\mathrm{Diff}}(t_m) + \sqrt{2Dh\Delta t}\zeta_m \ ,$$

where $m = 1, 2, \ldots (N+1)/h$ and ζ_m is standard Gaussian white noise.

We simulate three independent time-series, $\{y^{\mathrm{DNA}}(t_m)\}_{m=-Q/h}^{(N+1)/h}$, $\{z^{\mathrm{DNA}}(t_m)\}_{m=-Q/h}^{(N+1)/h}$, and $\{x^{\mathrm{Diff}}(t_m)\}_{m=-Q/h}^{(N+1)/h}$, where we set $x(t_{-Q/h}) = y(t_{-Q/h}) = z(t_{-Q/h}) = 0$, and Q is chosen such that the DNA has time to thermalize before we sample the time-series. We calculate the longitudinal positions of the DNA x^{DNA} from Eq. (B.4.1) and calculate the motion blurred positions $\{\overline{y}^{\mathrm{DNA}}(t_n)\}_{n=0}^{N}$, $\{\overline{x}^{\mathrm{DNA}}(t_n)\}_{n=0}^{N}$, and $\{\overline{x}^{\mathrm{Diff}}(t_n)\}_{n=0}^{N}$ using Eq. (B.4.5).

We finally sum the movement of the DNA and diffusive movement and add positional noise to obtain the "measured" positions,

$$x_n = \overline{x}^{\mathrm{Diff}}(t_n) + \overline{x}^{\mathrm{DNA}}(t_n) + \sigma\xi_n \ ,$$

[13] At least for the lowest mode.

where ξ_n is standard white noise. We also calculate the transversal positions,

$$y_n = y^{\mathrm{DNA}}(t_n) + \sigma\xi_n \ .$$

Table B.1 and Figs. 2.5-2.7 present numerical results for the performance of the MLE, which explicitly accounts for DNA fluctuations, and the CVE, which does not, for different values of the diffusion coefficient D.

B.4.2 Correlations between transversal and longitudinal DNA fluctuations

The optimal way to estimate the parameters from the data of a particle diffusing on DNA is to fit simultaneously to both $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$. To do this in practice, we assume that $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$ are independent. However, since the longitudinal motion of the DNA is dependent on its transversal motion, $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$ are not completely independent. We calculate the correlations between $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$ for an ensemble of 10,000 Monte Carlo generated time-series of length $N = 100$. As Fig. B.7 shows, the correlations between $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$ are in general small and can globally be ignored.

$D = 0.1\ \mu\text{m}^2/\text{s}$

$2\pi f_c$	**MLE w/ DNA**		**CVE w/o DNA**	
	Bias [μm^2/s]	Variance [μm^4/s^2]	Bias [μm^2/s]	Variance [μm^4/s^2]
20 Hz	0.004 ± 0.003 (0.0)	0.0086 ± 0.0007 (0.0100)	0.244 ± 0.004 (0.245)	0.018 ± 0.001
30 Hz	-0.001 ± 0.002 (0.0)	0.0045 ± 0.0003 (0.0053)	0.120 ± 0.002 (0.121)	0.0060 ± 0.0004
40 Hz	-0.003 ± 0.002 (0.0)	0.0028 ± 0.0002 (0.0037)	0.067 ± 0.002 (0.068)	0.0032 ± 0.0002

$D = 0.3\ \mu\text{m}^2/\text{s}$

$2\pi f_c$	**MLE w/ DNA**		**CVE w/o DNA**	
	Bias [μm^2/s]	Variance [μm^4/s^2]	Bias [μm^2/s]	Variance [μm^4/s^2]
20 Hz	-0.015 ± 0.006 (0.0)	0.045 ± 0.002 (0.051)	0.244 ± 0.005 (0.245)	0.027 ± 0.002
30 Hz	-0.018 ± 0.005 (0.0)	0.021 ± 0.001 (0.030)	0.120 ± 0.003 (0.121)	0.00129 ± 0.0008
40 Hz	-0.019 ± 0.004 (0.0)	0.0146 ± 0.0007 (0.0222)	0.068 ± 0.003 (0.068)	0.00092 ± 0.0005

$D = 0.5\ \mu\text{m}^2/\text{s}$

$2\pi f_c$	**MLE w/ DNA**		**CVE w/o DNA**	
	Bias [μm^2/s]	Variance [μm^4/s^2]	Bias [μm^2/s]	Variance [μm^4/s^2]
20 Hz	-0.039 ± 0.008 (0.0)	0.072 ± 0.003 (0.113)	0.245 ± 0.006 (0.245)	0.040 ± 0.002
30 Hz	-0.038 ± 0.007 (0.0)	0.044 ± 0.002 (0.070)	0.121 ± 0.005 (0.121)	0.024 ± 0.001
40 Hz	-0.039 ± 0.006 (0.0)	0.033 ± 0.002 (0.053)	0.068 ± 0.004 (0.068)	0.019 ± 0.001

Table B.1: Bias and variance of the MLE, which explicitly accounts for DNA fluctuations, and the CVE, which does not. Values are mean plus/minus s.e.m. obtained from Monte Carlo simulations of an ensemble of 1,000 time-series of length $N = 100$. Values inside parenthesis are theoretically expected values. The positional noise variance is $\sigma^2 = 1,500\ \text{nm}^2$ and the amplitude parameters for DNA fluctuations are $c_1\mathcal{X}_{1,1} = 2.1$ m/s and $c_1\mathcal{Y}_1 = 0.20\ \text{m}^2/\text{s}$.

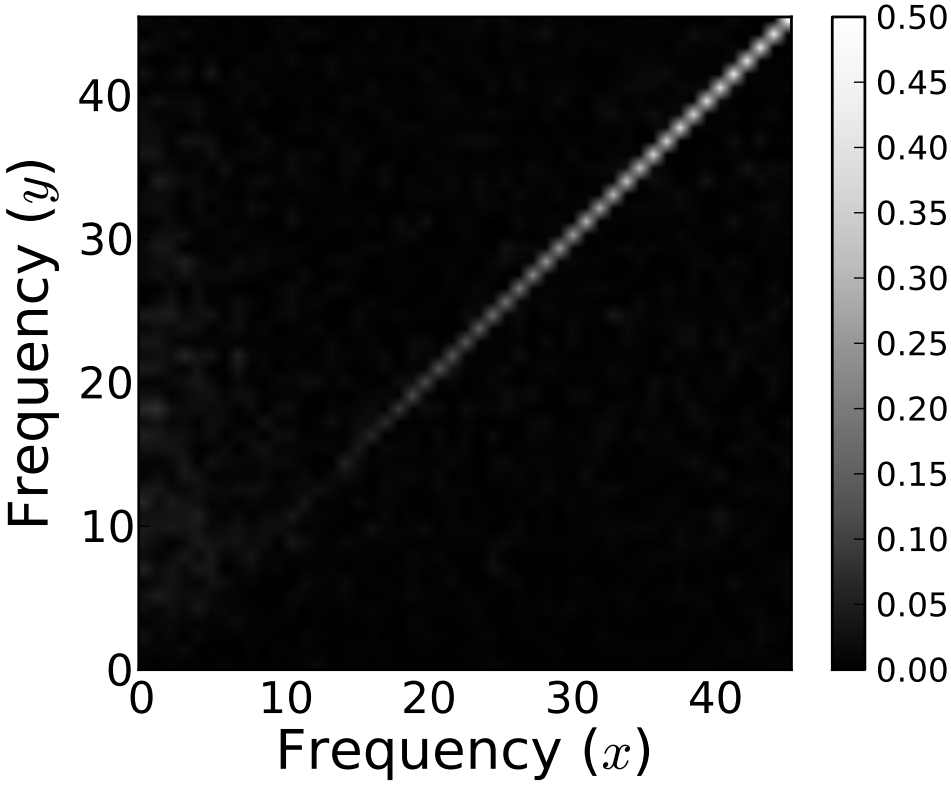

Figure B.7: Color-coded contour plot, which shows the correlations between $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$. From Monte Carlo generated data with $D = 0.3$ μm^2/s, $c_1 = 30$ Hz, $\mathcal{X}_{1,1} = 0.07$ μm, $\mathcal{Y}_1 = 0.08$ μm, and $\sigma^2 = 1500$ nm. The ensemble size is $M = 10,000$, time-series length is $N = 100$, and shutter-time is equal to time-lapse, $\tau = \Delta t$. The plot shows that we do not overall make a big error when assuming that $\widehat{P}_f^{(\Delta y)}$ and $\widehat{P}_f^{(\Delta x)}$ are uncorrelated. The maximal value of the correlations is $\max \rho_{y,d} = 0.47$ at the Nyquist frequency of both y and x, while the mean correlation coefficient is $\overline{\rho} = 0.0370 \pm 0.0004$.

Appendix C

hOgg1 repair proteins diffusing on flow-stretched DNA

C.1 Preliminary data analysis

Our analysis and modeling of the experimental data are based on several assumptions, namely that measured proteins undergo free diffusion on the DNA, that DNA fluctuations does not influence protein dynamics, and that drift is negligible. In this section we design and apply a series of simple tests designed to check if data agree with our assumptions, and procedures to deal with data that does not. In Sec. C.1.1 we test if the recorded transversal movement of individual proteins is consistent with the motion of a protein bound to a fluctuating DNA molecule. In Sec. C.1.2 we test the assumption that DNA fluctuations do not influence protein dynamics. We define an "experimental" window on the DNA, where this assumption is reasonable. Finally, in Sec. C.1.3 we check for drift in the longitudinal direction.

C.1.1 Checking individual time-series

C.1.1.1 Test for normality of transversal movement

We test the y-direction movement for normality using D'Agostino and Pearson's omnibus test [81]. This test generally has better power than Kolmogorov-Smirnov (KS) and chi-squared tests [82]. Normal distribution of the measured y-position is not the only relevant criteria for checking time-series, and the test is not very powerful for short time-series (we expect several good time-series

to show p-values smaller than 0.01, since we have recorded more than 500). So we combine the test with visual inspection of both the transversal trajectories and the distribution transversal positions.

Time-series, for which the test gives a p-value of less than 0.001 are discarded. (We expect a 50% chance for one valid time-series to give a p-value smaller than 0.001 and a 0.05% chance for two.) If the low p-value is due to a single outlier at either the start or the end of the time-series, the point is cut away from the time-series and the rest is used. This is done, since the single outlier is probably caused by the protein being recorded just before it binds to the DNA or right after it unbinds.

C.1.1.2 Visual inspection and short time-series

All time-series are inspected visually for proteins jumping in the y-direction (Fig. C.3), proteins getting stuck (Fig. C.4), and proteins showing drift in the y-direction (Fig. C.5). Time-series showing a jump at the start or the end are not discarded but are cropped as described above (Fig. C.2).

Furthermore, time-series shorter than $N_{\text{min}} = 12$ (corresponding to a residence time of $t_{\text{min}} = 0.132$ s) are discarded from the analysis, since it is impossible to reliably discern valid and invalid measurements when time-series are too short. The choice of N_{min} is somewhat arbitrary. However, the particular choice of N_{min} does not influence our results significantly (Figs. 2.11 and D.1).

In all, 33 out of 411 time-series longer than $N_{\text{min}} = 12$ are discarded, which amounts to eight percent.

C.1.2 Proteins on DNA

We perform a few simple tests investigate if the proteins interact with the DNA the way we expect them to. These are all based on the proteins' positions on the DNA.

We assume that DNA fluctuations do not affect the proteins' kinetics. So we expect, e.g., that the proteins' mean residence time is the same everywhere on the DNA. As Figs. C.6 and C.7 show, this is not the case. Upstream on the DNA, close to the tethering point, proteins have a high likelihood of getting stuck and then discarded. We thus tend to keep only short time-series, for which the protein unbinds before sticking to the coverslip. At the opposite end, near the free end of the DNA, proteins also have a shorter mean residence time. This is probably because the DNA motion increases appreciably near the

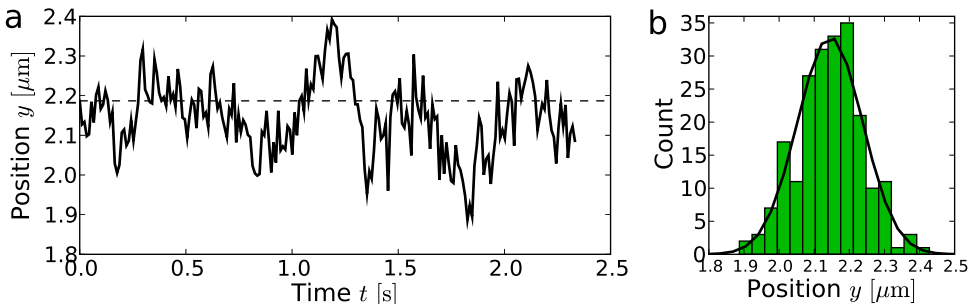

Figure C.1: An experimental trajectory which shows no anomalies. D'Agostino and Pearson's omnibus test gives a p-value of $p = 0.88$ and does not reject the hypothesis that the transversal positions are normal distributed. a) Measured trajectory of transversal movement. b) Distribution of measured transversal positions y. (Measured distribution: green bars; normal distribution with the same mean and variance as the experimental trajectory: black line.)

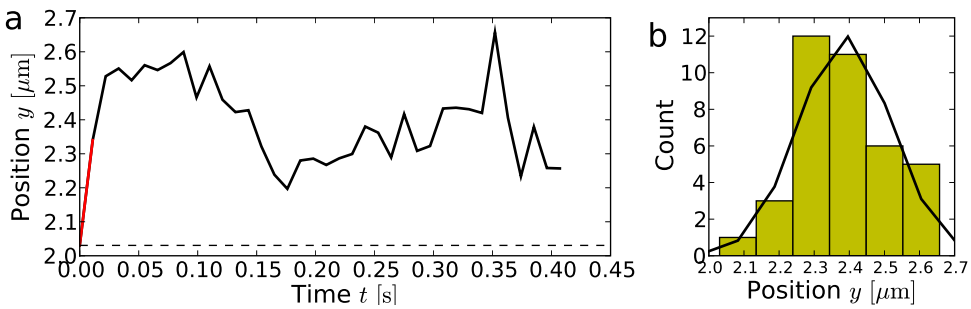

Figure C.2: An experimental trajectory showing an abnormal jump at the beginning of the time-series. The first point is discarded from the analysis. D'Agostino and Pearson's omnibus test gives $p = 0.85$. a) Measured trajectory of transversal movement. b) Distribution of measured transversal positions y. (Measured distribution: yellow bars; normal distribution with the same mean and variance as the experimental trajectory: black line.)

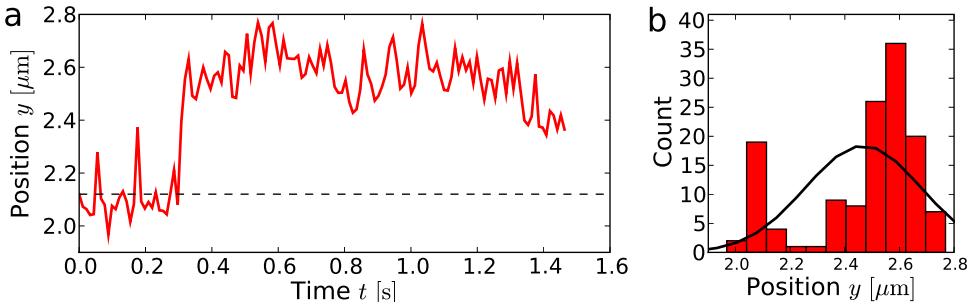

Figure C.3: An experimental trajectory that shows a jump in y-position. D'Agostino and Pearson's omnibus test gives $p = 0.0002$, i.e., it rejects the time-series. a) Measured trajectory of transversal movement. b) Distribution of measured transversal positions y. (Measured distribution: red bars; normal distribution with the same mean and variance as the experimental trajectory: black line.)

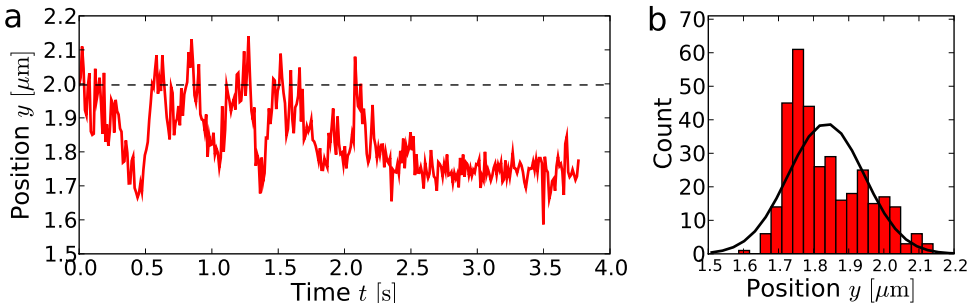

Figure C.4: An experimental trajectory of a protein that gets stuck. D'Agostino and Pearson's omnibus test gives $p = 3 \cdot 10^{-6}$. a) Measured trajectory of transversal movement. b) Distribution of measured transversal positions y. (Measured distribution: red bars; normal distribution with the same mean and variance as the experimental trajectory: black line.)

free end and thus shakes the proteins off[1]. Thus to avoid spurious results we only include time-series with a mean position on the DNA inside the interval $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$. For completeness, results for time-series outside of this "experimental window" are included in the following plots, marked by open symbols.

From our assumption that DNA fluctuations do not affect the proteins, we also expect to see a uniform distribution of the proteins along the DNA (Fig. C.8). This does seem to be the case of proteins inside the experimental window. For proteins near the tethering point, however, we clearly see that a lot of time-series have been discarded. Near the free end we do not see an effect of proteins being shaken off, since the lower density of proteins on the DNA is countered by an increase in the observed density of the DNA, due to an increase in the DNA's tortuosity.

We finally plot the variance of the transverse fluctuations as a function of the proteins' position on the DNA (Fig. C.9). We see that the variance increases along the DNA, as we would expect. This plot can be compared with the theoretical pictures for plug- and shear flows (Fig. B.3 and Fig. B.5 respectively). They can not be compared directly, however, since Fig. C.9 shows the variance as function of the x-position in the lab-frame, not the position s on the DNA contour as in Figs. B.3 and Fig. B.5.

C.1.3 Drift

To check for drift of the proteins, due to the flow, we have calculated the mean displacement during one time-lapse for the time-series included in the analysis, i.e., time-series inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$, which are longer than $t_{\text{min}} = 0.132$ s (Including all time-series does not change the results). The mean drift of the proteins is $\overline{d} = (0.0017 \pm 0.0004)\ \mu\text{m}$, clearly significant (Fig. C.10). However, the standard deviation of a displacement is $\overline{\text{std}\,(d)} = 0.099 \pm 0.003$, i.e., 60 times higher than the drift. Since the diffusion coefficient is calculated from a quadratic function of the displacements, the relative bias due to drift is of the order of $\overline{d}^2/\overline{\text{Var}\,(d)} = 1/4000$.

C.2 Hypothesis testing and model comparison

We use a few statistical tests in the analysis of the experimental data. These are briefly described here.

[1]Near the free end the DNA is not taut since the drag force tends to zero. This also means that our results for the DNA motion are not strictly valid here.

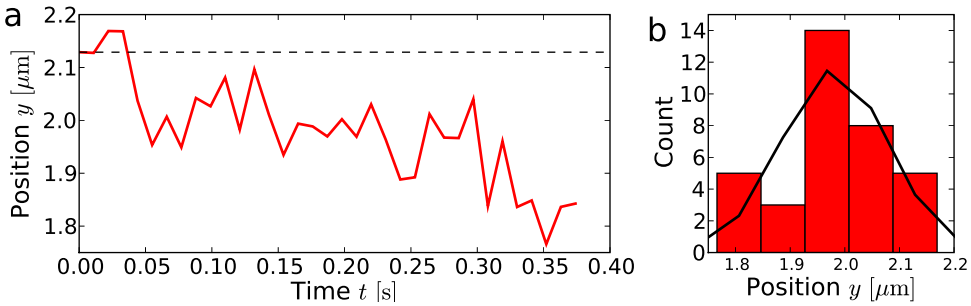

Figure C.5: An experimental trajectory of a protein showing drift in the y-direction. D'Agostino and Pearson's omnibus test gives $p = 0.98$, it does not reject the time-series. a) Measured trajectory of transversal movement. b) Distribution of measured transversal positions y. (Measured distribution: red bars; normal distribution with the same mean and variance as the experimental trajectory: black line.)

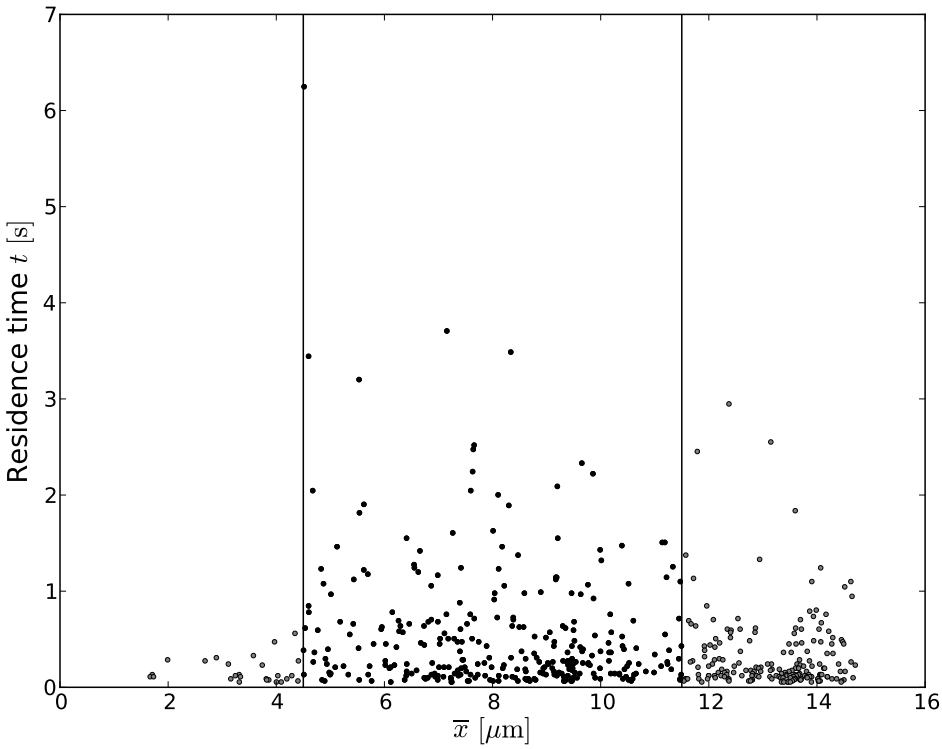

Figure C.6: Scatter-plot of residence time on DNA versus mean longitudinal position $\overline{x}$. Vertical lines mark the experimental window, $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$, where time-series are included in the following analysis.

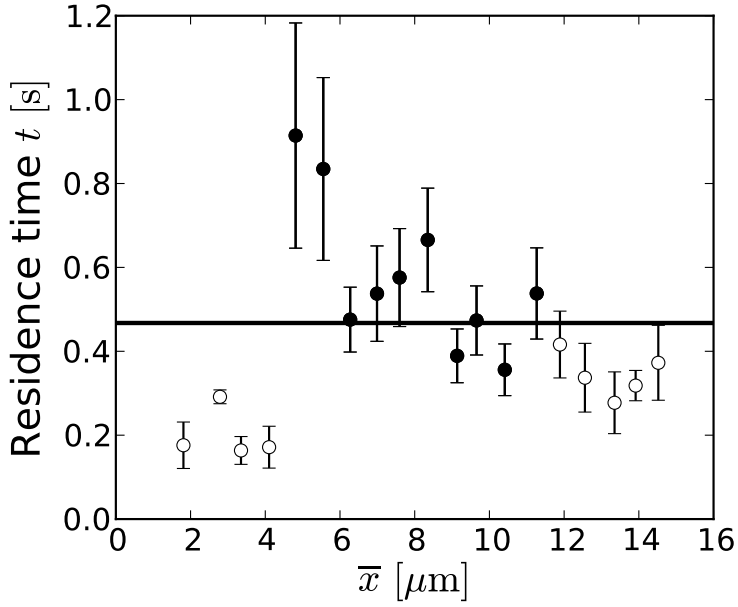

Figure C.7: Mean residence time on DNA plus/minus s.e.m. as a function of mean longitudinal position of the protein, $\overline{x}$. The average residence times on the DNA are shorter at the ends. (Time-series inside experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled circles; time-series outside experimental window: open circles.) Data support the hypothesis that mean time-series length is constant on the interval $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$. A chi-squared test for variance for constant mean gives $p = 0.10$. The mean time-series length is $\overline{N} = .462 \pm .03$ s.

C.2.1 Pearson's chi-squared test

Pearson's chi-square test, also known simply as the goodness of fit test, is used to compare a measured distribution to a theoretical distribution. We use it in particular to test our models for the distribution of residence times against experimental data.

For use in Pearson's chi-square test the measured data is divided into a number of bins q. The number of bins should be large enough such that the number of counts in a bin is Poisson distributed, while on the other hand, the number of counts per bin, O_i should be large enough such that the Poisson distribution is approximately Gaussian (as a rule of thumb, $O_i \geqslant 5$). A chi-square test statistic can then be computed as

$$\chi^2 = \sum_{i=1}^{q} \frac{(O_i - E_i)^2}{E_i} \quad ,$$

where O_i is the observed number of counts and E_i is the expected number of counts in bin i. This statistic follows a standard chi-squared distribution with $q-1$ degrees of freedom and can be used to calculate the support for a theory, i.e., a p-value, which gives the probability of randomly getting a χ^2-value equal to or higher than the one observed if the theory is true.

We design our test such that approximately 20 observed counts are in each bin and the number of bins is equal to $q = 11$.

C.2.2 Chi-squared test for variance

The chi-squared test for variance consist of estimating whether the dispersion of a number of observed data points $(\theta_i)_{i=0,1,\ldots,q}$ with known variances σ_i^2 is consistent with the assumption that the points are normally distributed around the same mean $\overline{\theta}$. The chi-square test statistic is given by

$$\chi^2 = \sum_{i=1}^{q} \frac{(\theta_i - \overline{\theta})^2}{\sigma_i^2} \quad ,$$

and follows a chi-squared distribution with $q-1$ degrees of freedom.

In practice we bin measurements and use the sample variances s_i^2 as estimates of the true variances σ_i^2 of the estimated means θ_i. Thus the test statistic is not exactly chi-squared distributed but follows something in between a Student's t-distribution and a chi-squared distribution. However, since the number of measurements in each bin is large (on average $n \approx 20$ or more) the distribution is much closer to the chi-squared distribution than to the Student's t-distribution, and the difference between the actual distribution and the chi-squared distribution is negligible.

C.2.3 Akaike's information criterion

Akaike's information criterion is an information-theoretic-based method for comparing multiple models against each other based on experimental data [83]. It compares how well different models describe data, while taking into account the number of free parameters in each model to avoid overfitting. It compares models based on the models' AIC values, which are defined as

$$\mathrm{AIC}_i = 2K_i - 2\ln \mathcal{L}_i\left(\widehat{\boldsymbol{\theta}}_i \middle| \mathbf{x}\right) \quad ,$$

where K_i is the number of free parameters in the ith model, $\widehat{\boldsymbol{\theta}}_i$ is the MLE of the model's free parameters and $\mathbf{x}$ is the experimental data.

AIC differences and Akaike weights Since the true underlying model is unknown, the AIC values are only relative. For a set of models $\{\mathcal{L}_i, \boldsymbol{\theta}_i\}_{i=1,\ldots,R}$ with associated AIC values $\{\mathrm{AIC}_i\}_{i=1,\ldots,R}$, AIC differences are defined as $\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - \min(\mathrm{AIC}_i)$, where the best model thus has $\Delta\mathrm{AIC} = 0$.

The Akaike weights assign relative probabilities to the individual models, given the data and the set of R models, and are calculated as

$$w_i = \frac{e^{-\Delta\mathrm{AIC}_i/2}}{\sum_j e^{-\Delta\mathrm{AIC}_j/2}} \quad .$$

C.3 Distribution of residence times

In this section we analyze the distribution of residence times of the hOgg1 proteins on DNA. In Sec. C.3.1 we show how a finite lower cutoff for measured residence times affects the measured residence-time distribution. In Sec. C.3.2 we show that if unbinding from the DNA is a Poisson process, bleaching of the fluorophores does not alter the form of the observed residence-time distribution. Finally, in Sec. C.3.3 we derive several candidate models for protein-DNA binding and compare these to the experimental data.

C.3.1 The problem of real data

We have in our analysis excluded time-series that are shorter than $N_{\min} = 12$ ($t_{\min} = 0.132$ ms). We need to take this into account when we analyze the residence-time distribution. Since the probability distribution must be normalized, the density of the residence-time distribution is

$$p(t|t \geqslant t_{\min}) = \begin{cases} 0 \ , & \text{for} \quad t < t_{\min} \\ p(t)/\mathcal{S}(t_{\min}) \ , & \text{for} \quad t \geqslant t_{\min} \end{cases} \quad ,$$

where the normalization constant $\mathcal{S}(t_{\min})$ is the survival function at time $t_{\min}$ $\mathcal{S}(t_{\min}) = \int_{t_{\min}}^{\infty} p(t)dt$.

C.3.2 The role of photobleaching

For a fluorescent molecule bound to DNA its observed residence time on DNA is given by the point in time when the molecule either unbinds from the DNA and diffuses out of the field-of-view, or when it bleaches. Photobleaching is a Poisson process and we assume that unbinding is as well. The probability that the protein disappears during the infinitesimal time-lapse $[t, t + dt[$ is then

$$p(t)dt = (k_{\text{off}} + k_{\text{bleach}})\mathcal{S}(t)dt \ ,$$

where the survival function $\mathcal{S}$ gives the probability that the protein is still bound to the DNA and visible at time t,

$$\begin{aligned} \mathcal{S}(t) &= \pi_0^t(1 - k_{\text{off}}dt')(1 - k_{\text{bleach}}dt') \\ &= \pi_0^t(1 - (k_{\text{off}} + k_{\text{bleach}})dt') \\ &= e^{-(k_{\text{off}}+k_{\text{bleach}})t} \ , \end{aligned}$$

where π is the *product integral*, defined as $\pi_0^t(1 + f(x)dx) = \lim_{\Delta x \to 0} \prod(1 + f(x_i)\Delta x)$. Thus the observed distribution is exponential, but with a higher rate than if no bleaching occurred $k = k_{\text{off}} + k_{\text{bleach}}$. Thus a residence-time distribution that is different from an exponential distribution cannot be explained by bleaching only.

C.3.3 Models for protein-DNA binding

C.3.3.1 One-state protein—Exponentially distributed residence times

A protein bound to a DNA with a constant binding energy will have a constant unbinding rate. The process of unbinding is thus a Poisson process and the proteins' residence times are exponentially distributed. When we take into account that we only observe proteins with a residence time longer than $t_{\min}$, the observed distribution is

$$p(t|t \geqslant t_{\min}) = \frac{1}{\tau} \ e^{-(t-t_{\min})/\tau} \ . \tag{C.3.1}$$

A maximum likelihood fit to the data gives $\tau = (0.54 \pm 0.03)$ s. The distribution is plotted along with the measured distribution in Fig. 2.9(a-b). A Pearson's chi-squared test gives a p-value of $p = 4 \cdot 10^{-6}$.

C.3.3.2 Two-state protein with fixed switching rates

A protein, which switches between two different states with constant rates while bound to the DNA, where it has different unbinding rates in each state is described by the differential equation

$$\frac{d}{dt}\begin{pmatrix} P_1 \\ P_2 \\ P_{\text{off}} \end{pmatrix} = \begin{pmatrix} -k_{12} - 1/\tau_1 & k_{21} & 0 \\ k_{12} & -k_{21} - 1/\tau_2 & 0 \\ 1/\tau_1 & 1/\tau_2 & 0 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ P_{\text{off}} \end{pmatrix} .$$

The residence time distribution is the probability density associated with P_{off},

$$p_t = \frac{dP_{\text{off}}}{dt} = \frac{1}{\tau_1}P_1 + \frac{1}{\tau_1}P_2 \ , \tag{C.3.2}$$

and is, as P_{off}, determined by P_1 and P_2, which control the dynamics.

$$\frac{d}{dt}\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} -k_{12} - 1/\tau_1 & k_{21} \\ k_{12} & -k_{21} - 1/\tau_2 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} . \tag{C.3.3}$$

The solutions to the characteristic equation for Eq. (C.3.3) are

$$\lambda_{\pm} = \frac{1}{2}\text{Tr} \pm \sqrt{\frac{1}{4}\text{Tr}^2 - \text{Det}} \ ,$$

where $\text{Tr} = k_{12} + k_{21} + 1/\tau_1 + 1/\tau_2$ and $\text{Det} = (k_{12} + 1/\tau_1)((k_{21} + 1/\tau_2) - k_{12}k_{21}$. Thus,

$$\frac{1}{4}\text{Tr}^2 - \text{Det} = \frac{1}{4}\left[(k_{12} - k_{21}) + \left(\frac{1}{\tau_1} - \frac{1}{\tau_2}\right)\right]^2 + k_{12}k_{21} \ ,$$

i.e., both roots of the characteristic polynomial are real and P_1 and P_2 are both a linear combination of two exponential distributions and by Eq. (C.3.2) so is p_t. That is, if the proteins switch between two states with constant rates, the residence time distribution is a sum of two exponentials, regardless of the rates. This is the same distribution as for a population of two different proteins with constant unbinding rates. The observed distribution of residence times is thus

$$p(t|t \geqslant T) = \frac{\alpha e^{-t/\tau_1}/\tau_1 + (1-\alpha)e^{-t/\tau_2}/\tau_2}{\alpha e^{-t_{\min}/\tau_1} + (1-\alpha)e^{-t_{\min}/\tau_2}} \ . \tag{C.3.4}$$

A maximum likelihood fit to the data gives $\tau_1 = 0.68$ s, $\tau_1 = 0.07$ s, and $\alpha = 0.37$. The distribution is plotted along with the measured distribution in Fig. 2.9(a-b). A Pearson's chi-squared test gives a p-value of $p = 0.04$.

C.3.3.3 Normal distributed binding energies

If we assume that the free energy of binding for the individual protein remains constant during the time the protein spends on the DNA and that the observed

dispersion in the residence times are due to different binding energies of the individual proteins, then it is natural to assume that these binding energies are normally distributed, $E_i \sim \mathcal{N}(\mu_E, \sigma)$. This leads to a lognormal distribution of the characteristic residence times,

$$\tau \sim \log \mathcal{N}(\mu, \sigma) \ ,$$

where $\mu = \mu_E + \ln \tau_0$, since $\tau = \tau_0 e^{-(E+U_i)}$. The measured residence times are then distributed as

$$\begin{aligned} p(t) &= \int_0^\infty p(t|\tau)p(\tau)d\tau \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma\tau^2} \ e^{-t/\tau-(\ln\tau-\mu)^2/2\sigma^2} d\tau \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^\infty e^{x-e^x t-(x+\mu)^2/2\sigma^2} dx \ , \end{aligned} \tag{C.3.5}$$

where we have defined $x = -\ln\tau$ The integrand falls off very quickly for positive x due to the exponential term, while the Gaussian term quickly suppresses the integrand for negative x. This makes it rather easy to approximate Eq. (C.3.5) by a finite integral, which can be calculated numerically. The survival function is

$$\begin{aligned} \mathcal{S}(T) &= \int_{t_{\min}}^\infty p(t)dt \\ &= \int_{t_{\min}}^\infty \int_{-\infty}^\infty \frac{e^{x-e^x t-(x+\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \ dxdt \\ &= \int_{-\infty}^\infty \frac{e^{-e^x t_{\min}-(x+\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \ dx \ , \end{aligned} \tag{C.3.6}$$

and the probability density is given by $p(t|t \geqslant t_{\min}) = p(t)/\mathcal{S}(t_{\min})$.

Maximum likelihood fitting to the experimental data gives $\mu = -1.15 \ k_B T$ and $\sigma = 0.82$. A Pearson's chi-squared test gives a p-value of $p = 0.02$.

C.3.3.4 Exponentially distributed characteristic residence times

We can model the characteristic residence time τ as exponentially distributed. This is the maximum entropy model assuming a distribution of τ-values and a given mean τ^* and has the advantage of being a model with only one parameter. However, we do not have an underlying physical explanation for the model.

The conditional probability density of t given τ^* is

$$p_\tau(\tau|\tau^*) = \frac{1}{\tau^*} \ e^{-\tau/\tau^*} \ .$$

So the distribution of residence times is

$$\begin{aligned} p_t(t|\tau^*) &= \int_0^\infty p_t(t|\tau)p_\tau(\tau|\tau^*)d\tau \\ &= \frac{1}{\tau_1}\int_0^\infty \frac{1}{\tau}\, e^{-t/\tau-\tau/\tau^*}d\tau \\ &= \frac{2}{\tau^*}\, K_0\left(2\sqrt{\frac{t}{\tau^*}}\right) \quad , \end{aligned}$$

where K_0 is the zeroth order modified Bessel function of the second kind. Thus,

$$\begin{aligned} p_t(t|\tau^*, t \geqslant t_{\min}) &= \frac{p_t(t|\tau^*)}{\mathcal{S}(t_{\min}|\tau^*)} \\ &= \frac{K_0\left(2\sqrt{t/\tau_1}\right)}{\sqrt{\tau_1 T}K_1\left(2\sqrt{T/\tau_1}\right)} \quad , \end{aligned} \tag{C.3.7}$$

Maximum likelihood fit to data gives $\tau^* = 0.38$ s and a Pearson's chi-squared test gives a p-value of $p = 0.06$.

C.3.3.5 Gamma distributed residence times

A final, simple phenomenological model for the residence times, which seems to correspond to the measured distribution, is $p(t) \propto t^{-\alpha}e^{-\beta t}$, i.e., a gamma distribution. This distribution is properly normalized by the condition

$$\int_{t_{\min}}^\infty p(t|t \geqslant t_{\min})dt = 1 \quad ,$$

thus for $p(t|t \geqslant t_{\min}) = C(t_{\min})t^{-\alpha}e^{-\beta t}$,

$$\frac{1}{C(t_{\min})} = \int_{t_{\min}}^\infty t^{-\alpha}e^{-\beta t}dt \tag{C.3.8}$$

$$= \beta^{-1+\alpha}\int_{\beta t_{\min}}^\infty t^{-\alpha}e^{-t}dt$$

$$= \frac{\Gamma(1-\alpha, \beta t_{\min})}{\beta^{1-\alpha}} \quad , \tag{C.3.9}$$

where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function. So the probability density is

$$p(t|t \geqslant t_{\min}) = \frac{\beta^{1-\alpha}}{\Gamma(1-\alpha, \beta t_{\min})}\, t^{-\alpha}e^{\beta t} \quad .$$

Finally, the survival function is

$$\mathcal{S}(t|t \geqslant t_{\min}) = \frac{\Gamma(1-\alpha, \beta t)}{\Gamma(1-\alpha, \beta t_{\min})} \quad .$$

Maximum likelihood fit to data gives $\alpha = 0.97$ and $\beta = 0.76$ s^{-1}, and a Pearson's chi-squared test gives a p-value of $p = 0.10$.

C.4 Parameter estimation

In this section we provide an in-depth description and analysis of the results of parameter estimation from experiments. In Sec. C.4.1 we describe how parameters are estimated from data. We analyze the results and compare them to a priori expectations we have about the data. In Sec. C.4.2 we investigate the effect of proteins crowding on the DNA. Finally, In Sec.C.4.3 we show that for a particle, which switches between fast and slow diffusion, the methods developed in the thesis estimate the mean diffusion coefficient of the particle, and that estimator statistics do not change.

C.4.1 Experimental results

We estimate diffusion coefficients D, positional noise s^2, and parameters describing DNA fluctuations $\boldsymbol{\phi} = (c_1, \mathcal{X}_{1,1}, \mathcal{Y}_1)^T$ using maximum likelihood estimation as described in Secs. 3.2.1.1 and 3.2.3.4. Fitting multiple modes does not alter the estimates of the diffusion coefficients (Fig. D.2). For time-series of $N = 35$ or shorter, the MLE algorithm does not always converge and the results of these fits are not valid. Simply discarding the time-series, for which the MLE does not converge may induce a bias. So for short time-series ($N \leqslant 35$) we estimate diffusion coefficients D and positional noise amplitudes σ^2 using the unbiased CVE as described in 3.2.1.2. Estimates of $\mathcal{X}_{1,1}$ and c_1 for use in calculating the bias are obtained from time-series longer than $N = 35$. Time-series are divided into bins of size 1 μm on the x-axis, and weighted averages of $\widehat{D}$, $\widehat{\sigma}^2$, and $\widehat{\boldsymbol{\phi}}$ are calculated as described in Secs. 3.2.1.2 and 3.2.3.4. Estimates of c_1, $c_1\mathcal{X}_{1,1}$, and $c_1\mathcal{Y}_1^2$ for different mean positions $\overline{s}$ on the DNA[2] are shown in Figs. C.11 and C.12. The correlation constant c_1 varies by a factor two over the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$ (Fig. C.11). This is probably due to the fact that we only include one mode in fits and that higher modes contribute significantly to the power near the tethering point and only the lowest mode contributes near the free end (Figs. B.5 and B.6). The positional noise is higher near the free end than near the tethered end of the DNA (Fig. C.13a). This agrees with our conception that the downstream DNA is farther from from the coverslip on average, where the evanescent field is less intense and the fluorescent molecules thus emit less photons. Comparing diffusion coefficient estimates with the proteins' positions on the DNA shows no correlations between positions and their estimated diffusion coefficients, even outside of the experimental window (Fig. C.14a). (If the DNA fluctuations

[2]We show $c_1\mathcal{X}_{1,1}$ and $c_1\mathcal{Y}_1^2$, since they are constant over the DNA, instead of $\mathcal{X}_{1,1}$ and $\mathcal{Y}_1$, which are not.

are not taken into account an apparent increase in diffusivity of the proteins is seen near the free end, Fig. C.14b.)

Experimental estimates of the bias of the CVE are calculated as the difference between the CVE and the MLE for time-series longer than $N = 35$ (Fig. C.15a). The bias varies along the DNA and is higher downstream where the DNA fluctuations are larger. Theoretical values of the bias of the CVE are calculated from Eq. (3.2.12) using average values of the estimates of c_1 and $\mathcal{X}_{1,1}$ (Figs. C.11 and C.12). The theoretical estimates of the bias, based on the estimates of c_1 and $\mathcal{X}_{1,1}$, agree with the experimentally measured bias, both locally on the DNA (Fig. C.15) and globally. The average value of the bias is $\overline{b(D)} = 0.104 \pm 0.014\ \mu\text{m}^2/\text{s}$ and the mean difference between the experimentally estimated and theoretically predicted bias is $\Delta b(D) = -0.014 \pm 0.012\ \mu\text{m}^2/\text{s}$.

For time-series longer than $N = 35$ recorded in the experimental window we observe a dispersion in estimated diffusion coefficients, which is two times larger than the standard deviation we would expect if all proteins had the same diffusion coefficient. The measured sample standard deviation is $\text{std}\left(\widehat{D}\right) = 0.29 \pm 0.03\ \mu\text{m}^2/\text{s}$, while the expected standard deviation due to stochastic estimation error is $\langle \text{std}\,(D) \rangle = 0.16\ \mu\text{m}^2\text{s}$.

C.4.2 Crowding on the DNA

Collisions between proteins on the DNA could possibly induce an increased dispersion of diffusion coefficient estimates, and could thus give a simple explanation of the dispersion we observe in the experimentally measured diffusion coefficients. When two proteins encounter each other on the DNA, we expect that they reflect if they are on the same strand while they may pass each other if they are on opposite strands. Besides the fluorescent proteins we record on the DNA an unknown number of bleached "dark" proteins will be bound to and diffusion on the DNA. So collisions may occur that we cannot see. The experiment is optimized to have the least dark proteins possible. So we expect that the number of dark proteins bound to the DNA is on the order, or smaller, than the number of visible proteins (5-10)[3]. Even at concentrations ten times higher than this dark proteins only induce a slight change in the mean and dispersion of estimated diffusion coefficients compared to their values for a free protein (Table C.1). This change is much too small to account for the high spread we see in the experimental data. Furthermore, collisions between proteins do not cause a correlation between measured diffusion coefficients and

[3]Since the sample is illuminated by TIRF, the exciting light intensity is low except very close to the coverslip surface, thus proteins in bulk are not exposed to much light before binding to the DNA, and not many proteins are expected to bleach before binding, which means the number of bleached proteins on the DNA is low, i.e., we expect it to be lower than the number of visible proteins.

residence times.

c	N	$\overline{D}$ [10^4 s^{-1}]	Var $\left(\overline{D}\right)$ [10^6 s^{-1}]	Var $(D)_0$ [10^6 s^{-1}]
0.001	10	1.984 ± 0.005	80.1 ± 0.8	78.7
	100	1.984 ± 0.005	8.2 ± 0.2	7.9
	1000	1.984 ± 0.006	0.09 ± 0.02	0.08
0.03	10	1.910 ± 0.005	75.5 ± 0.8	73.0
	100	1.910 ± 0.006	8.8 ± 0.3	7.3
	1000	1.910 ± 0.006	1.07 ± 0.09	0.73
0.01	10	1.719 ± 0.005	66.6 ± 0.07	59.1
	100	1.719 ± 0.005	8.9 ± 0.2	5.9
	1000	1.719 ± 0.007	1.4 ± 0.1	0.6

Table C.1: Results of Monte Carlo simulations to test the influence of molecular crowding on the estimates of diffusion coefficients. Performed on a population of random walkers distributed randomly on a lattice of 3000 sites with periodic boundary conditions. Diffusion coefficients were estimated from time-series measured with time-lapse $\Delta t = 11$ ms. The jump rate of the walkers was adjusted to match the experimental conditions, i.e., $\tau_0 = 1/(2D) = 2.5 \cdot 10^{-5}$ s, corresponding to a diffusion coefficient of $D = 0.5\ \mu\text{m}^2/\text{s}$. The number of sites was chosen to match the length of the DNA. The DNA length is $48,500$ bp long and each protein is approximately 15 bp wide, so the lattice size is $l = 48.500/15 \approx 3000$.

C.4.3 Estimation for two-state diffusion

The estimators presented in this thesis have been derived under the assumption that the diffusion coefficient of the measured particle remains constant throughout the measurement. Our analysis of the experimental data of hOgg1 proteins diffusing on DNA suggests that proteins switch stochastically between two or more states, in contradiction with our initial assumption. However, if the switching is independent of protein position the statistical properties of the estimators do not change. We only estimate the weighted mean of the diffusion coefficients instead of a single diffusion coefficient. This is due to the particular mathematical nature of Brownian motion (properties i) and ii) defined in Sec. A.1.1). For a diffusing particle switching between two states with different diffusion coefficients D_1 and D_2 the measured displacement during a time-lapse Δt is

$$\Delta x_n = \int_{t \in T_\alpha} \sqrt{2D_1}\eta_t dt + \int_{t \in T_{1-\alpha}} \sqrt{2D_2}\eta_t dt \ ,$$

where η_t is continuous-time standard white noise, T_α is the set of times, that the protein spends in state 1, $T_{1-\alpha}$ is the times it spends in state 2, and

$T_\alpha \cup T_{1-\alpha} = [t_n, t_n + \Delta t[$. The autocovariance of the displacements is then

$$\langle \Delta x_m \Delta x_n \rangle = 2\Delta t[\alpha D_1 + (1-\alpha) D_2]\delta_{m,n} \ ,$$

i.e, the same as for diffusion with a constant diffusion coefficient equal to the weighted average of D_1 and D_2.

C.5 Random walker in a quenched random potential landscape

We examine in this section whether a model for diffusion on DNA, which models the protein on the DNA as a random walker in a quenched random energy landscape, can reproduce the experimentally measured distributions of diffusion coefficients and residence times and the observed correlation between them.

The model, proposed by Slutsky and Mirny [20], assumes that the binding energies $E - U_i$ between the protein and the DNA are independent and normal distributed, where E is the unspecific mean binding energy between protein and DNA, and U_i is the specific part of the binding energy at site i (Fig. C.16). The rate for moving from site i to a neighboring site $i \pm 1$ is assumed to be

$$r_{i\pm 1} = 1/(2\tau_0) e^{-(U_{i\pm 1} - U_i)/k_B T} \ ,$$

where the characteristic trial time τ_0 is the time between successive attempts to move to a neighboring site, and the unbinding rate of a protein located at a site i on the DNA is

$$1/\tau = 1/\tau_0 e^{-(E-U_i)/k_B T} \ .$$

Approximate expressions for the mean diffusion coefficient and residence time of the proteins on DNA are derived in [20],

$$D \simeq \frac{1}{2\tau_0}\sqrt{1 + \frac{1}{2}\sigma^2} e^{-7\sigma^2/4} \ , \tag{C.5.1}$$

and

$$\langle t \rangle = \frac{1}{\tau_0} \ e^{E + \frac{1}{2}\sigma^2} \ . \tag{C.5.2}$$

If we know the values of the diffusion coefficient, the characteristic residence time, and the characteristic trial time τ_0 for a protein on DNA, Eqs. (C.5.1) and (C.5.2) give us approximate values for the mean binding energy E and the standard deviation σ of the specific energies. The characteristic trial time sets the time-scale of protein movement in the energy landscape on the DNA. An order-of-magnitude estimate of τ_0 can be found by assuming that the thermal

motion of the surrounding fluid is the same on the DNA surface as in bulk and that the protein does not experience any energy barriers when moving between sites. This gives $\tau_0 \sim 1$ bp$^2/D_{\text{free}} \sim 10^{-8}$ s [20]. This estimate is an approximative lower bound on the characteristic trial time, since we have a no-slip boundary condition on the DNA and thus that thermal fluctuations are smaller near the DNA than in bulk. Furthermore, if the diffusion involves barrier crossing between sites, the trial time will be increased further. An upper bound on τ_0 can be estimated from the measured mean diffusion coefficient, since for a completely flat landscape $D = 1/2\tau_0$, we must have $\tau_0 \geqslant 1/2D \approx 10^{-7}$ s. Thus for hOgg1 proteins diffusing on DNA $\tau_0 \sim 10^{-8} - 10^{-7}$ s.

We simulate an ensemble of random walkers moving in a quenched random energy landscape using the Gillespie algorithm [84] for $\tau_0 = 10^{-8}$ s, $\tau_0 = 2{\cdot}10^{-8}$ s, and $\tau_0 = 10^{-7}$ s. For each Monte Carlo experiment we choose E and σ such that mean diffusion coefficient and residence time are approximately the same as observed experimentally. In none of the cases do we see a deviation from exponentially distributed residence times or correlations between diffusion coefficients and residence times. The dynamics are fast enough that the system effectively self-averages during one time-lapse.

The Monte Carlo simulations show that Eqs. (C.5.1) and (C.5.2) are not accurate for all physically relevant parameter values. In particular they are inaccurate for parameter values corresponding to those measured experimentally for hOgg1 proteins diffusing on DNA. Equations (C.5.1) and (C.5.2) overestimate the diffusion coefficient by 75-200% and underestimate the mean residence time by 30-50%. In general the mean binding energy should be larger than $E \approx 14$ k$_B T$ and the standard deviation should be lower than $\sigma \approx 0.5$ k$_B T$ for the expressions to be accurate. A thorough study of when these expressions apply and if other more accurate expressions can be derived when Eqs. (C.5.1) and (C.5.2) do not apply, would be interesting, but is beyond the scope of this work.

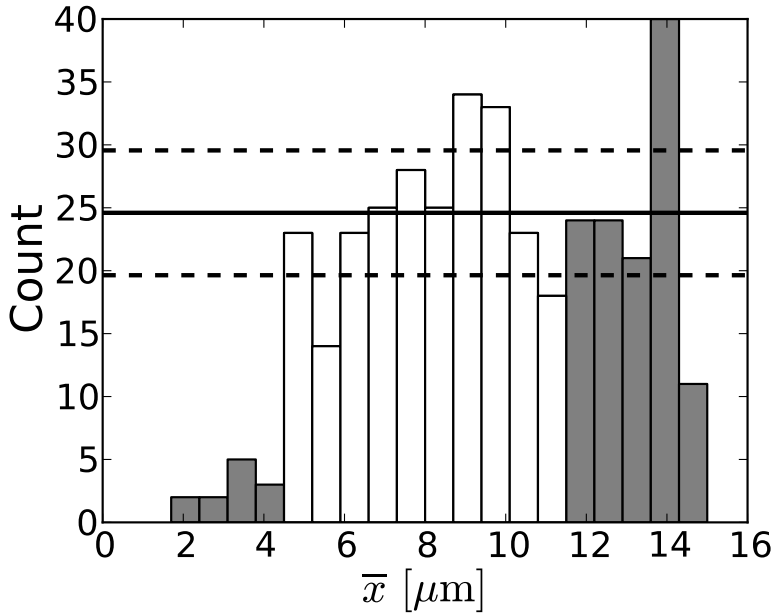

Figure C.8: Spatial distribution of proteins on the DNA. Time-series shorter than $t_{\min} = 0.132$ s (corresponding to $N_{\min} = 12$ measurements) have been discarded. A Pearson's chi-squared test for uniform distribution over the interval $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$ gives a p-value of $p = 0.09$.

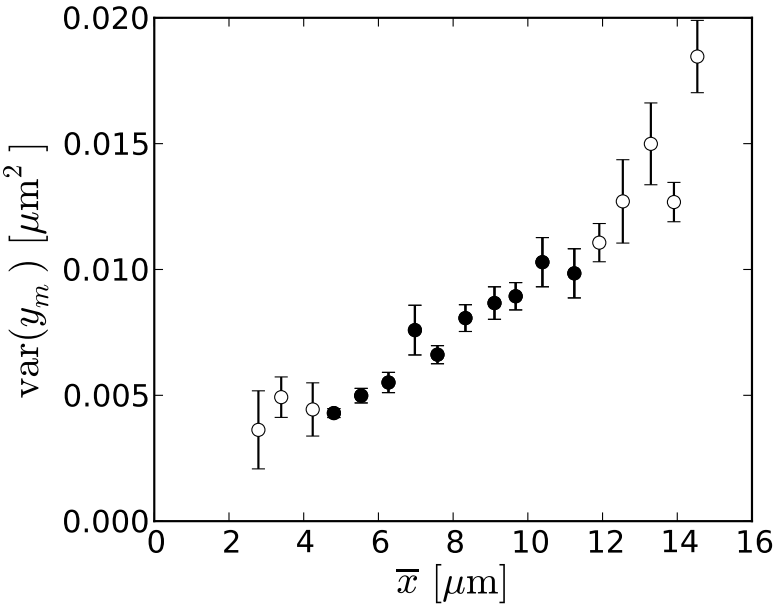

Figure C.9: Variance of the transversal (y-direction) position of the measured time-series versus mean position of the proteins on the DNA. Time-series shorter than $t_{\min} = 0.132$ s (corresponding to $N_{\min} = 12$ measurements) have been discarded. (Time-series inside experimental window $\overline{x} \in [4.5\ \mu\mathrm{m}, 11.5\ \mu\mathrm{m}]$: filled circles; time-series outside experimental window: open circles.) Perpendicular movement increases as one moves further downstream on the DNA and closer to the free end.

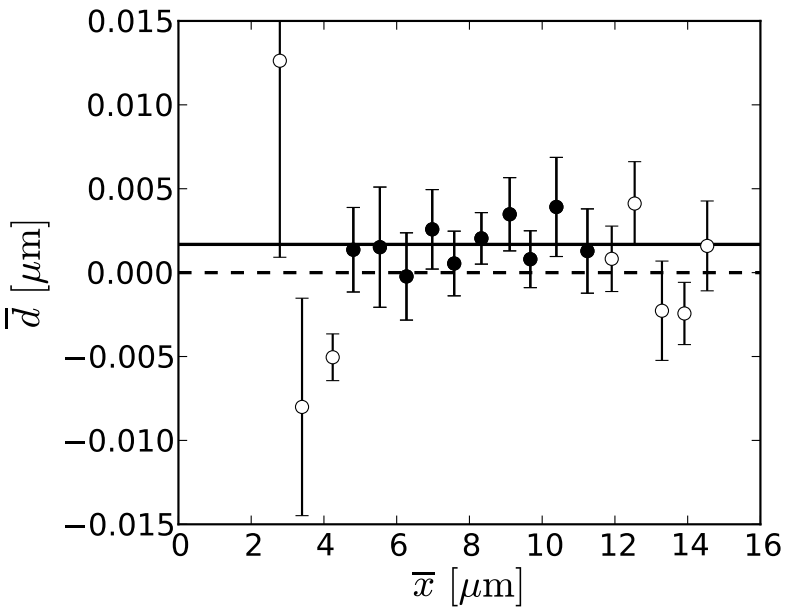

Figure C.10: Mean longitudinal displacement of proteins on DNA as a function of the proteins' mean position on the DNA. (Mean drift for time-series inside experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: closed circles; mean drift of time-series outside experimental window: open circles; average time-series inside experimental window: full line; zero: dashed line.)

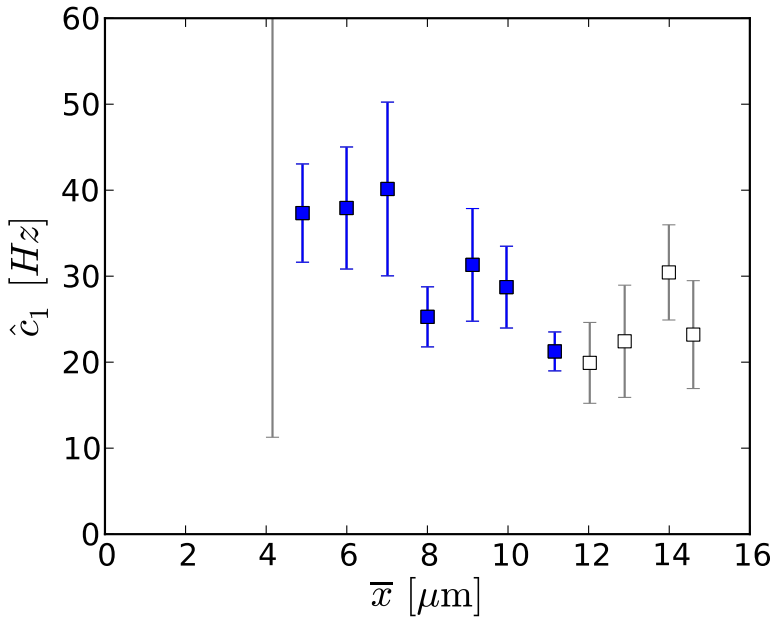

Figure C.11: MLE of c_1 for time-series longer than $N = 35$. (Estimates inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled squares, blue; estimates outside experimental window: open squares, grey.) Chi-square variance test for constant c_1 gives a p-value of $p = 4 \cdot 10^{-5}$. Error bars are estimated errors on the mean (s.e.m.) obtained from the weighted sample variances. The correlation time of DNA fluctuations is $\tau_1 = 1/c_1 \approx$ 40-50 ms for the transversal movement, which corresponds to a longitudinal correlation time of $2\tau_1 \approx$ 20-25 ms. In perfect agreement with experimental measurements of extension fluctuations of YO-YO dye stained λ-DNA, which reported a longitudinal relaxation time of 23 ± 4 ms [19].

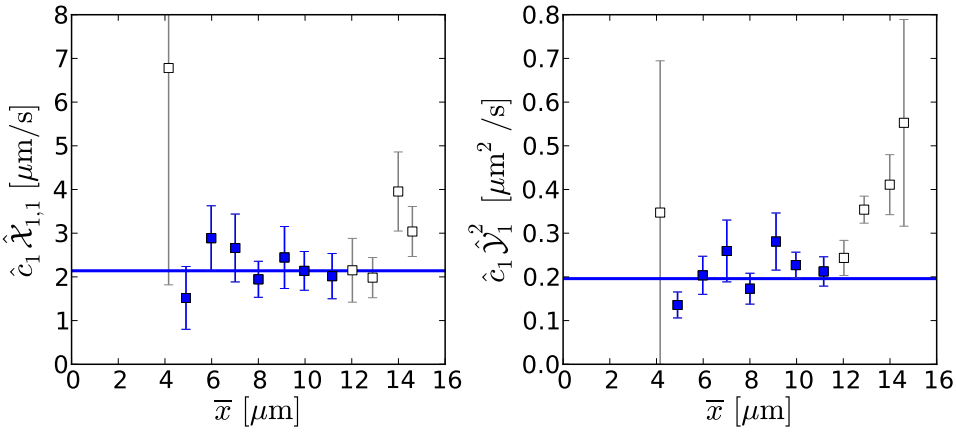

Figure C.12: a) MLE of $c_1\mathcal{X}_{1,1}$ for time-series longer than $N = 35$. Chi-square variance test for constant $c_1\mathcal{X}_{1,1}$ gives $p = 0.85$. The average value of $c_1\mathcal{X}_{1,1}$ is $\overline{c_1\mathcal{X}_{1,1}} = 2.1 \pm 0.1$ μm/s. b) MLE of $c_1\mathcal{Y}_1$ for time-series longer than $N = 35$. Chi-square variance test for constant $c_1\mathcal{Y}_1$ gives $p = 0.21$, and $\overline{c_1\mathcal{Y}_1} = 0.20 \pm 0.02$ μm^2/s. (Estimates inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled squares, blue; estimates outside experimental window: open squares, grey.) Error bars are estimated errors on the mean (s.e.m.) obtained from the weighted sample variances.

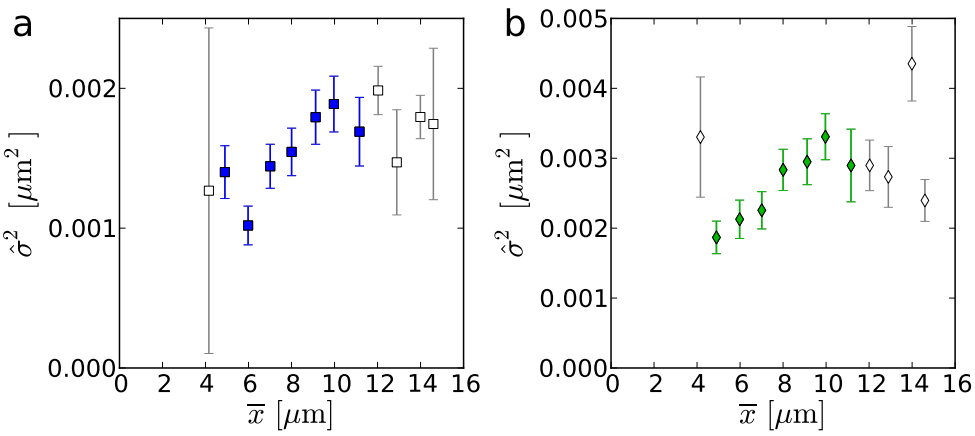

Figure C.13: a) MLE of positional noise variance σ^2, which explicitly accounts for DNA motion, for time-series longer than $N = 35$. (Estimates inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled squares, blue; estimates outside experimental window: open squares, grey.) Chi-square variance test for constant σ^2 gives $p = 0.004$ b) CVE of σ^2, which does not account for DNA motion, for time-series longer than or equal to $N = 12$. (Estimates inside the experimental window: filled diamonds, green; estimates outside experimental window: open diamonds, grey.) The DNA fluctuations induce a bias in the estimates of σ^2 of approximately a factor two. Chi-square variance test for constant σ^2 gives $p = 0.003$. Error bars are estimated errors on the mean (s.e.m.) obtained from the weighted sample variances.

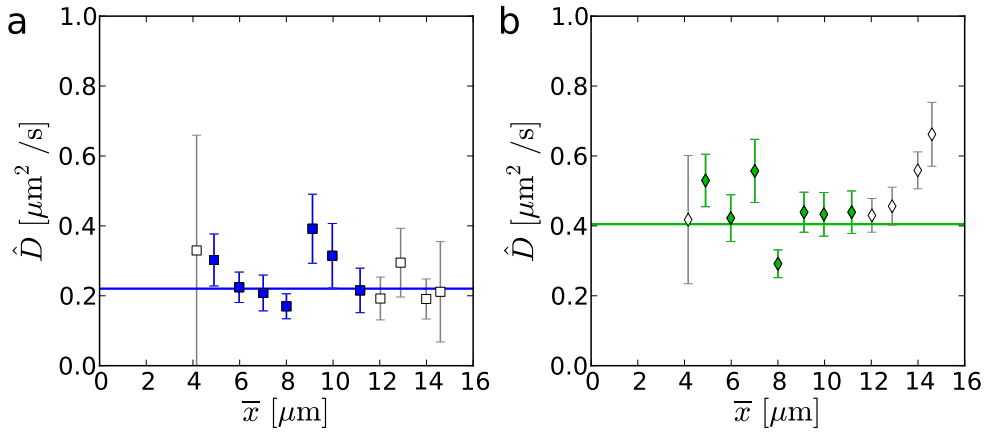

Figure C.14: a) MLE of diffusion coefficients D, which explicitly accounts for DNA motion, for time-series longer than $N > 35$. (Estimates inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled squares, blue; estimates outside experimental window: open squares, grey.) Chi-square variance test for constant mean D gives $p = 0.29$ b) CVE of D, which does not account for DNA fluctuations, for time-series longer than or equal to $N = 12$. (Estimates inside the experimental window : filled diamonds, green; estimates outside experimental window: open diamonds, grey.) Chi-square variance test for constant mean D gives $p = 0.02$. Error bars are estimated errors on the mean (s.e.m.) obtained from the unweighted sample variances. These underestimate the real dispersion, since errors on individual estimates differ due to differing time-series lengths. We cannot calculate the true error bars, but the errors shown are lower bounds. Chi-squared testing for variance thus underestimates the associated p-values for the hypothesis that mean diffusion coefficient is constant over the DNA.

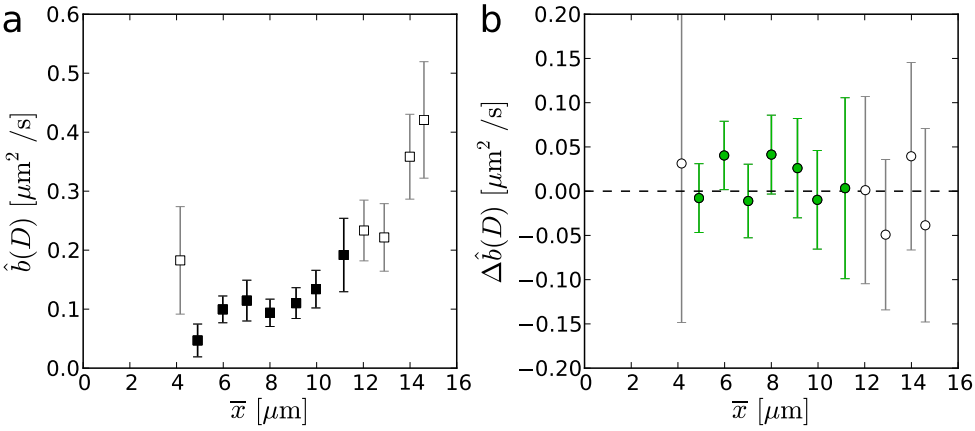

Figure C.15: Bias $b(D)$ of the CVE of the diffusion coefficient for time-series longer than $N = 35$. a) Experimental estimate of the bias, calculated as $\widehat{D}_{\text{cve}} - \widehat{D}_{\text{mle}}$. (Estimates inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled squares; estimates outside experimental window: open squares) b) Differences between the theoretical and the experimental estimates of the bias of the CVE. (Estimates inside the experimental window $\overline{x} \in [4.5\ \mu\text{m}, 11.5\ \mu\text{m}]$: filled circles, green; estimates outside experimental window: open circles, grey.) The theoretical bias is calculated using weighted means of the estimates of DNA parameters c_1 and $\mathcal{X}_{1,1}$.

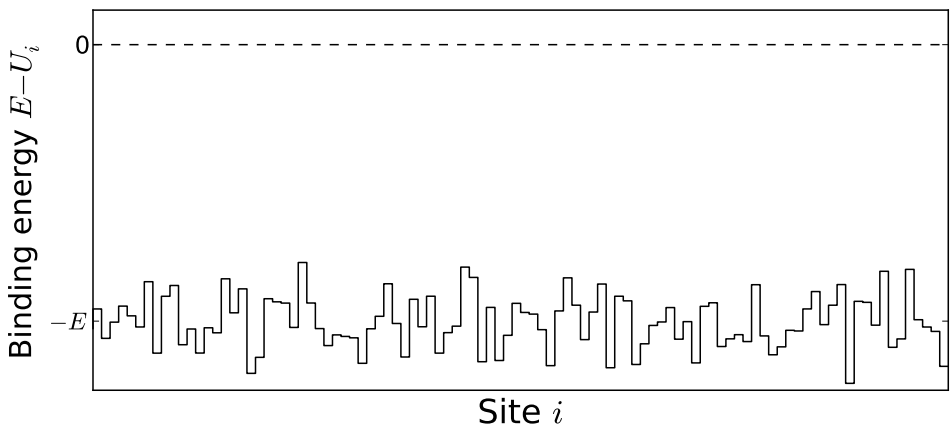

Figure C.16: An example of the energy landscape experienced by a protein diffusing on DNA. The protein on DNA is modeled as a random walker on a lattice with normal distributed binding energies $E_i = E - U_i$, where $U_i \sim \mathcal{N}(0, \sigma)$.

Appendix D

Supplementary figures

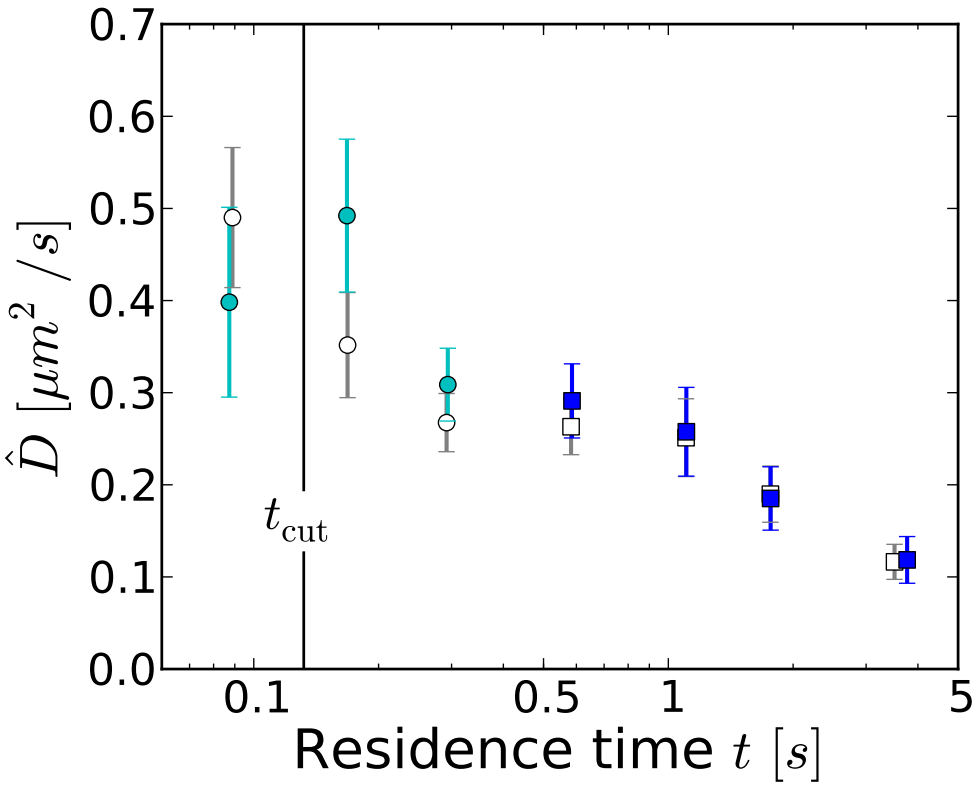

Figure D.1: Diffusion coefficient estimates $\widehat{D}$ versus protein residence time on DNA, t, for all time-series. (MLE, which explicitly takes DNA motion into account: blue squares; CVE where bias due DNA motion is subtracted: cyan circles; Estimates outside of the experimental window: open symbols, grey.) Including time-series outside of the experimental window does not alter results significantly.

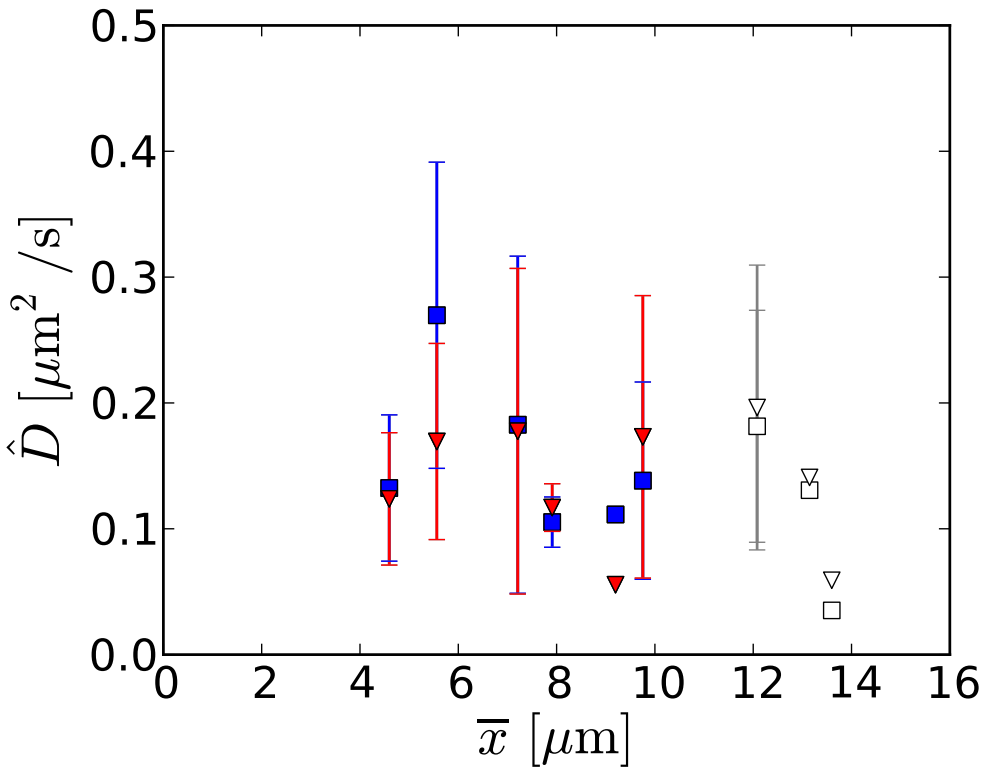

Figure D.2: Comparison of diffusion coefficient estimates from time-series longer than $N = 145$, estimated using an MLE where one DNA mode is included and an MLE where two DNA modes are included. (One-mode MLE: blue squares; two-modes MLE: red triangles; Estimates outside of the experimental window: open symbols, grey.) No difference between the estimates obtained using the one-mode and two-modes theories is observed. Average diffusion coefficient estimates are $\overline{D}_1 = 0.14 \pm 0.02$ μm^2/s for one-mode MLE and $\overline{D}_2 = 0.14 \pm 0.02$ μm^2/s for two-modes MLE.

Bibliography

[1] Anahita Tafvizi, Leonid A Mirny, and Antoine M van Oijen. Dancing on dna: kinetic aspects of search processes on dna. *Chemphyschem*, 12(8):1481–1489, Jun 2011.

[2] Jason Gorman and Eric C Greene. Visualizing one-dimensional diffusion of proteins along dna. *Nat Struct Mol Biol*, 15(8):768–774, Aug 2008.

[3] Stephen E Halford and John F Marko. How do site-specific dna-binding proteins find their targets? *Nucleic Acids Res*, 32(10):3040–3052, 2004.

[4] R. D. Vale, D. R. Soll, and I. R. Gibbons. One-dimensional diffusion of microtubules bound to flagellar dynein. *Cell*, 59(5):915–925, Dec 1989.

[5] Jonne Helenius, Gary Brouhard, Yannis Kalaidzidis, Stefan Diez, and Jonathon Howard. The depolymerizing kinesin mcak uses lattice diffusion to rapidly target microtubule ends. *Nature*, 441(7089):115–119, May 2006.

[6] Itsushi Minoura, Eisaku Katayama, Ken Sekimoto, and Etsuko Muto. One-dimensional brownian motion of charged nanoparticles along microtubules: a model system for weak binding interactions. *Biophys J*, 98(8):1589–1597, Apr 2010.

[7] Andreas W Sonesson, Ulla M Elofsson, Thomas H Callisen, and Hjalmar Brismar. Tracking single lipase molecules on a trimyristin substrate surface using quantum dots. *Langmuir*, 23(16):8352–8356, Jul 2007.

[8] Stefan Wieser and Gerhard J Schütz. Tracking single molecules in the live cell plasma membrane-do's and don't's. *Methods*, 46(2):131–140, Oct 2008.

[9] David Lasne, Gerhard A Blab, Stéphane Berciaud, Martin Heine, Laurent Groc, Daniel Choquet, Laurent Cognet, and Brahim Lounis. Single nanoparticle photothermal tracking (snapt) of 5-nm gold beads in live cells. *Biophys J*, 91(12):4598–4604, Dec 2006.

[10] Stefanie Y Nishimura, Samuel J Lord, Lawrence O Klein, Katherine A Willets, Meng He, Zhikuan Lu, Robert J Twieg, and W. E. Moerner. Diffusion of lipid-like single-molecule fluorophores in the cell membrane. *J Phys Chem B*, 110(15):8151–8157, Apr 2006.

[11] Matthew B. Smith, Erdem Karatekin, Andrea Gohlke, Hiroaki Mizuno, Naoki Watanabe, and Dimitrios Vavylonis. Interactive, computer-assisted tracking of speckle trajectories in fluorescence microscopy: application to actin polymerization and membrane fusion. *Biophys J*, 101(7):1794–1804, Oct 2011.

[12] Bornfleth, Edelmann, Zink, Cremer, and Cremer. Quantitative motion analysis of subchromosomal foci in living cells using four-dimensional microscopy. *Biophys J*, 77(5):2871–2886, Nov 1999.

[13] M. Goulian and S. M. Simon. Tracking single proteins within cells. *Biophys J*, 79(4):2188–2198, Oct 2000.

[14] H Qian, M P Sheetz, and E L Elson. Single particle tracking. analysis of diffusion and flow in two-dimensional systems. *Biophys. J.*, 60(4):910–921, 1991.

[15] X. Michalet. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E*, 82:041914, 2010.

[16] M. J. Saxton. Single-particle tracking: the distribution of diffusion coefficients. *Biophys J*, 72(4):1744–1753, Apr 1997.

[17] C. Radhakrishna Rao. *Linear Statistical Inference and It Applications.* Wiley Eastern, 2nd edition, 1973.

[18] A.J. Berglund. Statistics of camera-based single-particle tracking. *Phys. Rev. E*, 82:011917, 2010.

[19] Paul C Blainey, Antoine M van Oijen, Anirban Banerjee, Gregory L Verdine, and X. Sunney Xie. A base-excision dna-repair protein finds intrahelical lesion bases by fast sliding in contact with dna. *PNAS*, 103(15):5752–5757, Apr 2006.

[20] Michael Slutsky and Leonid A Mirny. Kinetics of protein-dna interaction: facilitated target location in sequence-dependent potential. *Biophys J*, 87(6):4021–4035, Dec 2004.

[21] Anna B Kochaniak, Satoshi Habuchi, Joseph J Loparo, Debbie J Chang, Karlene A Cimprich, Johannes C Walter, and Antoine M van Oijen. Proliferating cell nuclear antigen uses two distinct modes to move along dna. *J Biol Chem*, 284(26):17700–17710, Jun 2009.

[22] Albert Einstein. On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat. *Annalen der Physik*, 17:549–560, 1905.

[23] M. J. Saxton. Single-particle tracking: effects of corrals. *Biophys J*, 69(2):389–398, Aug 1995.

[24] J. Schuster, F. Cichos, and Ch Von Borczyskowski. Anisotropic diffusion of single molecules in thin liquid films. *Eur Phys J E Soft Matter*, 12 Suppl 1:75–80, Nov 2003.

[25] Annette Granéli, Caitlyn C Yeykal, Ragan B Robertson, and Eric C Greene. Long-distance lateral diffusion of human rad51 on double-stranded dna. *Proc Natl Acad Sci U S A*, 103(5):1221–1226, Jan 2006.

[26] Ji Hoon Kim and Ronald G Larson. Single-molecule analysis of 1d diffusion and transcription elongation of t7 rna polymerase along individual stretched dna molecules. *Nucleic Acids Res*, 35(11):3848–3858, 2007.

[27] Y.S. Jichul Kim Ju. Brownian microscopy for simultaneous in situ measurements of the viscosity and velocity fields in steady laminar microchannel flows. *Microelectromechanical Systems, Journal of*, 17:1135–1143, 2008.

[28] Ohad Givaty and Yaakov Levy. Protein sliding along dna: dynamics and structural characterization. *J Mol Biol*, 385(4):1087–1097, Jan 2009.

[29] Henrik Madsen. *Time Series Analysis*. Chapman and Hall, 2008.

[30] Söderström Torsten. Convergence properties of the generalised least squares identitication method. *Automatica*, 10:617–626, 1974.

[31] D. Montiel, H. Cang, and H. Yang. Quantitative characterization of changes in dynamical behavior for single-particle tracking studies. *J. Phys. Chem. B*, 110(40):19763–19770, October 2006.

[32] Paul C Blainey, Guobin Luo, S. C. Kou, Walter F Mangel, Gregory L Verdine, Biman Bagchi, and X. Sunney Xie. Nonspecifically bound proteins spin while diffusing along dna. *Nat Struct Mol Biol*, 16(12):1224–1229, Dec 2009.

[33] Marija Vrljic, Stefanie Y Nishimura, Sophie Brasselet, W. E. Moerner, and Harden M McConnell. Translational diffusion of individual class ii mhc membrane proteins in cells. *Biophys J*, 83(5):2681–2692, Nov 2002.

[34] Verena Ruprecht, Markus Axmann, Stefan Wieser, and Gerhard J. Schutz. What can we learn from single molecule trajectories? *Curr Protein Pept Sci*, 12(8):714–724, Dec 2011.

[35] J. F. Marko and E. D. Siggia. Stretching dna. *Macromolecules*, 28:8759–8770, 1995.

[36] S. R. Quake, H. Babcock, and S. Chu. The dynamics of partially extended single molecules of dna. *Nature*, 388(6638):151–154, Jul 1997.

[37] A. Crut, D. Lasne, J-F. Allemand, M. Dahan, and P. Desbiolles. Transverse fluctuations of single dna molecules attached at both extremities to a surface. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(5 Pt 1):051910, May 2003.

[38] Francoise Brochard-Wyart. Polymer chains under strong flows: Stems and flowers. *Europhys. Lett.*, 30:387–392, 1995.

[39] K. Laube, K. Günther, and M. Mertig. Analysis of the fluctuations of a single-tethered, quantum-dot labeled dna molecule in shear flow. *J Phys Condens Matter*, 23(18):184119, May 2011.

[40] J. P. Radicella, C. Dherin, C. Desmaze, M. S. Fox, and S. Boiteux. Cloning and characterization of hogg1, a human homolog of the ogg1 gene of saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 94(15):8010–8015, Jul 1997.

[41] H. M. Nash, R. Lu, W. S. Lane, and G. L. Verdine. The critical active-site amine of the human 8-oxoguanine dna glycosylase, hogg1: direct identification, ablation and chemical reconstitution. *Chem Biol*, 4(9):693–702, Sep 1997.

[42] Liwei Chen, Karl A Haushalter, Charles M Lieber, and Gregory L Verdine. Direct visualization of a dna glycosylase searching for damage. *Chem Biol*, 9(3):345–350, Mar 2002.

[43] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing.* Cambridge University Press, 2nd edition, 2002.

[44] Kim I Mortensen, L. Stirling Churchman, James A Spudich, and Henrik Flyvbjerg. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nat Methods*, 7(5):377–381, May 2010.

[45] Joost van Mameren, Erwin J G Peterman, and Gijs J L Wuite. See me, feel me: methods to concurrently visualize and manipulate single dna molecules and associated proteins. *Nucleic Acids Res*, 36(13):4381–4389, Aug 2008.

[46] J.H. Kim, V.R. Dukkipat, Pang S.W., and R.G. Larson. Stretching and immobilization of dna for studies of protein-dna interactions at the single-molecule level. *Nanoscale Res. Lett.*, 2:185–201, 2007.

[47] Philip R. Bevington and D. Keith Robinson. *Data Reduction an Error Analysis for the Physical Sciences.* WCB McGraw-Hill, 2nd edition, 1992.

[48] Andreas Biebricher, Wolfgang Wende, Christophe Escudé, Alfred Pingoud, and Pierre Desbiolles. Tracking of single quantum dot labeled ecorv sliding along dna manipulated by double optical tweezers. *Biophys J*, 96(8):L50–L52, Apr 2009.

[49] Y. Harada, T. Funatsu, K. Murakami, Y. Nonoyama, A. Ishihama, and T. Yanagida. Single-molecule imaging of rna polymerase-dna interactions in real time. *Biophys J*, 76(2):709–715, Feb 1999.

[50] Y. M. Wang, Robert H. Austin, and Edward C. Cox. Single molecule measurements of repressor protein 1d diffusion on dna. *Physical Review Letters*, 97(4):048302, 2006.

[51] Isabelle Bonnet, Andreas Biebricher, Pierre-Louis Porté, Claude Loverdo, Olivier Bénichou, Raphaël Voituriez, Christophe Escudé, Wolfgang Wende, Alfred Pingoud, and Pierre Desbiolles. Sliding and jumping of single ecorv restriction enzymes on non-cognate dna. *Nucleic Acids Res*, 36(12):4118–4127, Jul 2008.

[52] Jason Gorman, Arindam Chowdhury, Jennifer A Surtees, Jun Shimada, David R Reichman, Eric Alani, and Eric C Greene. Dynamic basis for one-dimensional dna scanning by the mismatch repair complex msh2-msh6. *Mol Cell*, 28(3):359–370, Nov 2007.

[53] Anahita Tafvizi, Fang Huang, Alan R Fersht, Leonid A Mirny, and Antoine M van Oijen. A single-molecule characterization of p53 search on dna. *Proc Natl Acad Sci U S A*, 108(2):563–568, Jan 2011.

[54] Anahita Tafvizi, Fang Huang, Jason S Leith, Alan R Fersht, Leonid A Mirny, and Antoine M van Oijen. Tumor suppressor p53 slides on dna with low friction and high stability. *Biophys J*, 95(1):L01–L03, Jul 2008.

[55] R. Lipowsky. The conformation of membranes. *Nature*, 349(6309):475–481, Feb 1991.

[56] M. J. Saxton and K. Jacobson. Single-particle tracking: applications to membrane dynamics. *Annu Rev Biophys Biomol Struct*, 26:373–399, 1997.

[57] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition, 2007.

[58] J. Schuster, F. Cichos, and C. von Borczyskowski. Diffusion measurements by single-molecule spot-size analysis. *J. Phys. Chem. A*, 106 (22):5403, 2002.

[59] Hendrik Deschout, Kristiaan Neyts, and Kevin Braeckmans. The influence of movement on the localization precision of sub-resolution particles in fluorescence microscopy. *J Biophotonics*, 5(1):97–109, Jan 2012.

[60] J. Weber. Fluctuation dissipation theorem. *Physical Review*, 101(6):1620, 1955.

[61] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 3rd edition, 2007.

[62] J.N. Pedersen. Internal communication. 2010.

[63] C. Bouchiat, M. D. Wang, J. Allemand, T. Strick, S. M. Block, and V. Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophys J*, 76(1 Pt 1):409–413, Jan 1999.

[64] J.W. Hatfield and S.R. Quake. Dynamic properties of an extended polymer in solution. *Physical Review Letters*, 82:3548, 1999.

[65] Tuan Q. Nguyen and Hans-Henning Kausch, editors. *Flexible Polymer Chains in Elongational Flow: Theory and Experiment*. Springer-Verlag New York, LLC, January 1999.

[66] S. R. Aragón and R. Pecora. Dynamics of wormlike chains. *Macromolecules*, 18:1868–1875, 1985.

[67] S. Yang, J. B. Witkoskie, and J. Cao. First-principle path integral study of dna under hydrodynamic flows. *Chemical Physics Letters*, 377:399–405, 2003.

[68] Christian Remling. Spectral analysis of higher-order differential operators ii: Fourth-order equations. *J. London Math. Soc.*, 59:188–206, 1999.

[69] W. N. Everitt, D. J. Smith, and M. van Hoeij. The fourth-order type linear ordinary differential equations. *arXiv:math*, 2006.

[70] J. J. Sakurai. *Modern Quantum Mechanics*. Addison Wesley Longman, revised edition, 1994.

[71] P. S. Doyle, B. Ladoux, and J. L. Viovy. Dynamics of a tethered polymer in shear flow. *Phys Rev Lett*, 84(20):4769–4772, May 2000.

[72] B Ladoux and P. S. Doyle. Stretching tethered dna chains in shear flow. *Europhys. Lett.*, 52:511–517, 2000.

[73] Charles M Schroeder, Rodrigo E Teixeira, Eric S G Shaqfeh, and Steven Chu. Characteristic periodic motion of polymers in shear flow. *Phys Rev Lett*, 95(1):018301, Jul 2005.

[74] C.A. Lueth and E. S. G. Shaqfeh. Experimental and numerical studies of tethered dna shear dynamics in the flow-gradient plane. *Macromolecules*, 42:9170–9182, 2009.

[75] Charles E. Sing and Alfredo Alexander-Katz. Theory of tethered polymers in shear flow: The strong stretching limit. *Macromolecules*, 44:9020–9028, 2011.

[76] Yu Zhang, Aleksandar Donev, Todd Weisgraber, Berni J Alder, Michael D Graham, and Juan J de Pablo. Tethered dna dynamics in shear flow. *J Chem Phys*, 130(23):234902, Jun 2009.

[77] F Brochard-Wyart. Deformations of one tethered chain in strong flows. *Europhys. Lett.*, 23:105, 1993.

[78] Thierry Savin and Patrick S Doyle. Static and dynamic errors in particle tracking microrheology. *Biophys J*, 88(1):623–638, Jan 2005.

[79] Wesley P Wong and Ken Halvorsen. The effect of integration time on fluctuation measurements: calibrating an optical trap in the presence of motion blur. *Opt Express*, 14(25):12517–12531, Dec 2006.

[80] K Berg-Sørensen and H Flyvbjerg. Power spectrum analysis for optical tweezers. *Review of Scientific Instruments*, 75(3):594–612, March 2004.

[81] R. B. D Agostino. An omnibus test of normality for moderate and large sample size. *Biometrika*, 58:341–348, 1971.

[82] R. D Agostino and E. S. Pearson. Testing for departures from normality. *Biometrika*, 60:613–622, 1973.

[83] P Burnham, K and R Anderson, D. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach.* Springer, 2nd edition, 2002.

[84] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, Volume 22(4):403–434, December 1976.