Dynamics of Genome Organization

by

Simon Benedikt Grosse-Holz

B.Sc., Friedrich-Alexander University Erlangen-Nuremberg (2015)
M.Sc., Friedrich-Alexander University Erlangen-Nuremberg (2017)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Author .....
Department of Physics
May 23, 2023

Certified by .....
Leonid A. Mirny
Distinguished Professor of Physics, and Medical Engineering and Science
Thesis Supervisor

Certified by .....
Mehran Kardar
Francis Friedman Professor of Physics
Thesis Co-Supervisor

Accepted by .....
Lindley Winslow
Associate Department Head of Physics

This page intentionally no longer blank.

Dynamics of Genome Organization

by

Simon Benedikt Grosse-Holz

Submitted to the Department of Physics
on May 23, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Physics

Abstract

A human cell contains about 2 m of DNA, packed into a nucleus with diameter $\sim$10 µm. The three-dimensional structure of this packing has been the subject of intense investigation essentially since the discovery of DNA itself, with an explosion of the field over the past 15 years, following the advent of chromosome conformation capture techniques. The fourth dimension—time—however, has remained elusive and the dynamics underlying the organization of the genome are much less known. In this thesis I present my contributions to our understanding of these dynamics, working towards a full four-dimensional characterization of genome organization. First, by pulling on a genomic locus in live cells, we revealed the rather liquid-like material properties of chromatin and dispelled the idea that chromatin in interphase forms a gel. Second, by tracking genomic elements known to act as boundary elements for loop formation, we quantified the dynamics of chromatin loops in live cells. My contribution to both projects lay in the development and application of novel data analysis, modeling, and inference methods, implementations of which have been made available to the community for future use. Finally, we devised a simple scaling argument to reconcile the orthogonal observations of chromosome structure, dynamics, and mechanics. In sum, these contributions further our understanding of the dynamical behavior of chromatin in living cells and provide valuable tools and directions for future research.

Thesis Supervisor: Leonid A. Mirny
Title: Distinguished Professor of Physics, and Medical Engineering and Science

Thesis Co-Supervisor: Mehran Kardar
Title: Francis Friedman Professor of Physics

Acknowledgments

Science is an inherently social endeavour; none of the work presented in this thesis would have been even remotely possible without the collaboration, support, encouragement, and mentorship of numerous people.

I want to extend my deepest gratitude to my advisor Leonid Mirny. His guidance (and sometimes the lack thereof) transformed me from a young graduate student to an independent researcher. Thank You, Leonid, for your unwavering support, the immense freedom and opportunities You create for our group, and the stimulating creative chaos that pervades it.

The MirnyLab is indeed a very special place, full of art, bikes, computers, espresso, and random friends associated with the lab with varying degrees of looseness. A few years ago we tried counting the number of lab members and ended up with an "effective $13\frac{1}{2}$". This rather unstructured cloud of fascinating people provides a wonderful environment for scientific discussion, non-scientific discussion, Cape retreats, game nights, movie nights, and late night programming marathons.

Much of this lab culture—and all of our computing infrastructure—carries the signature of Maksim Imakaev. Max taught me to ditch norms and think for myself—and that it is perfectly possible to maintain a fleet of 20 high-performance computers from a sail boat in the Bahamas. Johannes Nübler was a post-doc in the lab when I joined and supervised my first little rotation project; I am grateful for his initial guidance, introducing me to the lab and its culture, and taking a young grad student under his wings. Over the years since, multiple people have moved in and out of the lab: I overlapped to varying extents with Anton Goloborodko, Aafke van den Berg, George Spracklin, Martin Falk, Stuart Sevier, and Carino Gurjao, all of whom contributed their personal note to our group and made lab life fun. Kirill Polovnikov continues to teach me polymer physics. Nezar Abdennur brought a lot of rigor in methodology to the lab and serves as an idol with respect to production, sharing, and community-wide establishment of open-source software tools. Edward Bannigan supervised my own first attempts at polymer simulations and is an endless resource of dry humour. Sam Markson—though never formally associated with the lab—deserves an honorable mention for his "Kino cantabrigia" nights, screening movies from the earliest days of motion picture.

Sameer Abraham and Jeremy Owen shared much of the PhD journey in the MirnyLab with me. My desk being sandwiched between theirs, I could always rely on bugging either of them

with questions, vent about random coding problems, or infect them with my confusion about a particular issue. Sometimes those discussions would devolve into Friday afternoon philosophy on life, the universe, and everything. I am deeply grateful for their companionship during these years. Late at night, life in the lab tends to die down; for me, these have always been some of the most productive hours of the day (seeing as I am writing this at 2 am). Aleksandra Galitsyna is one of the very few people this time can be shared with, while preserving its fundamental character. And she lets me borrow the guitar that she keeps in the lab. Sofya Gaydukova and Yossef Finkelberg radiate a calm but fierce enjoyment of science and their hushed conversations across the wall separating their desks make you feel productive by just sitting next to it. Finally, with Emily Navarrete and Henrik Pinholt the next generation of lab folk is following suit; I am confident that they will continue to propagate, develop, and add their own twists to the MirnyLab spirit.

A close friend of the lab since before my time and by now a dear friend and colleague of mine, Laura Caccianini was my first collaborator. She gave me a theorist's crash course on fluorescence microscopy and was always patient when I would cook up some insane model to explain her data. She has become a pillar against the insanity of the world at large and without her I would have sunk into deep despair more than once. Our shared project has been reinvigorated thanks to the efforts of Luca Giorgetti and Pia Mach and I am excited to see where it will end up.

I have benefitted immensely from my collaboration with and mentorship by Anders Hansen. His structured and systematic thinking is second to none and provided a valuable counterpart to Leonid's hands-off approach. This combination was very productive, in that it allowed me to get the best of both worlds. Over the course of this collaboration I also learnt a lot from Michele Gabriele and specifically my close work with Hugo Brandão.

Christoph Zechner sparked my enthusiasm for rigorous statistics: squeezing out of the data what we can, while being mindful of what is not in there. Without his sustained mentorship and technical prowess, developing a technology like BILD would have been inconceivable; indeed, I had convinced myself that this would be impossible before we started working together.

Antoine Coulon adopted me into his group when I joined Leonid in Paris for half a year; he taught me how to actually do the microscopy experiments I had only been analyzing until then and pushed me to think about the connections between structure, dynamics, and mechanics of chromosomes. His ability to do a quick numerical estimate on whatever we happen to be discussing is inspiring. I am deeply indebted to him, his group, the UMR3664 unit, and Institut

Curie for making this visit possible, productive, and fun. Microscopy of empty slides is boring; Kyra Borgman and Julia Rönsch taught me how to culture cells and keep them happy. On the theoretical side I benefitted from many discussions with Vittore Scolari. Before all of that, Veer Keizer let me shadow her on the locus pulling experiments and even poke ("microinject") some of the cells myself; it increased my understanding of the whole system tremendously.

I started my forays into statistical and biological physics under the auspices of Mehran Kardar, who also serves as co-supervisor for this thesis. His understanding and teaching of statistical physics of all flavors is unparalleled and I dare say that without his classes in statistical physics (8.333) and statistical physics in biology (8.592) I would neither understand partition functions, nor be working on genome organization today.

I want to thank my thesis committee, Leonid Mirny, Mehran Kardar, Nikta Fakhri, and Julien Tailleur for their advice and guidance in preparing this dissertation.

Our group assistant Katrina Norman once told me that her job was to make sure that I have enough time in the day to actually do the science I am here for, instead of getting bogged down with administrative tasks; she does an absolutely amazing job at that. Above and beyond, she makes sure that everything in the lab runs smoothly. Such people are the reason that places like MIT work in the first place. In that spirit, I also want to extend my heartfelt gratitude to the Physics Department Administrative Team, specifically Catherine Modica and Sydney Miller; and Caroline Audouin and Marie Khuoy at Institut Curie.

Taking a somewhat larger step back, I also want to say thanks to Martin Dörfler, my highschool physics teacher. He showed me the fascination of the subject and how we can use mathematical tools to describe the world around us in astounding detail. I would not be here today, had he not told me to "go study physics, you'll like it".

Life—even though the habits of MIT students sometimes make you believe differently—does not consist of just science. I am infinitely grateful to friends in various corners of the world for keeping me sane when grad school (and/or the world) became insane; for hiking and biking; swimming in Loch Ness in March and Boston in February; trips to Maine, New Hampshire, New York, California; Boulder and Chicago; the Alps, the Jura and Chamonix; for infinite phone calls and online gaming during the pandemic; New Year's Eve; and countless others.

Finally, I thank my parents Michaela and Gerhard, and my sisters Friederike and Veronika, for their everlasting love and support.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

DNA as a physical object is quite a fascinating thing: a typical human chromosome—that is, a single *molecule* of DNA—is about 4 cm long [1]; this is a true *macro*-molecule. The structure of this molecule has been resolved to the famous double helix by Franklin, Watson, and Crick [2, 3], establishing pairs of the nucleobases Adenine, Thymine, Guanine, and Cytosine as its fundamental building blocks. In eukaryotic nuclei, stretches of about 150 of these base pairs then wrap around histone proteins, forming so-called nucleosomes—the lowest organizational level of *chromatin*. On the scale of tens of nucleosomes ($\sim$2 kb), chromatin structure seems to be quite disordered and is subject to active research [4, 5]. Moving on to yet larger scales of $\gtrsim$10 kb, however, we can mostly neglect these molecular details and consider chromatin as a continuous fiber, i.e. a polymer.

How this chromatin polymer is organized and structured in the nucleus has been the subject of intense research, recently fuelled by the development of chromosome capture methods such as Hi-C [6, 7]. These experimental techniques allow capturing physical contacts between distal genomic elements and thus provide a window into the structure of chromosomes all the way from the whole chromosome—tens to hundreds of millions of base pairs—down to the kb length scale (kb, "kilo-base", referring to 1000 base pairs).

One of the intriguing outcomes of Hi-C data analysis is the loop extrusion model [8, 9]: a protein complex acting as *loop extruding factor* (LEF; cohesin and condensin have been identified in this role) loads onto the chromatin fiber and starts to progressively grow a loop by reeling in material from both sides. This process stops when the LEF unbinds from chromatin, thus releasing the extruded loop, or when it stalls at a *boundary element* (identified as genomic sites

bound by the protein CTCF), thus stabilizing the loop in a fixed location. The presence of these boundary elements therefore biases the positioning of the loops, such that they appear as distinct features in ensemble averaged data like Hi-C—so-called TADs[1]. While loop extrusion was originally proposed purely on the basis of these observed Hi-C features, the process has since been observed *in vitro* [10–12] and is by now widely recognized as a plausible key mechanism for genome organization. Direct evidence *in vivo*, however, is lacking.

The prediction of loop extrusion from purely structural data is quite striking. Readout from Hi-C experiments is completely static: chromatin is crosslinked and digested, and detected contacts recorded in a contact matrix. This gives an unprecedented view of the higher order structure of chromatin, but of course leaves any dynamic processes out of the picture. TADs, however, cannot be explained by the presence of static loops (unless one considers a fine-tuned interaction landscape across the whole genome) [8]; the formation of TADs seems to require the dynamic process of loop *extrusion*. We are thus faced with the conceptual mismatch of studying an intrinsically dynamic process with a very static method.

Together with the group of Anders Hansen here at MIT and Christoph Zechner in Dresden we attempted to remedy this state of affairs through live-cell single particle tracking of two neighboring boundary elements [13]. We were able to infer the lifetime as well as absolute frequency of loops formed between these two boundary elements and found them to be rare and dynamic, quite in line with the loop extrusion model. While the precise dynamics of the extrusion process itself are still beyond the limits of resolution in this study, we hope to decipher these details with future studies in a refined experimental system.

A key challenge in these data sets is the very limited information obtained from the experimental system: we track two defined genomic loci, so the experimental readout is a 2×3D trajectory over time. Even to just extract the looping statistics highlighted above, i.e. detecting sustained contact between the two target elements, we had to resort to developing a full-blown Bayesian inference scheme that we term Bayesian Inference of Looping Dynamics (BILD), since simpler approaches failed our simulation benchmarks.

As opposed to the "few locus" tracking possible *in vivo*, recent progress in fixed cells has led to microscopy approaches with multiplexed reporters [14,15], where labels are attached to $\sim$100 individual genomic loci and read out by sequential rounds of imaging in a process referred to

[1]TAD as an acronym stands for *topologically associated domain*; however, "topological" in this context is somewhat misleading, since it refers simply to spatial association. I therefore prefer to use TAD as independent term describing features in Hi-C data.

as barcoding. The structures reconstructed from these multiplexed FISH[2] experiments overall match the inferences from ensemble Hi-C data, but of course allow more detailed analysis on the single molecule level. Developing a live-cell equivalent of these multiplexed methods would tremendously boost the study of genome dynamics.

The study of genome organization as described so far is an observational science: Hi-C and FISH probe the native structure of the chromatin polymer, while live-cell microscopy investigates its dynamics. While these data allow detailed *descriptive* study of nuclear organization, simple[3] questions about chromatin remain unanswered: what if we poke it with a stick? What kind of material are we dealing with? Should we think of chromatin as a gelatinous, relatively solid polymer mesh (as suggested by e.g. the observations in [16]), or is it more liquid-like? Does it yield to stress or does it resist deformation?

With these questions in mind, we embarked on a fascinating journey led by the group of Antoine Coulon in Paris. They had developed a magnetic micro-manipulation system to exert sustained force onto a defined genomic locus in living cells. And—contrary to the gel-hypothesis—it moved! Under the influence of sustained pN forces (comparable to the stall forces of various molecular motors), the targeted locus moved across the nucleus over tens of minutes. Quite interestingly, the observed motion was largely consistent with a simple Rouse model, the "vanilla" model of polymer dynamics. This is surprising, since the Rouse model neglects all but the one central component of a polymer system: the backbone connectivity of the chain. Everything else, like excluded volume, topological[4] interactions, non-specific interactions of the polymer with itself, or any interaction of the polymer with its environment (beyond the viscous solvent leading to overdamped dynamics) is left out of this model. Still, it seems to provide an (at least effectively) accurate description of the observed responses.

The point that chromosome mechanics seem to be well described by a Rouse model adds onto a curious conundrum in the chromosome organization literature: the structure observed by Hi-C and FISH is quite compact and often reported to adopt a *space-filling* conformation, whose signature is the scaling of the spatial distance R between two loci that are separated by a genomic distance s as $R(s) \sim s^{\frac{1}{3}}$; the dynamics observed by single particle tracking are routinely characterized by a mean-squared displacement (MSD) scaling as $\mathrm{MSD}(\Delta t) \sim (\Delta t)^{\frac{1}{2}}$. These

[2] "fluorescence *in situ* hybridization"

[3] "simple" from a macroscopic perspective

[4] here in a more mathematical sense: different sections of the fiber cannot pass through each other; however, strictly speaking this is still not a topological issue if we consider a finite linear polymer (as opposed to, e.g. a ring)

dynamics match the expectation from the Rouse model; the Rouse model, however, is based on an ensemble of equilibrium polymer conformations, where one would expect $R(s) \sim s^{\frac{1}{2}}$, a markedly looser packing. The observed $R(s)$ scaling, in turn, could be explained by the fractal globule model; but this would predict more recurrent dynamics, such that $\mathrm{MSD}(\Delta t) \sim (\Delta t)^{\frac{2}{5}}$ [17]. In this discrepancy between structure and dynamics, the above study on chromosome mechanics now sides with the dynamics, in that it seems consistent with a Rouse model. But it is still unclear how to reconcile all aspects of chromosome organization into a single, consistent model.

Pondering this issue, we realized that as long as we are interested in scale-free descriptions of the system (such that the above observables are governed by powerlaws), a simple dimensional argument gives constraints on the values of the different exponents and allows us to connect the seemingly orthogonal aspects of chromosome structure, dynamics, and mechanics. The key missing ingredient seems to be the dynamic behavior of finite size chromatin coils (as opposed to the point-like loci mostly studied to date); this of course directly suggests a direction for experimental investigation, which might be pursued in the not-too-distant future.

Outline of this thesis

The Rouse model in all forms and variants plays a central role in this thesis. A summary of my understanding and useful technical results are provided in chapter 2. As a direct application of the force inference developed at the end of that chapter, chapter 3 then describes the locus pulling project with Antoine Coulon. Zooming further into the nucleus, chapter 4 describes the looping inference based on single particle tracking with Anders Hansen and Christoph Zechner. My main contribution to this project was the development of Bayesian Inference of Looping Dynamics (BILD), described in detail in chapter 5. Originally a spin-off of the Rouse model calibration necessary for BILD, I developed a separate Bayesian MSD fitting scheme based on the theory of Gaussian processes, described in chapter 6. Then, having outlined my contributions to our understanding of chromosome dynamics and mechanics, I present the scaling argument making the connection to chromosome structure in chapter 7. Chapter 8 contains a formalized and extended version of the same argument, which I found useful for a systematic understanding of the underlying structure. Finally, chapter 9 provides my perspectives on what was achieved and where the field is going from here.

Chapter 2

The Rouse model

Much of the work presented in this thesis relies on different versions of the Rouse polymer model [18–20]. While this is a well-known model, I have found it useful to synthesize my understanding and specific results into a comprehensive overview, which I present in this chapter. As such, none of the results can be called new *per se*, though some specifics seemed to be absent from the literature. Two python modules are based on the treatment in this chapter: `rouse` [21] implements the discrete Rouse model described in section 2.2; the force inference described in section 2.4 is implemented in `rousepull` [22].

In its original formulation, the Rouse model describes a polymer as a collection of point particles ("monomers") connected by harmonic potentials [18]. Each monomer follows an overdamped Langevin equation, subject to thermal noise and the potententials connecting it to neighboring monomers. While this formulation in terms of a discrete chain undergoing continuous time evolution is the textbook default [19, 20], I find it useful mostly as common ground between two other formulations (table 2.1): when we are interested in the evolution only for discrete time steps, the model becomes fully discrete and thus very well suited for computational evaluation; conversely, taking a continuum limit along the chain removes the notion of discrete "monomers", giving rise to an analytically quite convenient framework.

This chapter is organized along these model variants (c.f. table 2.1). In section 2.1 we introduce the classical picture of discrete monomers evolving in continuous time. At its core, this is a general linear multivariate Langevin equation, such that the treatment does not only apply to a linear polymer, but in fact captures any harmonic spring network. Section 2.2 makes the transition to discrete time, thus introducing the model underlying the computations of e.g.

Time \ Chain	discrete	continuous
discrete	**Computational** most convenient for numerical computation; see section 2.2 and [21]	**Specialized** useful in special cases like force inference; see section 2.4 and chapter 3 and [22]
continuous	**Classical** original formulation; see section 2.1	**Analytical** most convenient for analytical treatment of infinite polymers; see section 2.3

Table 2.1: Variants of the Rouse model and their use cases.

chapter 5. We continue with the completely continuous model in section 2.3 and close in section 2.4 with a specific application of the latter: inferring the force acting on a single point on the chain, given its observed trajectory (cf. chapter 3).

Note: The Rouse model is completely isotropic, such that different spatial dimensions are fully independent. Below, we thus work in $d = 1$ spatial dimensions for notational convenience, and remark on generalization to higher dimensions where necessary.

2.1 Classical: discrete chain in continuous time

Consider N point particles at positions $x_1, \dots, x_N$, sequentially connected by harmonic springs of spring constant k and submersed in a viscous medium, such that they exhibit overdamped dynamics with friction constant γ and are driven by thermal noise $\xi_i(t)$. The equations of motion for this system read

$$\gamma \dot{x}_1(t) = k\,[x_2(t) - x_1(t)] + \xi_1(t)\,, \tag{2.1}$$

$$\gamma \dot{x}_i(t) = k\,[x_{i-1}(t) - x_i(t)] + k\,[x_{i+1}(t) - x_i(t)] + \xi_i(t) \qquad \forall\, i = 2, \dots N-1\,, \tag{2.2}$$

$$\gamma \dot{x}_N(t) = k\,[x_{N-1}(t) - x_N(t)] + \xi_N(t)\,, \tag{2.3}$$

with the thermal noise $\xi_i(t)$ a Gaussian random field satisfying

$$\langle \xi_i(t) \rangle = 0 \qquad \langle \xi_i(t) \xi_j(t') \rangle = 2\gamma k_\mathrm{B} T \delta_{ij} \delta(t - t') \qquad \forall i, j, t, t' \,, \tag{2.4}$$

where the noise amplitude $2\gamma k_\mathrm{B} T$ is dictated by the Einstein relation (fluctuation–dissipation theorem). We will consider a slight generalization of this system: namely, letting

$$B = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & & \ddots & & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 \end{pmatrix}, \qquad S = 2\gamma k_\mathrm{B} T \mathbb{1} \tag{2.5}$$

eqs. (2.1) to (2.4) constitute a special case of the linear multivariate Langevin equation

$$\gamma \dot{\boldsymbol{x}}(t) = -kB\boldsymbol{x}(t) + \boldsymbol{F} + \boldsymbol{\xi}(t) \,, \qquad \left\langle \boldsymbol{\xi}(t) \otimes \boldsymbol{\xi}^T(t') \right\rangle = S\delta(t - t') \,, \tag{2.6}$$

which will be the basis of our discussion in this section. Note that we introduced an external force $\boldsymbol{F}$ for generality. Note that while S in eq. (2.6) is symmetric positive definite by construction, no such constraint applies to B in principle. However, for simplicity (and relevance to our use cases) we will usually assume that B is symmetric, and thus orthogonally diagonalizable. The only exception to this is section 2.2.1, where we show that $SB^T = BS$ is a sufficient condition for the system to reach an equilibrium steady state.

As can easily be checked by differentiation, under eq. (2.6) an initial condition $\boldsymbol{x}(t_0) \equiv \boldsymbol{x}_0$ evolves as

$$\boldsymbol{x}(t) = \mathrm{e}^{-\frac{k}{\gamma} B(t - t_0)} \boldsymbol{x}(t_0) + \frac{1}{\gamma} \int_{t_0}^{t} \mathrm{d}\tau \, \mathrm{e}^{-\frac{k}{\gamma} B(t - \tau)} \left(\boldsymbol{F} + \boldsymbol{\xi}(\tau) \right) \,. \tag{2.7}$$

2.1.1 Steady state

If the connectivity matrix is positive definite, $B > 0$, then the exponentials in eq. (2.7) are all decaying. Long after the initialization (i.e. for $t_0 \to -\infty$) the system thus reaches steady state, where

$$\boldsymbol{x}(t) = \frac{\gamma}{k} B^{-1} \boldsymbol{F} + \frac{1}{\gamma} \int_0^\infty \mathrm{d}\tau \, \mathrm{e}^{-\frac{k}{\gamma} B\tau} \boldsymbol{\xi}(t - \tau) \,. \tag{2.8}$$

Mean and covariance in steady state are now easily identified as

$$\langle \boldsymbol{x} \rangle_{\mathsf{ss}} = \frac{\gamma}{k} B^{-1} \boldsymbol{F} \,, \tag{2.9}$$

$$\mathcal{J} \equiv \left\langle \boldsymbol{x} \otimes \boldsymbol{x}^T \right\rangle_{\mathsf{c,\ ss}} = \frac{1}{\gamma^2} \int_0^\infty \mathrm{d}\tau \, \mathrm{e}^{-\frac{k}{\gamma} B \tau} S \mathrm{e}^{-\frac{k}{\gamma} B^T \tau} \overset{*}{=} (2\gamma k B)^{-1} \, S \,, \tag{2.10}$$

where * signifies the condition $SB^T = BS$. This condition is an interesting special case, since it implies that the adopted steady state is in fact equilibrium (section 2.2.1).

Unfortunately, $B > 0$ is not the generic case: e.g. the connectivity matrix (2.5) for a free linear polymer is, in fact, not positive definite, but has a zero eigenvector $\boldsymbol{v}_0 = (1, 1, \ldots, 1)^T$. The corresponding degree of freedom $\frac{1}{N} \boldsymbol{v}_0^T \boldsymbol{x}(t)$ is the center of mass of the chain—which indeed we would not expect to reach steady state, but just keep diffusing freely. From eq. (2.7) it is clear that this is a general phenomenon: the degrees of freedom associated with zero eigenvectors of the connectivity matrix ("free" degrees of freedom) undergo drift–diffusion dynamics, instead of reaching steady state. However, we can still represent the steady state of the internal degrees of freedom (those corresponding to non-zero eigenvectors) in a form like eqs. (2.9) and (2.10) by replacing the inverse B^{-1} by the Moore-Penrose inverse[1] B^+. This effectively pins all "free" degrees of freedom to the origin, instead of allowing them to diffuse indefinitely, thus giving a well-defined steady state. Finally, the steady state distributions for the degrees of freedom that do reach steady state can always be calculated from eqs. (2.9) and (2.10), as long as we employ the Moore-Penrose inverse B^+ where necessary.

The conformational auto-correlation in steady state is given by

$$\left\langle \boldsymbol{x}(t + \Delta t) \otimes \boldsymbol{x}^T(t) \right\rangle_{\mathsf{c}} = \frac{1}{\gamma^2} \int_0^\infty \mathrm{d}\tau \, \mathrm{d}\tau' \, \mathrm{e}^{-\frac{k}{\gamma} B \tau} \left\langle \boldsymbol{\xi}(t + \Delta t - \tau) \otimes \boldsymbol{\xi}^T(t - \tau') \right\rangle \mathrm{e}^{-\frac{k}{\gamma} B^T \tau'} \tag{2.11}$$

$$= \frac{1}{\gamma^2} \int_0^\infty \mathrm{d}\tau \, \mathrm{e}^{-\frac{k}{\gamma} B (\tau + \Delta t)} S \mathrm{e}^{-\frac{k}{\gamma} B^T \tau} \tag{2.12}$$

$$\overset{(2.10)}{=} \mathrm{e}^{-\frac{k}{\gamma} B \Delta t} \mathcal{J} \,. \tag{2.13}$$

[1] The Moore-Penrose (pseudo-)inverse of a diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_N)$ is the diagonal matrix $D^+ = \mathrm{diag}(d_1^+, \ldots, d_N^+)$ with $d_i^+ = 1/d_i$ for all $d_i \neq 0$ and $d_i^+ = 0$ if $d_i = 0$. The Moore-Penrose inverse of a general diagonalizable matrix $M = ADA^{-1}$ is then defined as $M^+ = AD^+A^{-1}$.

2.1.2 Two-monomer MSD

As mentioned, the main use to which the classical (discrete chain, continuous time) formulation of the Rouse model will be put in this work is to make the connection between the computational (fully discrete) and analytical (fully continuous) formulations. We do so by studying the MSD for the relative coordinate of two fixed loci on the polymer.

We extend the model in eq. (2.6) to an infinite chain of discrete monomers, such that the connectivity matrix B becomes a Toeplitz operator with $B_{nn} = 2$, $B_{n,n\pm1} = -1$, and all other $B_{mn} = 0$. As such it is diagonalized by the Fourier transform, with eigenvalues given by the discrete cosine transform (DCTIII) of the diagonal entries:

$$B_{nm} = \int_{-\pi}^{\pi} \mathrm{d}\omega \, e^*(n,\omega) e(m,\omega) \lambda(\omega) \,, \tag{2.14}$$

with $\lambda(\omega) = 2(1-\cos\omega)$, the Fourier basis $e(n,\omega) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}n\omega}$, and * denoting complex conjugation. Note that while technically B is still not invertible (since $\lambda(0) = 0$), the infinite chain takes infinite time to reach the coil diffusion regime. Mathematically this has the consequence that $\{\omega \in [-\pi,\pi] : \lambda(\omega) = 0\} = \{0\} \subset [-\pi,\pi]$ is a subset of measure zero, such that the integrals below are well-defined, allowing us to mostly ignore this problem.

The system is driven by homogeneous thermal noise, $S = 2\gamma k_\mathrm{B} T \mathbb{1}$. For convenience, we introduce $D \equiv \frac{k_\mathrm{B} T}{\gamma}$.

We calculate the MSD $\mu(\Delta t) \equiv \left\langle (y(t+\Delta t) - y(t))^2 \right\rangle$ of a linear observable $y(t) \equiv \boldsymbol{w}^T \boldsymbol{x}(t)$:

$$\begin{aligned} \mu(\Delta t) = \boldsymbol{w}^T \Big\langle & \boldsymbol{x}(t+\Delta t) \otimes \boldsymbol{x}^T(t+\Delta t) - \boldsymbol{x}(t+\Delta t) \otimes \boldsymbol{x}^T(t) \\ & - \boldsymbol{x}(t) \otimes \boldsymbol{x}^T(t+\Delta t) + \boldsymbol{x}(t) \otimes \boldsymbol{x}^T(t) \Big\rangle \boldsymbol{w} \end{aligned} \tag{2.15}$$

$$\overset{(2.13)}{=} 2\boldsymbol{w}^T \left(1 - \mathrm{e}^{-\frac{k}{\gamma} B \Delta t}\right) \mathcal{J} \boldsymbol{w} \tag{2.16}$$

$$\overset{(2.10)}{=} \frac{2D\gamma}{k} \boldsymbol{w}^T \left(1 - \mathrm{e}^{-\frac{k}{\gamma} B \Delta t}\right) B^{-1} \boldsymbol{w} \tag{2.17}$$

$$\overset{(2.14)}{=} \frac{D\gamma}{k} \int_{-\pi}^{\pi} \mathrm{d}\omega \left| \sum_n w(n) e(n,\omega) \right|^2 \frac{1 - \mathrm{e}^{-\frac{2k}{\gamma} \Delta t (1-\cos\omega)}}{1-\cos\omega} \,. \tag{2.18}$$

We are interested in the relative position of two monomers a and b on the chain, meaning

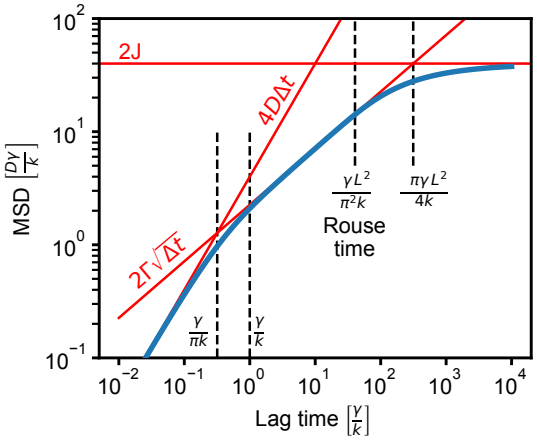


Figure 2.1: Three scaling regimes for the two-particle MSD of an infinite discrete Rouse polymer. Numerical evaluation of eq. (2.21) in blue, asymptotes (2.22), (2.30) and (2.34) in red. Note the width of the crossover regimes: the intersection of the diffusive and Rouse asymptotes is given by $\frac{\gamma}{\pi k}$, but the Rouse scaling is a good approximation only for $\Delta t > \frac{\gamma}{k}$. Similarly, Rouse scaling is a good approximation until the *Rouse time* $\frac{\gamma L^2}{\pi^2 k}$, but the asymptotes cross only at $\frac{\pi\gamma L^2}{4k}$; full equilibration takes yet another order of magnitude in time.

$w(n) = \delta_{an} - \delta_{bn}$. We thus find

$$\sum_n w(n)e(n,\omega) = \frac{1}{\sqrt{2\pi}}\left(\mathrm{e}^{\mathrm{i}\omega a} - \mathrm{e}^{\mathrm{i}\omega b}\right) = \sqrt{\frac{2}{\pi}}\mathrm{i}\mathrm{e}^{\mathrm{i}\omega\frac{a+b}{2}}\sin\omega\frac{a-b}{2}\,, \tag{2.19}$$

such that the expression for the MSD becomes

$$\mu(\Delta t) = \frac{2D\gamma}{\pi k}\int_{-\pi}^{\pi}\mathrm{d}\omega\,\frac{1-\mathrm{e}^{-\frac{2k}{\gamma}\Delta t(1-\cos\omega)}}{1-\cos\omega}\sin^2\frac{(a-b)\omega}{2} \tag{2.20}$$

$$\equiv \frac{2D\gamma}{\pi k}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\mathrm{d}z\,\frac{\sin^2 Lz}{\sin^2 z}\left(1-\mathrm{e}^{-\frac{4k}{\gamma}\Delta t\sin^2 z}\right)\,, \tag{2.21}$$

where we utilized $1-\cos\omega = 2\sin^2\omega$, substituted $z \equiv \frac{\omega}{2}$, and introduced the tether length $L \equiv a-b$.

Equation (2.21) exhibits three scaling regimes (fig. 2.1):

- at very short times we expand the exponential to first order and find

$$\mu(\Delta t \ll \frac{\gamma}{4k}) \approx \frac{2D\gamma}{\pi k}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\mathrm{d}z\,\frac{4k}{\gamma}\Delta t\sin^2 Lz = 4D\Delta t = 2\mu_{\text{single free monomer}}(\Delta t)\,, \tag{2.22}$$

  in accordance with the intuition that at short times the monomers do not feel their neighbors and thus diffuse freely.

- at long times, we have that $\mathrm{e}^{-\frac{4k}{\gamma}\Delta t\sin^2 z} \to 0\,\forall z \in [-\frac{\pi}{2},\frac{\pi}{2}]\setminus\{0\}$. We thus obtain

$$\mu(\Delta t\to\infty) = \frac{2D\gamma}{\pi k}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\mathrm{d}z\,\frac{\sin^2 Lz}{\sin^2 z} = 2\frac{D\gamma}{k}L\,, \tag{2.23}$$

where we prove the last equality by induction over $L \in \mathbb{N}_0$: the induction hypothesis $\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}z \, \frac{\sin^2 Lz}{\sin^2 z} = \pi L$ is trivially true for $L = 0$ and $L = 1$; we then use standard trigonometry to show that

$$\begin{aligned}
\sin^2(L+1)z + \sin^2(L-1)z = &\ (\sin Lz \cos z + \cos Lz \sin z)^2 \\
&+ (\sin Lz \cos z - \cos Lz \sin z)^2 && (2.24)\\
&= 2\sin^2 Lz \left(1 - \sin^2 z\right) + 2\cos^2 Lz \sin^2 z && (2.25)\\
&= 2\sin^2 Lz + 2\sin^2 z \cos 2Lz \,, && (2.26)
\end{aligned}$$

such that, using the induction hypothesis for L and $L-1$, we find

$$\begin{aligned}
\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}z \, \frac{\sin^2(L+1)z}{\sin^2 z} &= 2\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}z \, \frac{\sin^2 Lz}{\sin^2 z} - \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}z \, \frac{\sin^2(L-1)z}{\sin^2 z} \\
&\quad + 2\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}z \, \cos 2Lz && (2.27)\\
&= \pi(2L - (L-1)) && (2.28)\\
&= \pi(L+1) \,, && (2.29)
\end{aligned}$$

where the integral in the third term runs over full periods of the cosine and thus vanishes. Finally, we restate the result as

$$\mu(\Delta t \to \infty) = 2\frac{D\gamma}{k} L \equiv 2J \,. \tag{2.30}$$

- to find the scaling behavior at intermediate times, where the local chain connectivity is relevant, but the full tether has not equilibrated yet (such that the loci effectively "do not know that they are connected"), we take that tether to be infinitely long, $L \to \infty$. In this limit, $\sin^2 Lz$ oscillates arbitrarily fast, such that $\int \mathrm{d}z \, f(z) \sin^2 Lz = \frac{1}{2} \int \mathrm{d}z \, f(z)$ for

continuous $f(z)$. Consequently, eq. (2.21) becomes

$$\mu(\Delta t) = \frac{D\gamma}{\pi k}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}z\, \frac{1-\mathrm{e}^{-\frac{4k}{\gamma}\Delta t \sin^2 z}}{\sin^2 z} \tag{2.31}$$

$$= \frac{D\gamma}{\pi k}\int_{-1}^{1} \mathrm{d}\zeta\, \frac{1-\mathrm{e}^{-\frac{4k}{\gamma}\Delta t \zeta^2}}{\zeta^2\sqrt{1-\zeta^2}} \tag{2.32}$$

$$= 4D\Delta t\, \mathrm{e}^{-\frac{2k}{\gamma}\Delta t}\left[I_0\left(\frac{2k}{\gamma}\Delta t\right) + I_1\left(\frac{2k}{\gamma}\Delta t\right)\right], \tag{2.33}$$

where we substitute $\zeta := \sin z$ and $I_\alpha(z)$ are the modified Bessel functions of the first kind, which have the asymptotic expansion $I_\alpha(z) = \frac{\mathrm{e}^z}{\sqrt{2\pi z}}\left[1+\mathcal{O}\left(\frac{1}{z}\right)\right]$. Thus, for $\Delta t \to \infty$ we find

$$\mu(\Delta t) \approx 4D\sqrt{\frac{\gamma\Delta t}{\pi k}} \equiv 2\Gamma\sqrt{\Delta t}\,, \tag{2.34}$$

with $\Gamma \equiv 2D\sqrt{\frac{\gamma}{\pi k}}$. We will refer to this intermittent regime as the *Rouse regime*.

We can define the crossovers between the three regimes by equating the asymptotes, yielding

$$\tau_{D\to\Gamma} = \frac{\gamma}{\pi k} \qquad \text{and} \qquad \tau_{\Gamma\to J} = \frac{\pi\gamma}{4k}L^2 \tag{2.35}$$

for the diffusive-to-Rouse and Rouse-to-equilibrium transitions respectively. Note that $\tau_{\Gamma\to J}$ is a factor $\frac{\pi^3}{4} \approx 7.75$ greater than the commonly quoted Rouse time of $\frac{\gamma L^2}{\pi^2 k}$. The latter should not be interpreted as the location of the crossover, but as the time where the MSD starts deviating markedly from the Rouse scaling; similarly, while $\tau_{D\to\Gamma}$ marks the position of the crossover, numerical evaluation shows that the Rouse scaling remains a good approximation only for $\Delta t \gtrsim \frac{\gamma}{k} = \pi\tau_{D\to\Gamma}$ (fig. 2.1).

2.2 Computational: discrete chain in discrete time

From eq. (2.7) we can calculate

$$\boldsymbol{x}(t+\Delta t) = \mathrm{e}^{-\frac{k}{\gamma}B\Delta t}\boldsymbol{x}(t) + \int_0^{\Delta t}\mathrm{d}\tau\, \mathrm{e}^{-\frac{k}{\gamma}B\tau}\left[\boldsymbol{F} + \boldsymbol{\xi}(t+\Delta t-\tau)\right] \tag{2.36}$$

$$\equiv A\boldsymbol{x}(t) + \boldsymbol{G} + \boldsymbol{\eta}\,, \tag{2.37}$$

i.e. we can evolve the solution in discrete time steps. Note that this is not a discretization of the equations of motion, but of the solution to those. Therefore, the conformations $\boldsymbol{x}(t)$ we

obtain at discrete time points from this equation are simply samples from the exact solution to eq. (2.6), not an approximation.

To make the discrete nature of time more explicit, we rewrite:

$$\boldsymbol{x}_n = A\boldsymbol{x}_{n-1} + \boldsymbol{G} + \boldsymbol{\eta}_n\,, \qquad \langle \boldsymbol{\eta}_m \otimes \boldsymbol{\eta}_n \rangle = \Sigma\delta_{mn}\,. \tag{2.38}$$

By discretizing time, we have reformulated the stochastic differential equation eq. (2.6) as an AR(1) process ("auto-regressive of order 1"). The connection to the continuous-time formulation (2.6) is given by

$$A = \mathrm{e}^{-\frac{k}{\gamma}B\Delta t}\,, \tag{2.39}$$

$$\boldsymbol{G} = \int_0^{\Delta t} \mathrm{d}\tau\, \mathrm{e}^{-\frac{k}{\gamma}B\tau}\boldsymbol{F} \overset{*}{=} \frac{\gamma}{k}(\mathbb{1} - A)B^{-1}\boldsymbol{F}\,, \tag{2.40}$$

$$\Sigma = \int_0^{\Delta t} \mathrm{d}\tau\, \mathrm{e}^{-\frac{k}{\gamma}B\tau} S \mathrm{e}^{-\frac{k}{\gamma}B^T\tau} \overset{**}{=} \frac{\gamma}{2k}(\mathbb{1} - A^2)B^{-1}S\,, \tag{2.41}$$

where for * we assumed B to be non-singular, for ** we need B non-singular and $SB^T = BS$. Note that the integrals are finite in either case; the non-singularity condition is necessary only for the compact notation in terms of B^{-1}.

Again, we can solve the model starting from an initial condition x_{n_0}:

$$\boldsymbol{x}_n = A^{n-n_0}\boldsymbol{x}_{n_0} + \sum_{k=0}^{n-n_0-1} A^k(\boldsymbol{G} + \boldsymbol{\eta}_{n-k})\,. \tag{2.42}$$

Similarly to section 2.1.1, if all the eigenvalues of A are within the unit circle, a steady state exists and we can take $n_0 \to -\infty$:

$$\boldsymbol{x}_n = (\mathbb{1} - A)^{-1}\boldsymbol{G} + \sum_{k=0}^{\infty} A^k\boldsymbol{\eta}_{n-k}\,. \tag{2.43}$$

From this we can immediately read off the steady state distribution:

$$\langle \boldsymbol{x} \rangle_{\mathsf{ss}} = (\mathbb{1} - A)^{-1}\boldsymbol{G}\,, \tag{2.44}$$

$$\mathcal{J} \equiv \langle \boldsymbol{x} \otimes \boldsymbol{x}^T \rangle_{\mathsf{c,\,ss}} = \sum_{k=0}^{\infty} A^k\Sigma\left(A^T\right)^k \overset{*}{=} (\mathbb{1} - A^2)^{-1}\Sigma\,, \tag{2.45}$$

where for * we assume $\Sigma A^T = A\Sigma$. Since the model is driven by Gaussian noise, the steady

state distribution is Gaussian and thus fully determined by the first two moments, eqs. (2.44) and (2.45). Unsurprisingly, using eqs. (2.39) to (2.41) these expressions are identical to eqs. (2.9) and (2.10); similarly, eq. (2.13) translates as

$$\left\langle \boldsymbol{x}_m \otimes \boldsymbol{x}_n^T \right\rangle = A^{m-n} \mathcal{J} \,. \tag{2.46}$$

2.2.1 Entropy production

Since eq. (2.38) is a fully discrete model, we can calculate the expected entropy production per step in steady state. For now we specialize to the case $\boldsymbol{G} = 0$ (no external forcing). For a step from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$, the entropy production is given by

$$\Delta S(\boldsymbol{x}_1, \boldsymbol{x}_2) = \log \frac{P(\boldsymbol{x}_1 \to \boldsymbol{x}_2)}{P(\boldsymbol{x}_2 \to \boldsymbol{x}_1)} = \log \frac{P(\boldsymbol{\eta} = \boldsymbol{x}_2 - A\boldsymbol{x}_1)}{P(\boldsymbol{\eta} = \boldsymbol{x}_1 - A\boldsymbol{x}_2)} \tag{2.47}$$

$$= \operatorname{tr}\left[-\frac{1}{2}\left(\Sigma^{-1} - A^T\Sigma^{-1}A\right)\left(\boldsymbol{x}_2 \otimes \boldsymbol{x}_2^T - \boldsymbol{x}_1 \otimes \boldsymbol{x}_1^T\right) + \left(A^T\Sigma^{-1} - \Sigma^{-1}A\right)\boldsymbol{x}_2 \otimes \boldsymbol{x}_1^T \right], \tag{2.48}$$

where the distribution of the innovations $\boldsymbol{\eta}$ is a Gaussian with covariance Σ:

$$P(\boldsymbol{\eta}) = |2\pi\Sigma|^{-\frac{1}{2}} \, \mathrm{e}^{-\frac{1}{2}\boldsymbol{\eta}^T\Sigma^{-1}\boldsymbol{\eta}} \,; \qquad \log P(\boldsymbol{\eta}) = -\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\boldsymbol{\eta} \otimes \boldsymbol{\eta}^T\right] . \tag{2.49}$$

The expected entropy production in steady state is now found by taking an expectation value of eq. (2.48). Since we are in steady state, $\left\langle \boldsymbol{x}_2 \otimes \boldsymbol{x}_2^T \right\rangle = \left\langle \boldsymbol{x}_1 \otimes \boldsymbol{x}_1^T \right\rangle = \mathcal{J}$ and the first two terms cancel. For the third term we substitute $\left\langle \boldsymbol{x}_2 \otimes \boldsymbol{x}_1^T \right\rangle = A\mathcal{J}$ according to eq. (2.46) and finally find

$$\langle \Delta S \rangle_{\mathsf{ss}} = \operatorname{tr}\left(A^T\Sigma^{-1} - \Sigma^{-1}A\right)A\mathcal{J} \tag{2.50}$$

$$= \operatorname{tr}\left[\left(\mathbb{1} - A^2\right)\mathcal{J}\Sigma^{-1} - \mathbb{1}\right], \tag{2.51}$$

where the last step utilizes that the steady state covariance is stable under propagation: $A\mathcal{J}A^T + \Sigma = \mathcal{J}$. In eq. (2.45) we saw that if $\Sigma A^T = A\Sigma$ then $\mathcal{J} = (\mathbb{1} - A^2)^{-1}\Sigma$. In this case we clearly get $\langle \Delta S \rangle = 0$, i.e. we are dealing with an equilibrium system.

Note that $\Sigma A^T = A\Sigma$ (or, equivalently in terms of the continuous-time formulation of

section 2.1, $SB^T = BS$) is a sufficient, but not necessary condition for an equilibrium steady state.

2.2.2 Likelihood function for state trajectories in a multi-state Rouse model

One of the main uses of the fully discretized model (2.38) is its amenability to numerical computation, and the fact that the connectivity matrix B is not constrained to a linear polymer. This is thus the version of the Rouse model that we employ in chapter 4 for looping inference from super-resolution live-cell microscopy data. The core of this Bayesian inference algorithm is the calculation of the likelihood $\mathcal{L}(\theta) \equiv p\,(y \mid \theta)$ of observing a given trajectory y, given a *Loopingprofile* θ. The present section details these calculations.

Let us consider a collection of n discrete Rouse models that differ only by their connectivity matrices B. We will call these different models *looping states* (in the original work, this was just one open/linear, and one looped state) and label them with an index $\theta = 1, \ldots, n$. The continuous-time equation of motion (2.6) then becomes

$$\gamma \dot{\boldsymbol{x}}(t) = -kB(\theta(t))\boldsymbol{x}(t) + \boldsymbol{\xi}(t)\,, \qquad \langle \boldsymbol{\xi}(t) \otimes \boldsymbol{\xi}(t') \rangle = 2D\gamma^2 \mathbb{1}_N \delta(t - t')\,, \tag{2.52}$$

where $\boldsymbol{x}(t)$ is the N-dimensional vector of monomer positions, we introduce $D \equiv k_\mathrm{B}T/\gamma$ as before, $\mathbb{1}_N$ denotes the N-dimensional identity matrix, and $B(\theta(t))$ is the connectivity matrix pertaining to the state $\theta(t)$.

We constrain switches between states to occur only at discrete times $t = q\Delta t$, $q \in \mathbb{Z}$, for some fundamental time step Δt (for the application in chapter 4 this is the frame rate of the data). Over intervals $t \in [q\Delta t, (q+1)\Delta t)$ the coefficients of eq. (2.52) are then constant, such that its solution is given by eq. (2.7) as before. This allows us to discretize the propagation like in eq. (2.38), such that

$$\boldsymbol{x}_{q+1} = A(\theta(t_q))\boldsymbol{x}_q + \boldsymbol{\eta}_q\,, \qquad \langle \boldsymbol{\eta}_p \otimes \boldsymbol{\eta}_q \rangle = S(\theta(t_q))\delta_{pq}\,, \tag{2.53}$$

with $A(\theta)$ and $S(\theta)$ given in terms of $B(\theta)$ by eqs. (2.39) and (2.41), respectively:

$$A(\theta) = \mathrm{e}^{-\frac{k}{\gamma}B(\theta)\Delta t}\,, \qquad S(\theta) = 2D\gamma^2 \int_0^{\Delta t} \mathrm{d}\tau\, \mathrm{e}^{-\frac{k}{\gamma}B(\theta)\tau} \mathrm{e}^{-\frac{k}{\gamma}B^T(\theta)\tau}\,. \tag{2.54}$$

Note that in this section we use the symbol S instead of Σ for the discretized noise covariance.

We will generally assume that $B(\theta)$ is symmetric and positive semi-definite (as connectivity matrix of a passive spring network with unconstrained center of mass motion).

Equation (2.53) describes a linear model driven by Gaussian noise. As such, we can calculate the trajectory likelihood $p\left(y \mid \theta\right)$ efficiently via the Kalman filter equations [23, 24].

Let us assume that at time $t_q \equiv q\Delta t$ the system is described by a Gaussian ensemble with mean $\boldsymbol{\mu}_q$ and covariance Σ_q, which we write as $p(\boldsymbol{x}_q) = \mathcal{N}\left(\boldsymbol{x}_q; \boldsymbol{\mu}_q, \Sigma_q\right)$. We consider a linear observable $y_q \equiv \boldsymbol{w}^T \boldsymbol{x}_q$ (for our application in chapter 4 this will be the relative position of two monomers i and j on the chain: $(\boldsymbol{w})_k = \delta_{jk} - \delta_{ik}$), associated with a certain (Gaussian) measurement error σ^2; the probability density for y_q given the initial ensemble $p(\boldsymbol{x}_q)$ can then be expressed as

$$p\left(y_q \mid \boldsymbol{\mu}_q, \Sigma_q\right) = \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{x}_q \, p\left(y_q \mid \boldsymbol{x}_q\right) p\left(\boldsymbol{x}_q\right) \tag{2.55}$$

$$= \mathcal{N}\left(y_q; \boldsymbol{w}^T \boldsymbol{\mu}_q, \boldsymbol{w}^T \Sigma_q \boldsymbol{w} + \sigma^2\right) , \tag{2.56}$$

where the second line exploits the fact that the integral in the first line is a convolution of Gaussians.

Given an actual observation y_q now allows us to update our knowledge about the state ensemble. To that end, we calculate the optimal Kalman gain

$$\boldsymbol{k} = \frac{\Sigma_q \boldsymbol{w}}{\boldsymbol{w}^T \Sigma_q \boldsymbol{w} + \sigma^2} \tag{2.57}$$

in terms of which the Kalman update is expressed as

$$\boldsymbol{\mu}_q^{\mathsf{post}} = \boldsymbol{\mu}_q + \boldsymbol{k}\left(y_q - \boldsymbol{w}^T \boldsymbol{\mu}_q\right) \tag{2.58}$$

$$\Sigma_q^{\mathsf{post}} = \left(\mathbb{1}_N - \boldsymbol{k}\boldsymbol{w}^T\right)\Sigma_q \, . \tag{2.59}$$

This updated ensemble can now be propagated under eq. (2.53) to find

$$\boldsymbol{\mu}_{q+1} = A(\theta(t_q))\boldsymbol{\mu}_q \tag{2.60}$$

$$\Sigma_{q+1} = A(\theta(t_q))\Sigma_q A^T(\theta(t_q)) + S(\theta(t_q)) \, , \tag{2.61}$$

which then allows us to repeat the whole procedure for timepoint t_{q+1}.

The above procedure allows us to recursively compute $p\left(y_q \mid \boldsymbol{\mu}_q, \Sigma_q\right)$ for all q, once we

specify an initial ensemble to start from. Quite naturally, for this initial ensemble we choose the steady state associated with the initial looping state $\theta(0)$, which we can easily calculate as described in and below eq. (2.10). The prior ensemble $(\boldsymbol{\mu}_q, \Sigma_q)$ is then nothing but a convenient encoding of all the observations $\{y_p \mid p < q\}$ up to time t_q, as well as the assumed trajectory of state trajectories $\theta(t)$, such that we can rewrite the conditioning in the above expression as

$$p\left(y_q \mid \boldsymbol{\mu}_q, \Sigma_q\right) \equiv p\left(y_q \mid y_{q-1}, \ldots, y_1, \theta\right) . \tag{2.62}$$

The likelihood of observing the whole trajectory $\{y_q\}$ is now given by

$$\begin{aligned} \log \mathcal{L}(\theta) &\equiv \log p\left(y \mid \theta\right) && (2.63) \\ &= \sum_q \log p\left(y_q \mid y_{q-1}, \ldots, y_1, \theta\right) . && (2.64) \end{aligned}$$

This *Rouse likelihood* is the center piece of the Bayesian inference framework developed in chapter 5 and applied in chapter 4.

Note that instead of using the Kalman filter to calculate the Rouse likelihood in eq. (2.63), we could also have exploited the fact that the full ensemble of conformations $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ for a given looping profile $\theta(t)$ is always Gaussian, and just evaluate the likelihood directly. This however would require assembling the full $TN \times TN$ covariance matrix of that distribution and therefore scales quadratically in the trajectory length T. The calculation via the Kalman filter on the other hand scales only linearly in trajectory length and is thus computationally more efficient.

2.3 Analytical: continuous chain in continuous time

We now transition from the discrete to a continuous, infinitely long chain; this turns out to be more convenient for analytical treatment; the infinite chain is also a good approximation for our application to chromosomes. To that end, we replace the vector $\boldsymbol{x}(t)$ of N discrete monomer coordinates with conformations $x(s,t)$ for a continuous backbone coordinate $s \in \mathbb{R}$. In this continuum limit, the connectivity matrix (2.5) for the free linear polymer becomes a negative

Laplacian, $-\partial_s^2$, such that the equation of motion reads

$$\gamma \dot{x}(s,t) = \kappa \partial_s^2 x(s,t) + F(s,t) + \xi(s,t)\,, \tag{2.65}$$

$$\langle \xi(s,t)\xi(s',t')\rangle = 2\gamma k_{\mathsf{B}} T \delta(s-s')\delta(t-t')\,. \tag{2.66}$$

As before, $\xi(s,t)$ is a zero-mean Gaussian field representing the thermal noise, whose amplitude is given by the Einstein relation (fluctuation dissipation theorem). For convenience we introduce $D \equiv \frac{k_{\mathsf{B}}T}{\gamma}$.

Equation (2.65) is a heat equation, whose solution is given by a Weierstrass transform of the inhomogeneity $F + \xi$:

$$x(s,t) = \int_{\mathbb{R}} \mathrm{d}\sigma \int_0^t \mathrm{d}\tau \, \frac{\mathrm{e}^{-\frac{\gamma(\sigma-s)^2}{4\kappa(t-\tau)}}}{\sqrt{4\pi\gamma\kappa(t-\tau)}} \left(F(\sigma,\tau) + \xi(\sigma,\tau)\right) \tag{2.67}$$

$$= \frac{1}{\gamma} \int_{\mathbb{R}} \mathrm{d}\sigma \int_0^t \mathrm{d}\tau \, \mathcal{N}\left(\sigma; s, \frac{2\kappa}{\gamma}(t-\tau)\right) \left(F(\sigma,\tau) + \xi(\sigma,\tau)\right)\,, \tag{2.68}$$

where we assumed the collapsed initial condition $x(s,0) = 0\,\forall s$.

Clearly the expected response to an external force $F(s,t)$ is given by

$$\langle x(s,t)\rangle = \int_{\mathbb{R}} \mathrm{d}\sigma \int_0^t \mathrm{d}\tau \, \frac{\mathrm{e}^{-\frac{\gamma(\sigma-s)^2}{4\kappa(t-\tau)}}}{\sqrt{4\pi\gamma\kappa(t-\tau)}} F(\sigma,\tau)\,, \tag{2.69}$$

which we will further explore in sections 2.3.4 and 2.4.

2.3.1 Covariance structure

The solution (2.68) allows us to calculate the full covariance structure of $x(s,t)$ in real space. Without loss of generality we assume $t' \equiv t + \Delta t \geq t$, such that we can write

$$\begin{aligned}
\langle x(s',t')x(s,t)\rangle_{\mathsf{c}} &= \frac{1}{\gamma^2}\int \mathrm{d}\sigma'\mathrm{d}\sigma \int_0^{t'}\mathrm{d}\tau \int_0^{t}\mathrm{d}\tau' \mathcal{N}\left(\sigma'; s', \frac{2\kappa}{\gamma}(t'-\tau')\right) \\
&\qquad \times \mathcal{N}\left(\sigma; s, \frac{2\kappa}{\gamma}(t-\tau)\right)\langle \xi(\sigma',\tau')\xi(\sigma,\tau)\rangle \qquad (2.70)\\
&= 2D\int \mathrm{d}\sigma \int_0^t \mathrm{d}\tau\, \mathcal{N}\left(-\sigma; s'-s, \frac{2\kappa}{\gamma}(t-\tau)\right)\mathcal{N}\left(\sigma; 0, \frac{2\kappa}{\gamma}(t'-\tau)\right) \qquad (2.71)\\
&= 2D\int_0^t \mathrm{d}\tau\, \mathcal{N}\left(0; s'-s, \frac{2\kappa}{\gamma}(t'+t-2\tau)\right) \qquad (2.72)\\
&= 2D\int_0^t \frac{\sqrt{\gamma}\,\mathrm{d}\tau}{\sqrt{8\pi\kappa t\left(\frac{\Delta t}{2t}+1-\frac{\tau}{t}\right)}}\exp\left(-\frac{\gamma\Delta s^2}{8\kappa t\left(\frac{\Delta t}{2t}+1-\frac{\tau}{t}\right)}\right), \qquad (2.73)
\end{aligned}$$

where we first use the noise correlations $\langle \xi(\sigma',\tau')\xi(\sigma,\tau)\rangle = 2D\gamma^2\delta(\sigma'-\sigma)\delta(\tau'-\tau)$, then transform $\sigma \leftarrow s-\sigma$, and finally execute the integral over σ, which is a convolution of two Gaussians. In the last step, we expand the expression for the Gaussian and introduce $\Delta s \equiv s'-s$ and $\Delta t \equiv t'-t$. We now substitute $z \equiv 1-\frac{\tau}{t}$ and employ the incomplete Gamma function $\Gamma(\nu,z) \equiv \int_z^\infty t^{\nu-1}\mathrm{e}^{-t}\mathrm{d}t$ to rewrite the resulting integral as

$$\begin{aligned}
\langle x(s',t')x(s,t)\rangle_{\mathsf{c}} &= 2D\int_0^1 \frac{\sqrt{\gamma t}\,\mathrm{d}z}{\sqrt{8\pi\kappa\left(z+\frac{\Delta t}{2t}\right)}}\exp\left(-\frac{\gamma\Delta s^2}{8\kappa t\left(z+\frac{\Delta t}{2t}\right)}\right) \qquad (2.74)\\
&= 2D\,\frac{\gamma\,|\Delta s|}{8\kappa\sqrt{\pi}}\left[\Gamma\left(-\frac{1}{2}, \frac{\gamma\Delta s^2}{4\kappa(2t+\Delta t)}\right) - \Gamma\left(-\frac{1}{2}, \frac{\gamma\Delta s^2}{4\kappa\Delta t}\right)\right]. \qquad (2.75)
\end{aligned}$$

Ultimately we are interested in the equilibrium behavior of the chain. We therefore aim to expand eq. (2.75) for large t, while holding Δt and Δs constant. To that end we note the following representation of the incomplete Gamma function[2]: $\Gamma\left(-\frac{1}{2}, z\right) = \frac{2\mathrm{e}^{-z}}{\sqrt{z}} - 2\sqrt{\pi}\,\mathrm{erfc}\,\sqrt{z}$, which for small z expands as $\Gamma\left(-\frac{1}{2}, z\right) = \frac{2}{\sqrt{z}} - 2\sqrt{\pi} + \mathcal{O}(\sqrt{z})$. Expanding the first term and

[2] http://functions.wolfram.com/06.06.03.0006.01; also easily checked by differentiation

substituting the exact expression for the second one we find

$$\begin{aligned}\langle x(s',t')x(s,t)\rangle_{\mathsf{c}} = \quad & 2D\left[\sqrt{\frac{\gamma t}{2\pi\kappa}} \quad - \sqrt{\frac{\gamma\Delta t}{4\pi\kappa}}\exp\left(-\frac{\gamma\Delta s^2}{4\kappa\Delta t}\right) - \frac{\gamma\,|\Delta s|}{4\kappa}\operatorname{erf}\sqrt{\frac{\gamma\Delta s^2}{4\kappa\Delta t}}\right] \\ & + \mathcal{O}\left(\sqrt{\frac{\Delta t}{t}}, \sqrt{\frac{\gamma\Delta s^2}{\kappa t}}\right) \qquad (2.76) \\ \equiv \quad & A\sqrt{t} \qquad + C^0(\Delta s, \Delta t) + \mathcal{R}\,, \qquad (2.77)\end{aligned}$$

Note that the first term $A\sqrt{t}$ describes the continuing expansion of the chain, and accordingly diverges as $t \to \infty$. This means that the system as a whole actually never equilibrates, agreeing with the intuition that an infinite polymer with an initially completely collapsed conformation would not reach a steady state, but just keep expanding (note that since the chain is infinitely long, there is no whole coil diffusion at long times). However, for quantities that do not depend on the absolute position of the chain (like two-locus MSD) this term drops out, such that they do reach a steady state on time scales $t \gg \frac{\gamma}{\kappa}\Delta s^2$.

Two-point correlations in steady state (and thus, due to Gaussianity, the whole steady state distribution) are thus determined by

$$\begin{aligned}C^0(\Delta s, \Delta t) &\equiv \langle x(s',t')x(s,t)\rangle_{\mathsf{c}} - A\sqrt{t} \qquad (2.78) \\ &= -2D\left[\sqrt{\frac{\gamma\Delta t}{4\pi\kappa}}\exp\left(-\frac{\gamma\Delta s^2}{4\kappa\Delta t}\right) + \frac{\gamma\,|\Delta s|}{4\kappa}\operatorname{erf}\sqrt{\frac{\gamma\Delta s^2}{4\kappa\Delta t}}\right]. \qquad (2.79)\end{aligned}$$

We will demonstrate the utility of this expression in section 2.3.2, where we calculate the two-point MSD and compare it to the expressions obtained for the discrete chain in section 2.1.2.

2.3.2 MSD of linear observables in the continuous model

Consider the linear observable $y(t) \equiv \int \mathrm{d}s\, w(s)x(s,t)$. We can use eq. (2.77) to calculate its MSD in steady state:

$$\begin{aligned}
\mathrm{MSD}_y(\Delta t) &= \left\langle [y(t+\Delta t) - y(t)]^2 \right\rangle && (2.80)\\
&= \int w(s')\mathrm{d}s'\, w(s)\mathrm{d}s\, \left\langle [x(s',t+\Delta t) - x(s',t)]\, [x(s,t+\Delta t) - x(s,t)] \right\rangle && (2.81)\\
&\overset{(2.77)}{=} \int w(s')\mathrm{d}s'\, w(s)\mathrm{d}s\, \left[2C^0(\Delta s, 0) - 2C^0(\Delta s, \Delta t)\right] + \mathcal{R} && (2.82)\\
&\xrightarrow{t\to\infty} \int w(s')\mathrm{d}s'\, w(s)\mathrm{d}s\, \mathrm{MSD}^0(\Delta s, \Delta t)\,, && (2.83)
\end{aligned}$$

with

$$\begin{aligned}
\mathrm{MSD}^0(\Delta s, \Delta t) &= 2C^0(\Delta s, 0) - 2C^0(\Delta s, \Delta t) && (2.84)\\
&= D\sqrt{\frac{\gamma\Delta t}{\pi\kappa}}\, \mathrm{E}_{\frac{3}{2}}\left(\frac{\gamma\Delta s^2}{4\kappa\Delta t}\right) \equiv \frac{1}{2}\Gamma\sqrt{\Delta t}\, \mathrm{E}_{\frac{3}{2}}(z)\,. && (2.85)
\end{aligned}$$

The exponential integral $\mathrm{E}_{\frac{3}{2}}$ has the representation

$$\mathrm{E}_{\frac{3}{2}}(z) = 2\mathrm{e}^{-z} - 2\sqrt{\pi z}\,\mathrm{erfc}\,\sqrt{z}\,; \tag{2.86}$$

Since $\mathrm{E}_{\frac{3}{2}}(0) = 2$ we can immediately write the MSD of a single locus ($w(s) = \delta(s)$) as $\Gamma\sqrt{\Delta t}$.

In fact, eqs. (2.83) and (2.85) allow us to quickly calculate MSDs for a host of interesting observables:

- As noted above, for a single locus on the polymer, we simply have

$$\mathrm{MSD}_{\text{single locus}}(\Delta t) = \Gamma\sqrt{\Delta t} \tag{2.87}$$

  with $\Gamma = 2D\sqrt{\frac{\gamma}{\pi\kappa}} = \frac{2k_\mathrm{B}T}{\sqrt{\pi\gamma\kappa}}$.

- For the relative position of two particles separated by a backbone distance L (analogous to section 2.1.2) we set $w(s) = \delta(s-L) - \delta(s)$ and find (after some algebra)

$$\mathrm{MSD}_{\text{two loci}}(\Delta t) = 2\Gamma\sqrt{\Delta t}\left(1 - \mathrm{e}^{-\frac{\tau}{\Delta t}}\right) + 2J\,\mathrm{erfc}\,\sqrt{\frac{\tau}{\Delta t}}\,, \tag{2.88}$$

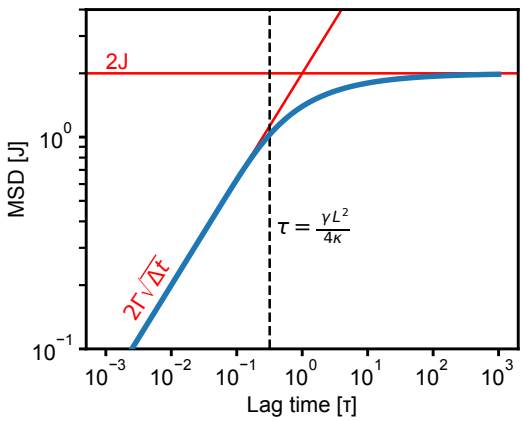


Figure 2.2: Two-particle MSD of the continuous Rouse model. Plot of eq. (2.88) in blue, asymptotes in red, with the time scale τ indicated. Note the absence of the third, early time scaling regime of the discrete model (cf. fig. 2.1).

with $\Gamma \equiv 2D\sqrt{\frac{\gamma}{\pi\kappa}}$ as before, $J \equiv \frac{D\gamma}{\kappa}L = \langle y^2 \rangle$ the mean squared distance between the two loci in steady state, and the crossover time scale $\tau \equiv \frac{\gamma L^2}{4\kappa} = \frac{1}{\pi}\left(\frac{J}{\Gamma}\right)^2$. Equation (2.88) and its asymptotes are shown in fig. 2.2

- The center of mass of of a stretch of polymer of length L is described by $w(s) = \frac{1}{L}\Theta(L-s)\,\Theta(s)$. For any function $f(z)$ of $z \equiv \frac{\Delta s^2}{4\kappa\Delta t}$ the integral in eq. (2.83) can then be transformed as

$$\int w(s')\mathrm{d}s'\, w(s)\mathrm{d}s\, f(z) = \frac{2\kappa\Delta t}{L^2}\int_0^\zeta \frac{\mathrm{d}a}{\sqrt{a}}\int_0^a \frac{\mathrm{d}z}{\sqrt{z}}\, f(z)\,, \tag{2.89}$$

with $\zeta \equiv \frac{L^2}{4\kappa\Delta t}$, analogous to the definition of z. Plugging into eqs. (2.83) and (2.85) then gives

$$\mathrm{MSD}_{\mathrm{COM}(L)}(\Delta t) = \frac{\kappa\Gamma}{L^2}(\Delta t)^{\frac{3}{2}}\int_0^\zeta \frac{\mathrm{d}a}{\sqrt{a}}\int_0^a \frac{\mathrm{d}z}{\sqrt{z}}\,\mathrm{E}_{\frac{3}{2}}(z) \tag{2.90}$$

$$= \frac{\kappa\Gamma}{L^2}(\Delta t)^{\frac{3}{2}}\left[\mathrm{E}_{\frac{3}{2}}(\zeta) - \mathrm{E}_{\frac{5}{2}}(\zeta) - \frac{4}{3} + 2\sqrt{\pi\zeta}\right]. \tag{2.91}$$

To make sense of the term in square brackets, note the series expansion

$$\mathrm{E}_{\frac{3}{2}}(z) - \mathrm{E}_{\frac{5}{2}}(z) = \frac{4}{3} - 2\sqrt{\pi z} + 4z - \frac{4}{3}\sqrt{\pi}z^{\frac{3}{2}} + \mathcal{O}(z^2) \tag{2.92}$$

for small z and asymptotic expansion $\mathrm{E}_\nu \to \frac{1}{z}\mathrm{e}^{-z}$ as $z \to \infty$. With these we find the

expected limiting behavior

$$\mathrm{MSD}_{\mathsf{COM}(L)}(\Delta t) \to \frac{2D}{L}\Delta t \qquad \text{as } \Delta t \to 0, \text{ and} \tag{2.93}$$

$$\mathrm{MSD}_{\mathsf{COM}(L)}(\Delta t) \to \Gamma\sqrt{\Delta t} \qquad \text{as } \Delta t \to \infty. \tag{2.94}$$

Note that this is (unsurprisingly) exactly the behavior of a single monomer in the discrete model, c.f. section 2.1.2. We can thus establish the discrete model as appropriate coarse-graining of the continuous one, by dividing the continuous chain into contiguous "blobs" and using their center of mass as monomer coordinate.

- So far we were concerned with tracer particles that are stationary on the polymer. In the context of molecular motors—such as cohesin—it might be interesting to consider a tracer that moves along the polymer backbone with some velocity v, such that $w(s) = \delta(s - vt)$. Note that now $w(s)$ has become time-dependent, such that the formulation (2.81) is not valid anymore; the appropriate modifications are, however, straightforward and one finds that in this case

$$\mathrm{MSD}_{\mathsf{motor}(v)}(\Delta t) = -2C^0(v\Delta t, \Delta t) \tag{2.95}$$

$$= \Gamma\sqrt{\Delta t}\,\mathrm{e}^{-\frac{\gamma v^2}{4\kappa}\Delta t} + \frac{D\gamma}{\kappa} v\Delta t \,\mathrm{erf}\sqrt{\frac{\gamma v^2}{4\kappa}\Delta t}\,. \tag{2.96}$$

Intuitively, at short times the polymer fluctuations dominate the tracer movement, such that it behaves just like a stationary tracer with $\mathrm{MSD}(\Delta t) = \Gamma\sqrt{\Delta t}$. At long times, the tracer outruns the polymer motion and moves linearly along the chain—which is ideal and thus adopts a random walk conformation. The tracer thus becomes effectively diffusive, since it "walks deterministically along a random path".

2.3.3 Correspondence between discrete and continuous chain

In section 2.1.2 we found three scaling regimes for the two-particle MSD $\mu(\Delta t)$ of an infinite, discrete Rouse chain: at very short times monomers diffuse freely, such that $\mu(\Delta t) = 4D\Delta t$. At intermediate times, monomers start feeling the local chain they are connected to, but the tether between the two loci under study has not yet equilibrated. Correspondingly, $\mu(\Delta t) = 2\Gamma\sqrt{\Delta t}$, with $\Gamma = 2D\sqrt{\frac{\gamma}{\pi k}}$ (eq. (2.34)). At long times, the chain between the two monomers under

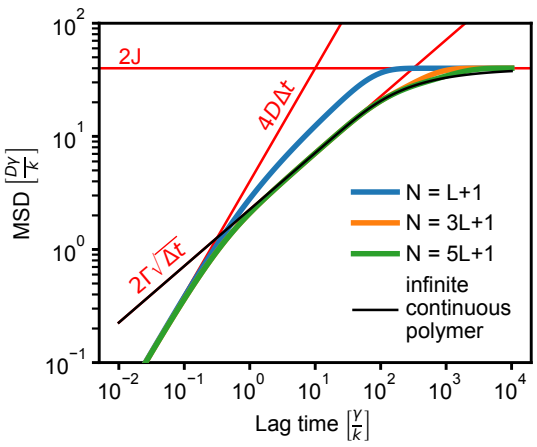


Figure 2.3: Comparison of continuous and discrete Rouse models. Two-particle MSD for analytical (thin black) and computational (thick colored; as indicated) Rouse models. While the analytical model considers a subchain embedded in an infinite polymer, in the computational model this embedding chain contains $N < \infty$ monomers. For this example, the subchain of interest (L bonds) is always embedded symmetrically, such that for $N = L + 1$ we are tracking the loose ends of the chain; clearly in this case the infinite chain of the analytical model is a bad approximation (blue vs. black curves). As the embedding chain grows longer, this finite-size effect diminishes. At early times, the monomers of the computational model are diffusive, while the continuous model maintains Rouse scaling.

study equilibrates, such that the MSD plateaus at $\mu(\Delta t) = 2J$ with $J = \frac{D\gamma}{k}L$ (eq. (2.30)).

For the continuous, infinite chain, section 2.3.2 provides a closed analytical expression for the two-particle MSD:

$$\mu(\Delta t) = 2\Gamma\sqrt{\Delta t}\left(1 - \mathrm{e}^{-\frac{\tau}{\Delta t}}\right) + 2J\,\mathrm{erfc}\sqrt{\frac{\tau}{\Delta t}} \tag{2.97}$$

with $\Gamma = 2D\sqrt{\frac{\gamma}{\pi\kappa}}$, $J = \frac{D\gamma}{k}L$, and $\tau \equiv \frac{1}{\pi}\left(\frac{J}{\Gamma}\right)^2$ (eq. (2.88)).

We can thus connect the completely discrete ("computational") model of section 2.2 to the completely continuous ("analytical") one studied in this section. There are precisely two differences between them, illustrated in fig. 2.3:

- the discrete chain has a third scaling regime at short times, where monomers diffuse freely. This is an artifact of the discrete model, owed to the precise microscopic dynamics assumed here (which in fact are quite unphysical; a real polymer is certainly not composed of point particles and harmonic springs). In applying the discrete model to physical situations, care should thus be taken to not consider time scales $\Delta t \lesssim \frac{\gamma}{k}$.

- the computational model necessarily has to work with a finite chain, while the analytical treatment (in continuous as well as discrete case) works in the limit of an infinitely long chain. Finite chain effects thus should always be considered; in the case of the looping inference of chapters 4 and 5 we do so by always considering a chain that is at least three times as long as the subchain of interest. This ensures that up to the equilibration time of that subchain the dynamics are close to those of an infinite chain; see fig. 2.3.

2.3.4 Pulling on a locus

Inspired by the experiments presented in chapter 3, let us consider an external force pulling on a specific genomic locus.

For a start, consider pulling with a constant point force $F(s,t) \equiv F\delta(s)$. Using the solution (2.68) we can immediately write the expected conformation of the polymer:

$$\langle x(s,t)\rangle = F\int_0^t \mathrm{d}\tau\, \frac{\mathrm{e}^{-\frac{\gamma s^2}{4\kappa(t-\tau)}}}{\sqrt{4\pi\gamma\kappa(t-\tau)}} \equiv \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}}\,\chi\left(\sqrt{\frac{\gamma}{\kappa t}}s\right), \tag{2.98}$$

where

$$\chi(z) \equiv \int_0^1 \mathrm{d}\tau\, \frac{\mathrm{e}^{-\frac{z^2}{4\tau}}}{\sqrt{4\tau}} = \mathrm{e}^{-\frac{1}{4}z^2} - \frac{\sqrt{\pi}}{2}\left(|z| - z\,\mathrm{erf}\left(\frac{z}{2}\right)\right) \tag{2.99}$$

is the characteristic shape of the conformation in terms of the rescaled backbone coordinate $z \equiv \sqrt{\frac{\gamma}{\kappa t}}s$.

Along the same lines, we can calculate the behavior of a finite size locus under uniform force application, i.e. $F(s,t) = \frac{F}{l}\Theta\left(\left|\frac{l}{2} - s\right|\right)$:

$$\langle x(s,t)\rangle = \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}}\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathrm{d}a\,\chi\left(\sqrt{\frac{\gamma}{\kappa t}}\,(s-al)\right) \equiv \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}}\,\chi_L(z)\,, \tag{2.100}$$

with the rescaled locus length $L = \sqrt{\frac{\gamma}{\kappa t}}l$ and backbone coordinate $z = \sqrt{\frac{\gamma}{\kappa t}}s$, and

$$\chi_L(z) \equiv \frac{1}{L}\left[\psi\left(\frac{L+2z}{4}\right) + \psi\left(\frac{L-2z}{4}\right)\right] - \frac{\sqrt{\pi}}{2}\begin{cases} \frac{L}{4} + \frac{z^2}{L}\,, & \text{for } |z| \le \frac{L}{2} \\ |z|\,, & \text{for } |z| > \frac{L}{2} \end{cases} \tag{2.101}$$

$$\psi(\zeta) = \zeta\mathrm{e}^{-\zeta^2} + \sqrt{\pi}\left(\zeta^2 + \frac{1}{2}\right)\mathrm{erf}\,\zeta\,. \tag{2.102}$$

We aim to understand this solution better by considering the long and short time limits. Note that z, L, and thus $\zeta_\pm \equiv \frac{L\pm 2z}{4}$ all scale as $t^{-\frac{1}{2}}$, such that they become small at long times. We then expand $\psi(\zeta) = 2\zeta + \mathcal{O}(\zeta^3)$ to find

$$\chi_L(z) = 1 - \frac{\sqrt{\pi}}{2}\begin{cases} \frac{L}{4} + \frac{z^2}{L}\,, & \text{for } |z| \le \frac{L}{2} \\ |z|\,, & \text{for } |z| > \frac{L}{2} \end{cases} + o\left(t^{-1}\right), \tag{2.103}$$

such that

$$\langle x(s,t)\rangle = \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}} - \frac{Fl}{2\kappa}\begin{cases}\frac{1}{4}+\frac{s^2}{l^2}\,, & \text{for } |s|\le\frac{l}{2}\\ \frac{|s|}{l}\,, & \text{for } |s|>\frac{l}{2}\end{cases} + o\left(t^{-\frac{1}{2}}\right)\,. \tag{2.104}$$

For short times, consider the asymptotic expansion

$$\psi(\zeta) = \sqrt{\pi}\left(\zeta^2+\frac{1}{2}\right)\operatorname{sign}\zeta + o\left(\zeta^{-1}\mathrm{e}^{-\zeta^2}\right) \tag{2.105}$$

and let $z \ge 0$ without loss of generality (since $\chi_L(z)$ is symmetric in z). Equation (2.101) then becomes

$$\chi_L(z) = \frac{\sqrt{\pi}}{L}\begin{cases}\zeta_+^2+\zeta_-^2+1-\frac{L^2}{8}-\frac{z^2}{2}\,, & \text{for } 0\le z<\frac{L}{2}\\ \zeta_+^2+\frac{1}{2}-\frac{L^2}{4}\,, & \text{for } z=\frac{L}{2}\\ \zeta_+^2-\zeta_-^2-\frac{Lz}{2}\,, & \text{for } z>\frac{L}{2}\end{cases} + \mathcal{O}\left(\sqrt{t}\mathrm{e}^{-\frac{1}{t}}\right) \tag{2.106}$$

$$= \frac{\sqrt{\pi}}{L}\begin{cases}1\,, & \text{for } 0\le z<\frac{L}{2}\\ \frac{1}{2}\,, & \text{for } z=\frac{L}{2}\\ 0\,, & \text{for } z>\frac{L}{2}\end{cases} + \mathcal{O}\left(\sqrt{t}\mathrm{e}^{-\frac{1}{t}}\right)\,. \tag{2.107}$$

Resolving $L \equiv \sqrt{\frac{\gamma}{\kappa t}}l$ and re-inserting into eq. (2.100), we find

$$\langle x(s,t)\rangle = \frac{Ft}{\gamma l}\begin{cases}1\,, & \text{for } |s|<\frac{l}{2}\\ \frac{1}{2}\,, & \text{for } |s|=\frac{l}{2}\\ 0\,, & \text{for } |s|>\frac{l}{2}\end{cases} + \mathcal{O}\left(\sqrt{t}\mathrm{e}^{-\frac{1}{t}}\right)\,. \tag{2.108}$$

Summarizing, at early times, the system behaves as if there was no connection between the forced locus and the rest of the chain: the locus itself moves uniformly with a velocity of $\frac{F}{\gamma l}$, while the rest of the chain is unperturbed (eq. (2.108)). Once the locus moves far enough, the attachment to the surrounding chain starts deforming it from the edges; furthermore, having to drag along increasing amounts of surrounding chain starts slowing down the motion. At long times, the conformation at the tip of the pulled chain reaches a steady state: the parts of the chain that are just pulled along become linearly stretched, while the locus adopts a parabolic configuration, thus achieving a linear increase in chain tension towards both edges (eq. (2.104)).

Similar to section 2.3.3, note how the locus with finite extent behaves exactly like a monomer in the discrete model would. Furthermore, note that after the initial phase (roughly: once the locus has translocated about its own size), the finite extent of the locus does not matter anymore and its motion is exactly the same as if we had exerted a point force. This is seen easily by evaluating eqs. (2.98) and (2.100) at $s = 0$: according to eq. (2.98) we have $\langle x(0,t)\rangle = \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}}$ while from the long time limit of eq. (2.100) we find

$$\langle x(0,t)\rangle = \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}} - \frac{Fl}{8\kappa} + o\left(t^{-\frac{1}{2}}\right) \xrightarrow{t\to\infty} \frac{F\sqrt{t}}{\sqrt{\pi\gamma\kappa}}\,. \tag{2.109}$$

Why does the finite size (and thus increased viscous drag) of the locus not matter at long times? Once a significant amount of surrounding chain gets dragged along, the pulling force is dissipated mainly there, not at the locus itself. It is thus largely irrelevant how large the locus is itself; once it starts moving significantly, its dynamics is determined by the surrounding chain.

2.4 Specialized: continuous chain in discrete time

In chapter 3 we analyze chromosome pulling experiments through the lens of the continuous-chain force response eq. (2.69), investigated in some detail for constant forces in section 2.3.4. The experimental data is based on movies with a fixed frame rate and as such inherently discrete in time; taking this into account in the analytical treatment leads us to the force inference method developed in this chapter.

2.4.1 Modeling the experimental system

The system in chapter 3 is a genomically defined 4 Mb locus embedded in human chromosome 1, which itself is 249 Mb long. The locus gets coated with nano particles which are

- fluorescent, such that the locus can be tracked in the microscope;
- magnetic dipoles, which enables force application; and
- multi-valent, which means that with the nano particles the locus should be understood as a big ($\sim$ 500 nm diameter), solid (i.e. crosslinked) ball.

The locus being such a big object, one might expect that its motion is affected by friction with the surrounding medium; in fact, however, we find this contribution to be negligible. This is in

line with a numerical estimate of the friction caused by this object: for a nucleoplasmic viscosity $\eta \lesssim 10\,\mathrm{cP}$ [25, 26], radius of $R \approx 250\,\mathrm{nm}$, and peak velocity of $v \approx 0.1\,\mu\mathrm{m/s}$ (at 2 pN of force), the Stokes friction due to the compacted locus itself is

$$F_{\mathrm{drag}} = 6\pi\eta R v \lesssim 5 \times 10^{-2}\,\mathrm{pN}\,, \tag{2.110}$$

i.e. negligible against the pN forces exerted by the pulls. In line with this estimate, we also do not observe the diffusive slowdown one would expect in the force free MSD of the locus.

In summary, the model we will consider in this section simply treats the locus as a single point on an infinitely long polymer, subject to a time-varying force $F(t)$. The system thus obeys eq. (2.65), with $F(s,t) = F(t)\delta(s)$.

2.4.2 Force inference with a compact locus

The expected trajectory of the locus ($s = 0$ on the polymer) under a time dependent force $F(s,t) = F(t)\delta(s)$ is given by eq. (2.69) as

$$\langle x(0,t)\rangle = \int_0^t \mathrm{d}\tau\, \frac{F(\tau)}{\sqrt{4\pi\gamma\kappa(t-\tau)}}\,, \tag{2.111}$$

Note that this formulation assumes that $x(0,0) = 0$, i.e. at the beginning of the experiment, where the chain is (assumed to be) in equilibrium, the locus is positioned at $x_0 = 0$.

Since the motion of the locus is measured directly in the experiments, we are interested in solving this relationship for $F(t)$; which is a functional degree of freedom, so this inference problem is underdetermined for any finite amount of experimental data, where we know the position of the locus only at discrete times $\{t_i\}_{i=0,\ldots,N}$. To make the problem well-defined, we assume that $F(t)$ is piecewise constant:

$$F(t) = f_i \qquad \forall t \in [t_{i-1}, t_i)\,, i = 1, \ldots, N\,, \tag{2.112}$$

where $t_0 \equiv 0$ is the beginning of the experiment and $f_0 \equiv F(t < t_0) = 0$ is the assumption that the system is in equilibrium before the experiment. Within the intervals of constant force,

eq. (2.111) can be integrated easily, from which we find

$$x(t_i) = \frac{1}{\sqrt{\pi\gamma\kappa}} \Re \sum_{j=1}^{N} \left(\sqrt{t_i - t_{j-1}} - \sqrt{t_i - t_j} \right) f_j \,. \tag{2.113}$$

The symbol $\Re$ signifies "real part", such that $\Re\sqrt{x} = 0$ for $x < 0$; this is nothing but a convenient encoding of the fact that the integral in eq. (2.111) is exactly such that negative radicands never appear. Now, introducing the full position and force trajectories $\boldsymbol{x} \equiv (x_1, \ldots, x_N)$, $\boldsymbol{f} \equiv (f_1, \ldots, f_N)$, respectively, eq. (2.113) simply amounts to a linear transformation between the two:

$$\boldsymbol{x} = \frac{1}{\sqrt{\pi\gamma\kappa}} M \boldsymbol{f} \,, \tag{2.114}$$

where

$$M = \begin{pmatrix} \sqrt{t_1} & & & \\ \sqrt{t_2} - \sqrt{t_2 - t_1} & \sqrt{t_2 - t_1} & & \\ \sqrt{t_3} - \sqrt{t_3 - t_1} & \sqrt{t_3 - t_1} - \sqrt{t_3 - t_2} & \sqrt{t_3 - t_2} & \\ \vdots & & & \ddots \end{pmatrix} . \tag{2.115}$$

Solving for the force profile in terms of the observed trajectory is now reduced to a straightforward matrix inverse:

$$\boldsymbol{f} = \sqrt{\pi\gamma\kappa} M^{-1} \boldsymbol{x} \,. \tag{2.116}$$

Finally, to apply this force inference in practice, we have to calibrate the prefactor $\sqrt{\pi\gamma\kappa}$, which determines the absolute magnitude of the force. Conveniently (and as expected by fluctuation–dissipation), eq. (2.87) shows that the same combination of constants governs the thermal fluctuations of the locus in the force free case: $\mathrm{MSD}(\Delta t) = \Gamma\sqrt{\Delta t}$ with $\Gamma = \frac{2dk_\mathrm{B}T}{\sqrt{\pi\gamma\kappa}}$. This allows us to calibrate

$$\sqrt{\pi\gamma\kappa} = \frac{2dk_\mathrm{B}T}{\Gamma} \,, \tag{2.117}$$

with d the number of spatial dimensions in the MSD measurement.

Interestingly, eq. (2.117) provides a good fit to the data (modulo the additional hindrance described below; see also chapter 3), using the incubation chamber temperature of 37 °C. Within the experiments presented in chapter 3 we thus find no evidence for an "effectively higher temperature" due to active fluctuations, as one might have expected within a living cell.

2.4.3 Uncertainty in force estimate

Note that eq. (2.111) relates the *expected* trajectory of the locus to the applied force; a point we mostly chose to neglect in the previous section. Furthermore, any real experiment is associated with some error, such that we do not measure x_i directly, but $y_i \equiv x_i + \xi_i$, with a given localization uncertainty $\langle \xi_i \xi_j \rangle = \sigma^2 \delta_{ij}$. This section gives a quick treatment of these second order effects, which will provide us with an uncertainty estimate for the inferred force.

Given an applied force profile $\boldsymbol{f}$, the trajectory of the locus is sampled from a Gaussian ensemble whose *mean* is given by eq. (2.114). The corresponding variance is given by thermal fluctuations and can be determined from eq. (2.79). Together with the localization error σ^2, one finds

$$\begin{aligned} S_{ij} &\equiv \langle y_i y_j \rangle && (2.118)\\ &= \langle [x(0,t_i) + \xi_i - x(0,0) - \xi_0]\,[x(0,t_j) + \xi_j - x(0,0) - \xi_0] \rangle && (2.119)\\ &= \frac{k_\mathrm{B}T}{\sqrt{\pi\gamma\kappa}} \left[\sqrt{t_i} + \sqrt{t_j} - \sqrt{|t_i - t_j|} \right] + \sigma^2 \left(\delta_{ij} + 1 \right) . && (2.120) \end{aligned}$$

For the purpose of inferring the force profile $\boldsymbol{f}$, the covariance matrix S should be considered as "measurement uncertainty" of the observed trajectory $\boldsymbol{y}$. The corresponding uncertainty in the inferred force profile $\boldsymbol{f}$ is then given by simple error propagation through the inference (eq. (2.116)):

$$\mathrm{Cov}(\boldsymbol{f}) = \pi\gamma\kappa M^{-1} \,\mathrm{Cov}(\boldsymbol{y}) M^{-T} = \pi\gamma\kappa M^{-1} S M^{-T} , \tag{2.121}$$

where M^{-T} is the inverse transpose of M.

In summary, given a sequence of observations $\{(y_i, t_i)\}_{i=0,\ldots,N}$ of the locus position at defined time points, the posterior distribution over the force profile $\boldsymbol{f} \equiv (f_1, \ldots, f_N)$ is a normal distribution with mean

$$\langle \boldsymbol{f} \rangle = \sqrt{\pi\gamma\kappa} M^{-1} \boldsymbol{y} \tag{2.122}$$

and covariance

$$\left\langle \boldsymbol{f} \otimes \boldsymbol{f}^T \right\rangle_\mathsf{c} = \pi\gamma\kappa M^{-1} S M^{-T} , \tag{2.123}$$

with M and S defined in eqs. (2.115) and (2.118), respectively.

2.4.4 Force inference with additional hindrance

In chapter 3 we find that, while the initial behavior of the locus is well captured by the approach outlined in section 2.4.2, the farther it moves through the nucleus, the more resistance it seems to encounter. We explore multiple possibilities for how this could come about, the technical details of which are treated here. Figure 2.4 illustrates the performance of each of these variations in the experimental context of chapter 3.

Viscoelastic medium

Instead of the purely viscous solvent that we assumed so far, it has been suggested that the nucleoplasm itself is a viscoelastic medium [25,27]. This can be incorporated in the Rouse model by replacing the time derivative in the equation of motion (2.65) by a fractional derivative, thus introducing an appropriate memory kernel. This model has been studied previously [28]; for our purposes, however, it turns out to be sufficient to apply dimensional analysis to generalize our force inference.

For reference, recall eqs. (2.87) and (2.111), giving the MSD in the absence of force, as well as the force response of our locus:

$$\mathrm{MSD}\,(\Delta t) = \frac{2k_\mathrm{B}T}{\sqrt{\pi\gamma\kappa}}\sqrt{\Delta t}\,, \tag{2.124}$$

$$\langle x(t)\rangle = \frac{1}{\sqrt{\pi\gamma\kappa}}\int_0^t \frac{F(t)\mathrm{d}\tau}{2\sqrt{t-\tau}}\,. \tag{2.125}$$

A viscoelastic medium is usually taken to cause subdiffusive motion of free tracer particles, i.e. $\mathrm{MSD} \propto (\Delta t)^\alpha$ with $0 < \alpha \leq 1$. For loci on a polymer, this translates as $\mathrm{MSD} \propto (\Delta t)^{\frac{\alpha}{2}}$ [28]. Now, the only model constant entering eqs. (2.124) and (2.125) is the combination $\lambda \equiv \sqrt{\pi\gamma\kappa}$; to achieve an MSD scaling of $\frac{\alpha}{2}$, the model should thus be adjusted such that λ has units of

$$[\lambda] = \frac{\mathsf{mass}}{(\mathsf{time})^{2-\frac{\alpha}{2}}}\,, \tag{2.126}$$

such that the MSD in eq. (2.124) can still have units of length. Since eq. (2.125) depends on this same model constant λ, it would now be dimensionally inconsistent, unless the square root kernel $\sqrt{t-\tau}$ in the integral is replaced by $(t-\tau)^{1-\frac{\alpha}{2}}$. We thus find that in a viscoelastic

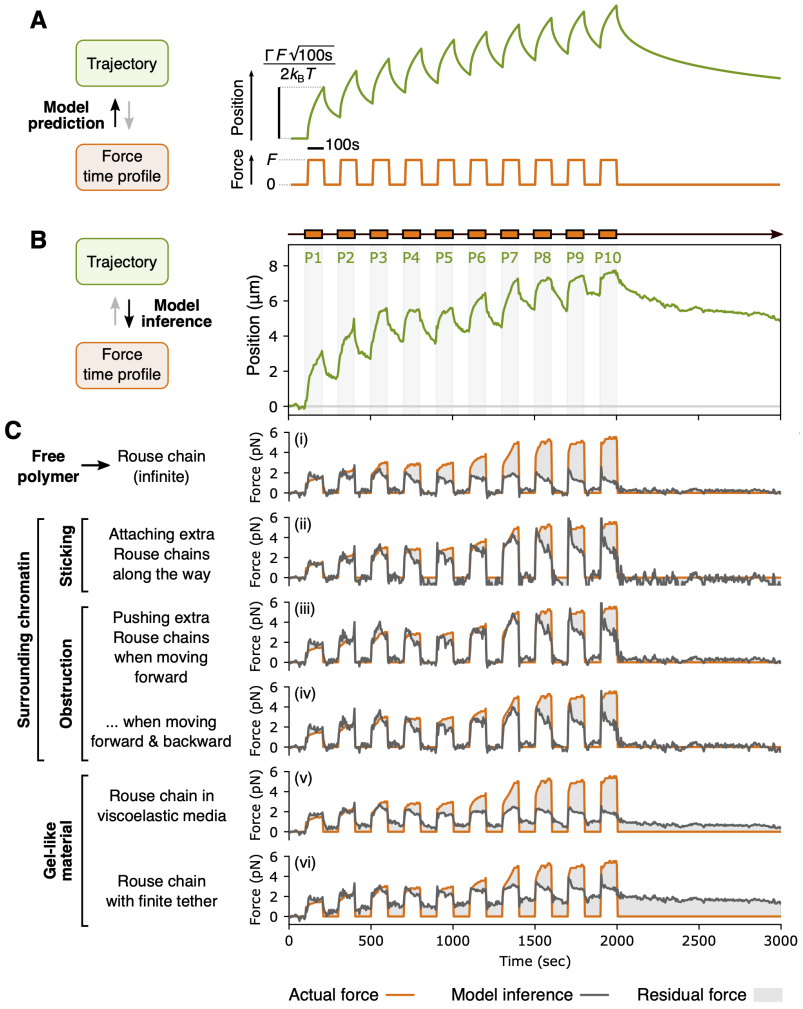


Figure 2.4: Variations of force inference for locus pulling. Equation (2.122) establishes a one-to-one correspondence between the observed trajectory (green) and the time profile of the applied force (orange). We can thus (**A**) predict the trajectory for a given force profile, or (**B**) infer the force profile from the observed trajectory. (**C**) Force inference results with different variations of the underlying model (gray), compared to the magnetically applied force (orange). Note how the latter increases over time, since the locus moves towards the magnetic pillar. (i) a pure Rouse model fits the initial behavior (first two pulls) and all releases, but does not capture the progressive increase in force. Incorporating interactions with the surrounding chromatin fibers through sticking ((ii); fig. 2.5, **C**) or obstruction (one- or two-sided "glove", (iii), (iv); fig. 2.5, **A, B**) provides a qualitatively improved fit. Additional elastic response in the system—either through a viscoelastic solvent ((v); eq. (2.128)) or by tethering one end of the chromosome to an immovable object ((vi); eq. (2.130))—does not qualitatively improve the fit over the vanilla Rouse model. On the contrary, now we are inferring a positive force upon release, meaning the observed recoil of the locus is slower than expected.

medium eqs. (2.124) and (2.125) become

$$\mathrm{MSD}\,(\Delta t) \propto \frac{k_\mathrm{B}T}{\lambda}\,(\Delta t)^{\frac{\alpha}{2}}\ , \tag{2.127}$$

$$\langle x(t)\rangle \propto \frac{1}{\lambda}\int_0^t \frac{F(\tau)\mathrm{d}\tau}{(t-\tau)^{1-\frac{\alpha}{2}}}\,. \tag{2.128}$$

Note that both equations contain numerical prefactors that might depend on the exponent α and can thus not be determined from this dimensional argument. However, up to this constant prefactor, we can run the inference scheme based on eq. (2.128) just like before (outlined in section 2.4.2). For the purposes of chapter 3 this heuristic treatment is sufficient, since the observed hindrance in the data is qualitatively not consistent with a viscoelastic medium, as illustrated in fig. 3.3.

Finite tether

In the vanilla inference scheme, we assume that the chromosome in which the locus is embedded stretches infinitely far in both directions. This seems to be a good approximation, as long as the whole chromosome indeed moves like a free polymer. However, this does not necessarily have to be the case; for example some parts of it might be fixed at the lamina (Lamina Associated Domains, LADs), or at other points in the nucleus. For this reason, we ask how our inference scheme changes if we assume the polymer on one side of the locus to have a finite extent, its end being fixed in space.

We still position the locus at $s = 0$ and assume the chain to extend to infinity for $s < 0$. In the positive direction, we add the boundary condition that $x(L,t) = 0\,\forall t$. Mathematically, we can explicitly take this boundary condition into account by modifying the fundamental solution according to the method of images. Equation (2.68) then becomes

$$\begin{aligned} x(s,t) = &\int_{-\infty}^{L}\mathrm{d}\sigma\int_0^t \mathrm{d}\tau\,\frac{1}{\sqrt{4\pi\gamma\kappa(t-\tau)}} \\ &\times\left[\exp\left(-\frac{\gamma(s-\sigma)^2}{4\kappa(t-\tau)}\right)-\exp\left(-\frac{\gamma(s-2L+\sigma)^2}{4\kappa(t-\tau)}\right)\right] \\ &\times\left(\xi(\sigma,\tau)+F(\sigma,\tau)\right)\ , \end{aligned} \tag{2.129}$$

which then for $F(s,t) = F(t)\delta(s)$ gives the analog of eq. (2.111):

$$\langle x(0,t)\rangle = \int_0^t \mathrm{d}\tau \, \frac{F(\tau)}{\sqrt{4\pi\gamma\kappa(t-\tau)}} \left[1 - \exp\left(-\frac{\gamma L^2}{\kappa(t-\tau)}\right)\right] . \tag{2.130}$$

Now following the same discretization scheme as before, we write

$$M_{ij} = \Theta(t_i - t_j) \int_{t_{j-1}}^{t_j} \frac{\mathrm{d}\tau}{2\sqrt{t_i - \tau}} \left[1 - \exp\left(-\frac{L^2}{\kappa(t_i - \tau)}\right)\right] \tag{2.131}$$

$$= \Theta(t_i - t_j)\left(\sqrt{t_i - t_{j-1}}\left[1 - \exp\left(-\frac{\pi^2\tau_L}{t_i - t_{j-1}}\right)\right] + \pi^{\frac{3}{2}}\sqrt{\tau_L}\ \mathrm{erfc}\sqrt{\frac{\pi^2\tau_L}{t_i - t_{j-1}}}\right., \tag{2.132}$$

$$\left. - \sqrt{t_i - t_j}\left[1 - \exp\left(-\frac{\pi^2\tau_L}{t_i - t_j}\right)\right] + \pi^{\frac{3}{2}}\sqrt{\tau_L}\ \mathrm{erfc}\sqrt{\frac{\pi^2\tau_L}{t_i - t_j}}\right)$$

where Θ is the Heaviside function (with the convention that $\Theta(0) = 1$) and we introduced $\tau_L \equiv \frac{L^2}{\pi^2\kappa}$ which is the Rouse time of a chain of length L. With this modified expression for the matrix M, we can then proceed to run the inference as outlined before (sections 2.4.2 and 2.4.3). Note that for an infinitely long tether—$\tau_L \to \infty$—eq. (2.132) reproduces M in eq. (2.115).

Dragging along surrounding chromatin

The Rouse model assumes that the pulled locus does not interact in any way with other parts of the chromosome (aside from the backbone connectivity) or other chromosomes in the nucleus. This is the phantom chain assumption; it is obviously a serious simplification. Here we will consider three ways that the pulled locus could be interacting with the surrounding chromatin, and how we can take that into account in the inference (fig. 2.5)

The idea behind the first type of models is that when moving, the locus has to push the surrounding chromatin out of the way, which will accumulate in front of the locus as shown in fig. 2.5, **A**. This is reminiscent of a baseball being caught in a catcher's glove, which is why we refer to these models as "glove models". In these models the locus is held back by the "glove" of accumulated chromatin in front of it, but is completely free to move backwards out of it (and will do so upon force release); the glove itself will then relax according to Rouse dynamics. This completely free recoil might appear reasonable, if we assume that the chromatin in the locus' path has already been pushed out of the way on the pull. If this is not the case, or the locus upon recoil takes a different path, we should expect a similar glove to build up behind the locus, as depicted in fig. 2.5, **B**.

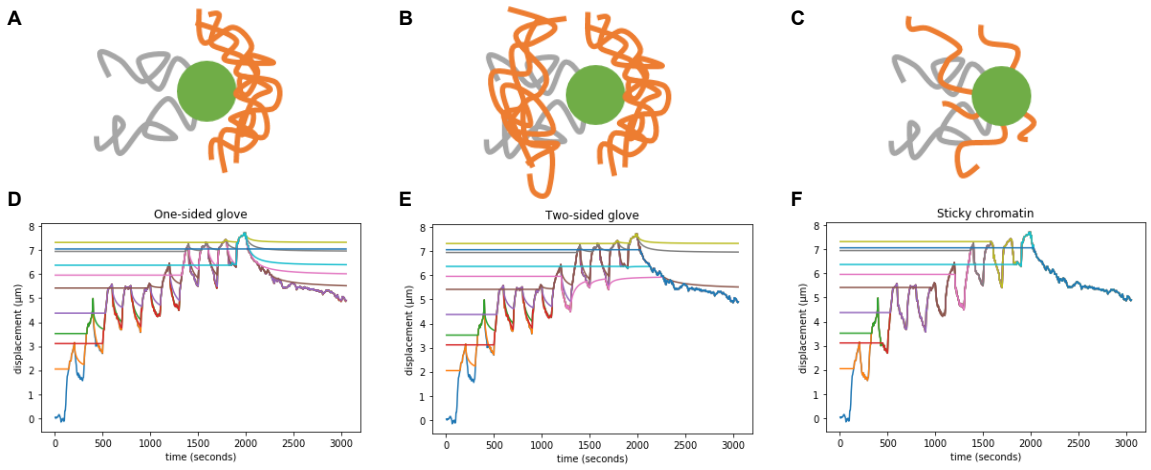


Figure 2.5: Three models for dragging surrounding chromatin. (**A–C**) Cartoons representing the one-sided glove, two-sided glove, and sticky model, respectively. (**D–F**) Corresponding illustrations of virtual particles attaching to the locus at various time points. A few virtual particles were chosen for representation; when running the actual inference we attach a virtual particle at every time point.

A different type of model assumes that chromatin has strong non-specific interactions, meaning it will stick to the locus as it moves through the nucleus (fig. 2.5, **C**). In this case we will simply assume that at each point in time, some of the surrounding chromatin attaches to the locus and then moves with it.

The additional restoring forces generated by all of these models are then calculated using the same inference scheme as for the main locus itself. For every point (x_i, t_i) in the given trajectory we generate a trajectory for a virtual particle that attaches to the locus at that point in time (fig. 2.5, **D–F**). Consequently, for $t < t_i$ this virtual particle is at rest at x_i, while for $t > t_i$ it follows the motion of the locus. How exactly this works depends on the model we use:

- for the sticky chromatin model (fig. 2.5, **F**), the trajectory for $t > t_i$ is simply exactly the one of the locus.[3]

- for the one-sided glove model (fig. 2.5, **D**), the virtual particle only stays attached as long as the locus is moving forward. As soon as it starts recoiling, we calculate the relaxation of the virtual particle according to the Rouse model. This relaxation proceeds until the virtual particle comes back into contact with the locus, at which point the procedure starts again. This ensures that the virtual particle always stays ahead of the locus, and relaxes according to Rouse dynamics if it loses contact.

[3]One could also imagine a finite lifetime for this stickiness, or a critical force that would break it, etc. For simplicity, we restrict ourselves to the basic model described.

- the two-sided glove model (fig. 2.5, **E**) works similarly to the one-sided one, except that particles that attach when the locus moves backwards will stay behind instead of ahead of it. Apart from that the procedure for obtaining their trajectories is exactly the same.

Finally, we use the force inference on these trajectories to infer the additional restoring forces exerted by the dragged chromatin. We weight the individual contributions by the local chromatin density at the point of attachment and introduce an overall prefactor that is used to adjust the overall strength of attachment.

Chapter 3

Live-cell micromanipulation of a genomic locus reveals interphase chromatin mechanics

This chapter was co-authored by Veer I. P. Keizer, Myself, Maxime Woringer, Laura Zambon, Koceila Aizel, Maud Bongaerts, Fanny Delille, Lorena Kolar-Znika, Vittore F. Scolari, Sebastian Hoffmann, Edward J. Banigan, Leonid A. Mirny, Maxime Dahan, Daniele Fachinetti*, Antoine Coulon*.

It was published in *Science* **377**, 2022, DOI:10.1126/science.abi9810.

*senior authors

3.1 Abstract

Our understanding of the physical principles organizing the genome in the nucleus is limited by the lack of tools to directly exert and measure forces on interphase chromosomes in vivo and probe their material nature. Here, we present a novel approach to actively manipulate a genomic locus using controlled magnetic forces inside the nucleus of a living human cell. We observe viscoelastic displacements over microns within minutes in response to near-picoNewton forces, which are well captured by a Rouse polymer model. Our results highlight the fluidity of chromatin, with a moderate contribution of the surrounding material, revealing the minor role of crosslinks and topological effects and challenging the view that interphase chromatin is a gel-like material. Our new technology opens avenues for future research, from chromosome mechanics to genome functions.

3.2 Introduction

Recent progress in observing and perturbing chromosome conformation has led to an unprecedented understanding of the physical principles at play in shaping the genome in 4 dimensions (4D) [29]. From genomic loops and topologically associating domains (TADs) to spatially segregated A/B compartments and chromosome territories, the different levels of organization of the eukaryotic genome are thought to arise from various physical phenomena, including phase separation [30–32], ATP-dependent motors [32, 33], and polymer topological effects [34]. Nonetheless, the physical nature of chromatin and chromosomes inside the nucleus and its functional implications for mechanotransduction remain an active area of investigation [35, 36]. Observation-based studies assessing the mobility of the genome in living cells, from single loci [37–39] and small regions [40] to large domains [16], underline the possible existence of different material states of chromatin (liquid, solid, gel-like). Extra-nuclear mechanical perturbations, including whole-nucleus stretching [41, 42], micro-pipette aspiration [43], and application of local pressures [43, 44] or torques [45] onto a cell, all affect the overall geometry of the nucleus and reveal global viscoelastic properties. Intra-nuclear mechanical manipulation of the genome, on the other hand, is rare and technically challenging [36]. Viscoelasticity measurements using a microinjected 1 µm bead suggested that interphase chromatin may be a crosslinked polymer network (i.e. gel) [46]. Recently, intra-nuclear mechanics was elegantly probed by monitoring the fusion of both artificial [47] and naturally occurring [48] droplet-like structures. Active mechan-

ical manipulation of an intra-nuclear structure was recently achieved using an optical tweezer to displace a whole nucleolus in oocytes [49] and using optically induced thermophoretic flows within prophase [50] or interphase nuclei [51]. However, these approaches are limited to the manipulation of large structures or do not apply forces directly on chromatin. These limitations have made it difficult to disentangle various effects (mechanical response of the nucleus vs. chromatin itself; hydrodynamics vs. polymer viscoelasticity), leading to contradictory results. Hence, an approach for the direct and active mechanical manipulation of specific genomic loci inside living cells is needed. To meet this need, we developed a technique for targeted micromanipulation of a specific genomic locus in the nucleus of a living cell, allowing us to probe the physical properties of an interphase chromosome by measuring its response to a controlled point force.

3.3 Results

Mechanical manipulation of a genomic locus in a living cell

Our approach relies on tethering magnetic nanoparticles (MNPs) to a genomic locus and applying an external magnetic field (fig. 3.1, **A**). We chose ferritin MNPs for their small size [53, 54]: 12 nm in diameter for ferritin (PDB 1GWG), 28 nm for the full MNP [54]. We produced ferritin MNPs by synthesizing in vitro recombinant eGFP-labeled ferritin cages and loading them with a magnetic core (see Methods). We microinjected MNPs into the nuclei of living human U-2 OS cells previously engineered to contain an artificial array inserted at a single genomic location in a subtelomeric region of chromosome 1 (band 1p36) [52]. This genomic array contains 200 copies of a 20 kb genetic construct, each including 96 tetO binding sites and a transgene. It has been extensively used in the past to study the function of several chromatin modifications, RNA polymerase II (Pol II) recruitment, and RNA synthesis during induction of the transgene [52, 55, 56]. Hence, although we used it here uninduced, this array can recapitulate functional chromatin-based processes, such as transcriptional activation. MNPs were targeted to the array using a constitutively expressed fusion protein (TetR, mCherry and anti-GFP nanobody) serving as a tether (fig. 3.1, **A**). Upon injection, MNPs diffused through the nucleus and accumulated at the array, forming a fluorescent spot in both eGFP and mCherry channels. Quantification of the fluorescence signals indicated that MNPs were at nanomolar concentrations in the nucleus following injection and accumulated at the genomic locus in the range of hundreds to thousands of

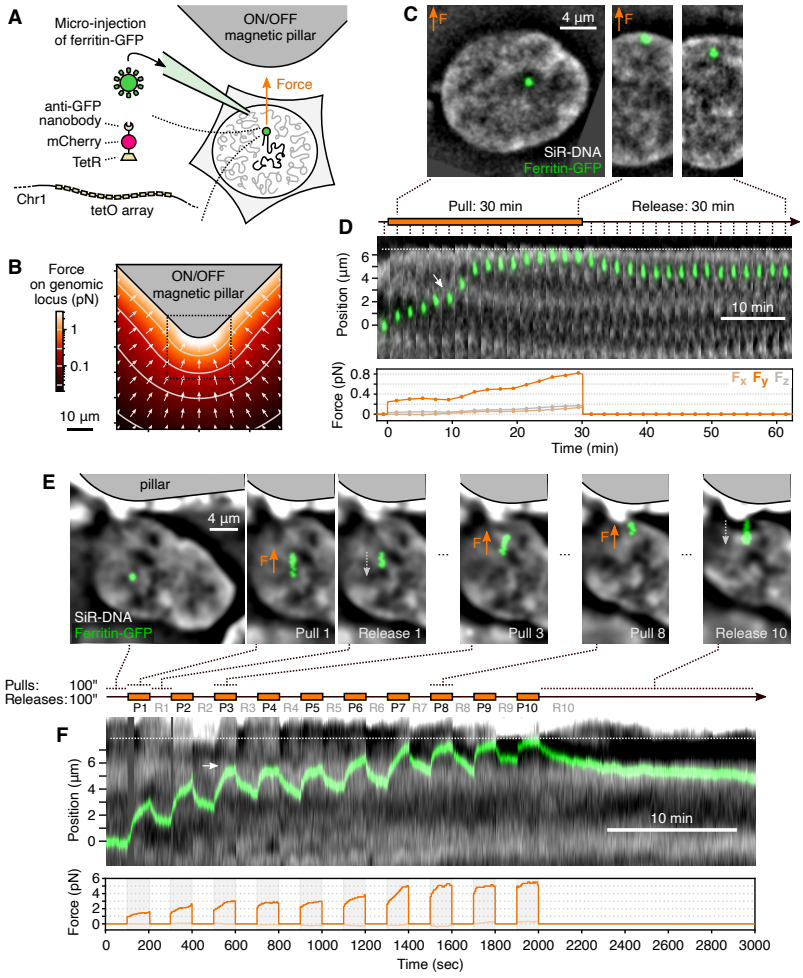


Figure 3.1: Mechanical micro-manipulation of a genomic locus in living cells. (**A**) Magnetic nanoparticles (MNPs) of GFP-labeled ferritin are microinjected into the cell nucleus and targeted to a genomic array containing ~19,000 tetO binding sites [52] with a linker protein. Cells are imaged on a coverslide with microfabricated magnetic pillars that produce a local magnetic field and attract the genomic locus. (**B**) The force exerted onto the locus depends on its position relative to the pillar and is characterized using a pre-calculated force map (see Methods), here shown for 1000 MNPs at the locus. (**C**) Example of a pull–release experiment showing the locus being displaced during the 30 min of force exertion and recoiling during the 30 min of force release (30'-PR scheme). (**D**) Kymograph of the same experiment showing each time frame, along with the force time profile calculated using the force map. (**E**) Experiment where pulls and releases are 100 s and the pull–release cycle is repeated 10× (100"-PR scheme). Images are time projections, i.e. showing in green all the positions of the center of mass of the locus over the periods represented on the timeline. The arrows indicate the direction of the motion. (**F**) Kymograph of the same experiment, showing the displacement of the locus and the spatial patterns of DNA density in the nucleus, along with the force time profile. All SiR-DNA images are band-passed (see Methods). On (D, F), dotted lines: nuclear periphery, white arrows: feature of interest in the spatial distribution of DNA density.

MNPs (median 1500 MNPs). The locus should be regarded as a condensed and heterochromatic 4 Mb region (1.6% of chromosome 1) residing in a euchromatic genomic context, as previously reported [52], with small MNPs (each being ~2-3 times [54] the size of a nucleosome) sparsely decorating chromatin (1 MNP per ~2.7 kb). Consistently, we observed that the locus typically resided in low to intermediate DNA density regions and is itself relatively condensed and that binding of MNPs to the locus did not substantially affect its morphology. Microinjection and attraction of unbound MNPs did not substantially alter chromatin distribution and densities inside the nucleus. Cells were imaged on a coverglass with custom-made microfabricated pillars [57, 58], which behave as local magnets only when subjected to an external magnetizing field. Hence, ON/OFF modulation of the local force field could be achieved while imaging by placing/removing an external magnet on the microscope stage. The shape and orientation of the pillars were chosen to maximize the magnetic field gradient and hence the force. We performed magnetic simulations and experimental calibrations using two independent methods to determine the magnitude and orientation of the force applied onto the genomic locus, as a function of the number of MNPs bound to it and its position relative to the magnetic pillar (fig. 3.1, **B**). The typical forces applied onto the locus were in the sub-picoNewton (pN) range, occasionally reaching a few pN (median force = 0.45 pN). These values are in the range of forces exerted by molecular motors in the nucleus, e.g. comparable to the stalling force of ~0.5 pN for the structural maintenance of chromosomes (SMC) complex condensin [10] and a few pN for RNA polymerase II (Pol II) [59].

Force induced movement of a genomic locus reveals viscoelastic properties of chromatin

We first applied the magnetic force for 30 min and released it for another 30 min, while performing low-illumination 3D imaging with a 2 min interval (30'-PR scheme). We observed a clear motion of the locus toward the magnet upon application of the force and a slow and partial recoil when the force stopped (fig. 3.1, **C–D**). This indicates that a sub-pN force, when applied in a sustained and unidirectional manner on a genomic locus, elicits a displacement of that locus by several microns over minutes. It also shows that the chromosomal locus can move across the nuclear environment, which is believed to be crowded and entangled. We also applied the force periodically—pulling for 100 s, releasing for 100 s and repeating this cycle 10 times (100"-PR scheme)—while performing fast 2D imaging with a 5 s interval (fig. 3.1, **E–F**). Several observations from these two experiments hinted at the material properties of chromatin. First,

the trajectories showed recoils during release periods and a gradual slow-down during both pulls and releases, characteristic of a viscoelastic material. Second, spatial heterogeneities in the trajectories were visible and appeared to relate to the spatial distribution of DNA density (fig. 3.1, **D,F**; white arrows; the motion of the locus was hindered where the DNA density varied). Third, recoil after force release was seen even after collision with the nuclear periphery, indicating that the material there (peripheral heterochromatin, nuclear lamina) was not sticky enough to fully retain the locus. Fourth, the spatial distribution of DNA density in the nucleus does not show large-scale deformations, indicating that the locus did not drag along large amounts of material. Together, the force-induced displacements we observed are consistent with viscoelastic and non-confining chromatin and constitute a basis to further develop and test physical models of interphase chromosomes.

Quantitative force response and scaling laws of interphase chromatin mechanics

To quantify viscoelastic properties of chromatin, we analyzed the trajectories of the locus in 35 cells undergoing the 30'-PR scheme (corrected for cell motion and force orientation, see Methods). We observed a range of behaviors in both pulls and releases, regarding initial speed, total distance travelled, and shape of time profiles (fig. 3.2, **A–B**). Most traces showed a displacement that was clearly distinguishable from diffusion (fig. 3.2, **A–B**, hatched areas; see Methods). Collision with the nuclear periphery (open symbols on fig. 3.2, **A–B**) happened in 9 out of 35 traces and hence the total displacement during the pull is most often not limited by the nuclear periphery. The initial force applied onto the locus largely predicted the variability seen in the initial motion (fig. 3.2, **C**). The recoil motion after force release was in part predicted by the total distance over which the locus had been displaced during the pull (fig. 3.2, **D**), with a simple linear relationship highlighting the elastic nature of chromatin. Deviations from these simple proportionality relationships indicate that the specific nuclear context or the state of the genomic locus might influence its response. In particular, we observed that when the locus moved slower than expected, it was less DNA dense, and when the locus moved faster than expected, it was more DNA dense, suggesting that the compaction state of the locus itself affected its response to the force. Absolute nuclear position of the locus did not correlate with its response to the force, but if the locus reached the periphery during the pull, it often recoiled slower than expected. Despite the variability between traces, double logarithmic plots of all the pulls and releases from the 30'-PR and 100"-PR trajectories, together with 3 additional high-framerate (dt = 0.5")

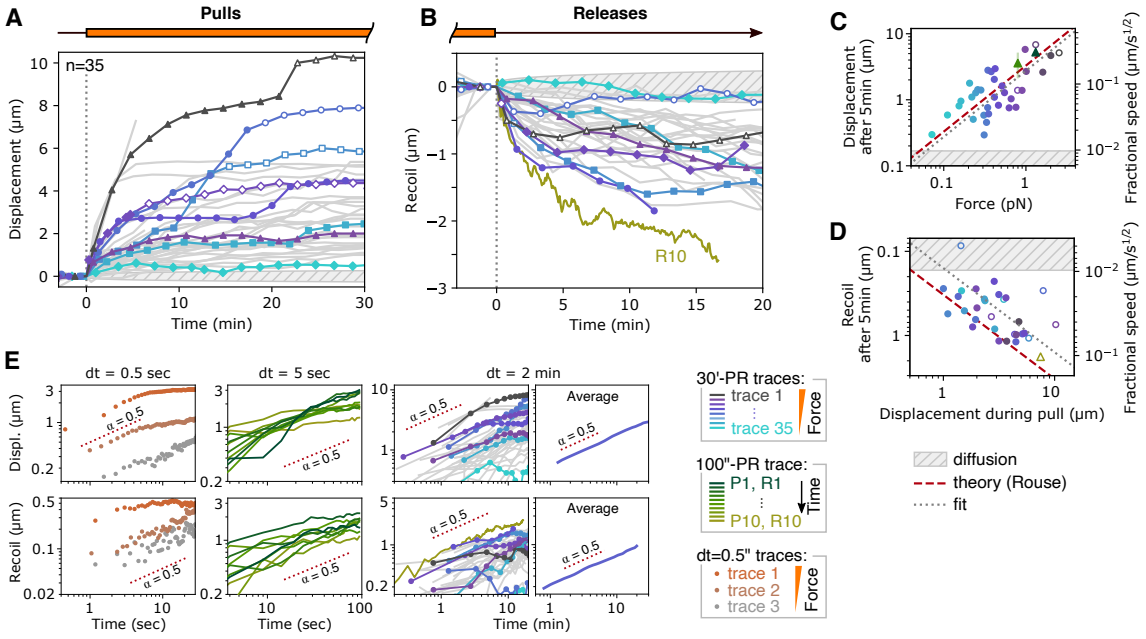


Figure 3.2: Quantitative analysis of locus movement in response to force. (**A**) Trajectories of the genomic locus in the direction of the applied force for 35 different cells during force exertion with the 30'-PR scheme. A selection of trajectories, representative of the breadth of observed behaviors, are highlighted and color-coded by force. Hatched areas in A to D correspond to the null model of pure diffusion based on MSD measurement (see Methods). (**B**) Recoil trajectories relative to the time and position at the start of the release are shown for the same loci as in A. Curve R10 is the last release of the 100"-PR trajectory. (**C**) Displacements measured at $\Delta t = 5\,\text{min}$ of force exertion on all the traces from A, plotted against the magnitude of the force. Coordinates are interpolated between the frames before/after Δt. The green line/triangle correspond to the envelope of the 100"-PR trajectory. Reported forces are the average over Δt. Displacements are also expressed in $\mu\text{m/s}^{0.5}$ (right axis), allowing us to place pull P1 from the 100"-PR trace, measured at $\Delta t = 100\,\text{s}$ (dark green triangle). The red line indicates the expected relationship from Rouse theory, solely based on an MSD measurement. (**D**) Recoil after $\Delta t = 5\,\text{min}$ of force release on all the traces from B and the last release of the 100"-PR trajectory (R10, green triangle), plotted against the total displacement during the pull. The red line indicates the expected relationship from Rouse theory. Open symbols on A–D indicate when loci are within 1.5 µm of the nuclear periphery (at the moment of measurement on A–C, at the moment of force release on D). (**E**) Displacement and recoil trajectories, aligned on the time and position at the moment of force switching, are represented as double logarithmic plots for pull–release experiments imaged with different frame intervals: $\Delta t = 0.5\,\text{s}$ (see Methods), $\Delta t = 5\,\text{s}$ (from the 100"-PR trace; fig. 3.1, **F**), and $\Delta t = 2\,\text{min}$ (all 30'-PR traces; panels A and B). For the latter, average trajectories (right plots) were calculated over all displacements where the applied force remains $\leq 2\,\text{pN}$ (28/35 traces) and over all the recoils after a displacement of $\leq 5\,\mu\text{m}$ (21/35 traces). Red dotted lines indicate the power-law behavior, with exponent 0.5, predicted by Rouse theory.

pull–release trajectories, revealed linear portions in the curves with a slope of 0.5, over more than three orders of magnitude in time (fig. 3.2, **E**). This behavior suggests that the different levels of the hierarchical genome organization are not characterized by vastly distinct mechanical properties. In addition, displacements that scale with time as $t^{0.5}$ can be empirically described by a 'fractional speed', i.e., a single value in $\mu m/s^{0.5}$ capturing how the motion evolves over time (fig. 3.2, **C–D**). The first pull of the 100"-PR trace, represented in this unit, indeed follows the same relationship as the 30'-PR traces (dark green triangle on fig. 3.2, **C**) and the slope of the resulting force–displacement plot yields a unique factor of $0.158 \pm 0.014\,\mu m/s^{0.5}/pN$, characterizing the dynamic response of chromatin to force. Together, these results indicate that a large part of the response of chromatin to force can be described by simple laws.

Chromatin force response is well described by a free polymer model (Rouse chain)

We then sought a model of chromatin that best explains our quantitative measurement of force-induced locus displacement. Several features in our data suggest a classical polymer model known as a Rouse polymer [18] as a first approximation to describe the response of chromatin to forces. The Rouse model represents a polymer in which each monomer diffuses by thermal motion in a viscous medium and is connected to its two neighbors by elastic bonds. Importantly, Rouse ignores steric effects (contact, hindrance), crosslinks (affinity, stickiness), and topological effects (fibers can pass through each other). This model is frequently invoked for chromatin dynamics since it predicts the characteristic power-law scaling—i.e. a linear relationship on a log–log plot—of the mean squared displacement (MSD) vs. time, with exponent 0.5, as observed here and for other genomic loci in eukaryotes [60–62]. We extended Rouse theory to study how a polymer responds to a point force (section 2.4). Our calculations predict a power-law behavior with exponent 0.5 for displacements and recoils in response to force, consistent with our experimental observations (fig. 3.2, **E**). These two power laws have the same physical origin, so the diffusion coefficient obtained independently from the MSD ($1627 \pm 19\,nm^2/s^{0.5}$) directly relates to—and predicts—the slope of the force–displacement plot (fig. 3.2, **C**, red line), i.e. $1627\,nm^2/s^{0.5} \,/\, 2k_BT = 0.190 \pm 0.003\,\mu m/s^{0.5}/pN$. This agreement between two independent passive and active measurements (diffusion and force response, i.e. red vs. gray lines on fig. 3.2, **C**) supports the Rouse model to explain our chromatin dynamics data. Inspected on a cell-by-cell basis, the force-free MSD of the locus before and after the pull–release experiments appears very moderately reduced in most cases. Its natural variability between cells does not

appear to explain the variability of the response to force. After force release, the Rouse model also predicts a recoil proportional to the total displacement during the pull. However, in many cases, the locus recoiled somewhat slower than predicted by the Rouse theory. Instead, the theoretical prediction appears to define an upper bound for the recoils (fig. 3.2, **D**, red line) and deviations from Rouse theory are more pronounced at the nuclear periphery. Together, this analysis suggests that the dynamic response of the chromosome to the force can be described by the Rouse polymer model, with additional effects from the nuclear environment.

Model-based trajectory analysis reveals moderate hindrance by surrounding chromatin

To further understand the physical nature of chromatin, we asked how alternative polymer models are able to capture the 100"-PR trace (fig. 3.3, **A–D**). The approach is to use the displacement trajectory and infer, assuming a given polymer model, the time profile of the force that produced the measured trajectory (fig. 2.4). Disagreement between predicted and actual force profiles indicates when models are incorrect or incomplete, allowing one to select and refine the best model(s). With this approach, we compared a series of models (fig. 2.4, **C**). First, a simple Rouse model without any adjustable parameters (i.e. calibrated using the MSD vs. time plot) predicts well the first pull and all the release periods (fig. 3.3, **B**). However, the prediction leaves some of the applied force unexplained (gray area between curves, fig. 3.3, **B**), suggesting a missing component in the model that would additionally slow down or hinder the progression of the locus. This residual unexplained force did not scale with speed and hence could not be explained as an additional viscous drag on the locus. Instead, it increased progressively across successive pulls, suggesting an accumulation of hindrance as the locus moved through the nucleus. To represent this, we added a capacity for the locus to interact with the surrounding chromatin, represented as extra Rouse chains that are either attached to or pushed by the locus along its path (fig. 3.3, **C** and fig. 2.4, **C**). These models better predicted the force profile throughout the trajectory compared to a pure Rouse model. The only free parameter in fig. 3.3, **C** is the frequency at which the locus interacts with other polymers, which we found very low (fig. 2.4, **C**), indicating that the interaction with the surrounding chromatin was moderate. This is also consistent with the small but detectable reduction in mobility of the locus before and after pull–release experiments and the subtle redistribution of DNA densities around the pulled locus. Taken together, these modeling results suggest that, upon force application and release on our genomic locus, chromatin is well described as a Rouse polymer—i.e. a free polymer in

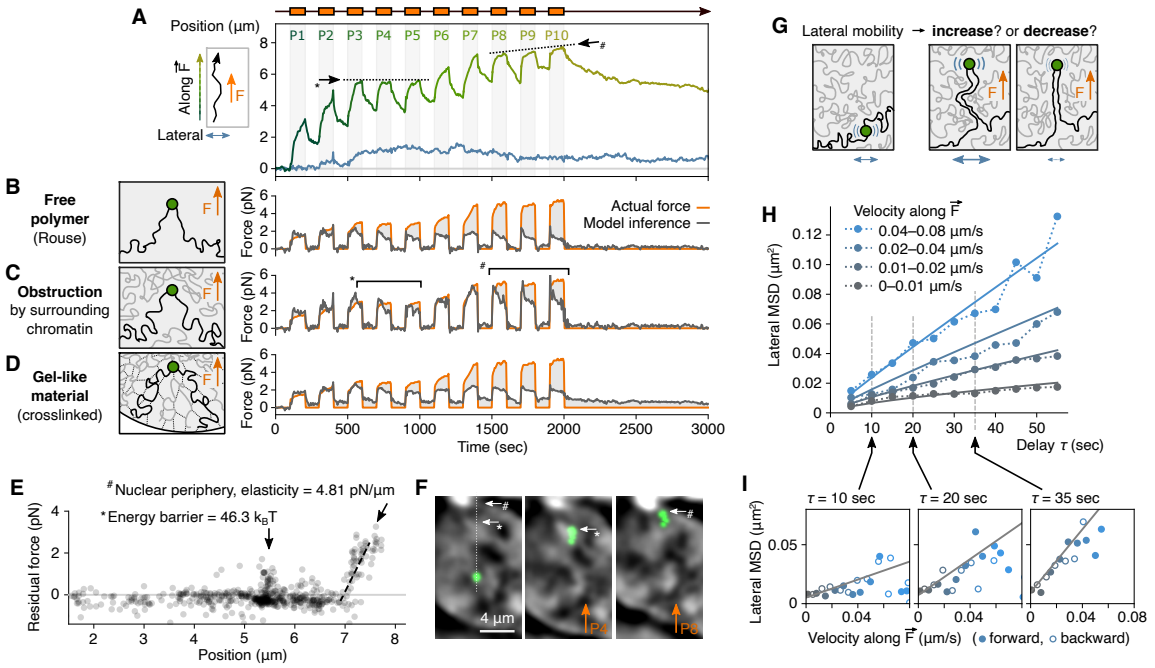


Figure 3.3: Model-based analysis and hypothesis testing. (**A**) Trajectory of the locus shown in the direction of the force (green curve) and orthogonal to the force in the imaging plane (blue curve) for the 100"-PR experiment. Arrows indicate apparent obstacle (see also * and # in panels C, E and F). (**B–D**) Evaluation of different models in their capacity to reproduce the experimentally measured force time profile (orange curve) by inferring it from the trajectory (gray curve). Models shown here are (**B**) a simple Rouse polymer [18], (**C**) the same model with extra polymer chains being pushed by the locus to represent the surrounding chromatin, and (**D**) a gel-like material, represented as a Rouse polymer in a viscoelastic environment. See full list in fig. 2.4. (**E**) The residual unexplained force from the second model (area between curves in C) is plotted along the trajectory of the locus, highlighting an obstacle (*) and an elastic region near the nuclear periphery (#) for which physical parameters are measured and which are visible in panels A, C and F. (**F**) Time projection images, respectively before the first pull and during pulls P4 and P8, showing how the spatial distribution of DNA density in the nucleus relates to the identified obstacles. SiR-DNA images are band-passed (see Methods). (**G**) Hypotheses on how the lateral mobility of the locus may change depending on its force-induced displacement. (**H–I**) Mean square displacement (MSD) of the lateral movement of the locus, calculated as a function of both time delay and velocity in the direction of the force. Solid lines on both the MSD–delay (H) and the MSD–velocity (I) representations correspond to a single-parameter fit describing how lateral mobility increase with velocity in the direction of the force.

a viscous environment—with moderate interactions from the surrounding chromatin, indicating that hindrance, crosslinks, and topological effects play a minor role.

Interphase chromatin does not behave as a gel in force-response experiments

Interphase chromatin was proposed to be a gel-like material [39, 40, 46]. A gel is a highly crosslinked polymer, i.e. unlike a linear polymer where monomers are linked to two neighbors, extra links between non-adjacent monomers form an interconnected mesh, giving the gel solid-like properties. For chromatin, this could in principle arise from affinity between nucleosomes, as well as loops/bridges formed by proteins/complexes/condensates and topological entanglement between chromatin fibers. First, in such an interconnected mesh structure, short paths effectively linking the pulled locus to all other loci in the nucleus would result in long-range deformation of the spatial pattern of DNA density, which we do not observe (fig. 3.1, **D,F**). Second, if the chromatin surrounding the locus were gel-like, it would effectively act as a viscoelastic medium. This assumption does not recapitulate well the experimental data (even with two free parameters, fig. 3.3, **D** and fig. 2.4, **C**) and is inconsistent with the observed scaling of 0.5 in the MSD, which argues for a simply viscous and non-elastic medium. Finally, if the locus was part of an interconnected mesh, short series of links would tether it to large structures (e.g. periphery, nucleoli). A Rouse model that includes a finite tether does not recapitulate the experimental data (fig. 2.4, **C**) and is inconsistent with the linear behavior observed in fig. 3.2, **E** up to several microns. These results again suggest minor effects of crosslinks and topological constraints and argue against the view that interphase chromatin behaves like a gel at the spatial and temporal scale of our observations.

Heterogeneities in the trajectory reveal obstacles in the nuclear interior and a soft elastic material at the nuclear periphery

Even the models that best capture the data leave part of the force unexplained (fig. 3.3, **C**, gray area). We thus plotted this residual unexplained force as a function of spatial position (fig. 3.3, **E**). This revealed an accumulation of non-null residual forces at specific locations, matching visible features in the trajectory and in spatial distribution of DNA density in the nucleus. First, the residual force in pulls P3 to P5 corresponds to an apparent obstacle in the trajectory (* on fig. 3.3, **A,C**) occurring at a high-to-low transition of DNA density (fig. 3.3, **F** and fig. 3.1, **F**). It appears as a spatially defined barrier of residual force (fig. 3.3, **E**), requiring

an energy of $\sim 46k_{\mathrm{B}}T$ to overcome. This suggests that, while DNA dense regions are not obstacles per se, the interface between high- and low-density regions may constitute a barrier. The energy we estimated suggests that such barriers may be overcome by ATP-dependent molecular motors [10,59] but unlikely by spontaneous thermal fluctuations. Second, the residual force in pulls P8 to P10 (# on fig. 3.3, **A,C**) corresponds to the collision with structures near the nuclear periphery (fig. 3.1, **F** and # on fig. 3.3, **F**). The observed linear force–distance relationship (fig. 3.3, **E**) indicates a solid-like elastic behavior for these structures, over at least 600 nm and with a spring constant of 4.81 pN/µm. This is much softer than what was measured by whole-nucleus stretching experiments [41, 42], which could be explained by the small size of the locus and/or the existence of a soft layer of elastic peripheral components (heterochromatin, nuclear lamina) rather than the material directly contributing to the structural rigidity of the nucleus.

Lateral mobility of the locus reflects transient collisions with obstacles in the nucleoplasm

To further investigate the material encountered by the locus, we analyzed the lateral motion of the locus as it was pulled and released (fig. 3.3, **A**, blue curve). We hypothesized that, on one hand, collisions with obstacles could increase lateral mobility or, on the other hand, the locus being dragged into a more constraining and entangled environment could result in a reduction of its mobility (fig. 3.3, **G**). After computing the MSD of the lateral motion as a function of both time delay t and velocity v_y along the direction of the force (fig. 3.3, **H–I**), we observed a clear increase of lateral mobility when the locus moved (for both forward and backward movements; fig. 3.3, **I**), suggesting the existence of obstacles that deflected the motion. This additional mobility in the MSD is captured by a term proportional to v_y, as expected for collisions, and proportional to t (not $t^{0.5}$), as expected if the force due to the collision with obstacles persists in the same direction across several frames, indicating the existence of large obstacles. Indeed, in P3 for instance, the lateral motion clearly shows a directional behavior (fig. 3.1, **E** and fig. 3.3, **A**). However, the relationships we observed on fig. 3.3, **H–I** held even when excluding all the timepoints before P4, indicating that the collision with obstacles was widespread throughout the nucleus. These results, together with our observation that very few chromatin fibers appeared to be carried along with the locus, indicate that obstacles are frequently encountered by the locus, but most interactions are weak and transient.

3.4 Discussion

Our measurements of how a genomic locus inside the nucleus of a living cell responds to a point force indicates that interphase chromatin has fluid-like properties and behaves as a free polymer. This contrasts with previous studies depicting chromatin as a stiff, crosslinked polymer gel with solid-like properties [39, 40, 46]. Our observation that near-pN forces can easily move a genomic locus across the nucleus over a few minutes (fig. 3.1, **D,F**) also contrasts with a previous study reporting confined sub-micron displacements over seconds upon application of 65 to 110 pN forces to a 1 µm bead [46]. We propose that our results may be reconciled with previous experiments in several ways. First, unlike a micron-size bead, the locus in our experiments is small and may be deformable enough to pass through the surrounding chromatin. Second, chromatin may contain many small, gel-like patches, embedded in a structure with liquid, Rouse-like properties at a larger scale. This is also in line with our observation that the transiting locus frequently encounters obstacles. Third, chromatin may be a weak gel, i.e. with short-lived crosslinks [39]. Such a gel could continuously maintain a stiff, globally connected network that resists stresses over large length scales, while permitting fluid-like motions at smaller scales. Future experiments perturbing chromatin state and chromatin associated proteins will be important to reconcile observed micro- and mesoscale mechanics.

Organization of chromosomes that allows movement of genomic loci across large distances by weak forces could have implications for a range of genome functions. Large-scale movements of chromosomes occur during nuclear inversion in rod cell differentiation for nocturnal mammals [63]. Specific genes undergo long-range directional motion upon transcriptional activation [64, 65]. Long and highly transcribed genes can form ~5 µm giant loops, believed to be due to chromosome fiber stiffening [66]. Certain double strand break sites undergo large-scale, nuclear F-actin dependent relocation to the nuclear periphery [67]. These DNA-based biological processes require a nuclear organization in which such movements are possible. Our results reveal mechanical properties of chromatin where such large-scale movements would only require weak (near pN) forces. Although sustained unidirectional forces are unlikely to occur naturally in the nucleus, the magnitude of the forces and the timescale of force exertion in our experiments are comparable to those of molecular motors like SMC complexes and Pol II—i.e. in the sub-pN [10] or low-pN [59] ranges and applied over minutes (e.g. 10 min for Pol II to elongate through a 25 kb gene, 5-30 min for SMC complexes). Hence, some molecular motors

in the nucleus operate in a force range that is sufficient to substantially reorganize the genome in space.

Future work will be important to expand and complement our results. Although the genomic array we used here is known to be chromatinized and has been used extensively to recapitulate and study functional chromatin-based processes [52,55,56], we cannot exclude that its repetitive and artificial nature might prevent some of our measurements from being applicable to non-repetitive and native regions. Manipulating loci other than a subtelomeric locus on the longest chromosome (chromosome 1), in other genomic contexts (hetero-/euchromatin), and in different cell types will be important to assess the generalizability of our findings in various biological contexts.

Our approach to mechanically manipulate and relocate genomic loci in the nuclear space opens many avenues for future research, from the study of interphase chromosome mechanics to the perturbation of genome functions, including transcription, replication, DNA damage repair and chromosome segregation. By giving access to physical parameters and revealing fundamental scaling laws to describe chromatin mechanics, our work provides a foundation for future theories of genome organization.

Chapter 4

Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging

This chapter was co-authored by Michele Gabriele[†], Hugo B. Brandão[†], Myself[†], Asmita Jha, Gina M. Dailey, Claudia Cattoglio, Tsung-Han S. Hsieh, Leonid Mirny*, Christoph Zechner*, and Anders S. Hansen*.

It was published in *Science* **376**, 2022, DOI:10.1126/science.abn6583.

[†]equal contribution
*senior authors

4.1 Abstract

Animal genomes are folded into loops and topologically associating domains (TADs) by CTCF and loop extruding cohesins, but the live dynamics of loop formation and stability remain unknown. Here, we directly visualize chromatin looping at the *Fbn2* TAD in mouse embryonic stem cells using super-resolution live-cell imaging and quantify looping dynamics by Bayesian inference. Surprisingly, the *Fbn2* loop is both rare and dynamic, with a looped fraction of ~3-6.5% and a median loop lifetime of ~10-30 minutes. Our results establish that the *Fbn2* TAD is highly dynamic, where ~92% of the time cohesin-extruded loops exist within the TAD without bridging both CTCF boundaries. This suggests that single CTCF boundaries rather than the fully CTCF–CTCF looped state may be the primary regulators of functional interactions.

4.2 Main Text

Mammalian genomes are folded into loops and domains known as topologically associating domains (TADs) by the proteins CTCF and cohesin [68]. Mechanistically, cohesin is thought to load on DNA and bidirectionally extrude loops until it is blocked by CTCF such that CTCF establishes TAD boundaries [8, 10, 69–71]. Functionally, CTCF- and cohesin-mediated looping and TADs play critical roles in multiple nuclear processes including regulation of gene expression, somatic recombination, and DNA repair [68]. For example, TADs are thought to regulate gene expression by increasing the frequency of enhancer–promoter interactions within a TAD, and decreasing enhancer–promoter interactions between TADs [72]. However, to understand how TADs and loops are formed and maintained, and how they function, it is necessary to understand the dynamics of CTCF/cohesin-mediated loop formation and loop lifetime.

Though recent advances in single-cell genomics and fixed-cell imaging have made it possible to generate static snapshots of 3D genome structures in single cells [15, 73–77], live-cell imaging is required to understand the dynamics of chromatin looping [78]. Furthermore, previous studies have yielded conflicting results as to whether loops are well-defined in single cells [15, 73–77], perhaps due to the difficulty associated with distinguishing *bona fide* CTCF- and cohesin-mediated loops from mere proximity that emerges stochastically [78]. Recent pioneering work has visualized enhancer–promoter interactions [79–81] and long-range V(D)J–chromatin interactions [39] in live cells. However, the dynamics of loop formation and lifetime of CTCF/cohesin loops have not yet been quantified in living cells.

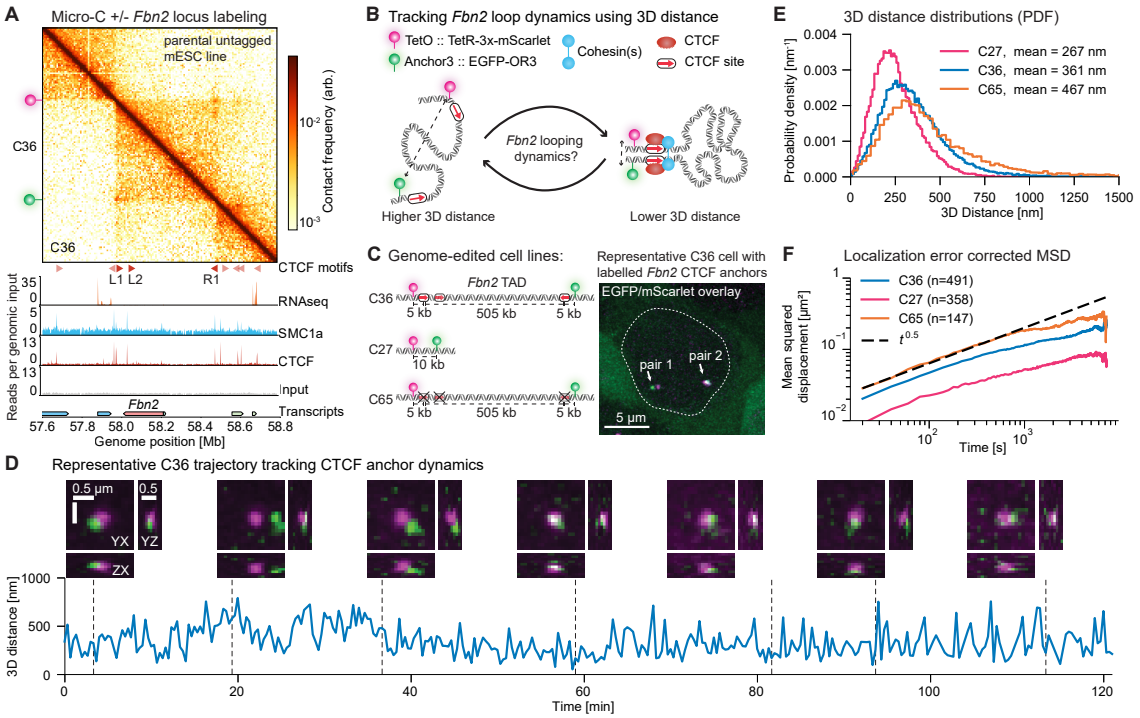


Figure 4.1: Endogenous labeling and tracking of the *Fbn2* loop with super-resolution live cell imaging. (**A**) Fluorescent labeling of *Fbn2* loop anchors does not perturb the *Fbn2* TAD. Micro-C contact map comparing the parental untagged (C59, top left) and tagged (C36, bottom right) cell lines. Red triangles: CTCF motifs with orientation. C36 ChIP-seq shows CTCF (GSM3508478) and cohesin (SMC1A; GSM3508477) binding as compared to Input (GSM3508475). *Fbn2* is not expressed (RNA-seq GSE123636; annotation: GRCm38). Genome coordinates: mm10. (**B**) Overview of tagging and readout using 3D distance. (**C**) Overview of the genome-edited cell lines (left) and a representative maximum intensity projection (MIP) of a cell nucleus showing two pairs of "dots" (right). (**D**) Representative 3D trajectory over time of a dot pair. MIPs of the 3D voxels centered on the mScarlet dot (top) and 3D distances between dots (bottom) are shown. (**E**) 3D distance probability density functions of dot pairs (n=32,171; n=46,163; n=13,566 distance measurements for C27, C36, C65 respectively) (**F**) Localization error corrected 2-point mean squared displacement (MSD) plots (n=358; n=491; n=147 trajectories in C27; C36; C65 respectively).

To visualize the dynamics of CTCF/cohesin looping, we chose as our model system the loop holding together the two CTCF-bound boundaries of the 505 kb *Fbn2* TAD in mouse embryonic stem cells (mESCs). This TAD is verified to be CTCF dependent [82] and relatively simple as it contains a single gene, *Fbn2*, which is not expressed in mESCs (fig. 4.1, **A**). We used genome-editing to homozygously label the left and right CTCF sites of the *Fbn2* TAD with TetO and Anchor3 arrays, which we then visualized by co-expressing the fluorescently tagged binding proteins TetR-3x-mScarlet and EGFP-OR3 [83] (clone C36) (fig. 4.1, **B–D**). We developed a comprehensive image analysis framework, *ConnectTheDots*, to extract trajectories of 3D loop anchor positions from the acquired movies. By optimizing 3D super-resolution live-

cell imaging conditions [78], we could track *Fbn2* looping dynamics at 20 second resolution for over 2 hours (fig. 4.1, **D**). After DNA replication in S/G2 phase, it is no longer possible to reliably distinguish intrachromosomal from sister-chromosomal interactions [78]. We therefore filtered out replicated and low-quality dots using a convolutional neural network. Thus, we only consider G1 and early S-phase cells.

To validate our system for tracking *Fbn2* loop dynamics, we carried out a series of control experiments. First, we confirmed using Micro-C [84, 85] that our locus labeling approach did not measurably perturb the *Fbn2* loop (fig. 4.1, **A**). Second, to 'mimic' the looped state, we deleted the 505 kb between the CTCF sites, generating clone C27 ("ΔTAD"; fig. 4.1, **C**). As expected, this significantly reduced the 3D distance (fig. 4.1, **E**; the non-zero 3D distance distribution for C27 is expected due to localization noise and the 5 kb tether between CTCF sites and fluorescent labels). Third, as a negative control for CTCF-mediated looping, we generated clone C65 ("ΔCTCFsites"; fig. 4.1, **C**) by homozygously deleting the 3 CTCF motifs in the *Fbn2* TAD (L1, L2, R1; fig. 4.1, **A**) and validated that this resulted in loss of CTCF binding and cohesin co-localization by ChIP-Seq. As expected, the 3D distance was significantly increased in C65 (fig. 4.1, **E**). Next, we calculated mean-squared displacements (MSDs) of the relative position of the two loci (2-point MSD), which is unaffected by cell movement. Chromatin dynamics was consistent with Rouse polymer dynamics with a scaling of $\mathrm{MSD} \sim t^{0.5}$ for all three clones [62] (fig. 4.1, **F**). We conclude that our approach faithfully reports on CTCF looping dynamics in live cells without noticeable artifacts.

To elucidate the specific roles of CTCF and cohesin, we endogenously tagged RAD21, CTCF, and the cohesin unloader WAPL with mAID in the C36 line, allowing for degradation with Indole-3-acetic acid (IAA) [86]. For RAD21 and CTCF, we achieved near-complete depletion in 2 hours, long-term depletion led to cell death, Micro-C analysis revealed loss of the *Fbn2* loop or corner peak as expected [87–90] (fig. 4.2, **A**), and ChIP-Seq analysis showed loss of chromatin binding. For WAPL, depletion took 4 hours and was less complete, long-term depletion occasionally yielded visibly compacted "vermicelli" chromosomes [91], and Micro-C analysis revealed increased corner peak strength (27, 28, 30) (fig. 4.2, **A**). All three AID lines exhibited lower protein abundance likely due to leaky protein depletion.

Having validated the AID cell lines, we next performed live-imaging to study the specific roles of RAD21, CTCF, and WAPL in loop extrusion *in vivo*. Consistent with RAD21 being required for loop extrusion, RAD21 depletion strongly increased the 3D distances (fig. 4.2, **B–C**).

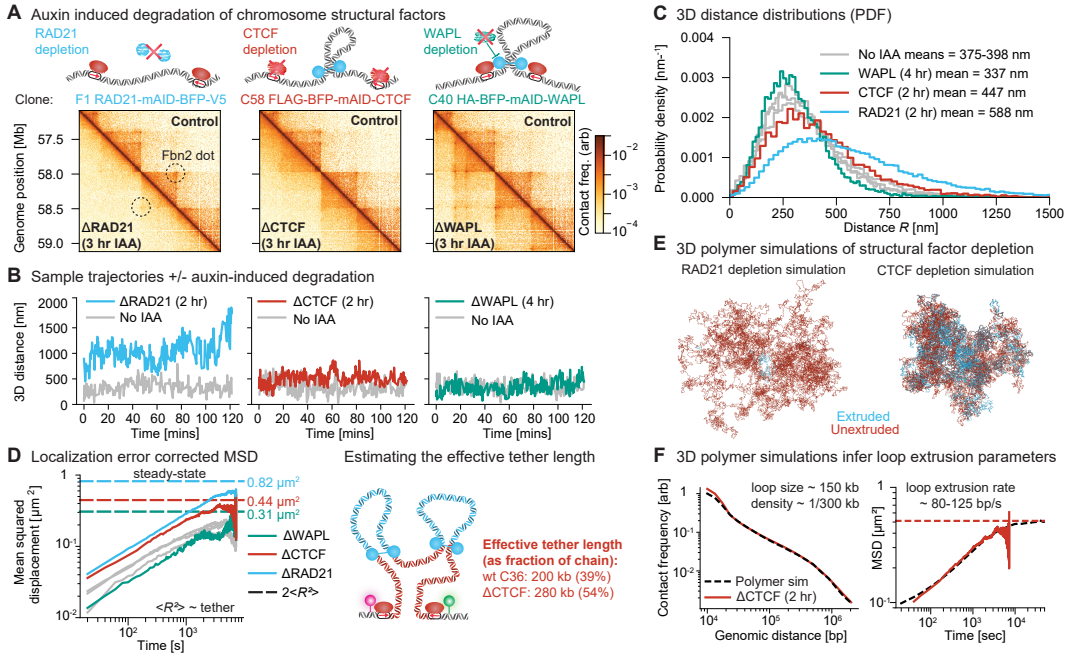


Figure 4.2: Degradation of CTCF, cohesin, and WAPL reveal their role in loop extrusion and looping-mediated chromosome compaction. (**A**) Micro-C data for the AID-tagged clones for RAD21 (left), CTCF (middle), and WAPL (right), showing control data (no IAA treatment; top half) and protein degradation data (3 hours post IAA; bottom half) with schematics illustrating the expected effect. (**B**) Representative trajectories with (colored lines) or without IAA treatment (gray lines) for each AID-tagged clone. (**C**) 3D distance probability density functions of dot pairs (n=45,379; n=10,469; n=18,153 distance measurements for ΔRAD21 (2 hr), ΔCTCF (2 hr), ΔWAPL (4 hr) depletion conditions respectively, and n=17,605; n=11,631; n=21,001 for the same clones without treatment). (**D**) Localization error corrected 2-Point MSD plots for the AID-tagged clones (left) (n=537; n=137; n=215 trajectories in ΔRAD21 (2 hr), ΔCTCF (2 hr), ΔWAPL (4 hr) depletion conditions respectively, and n=183; n=151; n=257 without treatment (gray lines)). The effective tether length is obtained by computing the ratio of the steady-state variance of each clone to the value in the RAD21-depletion condition (note that $2\langle R^2 \rangle$ is also the asymptotic value of the MSD; cf. eq. (2.88)). (**E**) Representative 3D polymer conformation from simulations mimicking the ΔRAD21 (95% cohesin depletion) (left) and ΔCTCF (100% CTCF depletion) (right) depletion conditions. Red: unextruded segment; Blue: extruded segment. (**F**) Matching simulations to the data to obtain loop extrusion parameters (3 fit parameters). The extrusion rate is for two-sided extrusion.

Consistent with CTCF being the boundary factor required for *Fbn2* loop formation (fig. 4.1, **B**), but not required for loop extrusion, CTCF depletion increased 3D distances albeit significantly less than RAD21 depletion [8]. Finally, consistent with prior observations that WAPL depletion increases cohesin residence time and abundance on chromatin [91] potentially allowing it to extrude longer and more stable loops [89, 92], WAPL depletion decreased the 3D distances (fig. 4.2, **B–C**).

To quantify the extent of loop extrusion of the *Fbn2* TAD, we turned to polymer physics theory. The Rouse model predicts a linear relationship between chain length and mean squared distance ($\langle R^2 \rangle$) between the fluorescent labels (dashed lines in fig. 4.2, **D**; see eq. (2.88)). By assuming that ΔRAD21 represents the fully unextruded state with a genomic separation of 515 kb (fig. 4.1, **C**), we can then assign an "effective tether length" (i.e. the unextruded fraction) to each condition. We find an effective tether of ~200 kb in C36 (WT) and ~280 kb in ΔCTCF, corresponding to ~39% and ~54% of the full genomic separation, respectively. By subtraction, the genomic separation between the two labels shortens by ~46% due to extrusion alone (ΔRAD21 vs. ΔCTCF) and ~61% due to extrusion with boundaries (ΔRAD21 vs. C36). This shows that by blocking extruding cohesins, CTCF increases the fraction extruded between the two CTCF boundaries. Overall, we estimate that on average just over half of the *Fbn2* TAD is extruded.

By combining these measurements (fig. 4.2, **B–D**) with our Micro-C data (fig. 4.2, **A**), we were then able to determine dynamic parameters of our polymer model of loop extrusion (fig. 4.2, **E,F**): the spacing between cohesins and their processivity, as well as the total strength of the CTCF boundaries. Consistent with our ΔRAD21 data, our polymer simulations resulted in chromosome decompaction after near-complete RAD21 depletion (fig. 4.2, **E**) and accurately matched our experimental data (fig. 4.2, **F**).

Next, we sought to identify where and when CTCF–CTCF loops occur in our trajectories. Due to localization noise and substantial temporal correlations in the data, simple analysis methods failed when benchmarked on simulations (section 5.1). We thus developed Bayesian inference of looping dynamics (BILD; chapter 5). In BILD, we coarse-grain the possible conformations of the TAD (fig. 4.3, **A**) into two states: 1) a state of sustained contact between the CTCF sites, presumably mediated by cohesin (the 'looped state') and 2) all other possible conformations, including partial extrusion, random contacts, and the fully unlooped conformation (the 'unlooped state'). While the looped state relies on CTCF activity, the unlooped state

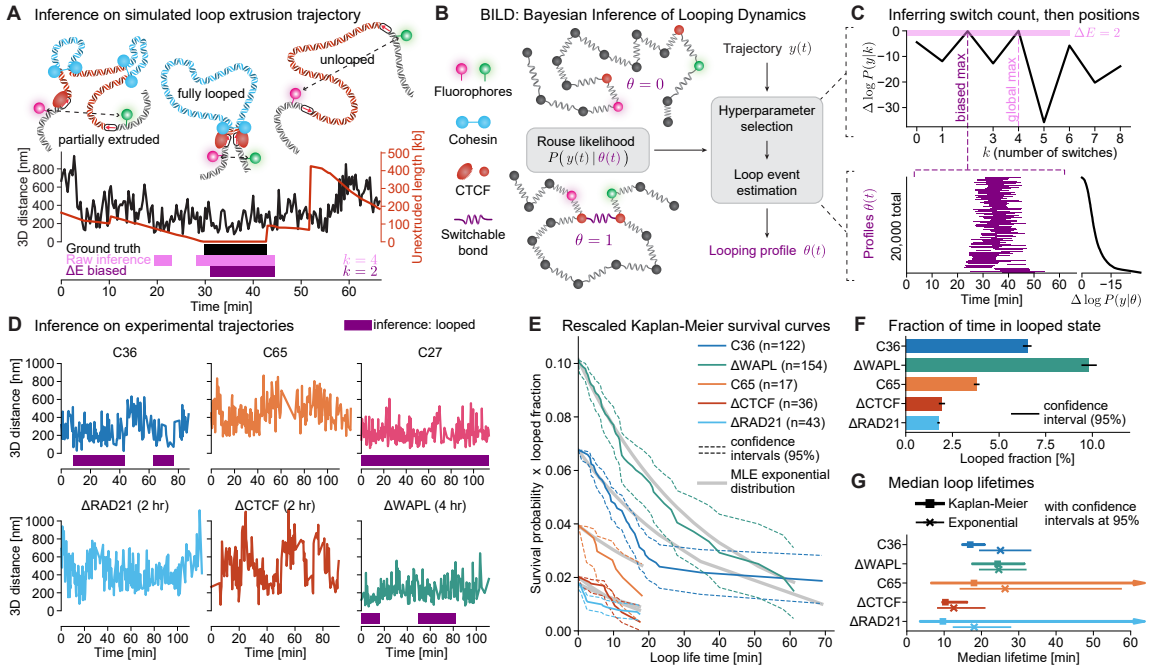


Figure 4.3: Bayesian inference of looping dynamics (BILD) reveals rare and dynamic CTCF loops. (**A**) Example trajectory from polymer simulations with loop extrusion. Extrusion shortens the effective tether (red: unextruded length, ground truth from simulations) between the CTCF sites. A ground truth loop is formed when the tether is minimal and cohesin is stalled at both CTCF sites (black bar). BILD captures these accurately (purple bar). (**B**) Schematic overview of BILD. Building on the analytical solution to the Rouse model, we employ hierarchical Bayesian model to determine the optimal looping profile for single trajectories. (**C**) Illustrative examples of inferred profiles on real trajectory data. (**D**) Kaplan-Meier survival curves rescaled by the inferred looped fraction. Gray lines are maximum likelihood fits of a single exponential to the data, accounting for censoring. (**E**) Fraction of time the *Fbn2*-locus spends in the fully looped conformation. Error bars are bootstrapped 95% confidence intervals. (**F**) Median loop lifetimes from the Kaplan-Meier survival curves (squares) or exponential fits (crosses). Confidence intervals are determined from the confidence intervals on the Kaplan-Meier curve and the likelihood function of the exponential fit, respectively. Where the upper confidence limit on the survival curve did not cross below 50% an arrowhead indicates a semi-infinite confidence interval.

reflects extrusion without bridged CTCF boundaries, resembling the ΔCTCF condition. Based on the $\mathrm{MSD} \sim t^{0.5}$ scaling observed in fig. 4.2, **D** we model the unlooped state as a free Rouse chain calibrated to the ΔCTCF data (fig. 5.4). To model the looped state, we introduce an additional bond between the two CTCF sites (fig. 4.3, **B**); this bond is switchable allowing transitions between the looped and the unlooped states. The length of the bond is set to reproduce the 10 kb distance between the fluorophores, using ΔRAD21 as reference for a free 515 kb chain (section 5.8). Finally, by employing a hierarchical Bayesian model [24], BILD then uses the different spatiotemporal dynamics of the looped state to infer which segments of each trajectory were in the looped state (purple segment in fig. 4.3, **A**). When tested on 3D polymer simulations with experimentally realistic noise, BILD accurately inferred both the looped fraction and loop lifetime (fig. 4.2, **E–F**; see also figs. 5.5 and 5.6). In summary, BILD allows us to distinguish CTCF-/cohesin-mediated looping from mere proximity.

We next used BILD to infer looping in our experimental trajectory data (fig. 4.3, **C–F**). BILD revealed that the *Fbn2* TAD is fully looped ~6.5% (~3%) of the time, but spends ~93.5% (97%) of the time in a fully unlooped or partially extruded conformation (fig. 4.3, **E**). We use brackets to indicate the looped fraction after false positive correction (fig. 5.6; the corrected looped fraction is ~6% if we calibrate BILD using a 15 kb fluorophore distance). In contrast, we observed a minimal looped fraction of ~2% (~0%) in ΔRAD21 and ΔCTCF, and ~4% (~1%) in C65 (ΔCTCFsites), whereas the looped fraction was significantly increased to ~10% (~6%) in ΔWAPL, consistent with WAPL unloading cohesin from chromatin [91]. Finally, we estimated the lifetime of the looped state (fig. 4.3, **D,F**). Accurate measurement of loop lifetimes from finite trajectories can be challenging when trajectories begin or end in the looped state, such that it is unclear how long the looped period truly lasts (e.g. the looped state in ΔWAPL trajectory in fig. 4.3, **D** existed an unknown time before the start of the movie). This problem, known in medical statistics as "censoring", can be solved using the Kaplan-Meier estimator. Using this approach, we obtained censoring-corrected survival curves (fig. 4.3, **D**) of the looped state, from which we estimated the median loop lifetime (fig. 4.3, **F**). Orthogonal to this non-parametric analysis, we also fitted an exponential model, yielding similar estimates. Together, these give an estimate of the median loop lifetime of ~10-30 min in C36 (WT) (fig. 4.3, **F**; fig. 5.6, **D**). These results reveal the fully looped state to be both rare (~3-6%) and quite dynamic (median ~10-30 min; mean ~15-45 min). Thus, during an average ~12 hour mESC cell cycle, the looped state will occur ~1-2 times lasting cumulatively ~20-45 min, but the remaining ~11.5 hours will

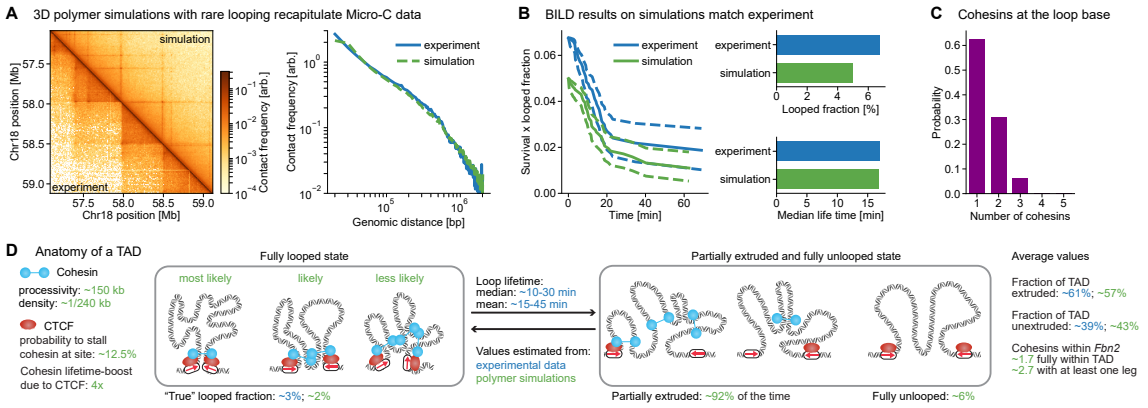


Figure 4.4: Comprehensive picture of the *Fbn2* TAD. (**A**) Comparison of Micro-C data for the C36 line to in silico Micro-C of our best-fit simulation, map (left) and contact probability scaling (right). (**B**) BILD applied to the same simulation (green), comparing to C36 (WT) experimental data (blue). (**C**) Number of cohesins forming the looped state in simulations (n = 18,789). (**D**) "Anatomy" of the *Fbn2* TAD. Quantitative description of the *Fbn2* TAD using both real data (blue) and our best-fit simulation (green). Cohesin processivity and density and CTCF stalling probability and lifetime boost are simulation parameters. Fraction of time in different conformations was extracted from simulation ground truth, using effective tether lengths of 1.1 kb and 505 kb as cutoffs to define "fully looped" and "fully unlooped", respectively. Fraction of TAD unextruded was calculated using the mean tether length over the full simulation.

be in the partially extruded or fully unlooped conformations.

To understand if a low looped fraction of ~3% is consistent with a clear and strong corner peak in the Micro-C map, we set up polymer simulations with loop extrusion. Consistent with recent reports [93, 94], we found that CTCF-mediated stabilization of cohesin was necessary to reproduce both these features in our simulations (fig. 4.4) [93,94]. We confirmed this effect using iFRAP of cohesin, finding that CTCF depletion decreases cohesin residence time. Incorporating this effect, we then simulated loop extrusion with a cohesin density of 1/240 kb and processivity of 150 kb (processivity = lifetime x extrusion speed). When cohesin reaches a CTCF site, it has a probability of 12.5% to stall, which, using the estimate of 50% CTCF occupancy [95], translates to a ~25% capture efficiency of CTCF. Once stalled on one side by CTCF, cohesin is stabilized 4-fold beyond its base lifetime of ~20 min [96], facilitating the formation of longer loops since the other side of cohesin may continue to extrude. These simulations reproduced both our experimental Micro-C maps (fig. 4.4, **A**) and the median loop lifetime and low looped fraction (fig. 4.4, **B**).

Together, these results allow us to paint a comprehensive mechanistic picture of the *Fbn2* TAD (fig. 4.4, **C–D**): most of the time (~92%), the TAD is partially extruded, with ~57-61% of the *Fbn2* region captured in ~1-3 extruding cohesin loops, while ~39-43% remain unextruded.

The fully unlooped conformation, as it would be found in the absence of cohesin, occurs only ~6% of the time, while the fully looped state is even more rare at ~3% (~2% in simulations) and has a median lifetime of ~10-30 min. Interestingly, our simulations reveal that the looped state is sometimes held together by multiple cohesins (fig. 4.4, **C**), which also explains why the loop lifetime can be substantially shorter than the CTCF-stabilized cohesin lifetime. Nevertheless, we stress that both the mechanistic assumptions of our polymer simulations and the experimental data constraining them are associated with uncertainty, resulting in uncertainty of the inferred parameters (fig. 4.4, **D**). For example, if we allow extruding cohesins to bypass each other in our simulations [12, 97], our estimates of the fold-stabilization of cohesin by CTCF would change from ~4 to ~2, the CTCF stalling probability from 12.5% to 25%, and the looped state would now be held together by a single cohesin. We also note that TADs smaller than the 505 kb *Fbn2* TAD as well as TADs with stronger CTCF boundaries may have a higher looped fraction [98]. Furthermore, we propose that our absolute quantification of the *Fbn2* looped fraction may now allow calibrated inference of absolute looped fractions genome-wide, based on Micro-C [77].

Our findings reveal the CTCF/cohesin-mediated looped state that holds together CTCF boundaries of TADs to be rare, dynamic, and transient. A key limitation of our study is that it represents just one loop in one cell type. Nevertheless, the *Fbn2* loop is among the strongest quartile of "corner peaks" in Micro-C maps, suggesting that most other similarly sized loops in mESCs are likely weaker than *Fbn2*. Our results thus rule out static models of TADs, where TADs exist in either a fully unlooped state or a fully looped state stably bridged by one cohesin (fig. 4.1, **B**). Instead, we show that the *Fbn2* TAD most often exists in a partially extruded state formed by a few cohesins in live cells (~92%; fig. 4.4, **D**), and that when the rare looped state is formed, it is transient (~10-30 min median lifetime; fig. 4.4, **B**). Since the partially extruded state dominates, this may be the functionally important TAD state. Thus, we suggest that CTCF-mediated transcriptional insulation may be more mediated by individual extrusion-blocking CTCF boundaries rather than the rare fully looped state. Similarly, this suggests that frequent cohesin-mediated contacts within a TAD rather than rare CTCF–CTCF loops may therefore be more important for regulatory interactions, such as those between enhancers and promoters. This dynamic picture of TADs in live cells (fig. 4.4, **D**), may also help explain cell-to-cell variation in 3D genome structure, and consequently stochasticity in downstream processes such as gene expression and cell differentiation.

Chapter 5

Bayesian Inference of Looping Dynamics (BILD)

BILD was designed to call chromatin looping from fluorescence microscopy data (fig. 5.1). We employed it in chapter 4 to infer the looping dynamics of the *Fbn2* TAD in mouse embryonic stem cells. This chapter contains a detailed treatment of the method; an implementation in python is available in the `bild` package [99].

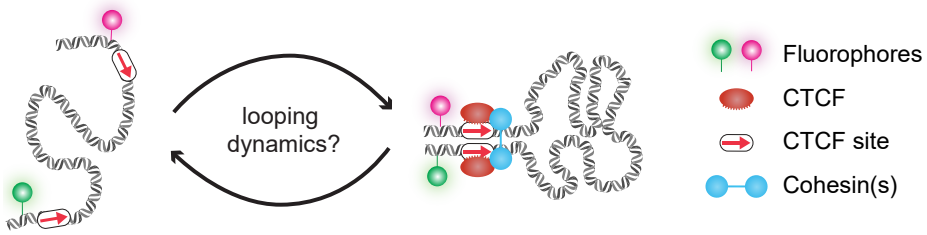


Figure 5.1: Identifying chromatin looping by fluorescence microscopy.

This chapter builds directly on the experimental data presented in chapter 4 and thus uses the terminology of that specific system.

5.1 Thresholding and mixture models fail to quantify looping

The question that ultimately prompted us to develop BILD is how to accurately quantify both the fraction of time spent in the looped state and the duration of individual looping events (fig. 5.1). Before developing the Bayesian inference method BILD (section 5.3) to achieve these goals, we initially explored a range of simpler and more direct analysis methods, including popular

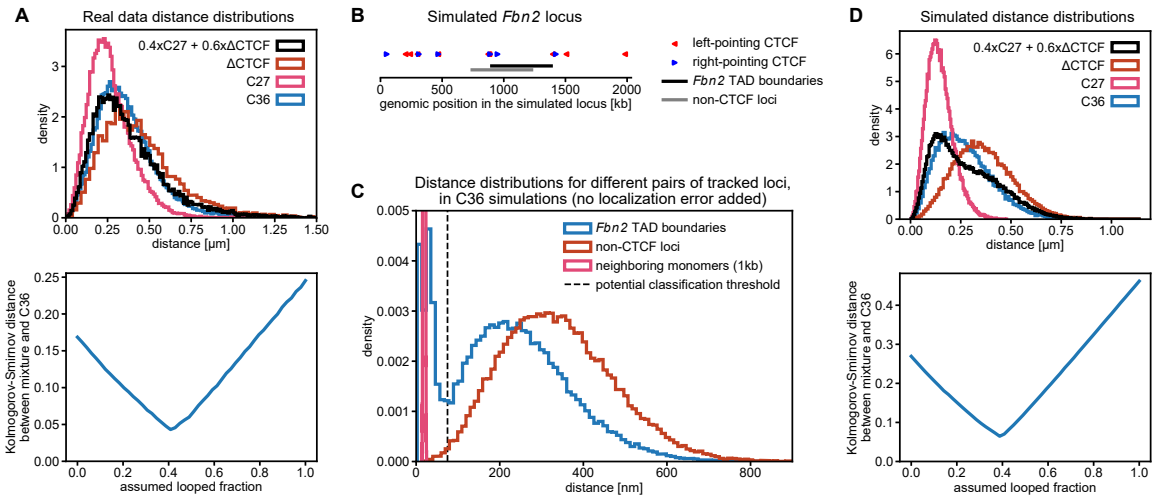


Figure 5.2: Mixture modelling is not adequate to infer looped fractions. (**A**) Trying to represent the distribution of distances observed in the WT (C36) cell line as a mixture of the positive (ΔTAD; C27) and negative (ΔCTCF) looping controls, one finds a best fit looping fraction of ~40 %. To judge whether this estimate is reliable, we simulated this procedure; (**C**) gives an overview over the simulated 2034 kb stretch of chromatin that we simulated and the CTCF sites located within. Black bar indicates the simulated *Fbn2* TAD, whose boundaries are tracked to obtain the simulated "WT". Note that we include neither the 5 kb offset between TAD boundaries and fluorophores, nor localization error. Grey bar indicates the loci used to emulate the ΔCTCF control, by choosing loci that are not in the vicinity of CTCF sites. (**C**) shows the distance distributions obtained from these simulations. Due to the idealized simulation setting, the "WT" distribution (blue) is bimodal and appears amenable to a mixture model. Note, however, that the negative looping control (red) does not match the second lobe of the WT distribution (see text). (**D**) adding realistic tether length and localization error to the simulation system recapitulates the 40 % looping fraction estimate from the real data. The simulations have a ground truth looped fraction of 14 %.

choices such as thresholding and mixture modeling. Both of these failed when we benchmarked them on 3D polymer simulations and thus do not provide adequate means to measure looping in real trajectories; this section explains why that is.

We first consider a mixture model to measure the looped fraction (fraction of time spent in the looped state). Given the experimentally observed distance distributions (fig. 5.2, **A**) in the WT (C36) cell line, the positive looping control ΔTAD (C27), and the negative looping control ΔCTCF, one might be tempted to model the wild-type (C36) data (whose trajectories we assume to be composed of looped and unlooped segments) as a linear mixture of ΔTAD and ΔCTCF, according to

$$p_{\mathrm{WT}}(r) = f_{\mathrm{looped}} p_{\Delta\mathrm{TAD}}(r) + (1 - f_{\mathrm{looped}}) p_{\Delta\mathrm{CTCF}}(r)\,. \tag{5.1}$$

The best-fit looped fraction can then be found by minimizing a suitable metric, such as the sum

of squared residuals or the Kolmogorov-Smirnov distance between the mixture distribution and the observed WT (C36) distribution. This approach yields a looped fraction of $\sim 40\,\%$ on our real data (fig. 5.2, **A**).

To illustrate why this analysis approach fails, we turn to polymer simulations where we know the ground truth looping behavior (fig. 5.2, **B,C**). We consider a simulation that resembles WT, but exhibits a ground truth looped fraction of $\sim$14 %. We consider an idealized setting where there is no localization uncertainty and the tracked loci are exactly the CTCF sites (no tether length between CTCF sites and fluorescent labels). As negative looping control, we consider loci far from the CTCF sites; as positive looping control, we consider simply two neighboring monomers (i.e. 1 kb separation). With this setup (no localization error, no tether length between CTCF sites and fluorescent labels, and looping control from the same simulation) we should be as close as possible to the ideal situation for the mixture analysis approach of eq. (5.1). We obtain the distance distributions shown in fig. 5.2, **C**.

Due to the lack of localization error and finite tether (see also fig. 5.3), we obtain a clearly bimodal distribution for the distances of the TAD boundaries. However, this distribution is *not* a mixture of the positive and negative looping controls, as the "unlooped" mode is clearly at significantly shorter distances than the negative looping control. This is due to strong temporal correlations in the data, introduced by polymer dynamics. The presence of the looped state in WT (C36) does not only affect the distances measured while the loop is present, but also means that during unlooped periods the loci will on average be closer together than in ΔCTCF, because the last looping event happened a finite time ago. This memory is stored in the conformation of the polymer, which takes considerable time to relax. This is why a mixture analysis fails here.

For completeness, we note that if we add localization error and finite tether, the mixture analysis recapitulates the $\sim 40\,\%$ looped fraction found with this method in the real data (fig. 5.2, **D**). This is of course a significant overestimate of the 14 % true looped fraction in this simulation.

Finally, if we were faced with a bimodal distribution like in the idealized simulation setting illustrated in fig. 5.2, **C**, we might consider a simple distance threshold to discriminate between the looped and unlooped states, e.g. by matching the minimum between the two modes in the distribution. In the shown example we would use the threshold shown by the dashed line at 75 nm, finding a looped fraction of 19 %, somewhat more in line with the ground truth. This approach, however, fails in the presence of localization error and finite tether, where we expect

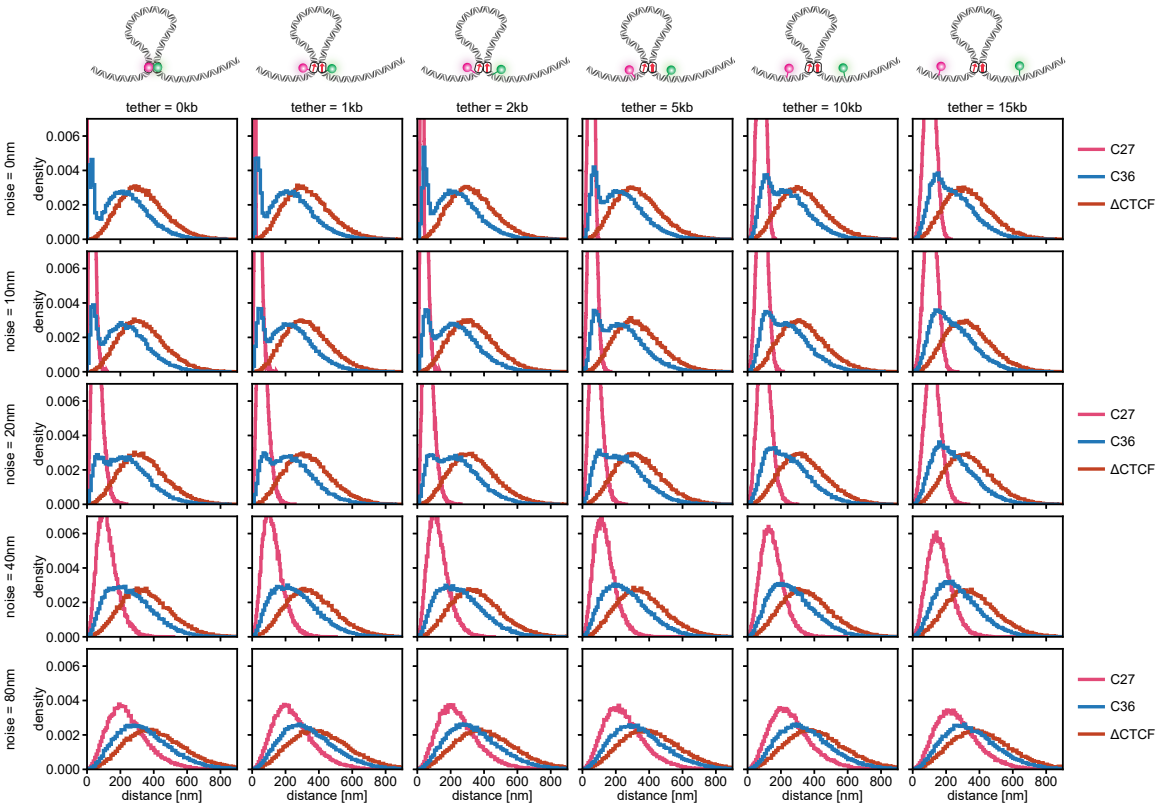


Figure 5.3: Simulated distance distributions illustrate the effects of label placement and localization error. Simulations were performed with a ground truth looped fraction of 14 % for illustrative purposes. Left to right: increasing the separation between the CTCF sites and tracked loci. The tracer particles (fluorescent labels) on both sides of the TAD were each offset from the true CTCF site by half the indicated total distance. Top to bottom: increasing localization error. Gaussian noise with a standard deviation indicated as "noise" was added to x and y components of each tracked locus; for z the standard deviation was twice the indicated value, reflecting experimental reality. The experiments in chapter 4 were performed with a tether length of 10 kb and localization error of ~40 nm.

this distribution to become unimodal, as illustrated in fig. 5.3.

In summary, simple analysis approaches to infer the looped fraction from real data fail due to strong correlations in the data, as well as localization error. Accurately measuring the lifetime of single looped intervals is even harder, facing the problem of splitting intervals due to fluctuations (issue 2 in section 5.2). These difficulties prompted us to develop BILD, a rigorous Bayesian loop inference method.

5.2 Overview

BILD aims to infer looping in experimental 3D distance trajectories. To that end, we partition the possible conformations of the TAD into two states:

- the *looped state*, where presumably cohesin bridges the two CTCF sites, thereby forming a loop complex
- the *unlooped state*, which contains all other possible conformations. This includes the fully unlooped conformation (no cohesins within the TAD), random (diffusion-mediated) contacts between the CTCF sites, as well as the partially extruded conformations where cohesin(s) is/are present in the TAD, but not specifically bridging the CTCF boundaries. This state should thus not be thought of as just the fully unlooped conformation, but rather as "all conformations *minus* the looped state". As such, it is best captured experimentally by the ΔCTCF condition, which exactly abolishes stabilized CTCF/cohesin-mediated interactions between the CTCF sites. Meanwhile, extrusion itself is still active, giving rise to partially extruded conformations just like in the wild-type (C36) cell line. Note that the 2-point MSD in ΔCTCF still follows the characteristic $\Delta t^{0.5}$ scaling, such that we can capture the temporal correlation structure in these data by modelling the chain as an effectively free Rouse polymer.

With each trajectory of 3D distances $\boldsymbol{y}(t)$ we now associate a binary trajectory $\theta(t)$, where $\theta = 1$ indicates the looped state, and $\theta = 0$ indicates the unlooped state (fig. 5.1 and fig. 4.3, **B**). We will refer to such a time series of loop states as a *looping profile*. The inference task is now to find an accurate looping profile $\theta(t)$ for a given trajectory $\boldsymbol{y}(t)$, thus segmenting trajectories into looped and unlooped periods. In doing so, we face two main challenges:

1. Strong temporal correlations in the data, due to the polymeric nature of chromatin, render purely data-driven approaches unreliable, as we discuss in section 5.1. For example, one might expect that unlooped periods in the WT (C36) data should be statistically similar to ΔCTCF (which serves as our definition of the unlooped state afterall). This is not true, since the polymer chain needs time to relax after a looping event.
2. Due to random fluctuations in the data (true polymer motion or localization error), even in a *bona fide* looped state we may measure relatively large distances, which might look like a loop breakage. This is very problematic, since we are trying to measure loop lifetimes; if we incorrectly infer a loop breakage in the middle of a long looped period, we obtain two intervals of half the original length, even though as little as a single data point might have been inferred wrongly. This would clearly constitute a strong bias towards shorter lifetimes.

We employ a Rouse model (chapter 2) to capture the temporal correlation structure of our experimental data, thus addressing the first issue. The Rouse model being analytically solvable also conveniently provides us with a likelihood function $p(y \mid \theta)$ over possible looping profiles $\theta(t)$, given a trajectory $y(t)$ (section 2.2.2; specifically eq. (2.63)). The problem in point 2 above can then be understood as overfitting: in absence of any penalization of profile complexity, the "best" profile (in the maximum likelihood sense) will be one that captures every fluctuation in the data, regardless of whether it is due to random fluctuations or an actual change in the state of the system. This is very similar to the problem of fitting a polynomial curve to n data points: as long as we do not constrain the degree of the polynomial, we can always find one of degree at most $n-1$ that matches all data points perfectly. Similarly, if we allow arbitrary switching between looped and unlooped states in the looping profile inference, profiles can become arbitrarily complex. We will thus solve point 2 above by running a Bayesian model selection scheme over the number of state changes (*switches*) in the profile, as described in more detail in section 5.3. These two points—capturing correlations by use of the Rouse model and penalizing profile complexity by Bayesian model selection—constitute the core of the BILD method.

The method as outlined above infers looping with high recall, but low precision (i.e. many false positive detections; fig. 5.5, **B**). We thus introduced an additional parameter ΔE (see eq. (5.5) and surrounding text for details), allowing us to tune the precision/recall trade-off (fig. 5.5, **J**). Throughout this work, if not stated otherwise, we set $\Delta E = 2$, a value that we found to give accurate inferences in simulations (fig. 5.5, **K**).

The remainder of this section is organized as follows: with the conceptual approach outlined above, section 5.3 provides technical details on the method. Section 5.4 describes how we calibrated the Rouse model that the inference is based on to our experimental data. Section 5.5 then displays our benchmarks of the method on simulated data, showing that we can indeed reliably infer loop lifetime and looped fraction. Finally, section 5.6 describes downstream processing of the inference output to generate the final results in fig. 4.3, **D,E,F**.

5.3 Method

To set up a well-defined inference problem, we parametrize the space of looping profiles $\Theta \equiv \{\theta : [0,T] \mapsto \{0,1\}\}$ as follows:

- $\theta_0 \equiv \theta(0) \in \{0, 1\}$ is the initial state,
- $k \in \mathbb{N}_0$ counts the number of switches (0 to 1 or *vice versa*),
- $0 < s_1 < s_2 < \ldots < s_k < T$ are the positions of the k switches.

These parameters fully and uniquely describe any possible looping profile. Clearly the number of parameters needed to describe a given profile depends on the number of switches k. It therefore appears natural to regard k as a hyperparameter and define the one-parameter model family

$$M_k \equiv (\Theta_k, \mathcal{L}_k, \pi_k) \ , \tag{5.2}$$

where $\Theta_k \subset \Theta$ is the subspace of profiles with k switches, $\mathcal{L}_k(\theta, y) = \mathcal{L}(\theta, y) \equiv p\,(y \mid \theta)$ is the Rouse likelihood from section 2.2.2, and $\pi_k(\theta) = \text{Uniform}_{\Theta_k}(\theta) = \frac{k!}{2T^k}$ is a uniform prior over profiles with k switches.

We have thereby set up the problem as a hierarchical Bayesian model, which can be inferred by the evidence approximation [24,100,101]: first we fix the hyperparameter (number of switches k) by maximizing the evidence, then we find the posterior distribution for the actual profile $\theta(t)$ with that fixed number of switches. Directly following this procedure, we infer looping with high recall, but low precision (section 5.5). In order to control this precision–recall trade-off, we introduce the *evidence tolerance* ΔE as described below. We find that a small $\Delta E > 0$ increases precision, while only marginally decreasing recall (fig. 5.5, **J**), and set $\Delta E = 2$ throughout the analysis presented in the main text (cf. section 5.5). A complete overview over our inference scheme follows.

The inference task is to find the best profile θ for a given observed trajectory y. To estimate the hyperparameter k we calculate the log-evidences

$$E_k \equiv \log p\,(y \mid k) = \log \int_{\Theta_k} \mathrm{d}\theta\, p\,(y \mid \theta)\, \pi_k(\theta) \tag{5.3}$$

and maximize, subject to the tolerance ΔE. Specifically, we find the minimal $\hat{k}$ such that the log-evidence is within ΔE of the true maximum:

$$E^* := \max_k E_k = \max_k p\,(y \mid k) \tag{5.4}$$

$$\hat{k} := \min \{k \,:\, E^* - E_k \leq \Delta E\} \ . \tag{5.5}$$

Having thus identified the appropriate model $M_{\hat{k}}$, we then calculate the posterior distribution over looping profiles under $M_{\hat{k}}$

$$p_{\hat{k}}(\theta \mid y) = \frac{p(y \mid \theta)\, \pi_{\hat{k}}(\theta)}{p\left(y \,\middle|\, \hat{k}\right)} \tag{5.6}$$

and pick the maximum a posterior (MAP) profile as a point estimate of the best profile for the given trajectory y:

$$\hat{\theta}(y) = \underset{\theta}{\operatorname{argmax}}\, p_{\hat{k}}(\theta \mid y) \ . \tag{5.7}$$

The remainder of this section describes our technical implementation of eqs. (5.3) and (5.7).

We estimate the integral (5.3) by Adaptive Multiple Importance Sampling (AMIS; [102]). In the following, k remains fixed and is often suppressed to declutter notation. AMIS utilizes a family of proposal distributions $q_\psi(\theta)$, parametrized by some set of parameters ψ, and then alternates between evaluating the target function $f(\theta) \equiv p(y \mid \theta)\, \pi_k(\theta)$ at points sampled from the proposal, and updating the proposal based on past samples. The key advantage of this approach is that samples from past steps are reweighted appropriately instead of discarded, such that we build up a properly weighted sample relatively quickly, without discarding any of the evaluations (which greatly contributes to computational feasibility). The n-th step in this sampling scheme is described in algorithm 5.1, while we refer to the original work [102] for more details.

To implement this scheme for our problem, we recall that since the number of switches k is fixed, a looping profile is parametrized by its initial value θ_0 and the switch positions $s_1, \ldots, s_k$. We rewrite the latter in terms of the fractions $u_a \equiv \frac{s_{a+1} - s_a}{T} \in [0, 1]$ with $s_{k+1} \equiv T$ (the total trajectory length) and $s_0 \equiv 0$, and write them collectively as $\boldsymbol{u} \equiv (u_0, \ldots, u_k)$. This allows us to employ the proposal distribution

$$q_{m,\boldsymbol{\alpha}}(\theta_0, \boldsymbol{u}) := \text{Bernoulli}_m(\theta_0) \otimes \text{Dirichlet}_{\boldsymbol{\alpha}}(\boldsymbol{u}) \ , \tag{5.11}$$

where $\text{Bernoulli}_m(\theta_0) \equiv m^{\theta_0}(1-m)^{1-\theta_0}$ and $\text{Dirichlet}_{\boldsymbol{\alpha}}(\boldsymbol{u}) \equiv \frac{1}{B(\boldsymbol{\alpha})} \prod_{a=0}^{k} u_a^{\alpha_a - 1}$ are, respectively, a Bernoulli distribution with mean m and a Dirichlet distribution with concentration parameters $\boldsymbol{\alpha}$. For initialization of the recursive AMIS scheme we choose the uniform distribution, given by the parameter values $m_0 = \frac{1}{2}$ and $\alpha_a = 1 \, \forall a$. In the sampling step, we choose $R_n = 100 \, \forall n$. For the updating step we estimate the proposal parameters by the method of

Input: the sample size R_n, proposal parameters ψ_n

1 draw R_n new samples $\{\theta_r^n\}_{r=1,\dots,R_n}$ from the current proposal $q_{\psi_n}(\theta)$

2 let $N_{\text{sample}} \equiv \sum_{i=0}^{n} R_n$ (the total number of samples)

3 evaluate on *all* samples $\{\theta_r^i\}_{\substack{r=1,\dots,R_i \\ i=1,\dots,n}}$ the importance weights

$$\omega_r^i = \frac{N_{\text{sample}} f(\theta_r^i)}{\sum_{j=0}^{n} R_j q_{\psi_j}(\theta_r^i)} \tag{5.8}$$

4 estimate the evidence as

$$P(y \mid k) = \int_{\Theta_k} \mathrm{d}\theta\, f(\theta) \approx \frac{1}{N_{\text{sample}}} \sum_{i=0}^{n} \sum_{r=1}^{R_i} \omega_r^i \equiv \bar{\omega} \tag{5.9}$$

5 calculate the estimates for log-evidence $\hat{E}_k^n$ and its standard error $\Delta\hat{E}_k^n$ as

$$\hat{E}_k^n = \log \bar{\omega}\,, \qquad \Delta\hat{E}_k^n = \frac{\Delta\bar{\omega}}{\bar{\omega}} = \frac{1}{\bar{\omega} N_{\text{sample}}} \sqrt{\sum_{i=0}^{n} \sum_{r=1}^{R_i} (\omega_r^i - \bar{\omega})^2} \tag{5.10}$$

6 find the new proposal parameters ψ_{n+1} by fitting the proposal to the current sample $\{(\theta_r^i, \omega_r^i)\}$

Algorithm 5.1: The n-th step in AMIS. See [102] for more details. Note that at each iteration, on top of the actual evidence we also get an estimate of the standard error.

moments, setting

$$m_{n+1} = \langle \theta_0 \rangle_n \ , \tag{5.12}$$

$$\boldsymbol{\alpha}_{n+1} = \frac{\langle \boldsymbol{u} \rangle_n}{k} \left(\sum_{a=0}^{k} \frac{\langle u_a \rangle_n \left(1 - \langle u_a \rangle_n\right)}{\mathsf{Var}_n \, u_a} - k \right) \ , \tag{5.13}$$

where $\langle \cdot \rangle_n$ are expectation values under the sample at step n, and similarly

$$\mathsf{Var}_n \, u_a \equiv \frac{1}{\sum_{i,r} \omega_r^i} \sum_{i=0}^{n} \sum_{r=1}^{R_i} \omega_r^i \left((u_a)_r^i - \langle u_a \rangle_n \right)^2 \tag{5.14}$$

is the variance of u_a at step n. To prevent overfitting of the proposal distribution to small samples at the beginning of the iterative scheme (the first proposal update happens after $R_0 = 100$ samples are drawn), we introduce two *braking parameters* b_m and b_α. Whenever parameters are updated, these limit the difference to the previous value:

$$|m_{n+1} - m_n| \leq R_n b_m \, , \tag{5.15}$$

$$\left| \log \frac{|\boldsymbol{\alpha}_{n+1}|_1}{|\boldsymbol{\alpha}_n|_1} \right| \leq R_n b_\alpha \, , \tag{5.16}$$

$$\tag{5.17}$$

where $|\alpha|_1 \equiv \sum_a \alpha_a$ is the total concentration. In practice we use $b_m = 10^{-3}$ and $b_\alpha = 10^{-2}$. With this, we now have all the ingredients to initialize AMIS, run individual sampling steps using algorithm 5.1, and perform well-regulated updates of the proposal parameters.

Finally, we have to provide a stopping criterion for the iterative AMIS scheme. To that end, we take a step back and consider the problem of k selection as a whole. For $k = 0$ there are only two possible profiles (completely looped or completely unlooped), so we can calculate the evidence exactly with two likelihood evaluations. For $k = 1$, taking into account our constraining switch positions to integer frames (section 2.2.2), we can still completely enumerate all possible profiles with $2T \sim 10^2$ evaluations of the likelihood. For $k = 2$ we would have to enumerate $2T(T-1) \sim 10^4$ profiles, while the sampling approach usually converges with $\sim 10^3$ evaluations, so for $k \geq 2$ we resort to sampling. However, taking this iterative approach, when sampling at k we can assume to already have an estimate or exact value for $E_{\tilde{k}}$ for all $\tilde{k} < k$. Importantly, the sampling also gives us estimates of the standard error $\Delta \hat{E}_k$, so after each AMIS iteration we

can check which of the estimates still has a realistic chance of being the best one. We formalize this approach in algorithm 5.2, thus ultimately obtaining all relevant estimates $\hat{E}_k$.

Input: $k_{\text{max}} = 10$, $n_{\text{init}} = 20$, $C_{\text{close}} = 5$, $\Delta k_{\text{lookahead}} = 2$
Output: posterior samples and evidence estimates $\hat{E}_k$

calculate E_0 and E_1 by complete enumeration
for $k \leq k_{\text{max}}$ **do**
  run n_{init} AMIS samples to get initial estimate of E_k
  repeat `// relevance resolution`
    find $k^* \equiv \text{argmax}_k \hat{E}_k$
    find all k that are reasonably close to this maximum:

$$I \equiv \left\{ k : \frac{\hat{E}_{k^*} - \hat{E}_k}{\sqrt{\left(\Delta \hat{E}_k\right)^2 + \left(\Delta \hat{E}_{k^*}\right)^2}} \leq C_{\text{close}} \right\} \tag{5.18}$$

    find the $k_\Delta \in I$ where the estimated error on the evidence is highest
    run one more AMIS iteration for $k = k_\Delta$
  until $|I| = 1$ **or** $k + 1 - \Delta k_{\text{lookahead}} \in I$
end
run through the relevance resolution once more

Algorithm 5.2: Scheme for successive AMIS sampling, focussed on sampling the relevant k values.

Having estimated the evidences E_k and found the optimal number of switches k via eq. (5.5), we now turn to the question of finding the optimal profile θ with that given number of switches. To that end we note that in calculating the evidences via AMIS, we already generated extensive posterior samples for all relevant k. We can therefore simply pick the profile θ_r^i with the highest posterior weight ω_r^i as point estimate. We note that on top of this point estimate we actually obtain a full (weighted) posterior sample

$$S \equiv \left\{ \left(\theta_r^i, \omega_r^i \right) \right\} , \tag{5.19}$$

which will be used in some of the downstream analysis in section 5.6.

Summarizing, we have shown in this section how we infer the looping profile θ from an observed trajectory, using the Rouse likelihood (2.63), hierarchical Bayesian inference, and Adaptive Multiple Importance Sampling (AMIS) [102]. We refer to this approach as Bayesian Inference of Looping Dynamics (BILD).

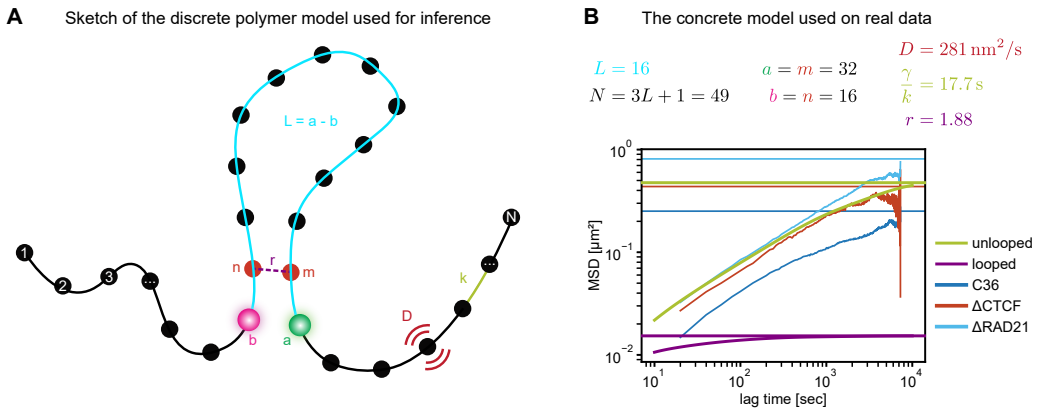


Figure 5.4: Overview of the model used for Bayesian Inference of Looping Dynamics (BILD). (**A**) Cartoon of the two-state Rouse model used by BILD. Black circles represent monomers with individual diffusivity D, connected by springs with constant k. Tracer particles are at positions a and b, bounding the "chain of interest" of length $L = a - b$ (wlog $a > b$). In the looped state an additional bond of relative strength r is introduced between monomers m and n. (**B**) Top: parameter values obtained by following the calibration scheme outlined in the text. Bottom: MSDs of the looped and unlooped model states overlaid on top of experimental data. MSDs for the model states are obtained from eq. (2.16).

5.4 Calibration of the inference model

To run the inference scheme described in section 5.3, we have to make sure that the underlying model (section 2.2.2) accurately captures the behavior we expect for the looped and unlooped states, i.e. we have to find numerical values for the constants such that the model captures our data most accurately (fig. 5.4, **B**).

In a first step, we calibrate the model for the unlooped state. Following fig. 4.3, **A**, the unlooped state should capture the behavior of the chain in the absence of sustained looping. It is thus a coarse-grained representation, capturing not only the fully unlooped conformations, but also partial extrusion and random transient contacts ("everything except the looped state"). Our experimental realization of this situation is the CTCF-AID cell line, in which the dynamics of the *Fbn2* loci should be mostly the same as in the wild-type cell line, except for the absence of sustained looping. We find that we can capture the dynamics of this condition well with an MSD of the form (2.88), our expectation for two loci on a linear polymer. We stress that this does not amount to the assumption that chromatin in the absence of CTCF is unlooped, but we are simply subsuming all the conformations associated with the unlooped state into an "effectively free" chain. One important consequence is that from this calibration of the unlooped state we cannot yet assemble the looped state (see below). To capture the physical parameters of this

coarse-grained model, we use `bayesmsd` (chapter 6) to fit the MSD (2.88) to the CTCF-AID data, thus obtaining numerical values for the phenomenological parameters Γ and J. We now utilize the correspondence between the continuous and discrete Rouse model (section 2.3.3) to assemble the discrete model needed for BILD from the continuous model we use to capture the dynamics of the calibration data.

The model for the unlooped state is specified by eqs. (2.38) and (2.56), from which we identify the following parameters: diffusivity D and friction constant γ of individual monomers; the spring constant k determining the strength of the backbone bonds; the number N of monomers, and positions a and b of the tracked loci on the chain (fig. 5.4, **A**). First of all, we note that eq. (2.38) can always be rescaled by $\frac{1}{\gamma}$, such that our effective degrees of freedom are $D, \frac{k}{\gamma} \in \mathbb{R}^+$, $N \in \mathbb{N}$, and $a, b \in \{1, \dots, N\}$ (wlog $a > b$). We also use the auxiliary variable $L \equiv a - b$. From the fits to the experimental data we know

$$\Gamma = 2D\sqrt{\frac{\gamma}{\pi k}} \qquad \text{and} \qquad J = \frac{D\gamma}{k}L\,. \tag{5.20}$$

Together with the constraint that we should not observe the freely diffusive regime of the Rouse monomers (fig. 2.1 and section 2.3.3), this reads

$$\frac{J}{\Gamma} = \sqrt{\frac{\pi\gamma}{4k}}L\,, \qquad \frac{\Gamma^2}{J} = \frac{4D}{\pi L}\,, \qquad \text{and} \qquad \Delta t_{\text{frame}} > \frac{\gamma}{k}\,. \tag{5.21}$$

Using only the first two equations to fix the microscopic parameters, we would have one degree of freedom left, since we can always add more monomers to the chain (increase L), if we rescale D and k appropriately. For computational efficiency we prefer L to be as small as possible, which in this context means that we aim to satisfy the bound provided by the last inequality as tightly as possible (with integer L). We therefore find

$$L = \left\lceil \sqrt{\frac{4}{\pi}} \frac{J}{\Gamma\sqrt{\Delta t_{\text{frame}}}} \right\rceil\,, \qquad D = \frac{\pi L \Gamma^2}{4J}\,, \qquad \frac{k}{\gamma} = \frac{\pi}{4}\left(\frac{L\Gamma}{J}\right)^2\,, \tag{5.22}$$

where $\lceil x \rceil$ indicates the smallest integer larger than x. The remaining parameters for the unlooped state are now the total length of the chain N and the positions a, $b \equiv a - L$ of the tracked monomers. We choose

$$N = 3L + 1\,, \qquad a = 2L + 1\,, \qquad b = L + 1\,, \tag{5.23}$$

such that on both ends of the "chain of interest" of length L we have chains of equal length L, resulting in a total chain long enough to emulate an infinite polymer up to the relaxation time scale of the chain of interest (fig. 2.3). This concludes our calibration of the unlooped state, which finally is determined by eqs. (5.22) and (5.23), in terms of Γ and J of eq. (5.20), which we find by fitting the CTCF-AID data (fig. 5.4, **B**).

For the looped state of our inference model we now have only the parameters associated with the extra bond left to determine. These are its strength r relative to a backbone bond (such that its spring constant is rk) and the two monomers m and n it connects to (fig. 5.4, **A**). Assuming the chain in the loop to be much longer than the single extra bond (thus not contributing to the steady state variance in the looped state), the effective tether length in the looped state is given by $L_{\text{looped}} = a - m + \frac{1}{r} + n - b$ (fig. 5.4, **A**), such that we find the steady state variance as

$$J_{\text{looped}} = \frac{D\gamma}{k} L_{\text{looped}} \,. \tag{5.24}$$

Experimentally, we can access J_{looped} via our RAD21-AID cell line and genomic information. For RAD21-AID (and only there) we can assume that the chain length $L_{\Delta\text{RAD21}}$, which we can extract from fitting the steady state variance $J_{\Delta\text{RAD21}}$, matches the known genomic separation of our tracers of 515 kb. Under the Rouse model (e.g. eq. (2.88)) $J \propto L$, such that we get

$$J_{\text{looped}} = \frac{10\,\text{kb}}{515\,\text{kb}} J_{\Delta\text{RAD21}} \,, \tag{5.25}$$

where 10 kb is the combined distance of the tracked loci from the CTCF sites, i.e. the effective tether length in the looped state. This allows us to calculate L_{looped} from eq. (5.24) and the previously determined constants D and $\frac{k}{\gamma}$, as well as the ΔRAD21 steady state variance $J_{\Delta\text{RAD21}}$. We then convert the real-valued L_{looped} into the parameters of the extra bond as outlined in algorithm 5.3, which is designed to achieve two goals: first, place the extra bond as symmetrically as possible between the tracked monomers. Second, keep r and $\frac{1}{r}$ as close to 1 as possible, such that the extra bond is as similar to a backbone bond as possible. We realize this latter goal by choosing $r \in \left[\frac{1}{\varphi}, \varphi\right]$, with the golden ratio $\varphi = 1 + \frac{1}{\varphi} = \frac{1+\sqrt{5}}{2}$.

In summary, to calibrate our inference model to experimental data we

- fit the unlooped state to our CTCF-AID data, then
- match strength and position of the extra bond to the RAD21-AID steady state, rescaled

Input: $L_{\text{looped}} \in \mathbb{R}^+$, positions a, b of the tracked monomers
Output: parameters m, n, r of the extra bond

if $L_{\text{looped}} \geq \frac{1}{\varphi}$ **then**
  find $\bar{L} \in \mathbb{N}_0$, $r \in \left[\frac{1}{\varphi}, \varphi\right]$ such that $L = \bar{L} + \frac{1}{r}$
else
  $\bar{L} \leftarrow 0$
  $r \leftarrow L_{\text{looped}}^{-1}$
end
$m \leftarrow a - \left\lceil \frac{\bar{L}}{2} \right\rceil$
$n \leftarrow b + \left\lfloor \frac{\bar{L}}{2} \right\rfloor$

Algorithm 5.3: Splitting the real-valued L_{looped} into proper parameters for the extra bond.

by the known genomic separations (see also section 5.8).

This allows us to fully calibrate the model (fig. 5.4, **B**), and thus run BILD (section 5.3).

5.5 Benchmarking BILD on simulations

We benchmarked BILD by comparing its output on simulated data (for details on the simulations see [13]) to the corresponding ground truth, i.e. the intervals where the (simulated) CTCF sites were truly bridged by (simulated) cohesin (fig. 5.5, **A**). We begin by investigating the performance of the "raw inference", setting the evidence bias parameter ΔE (introduced in eq. (5.5)) to zero. To combat the high false positive rate uncovered through this analysis, we then proceed to study $\Delta E > 0$ (see also section 5.7). Where appropriate below, we use standard terminology for binary classification tasks, including *true positive* (TP), *false positive* (FP; type I error), *false negative* (FN; type II error), the performance scores

$$\text{Recall} \equiv \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{and} \qquad \text{Precision} \equiv \frac{\text{TP}}{\text{TP} + \text{FP}}\,, \tag{5.26}$$

and the *prevalence*, which is defined as the fraction of ground truth positive data in the whole data set. In the context of this study, we also refer to the prevalence as the *looped fraction*, the fraction of time the locus spends in the looped state. We now study the inference performance from two perspectives

- First, the point-wise view: at each time point in each trajectory, the inference is a binary

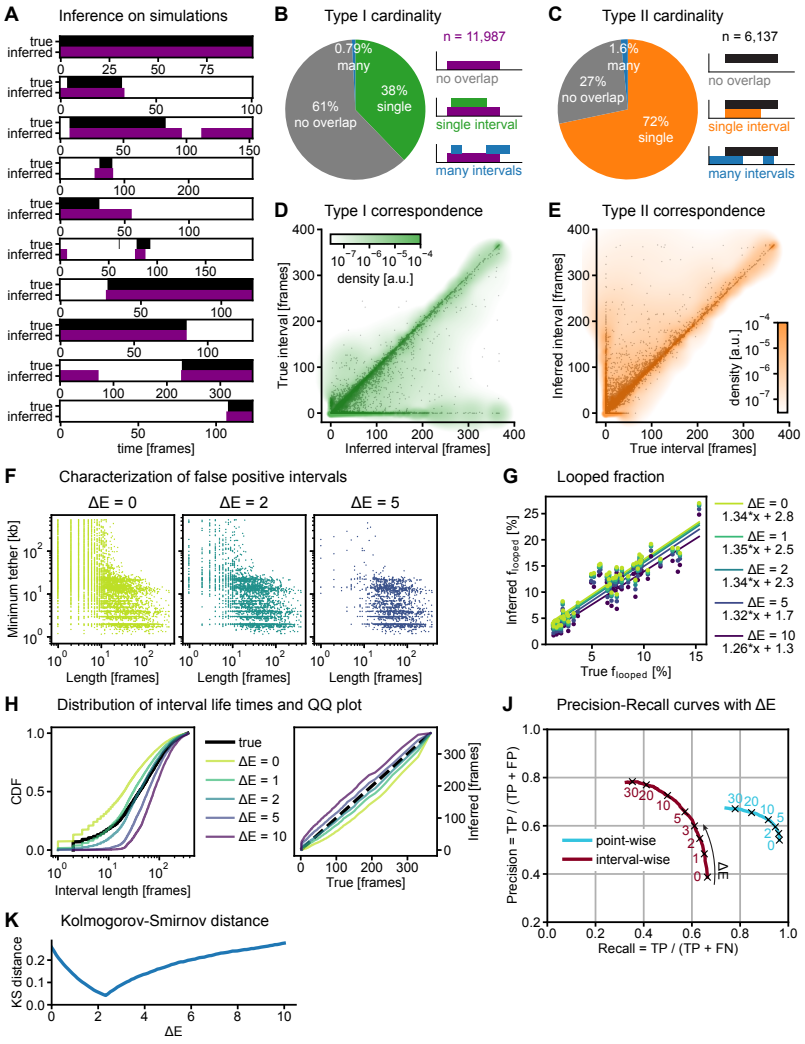


Figure 5.5: Validation of BILD on simulated trajectories. (**A**) Example inferences on randomly selected trajectories. Trajectories were randomly selected from all that had either at least 10 frames truly looped, or at least 10 frames inferred looped. (**B**) Counting true intervals overlapping each inferred interval. Counts higher than one are collectively labelled "many". (**C**) Counting inferred intervals overlapping each true interval. Counts higher than one are collectively labelled "many". (**D**) For each inferred interval, the length of the overlapping true interval, or zero if no true interval overlaps. In the rare cases where multiple intervals overlap, their length is summed. Green overlay is an adaptive Gaussian kernel density estimate, where each scatter point supports a Gaussian density with standard deviation equal to the distance to the tenth nearest neighbor or 2, whichever is greater. (**E**) Same as (D), but applied to each true interval, showing the length of the overlapping inferred interval. (**F**) Scatter plots of false positive intervals (i.e. inferred intervals not overlapping any true interval) at different settings of ΔE. The axes are length of the interval (horizontal) and minimum value of the ground truth effective tether length attained during the interval. (**G**) True vs. inferred looped fraction. Linear regression curves were fit by least squares. (**H**) Distribution of lengths of true and inferred intervals (left) and quantile–quantile plot (right). (**J**) Precision–Recall curves for the inference from point-wise and interval-wise points of view. See text for details. (**K**) Kolmogorov-Smirnov distance (maximum difference between the cumulative distributions) of true and inferred lifetime distributions, at different ΔE.

classification task, calling "looped" or "unlooped" states. Overall we find a recall of 96%, at 54% precision (cf. fig. 5.4, **J**), meaning the inference reliably identifies looping, but roughly half of the data that is inferred as looped are false positives. We thus expect to overestimate the prevalence (looped fraction) by a factor Recall/Precision ≈ 1.8. In fact, however, we find (fig. 5.4, **G**)

$$f_{\text{looped}}^{\text{inferred}} \approx 1.34 f_{\text{looped}}^{\text{true}} + 2.8\% \,, \tag{5.27}$$

leading us to assume two distinct error mechanisms: any true looping in the trajectory will be overestimated by a factor ~ 1.34, and on top of that $\sim 2.8\%$ of the trajectory will be called "looped" just due to random fluctuations. Importantly, while the former is stable under our evidence bias parameter ΔE, the second effect shows some sensitivity to ΔE (fig. 5.4, **G,J**), which we can therefore use to control this error contribution. In summary the point-wise perspective shows that most looping is indeed detected, but there is a significant number of false positives. We can quantify the overall effect on the inferred looped fraction relatively well and thus correct for this overestimate.

- Second, the interval-wise view (based on [103]): since ultimately we are most interested in the lifetime of the looped state, we have to ensure that whole looping intervals are inferred correctly. The biggest concern here is that of cardinality: a single true interval might be split in the inference, leading to (say) three different inferred intervals that, even though they might collectively cover the true interval completely and accurately, by themselves would have a lifetime approximately three times shorter than the true interval. Conversely, a single inferred interval might cover multiple true intervals and thus have a much longer inferred lifetime than the individual true intervals. We term the number of true intervals overlapping with each inferred interval the *type I cardinality*; similarly, the *type II cardinality* labels the number of inferred intervals overlapping with each true interval. We find both cardinalities to be 0 or 1 for almost all intervals, with cardinality > 1 occuring in only 0.79 % and 1.6 % of the cases for type I and type II respectively (fig. 5.4, **B,C**). We therefore conclude that cardinality is not a significant problem in this study.

  Having established that most inferred intervals correspond to at most one true interval, we can immediately study how well we capture the lifetime of these true intervals; we

generally find that if the inferred interval corresponds to a true one, it also captures the lifetime well (fig. 5.4, **D**). We therefore arrive again at the picture of a binary classifier, but now at the interval level: generally, an inferred interval will either accurately capture an interval in the true profile, or it will be a false positive altogether. This binary picture is reinforced by fig. 5.4, **E** showing that each true interval is either inferred correctly, or missed completely.

On the interval level, we find a high false positive rate of 61 %. These false positive inferences are composed of two populations (fig. 5.4, **F**): first, we find intervals that correspond to "almost looped" events, where the effective tether length between the CTCF sites becomes very short, but does not quite cross the threshold of 1.1 monomers that we use to define the looped state. This population is robust under variation of ΔE. Second, we find a population of short ($\lesssim 10$ frames) false positive detections that are independent of the extrusion state of the locus. These are random fluctuations of the polymer/noise that happen to resemble the looped state and show a strong dependence on ΔE, which we can thus use to control these erroneous inferences.

Having seen that we can use our evidence bias ΔE to control the false positive inferences, we now study how it affects the lifetime distribution we infer from our data (fig. 5.4, **H**). As expected, high ΔE mostly removes short intervals from the inference, thus reducing the false positive fraction (increasing precision; fig. 5.4, **J**). At the same time though, recall is reduced as some truly short intervals are not found anymore. It is *a priori* not clear how best to balance these effects, so ultimately ΔE remains a free parameter. From the precision–recall curves in fig. 5.4, **J** we conclude that $0 < \Delta E < 5$ improves overall inference quality (higher increase in precision than decrease in recall). A reasonable value within that range seems to be $\Delta E \approx 2$, for which the Kolmogorov-Smirnov distance between the true and inferred lifetime distributions on our validation data set becomes minimal (fig. 5.4, **K**). We emphasize, however, that this is essentially a rule of thumb, and any final results of the inference should be checked for their variation with ΔE. Corresponding analysis of our results from chapter 4 (fig. 4.3, **E,F**) can be found in fig. 5.6.

5.6 Downstream processing: estimation of looped fraction and loop lifetime

In this section we describe how we use the results of the inference (section 5.3) to measure looped fraction and loop lifetimes, the latter either non-parametrically with the Kaplan-Meier estimator, or with an exponential fit.

We define the looped fraction f_{looped} as

$$\text{looped fraction} = \frac{\text{time spent in looped state}}{\text{total trajectory length}}, \tag{5.28}$$

which is straight-forward to calculate from the inferred looping profiles $\theta(t)$. To also obtain an error estimate, we bootstrap an ensemble of mean looped fractions as described in algorithm 5.4. We finally report the mean of this bootstrapped ensemble as point estimate and its 2.5th and 97.5th percentiles as 95% confidence interval.

Input: N_{traj}, the number of trajectories in the data set
for each trajectory y_n, the posterior ensemble S_n of eq. (5.19)
$N_{\text{bootstrap}} = 1000$,
Output: a list F of $N_{\text{bootstrap}}$ evaluations of the mean looped fraction $\langle f_{\text{looped}} \rangle$

initialize an empty array F, of length $N_{\text{bootstrap}}$
for $i \in \{1, \ldots, N_{\text{bootstrap}}\}$ **do**
  $L \leftarrow 0$ `// counts number of (L)ooped frames`
  $T \leftarrow 0$ `// counts (T)otal number of frames`
  for N_{traj} *repeats* **do**
    draw $n \in \{1, \ldots, N_{\text{traj}}\}$ uniformly at random
    draw a profile from S_n
    $L \leftarrow L +$ number of looped frames in profile
    $T \leftarrow T +$ total number of frames in profile
  end
  $F[i] \leftarrow L/T$
end
return F

Algorithm 5.4: Bootstrapping the distribution of mean looped fractions

In calculating a characteristic lifetime of the looped state, we encounter the problem of censoring (cf. fig. 5.5, **A**). This means that many of the inferred looping intervals start or end with the trajectory, such that we do not observe them fully. Simply calculating mean lifetimes from the inferred intervals in this scenario would provide a heavy underestimate of the

true lifetime. This censoring problem is well-known in the medical literature and customarily addressed by using the Kaplan-Meier estimator for the survival function [104]. We follow that standard procedure: from the MAP looping profiles provided by BILD (section 5.3) we compute the set $\{(t_i, c_i)\}$ of loop lifetimes t_i and corresponding boolean variables c_i indicating whether the observation t_i was censored or not. We then use the Kaplan-Meier estimator

$$\hat{S}(\tau) = \prod_{t_i<\tau} \left(1 - \frac{d_i}{N_i}\right) \tag{5.29}$$

for the survival function $S(\tau) := P(t \geq \tau)$. Here, d_i is the number of uncensored events of length t_i, while N_i is the total number of events of length greater than t_i, censored or uncensored. Confidence intervals (at confidence level $1 - \alpha$) for this estimator are given by the exponential Greenwood formula

$$\mathrm{e}^{z_{\alpha/2}\sqrt{V(t)}} \log \hat{S}(t) \quad < \quad \log S(t) \quad < \quad \mathrm{e}^{-z_{\alpha/2}\sqrt{V(t)}} \log \hat{S}(t)\,, \tag{5.30}$$

$$V(t) \equiv \left(\log \hat{S}(t)\right)^{-2} \sum_{t_i<t} \frac{d_i}{N_i\,(N_i - d_i)}\,, \tag{5.31}$$

with $z_{\alpha/2}$ the $\frac{\alpha}{2}$th quantile of the normal distribution [105, 106]. Note that generally $z_{\alpha/2} < 0$ and specifically for 95% confidence intervals we have $z_{0.025} = -1.96$.

Finally, we also provide an estimate of the median lifetime from an exponential fit to the lifetime distribution. Starting from the same set $\{(t_i, c_i)\}$ as for the Kaplan-Meier estimate, we aim to fit the distribution

$$p(t) = \frac{1}{\tau}\mathrm{e}^{-\frac{t}{\tau}}\,. \tag{5.32}$$

The likelihood contributions for uncensored ($c_i = 0$) and censored ($c_i = 1$) intervals are respectively

$$p\left(t_i \mid \tau, c_i = 0\right) = \frac{1}{\tau}\mathrm{e}^{-\frac{t_i}{\tau}}\,, \qquad p\left(t_i \mid \tau, c_i = 1\right) = \int_{t_i}^{\infty} \mathrm{d}t\, p(t) = \mathrm{e}^{-\frac{t_i}{\tau}}\,. \tag{5.33}$$

The total log-likelihood of observing the given data from a model with mean lifetime τ is then the sum of these contributions:

$$\log p\left(\boldsymbol{t} \mid \tau, \boldsymbol{c}\right) = -N_{\text{uncensored}} \log \tau - \frac{1}{\tau}\sum_i t_i\,, \tag{5.34}$$

where we introduce the number of *uncensored* observations $N_{\text{uncensored}}$. Taking the derivative

and finding its root gives the MLE point estimate

$$\hat{\tau} = \frac{1}{N_{\text{uncensored}}} \sum_i t_i \,. \tag{5.35}$$

To find confidence intervals (at confidence level $1-\alpha$) on this estimate, we numerically find the roots $\tau_\pm$ of

$$\log p\left(\boldsymbol{t} \mid \tau_\pm, \boldsymbol{c}\right) \stackrel{!}{=} \log p\left(\boldsymbol{t} \mid \hat{\tau}, \boldsymbol{c}\right) - \frac{1}{2}\chi^2_{1,1-\alpha} \,, \tag{5.36}$$

where $\chi^2_{n,p}$ is the pth percentile of the χ^2 distribution with n degrees of freedom. Specifically, $\chi^2_{1,0.95} \approx 3.84$.

Note that $\hat{\tau}$ is an estimate for the mean of the distribution, but from our non-parametric approach via the Kaplan-Meier curves we can estimate only medians. We thus finally also report not the mean $\hat{\tau}$ of the fitted exponential, but its median $\hat{\tau}\log 2 \approx 0.7\hat{\tau}$.

5.7 Variation in inference results with evidence bias

As shown in section 5.5, the final results of our looping inference depend on the free parameter ΔE. We found from simulations that $\Delta E > 0$ helps to combat false positives, but what exactly this parameter should be remained unclear. Based on those simulations, a reasonable range is $0 < \Delta E < 5$, with $\Delta E \approx 2$ performing well overall. We proceed in a manner similar to section 5.5, first taking a point-wise perspective and studying the looped fraction, then moving on to an interval-based point of view and loop lifetimes.

In fig. 5.5, **G** we saw that the relationship between true looped fraction $f^{\text{true}}_{\text{looped}}$ and inferred looped fraction $f^{\text{inferred}}_{\text{looped}}$ is captured well by a linear relationship, whose prefactor (relative over-estimation of looped fraction) is nearly independent of ΔE, while the offset (random inference of looping without correspondence to ground truth) decreases with ΔE. A fuller picture of this relationship is shown in fig. 5.6, **A**, left panel, where we perform that linear regression for $\Delta E \in [0, 10]$. This relationship being comparatively robust allows us to use it for correction of looped fractions at all ΔE. On our simulated data set, this leads to a corrected inferred looped fraction $f^{\text{corrected}}_{\text{looped}}$ that is now nearly independent of ΔE and deviates from the ground truth looped fraction by a few percentage points at most (fig. 5.6, **A**, middle). This corrected looped fraction now shows a strong correlation with the ground truth value (fig. 5.6, **A**, right). We then applied this same correction to our real data. We find that we do not abolish all vari-

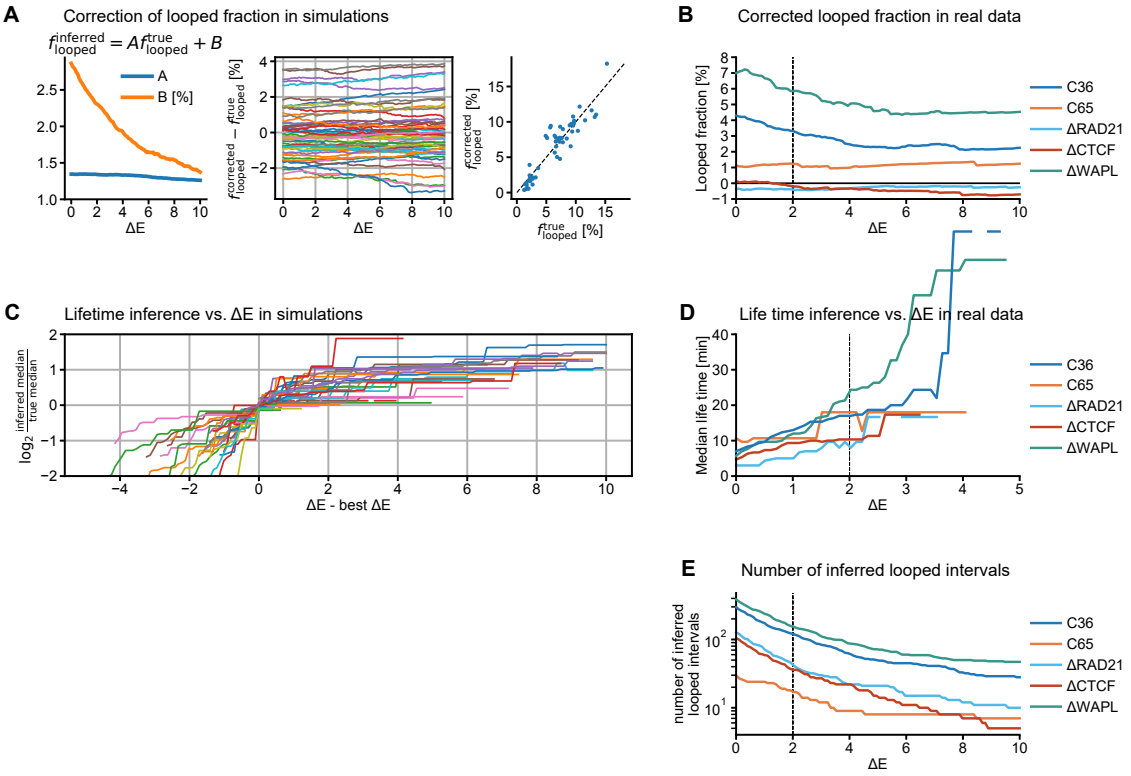


Figure 5.6: Variation of inference results with ΔE. (**A**) Left: quantifying the relationship between true and inferred looped fraction at different ΔE; cf. fig. 5.5, **G**. Middle & Right: using the results displayed on the left to correct the inferred looped fractions gives stable values over a wide range of ΔE, reproducing the ground truth to within two percentage points. "Corrected" values on the right are means over the curves shown in the middle, dashed line indicates identity. (**B**) Applying the same correction to our experimental data. (**C**) Inferred median lifetime vs. ΔE on all simulations in the validation data set. We normalize to the corresponding ground truth median lifetime, defined as the median of the Kaplan-Meier survival curve of the true intervals. Similarly, we match the horizontal coordinates to the ΔE value reproducing that true median lifetime for each simulation. (**D**) Variation of inferred median lifetimes with ΔE on the experimental data shown in fig. 4.3. Note that for C65, ΔRAD21, and ΔCTCF there are very few intervals inferred as looped in the first place, such that the inferred lifetimes are not necessarily particularly meaningful (cf. (B), (E)). (**E**) Number of inferred looped intervals for each condition at different ΔE. This plot shows the total number of looped intervals for each dataset without normalizing for the total number and length of trajectories.

ation with ΔE, hinting that the real data contains errors not accounted for in the simulations. Nevertheless, we find a corrected looped fraction for our WT (C36) cell line of $2-4\%$, with the WAPL degron generally exhibiting an increased looped fraction of three percentage points above WT. The correction also places the looped fractions for the CTCF and RAD21 degrons close to zero, which shows that indeed in these degron conditions there is very little looping, if any at all (fig. 5.6, **B**).

We then investigated the lifetime inference outlined in section 5.6. Specifically, we asked how ΔE affects the median lifetime as estimated from the Kaplan-Meier survival curves, which is our main estimate of loop life time. For each simulation in our validation data set we find the value of ΔE that makes the inferred median lifetime match the ground truth median, measured from the Kaplan-Meier survival curve of the true intervals in the data set. In fig. 5.6, **C** we then show how the inferred lifetime deviates from the true one if ΔE deviates from its optimal value. We find that while too small ΔE can lead to heavy underestimation of the lifetime, with a ΔE somewhat higher than optimal we estimate a lifetime that is at most a factor 2 higher than the true one. We therefore aim to err on the side of high ΔE.

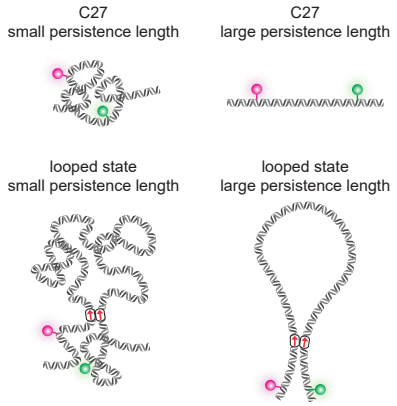
Plotting the inferred lifetime of the real data against ΔE reveals the change to be modest over the region $1 \leq \Delta E \leq 3$ (fig. 5.6, **D**), matching our prior expectation of $\Delta E \approx 2$ (fig. 5.5, **K**). Within that range we find a median loop lifetime in WT (C36) of 10-20 min.

5.8 Epilogue: Choosing a positive looping control to calibrate BILD

To calibrate BILD to infer looped states in trajectories, we need two controls. First, we used ΔCTCF to calibrate the unlooped state: while the looped state is by definition absent, the partially extruded, fully unloped, and transient stochastic proximity states remain. Second, we considered two methods of defining and calibrating the looped state for BILD. We considered using the ΔTAD (C27) control cell line, but ultimately employed a rescaling argument based on the ΔRAD21 condition as mentioned in section 5.4. Here we describe why we chose the rescaling approach over C27 (ΔTAD).

C27 is ill-suited as looping control

The Root-Mean-Square (RMS) distance between the two loci in our C27 data, after subtracting localization error, is 214 nm). This is a considerable distance, given the 10 kb tether between the two loci: using 50 bp/nm as a low estimate for chromatin compaction [107], the tether in C27 would have a contour length of only 200 nm, such that even fully extended it would not reach the observed separation between the probes.



Even if the fiber was locally somewhat decompacted, the large separation observed in C27 would still imply a large persistence length. Critically, this by itself invalidates the use of C27 to measure separation between the probes in the looped state, as illustrated by the sketch on the right.

In designing C27 as an approximation for the looped state we assumed a small persistence length, such that the spatial separation of the two loci is independent of whether they are 10 kb apart on a continuous chromatin fragment or whether there is actually a loop complex in the middle of this tether, to which the 505 kb *Fbn2* loop is "attached" (as in the true looped state; see left column of the sketch). However, if the persistence length were large, the CTCF/cohesin loop complex can affect separation between the probes by determing the direction in which chromatin fibers emanate from the complex. This would further make C27—with its single 10 kb fiber—a problematic proxy for the looped state.

While this argument shows that C27 could not be used to estimate the separation between the probes in the true looped state, even in the absence of other confounding factors, we note that there might also be biological reasons for the large distance in C27, such as epigenetic modifications due to the two fluorescent arrays being in permanent proximity, or detrimental downstream effects of the excision of the whole *Fbn2* TAD.

In conclusion, we found that C27 does not provide a good control for the actual looped state we should expect to see in other cell lines, where the *Fbn2* TAD has not been excised. We therefore calibrated our inference method as described in section 5.4, following the considerations below, and ultimately work with a looped state that has reduced RMS distance compared to C27.

Finding another proxy for the looped state

In determining how to calibrate the looped state in our inference, if not by C27, we followed two orthogonal approaches and found compatible estimates for the spatial separation of the two probes in the looped state.

First, one of the most recent estimates of chromatin fiber structure in eukaryotes (budding yeast) found plausible ranges of 52-82 bp/nm and 53-65 nm for compaction ratio C and persistence length l_p, respectively, given available live- and fixed-cell imaging, as well as Hi-C data [107]. Within the worm-like chain model (WLC; the model in which the quoted parameters were estimated), the mean square spatial separation $R^2(s)$ for some loci separated by the genomic distance s is given by

$$R^2(s) = 2l_p^2 \left[\frac{s}{\rho} - 1 + \mathrm{e}^{-\frac{s}{\rho}} \right], \tag{5.37}$$

where $\rho \equiv C l_p$ is the persistence length in bp. Following the argument outlined in the previous section we should take into account the presence of the loop complex if we want to estimate the distance in the looped state, as opposed to a continuous chromatin fiber (C27-like conformation). The mean square separation for the two loci, given that the chain has a sharp pinch in the middle (cf. sketch above, right column) is given by

$$R_\theta^2(s) = 2l_p^2 \left[\frac{s}{\rho} - 2 - \cos\theta + 2(1+\cos\theta)\mathrm{e}^{-\frac{s}{2\rho}} - \cos\theta \mathrm{e}^{-\frac{s}{\rho}} \right], \tag{5.38}$$

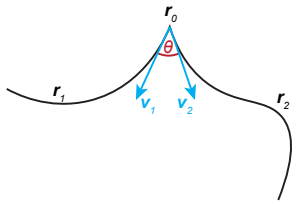
where $\theta \in [0, \pi]$ denotes the pinch angle, i.e. the angle between the two tangent vectors at the pinch / between two fibers emanating from the CTCF/cohesin complex (note how $\theta = \pi$ reproduces the un-pinched fiber of eq. (5.37)); for a derivation of this expression see below. Using the ranges for compaction and persistence quoted above, we find RMS distances $\sqrt{R^2_{\theta=0}(10\,\mathrm{kb})} = 62 - 82\,\mathrm{nm}$ for maximal pinching ($\theta = 0$) and $\sqrt{R^2_{\theta=\pi}(10\,\mathrm{kb})} = 104 - 137\,\mathrm{nm}$ without pinching (i.e. straight 10 kb tether like in C27). We conclude that 70-100 nm should be a reasonable range for the RMS separation in the looped state given this model, taking into account that the pinch due to the looped state is likely not completely sharp.

Second, we used the Rouse model to estimate the RMS distance across a 10 kb tether from the measured distance between the loci in the ΔRAD21 condition. ΔRAD21 is the only condition where this is a viable line of argument, since loop extrusion shortens the chain in all

other conditions. The measured RMS distance of 637 nm for the 515 kb tether in ΔRAD21 rescales to 89 nm for 10 kb. Scaling down even further, we find an RMS separation of 12.6 nm for a tether of 200 bp, corresponding very closely to one nucleosome. Our measurements in the ΔRAD21 condition are thus consistent with chromatin as a random walk of nucleosomes, yielding an estimate for the looped state that is comparable to the WLC estimate in the previous paragraph.

Finally, we benchmarked the choice of looped state on our simulation data. As shown in fig. 4.2, **F**, our 3D polymer simulations faithfully reproduce the experimentally observed contact scaling and can thus serve as a proxy for chromatin structure in the real data. We found that with the looped state calibrated by rescaling the probe distance for simulated ΔRAD21 condition, we were able to capture the true time scale of looping in the simulated WT. On the other hand, calibrating e.g. to C27 lead to significant overestimates of both the looped fraction and loop lifetime. We thus conclude that the calibration based on rescaling ΔRAD21 is an appropriate representation of the looped state, given our (lack of) knowledge of chromatin structure on the 10 kb scale.

Derivation of eq. (5.38)



We aim to derive the mean squared distance between two loci $\boldsymbol{r}_1(\sigma)$ and $\boldsymbol{r}_2(\sigma)$ on a worm-like chain with compaction C (in bp/nm) and persistence length ρ (in bp), given that they are equidistant (but on opposite sides) from the pinch at $\boldsymbol{r}_0$. At the pinch, the unit length tangent vector $\hat{\boldsymbol{v}}_1$ pointing towards $\boldsymbol{r}_1$ stands at a fixed angle θ to the tangent vector $\hat{\boldsymbol{v}}_2$ pointing towards $\boldsymbol{r}_2$, and the chains on both sides of the pinch are independent from each other.

By definition of the WLC, the tangent vectors along the chain are exponentially correlated with correlation length ρ. Thus, given the tangent vector $\hat{\boldsymbol{v}}$ at a position $\boldsymbol{r}_0$, the mean position for a locus $\boldsymbol{r}$ a distance σ further along the chain is

$$\langle \boldsymbol{r}(\sigma) - \boldsymbol{r}_0 \rangle = \frac{\rho}{C} \left(1 - \mathrm{e}^{-\frac{\sigma}{\rho}}\right) \hat{\boldsymbol{v}} \,. \tag{5.39}$$

We then find the mean squared distance MS_{12} between loci 1 and 2 as

$$\mathrm{MS}_{12} = \left\langle (\boldsymbol{r}_1(\sigma) - \boldsymbol{r}_0 + \boldsymbol{r}_0 - \boldsymbol{r}_2(\sigma))^2 \right\rangle \tag{5.40}$$

$$= \left\langle (\boldsymbol{r}_1(\sigma) - \boldsymbol{r}_0)^2 \right\rangle + \left\langle (\boldsymbol{r}_2(\sigma) - \boldsymbol{r}_0)^2 \right\rangle - 2 \left\langle (\boldsymbol{r}_1 - \boldsymbol{r}_0)(\boldsymbol{r}_2 - \boldsymbol{r}_0) \right\rangle \tag{5.41}$$

$$= \frac{2\rho^2}{C^2} \left[\frac{2\sigma}{\rho} - 2 - \cos\theta + 2(1+\cos\theta)\mathrm{e}^{-\frac{\sigma}{\rho}} - \cos\theta \mathrm{e}^{-\frac{2\sigma}{\rho}} \right] . \tag{5.42}$$

In the last line we utilized that the mean square separation between $\boldsymbol{r}_1(\boldsymbol{r}_2)$ and $\boldsymbol{r}_0$ follows the usual formula (5.37) for the WLC in the first two terms, and the expectation value in the third term factorizes into the product of mean positions since the conformations on the two sides of the pinch are independent; the two factors are then given by eq. (5.39). Substituting $\hat{\boldsymbol{v}}_1 \cdot \hat{\boldsymbol{v}}_2 \equiv \cos\theta$ and simplifying yields the stated result. Finally, we substitute $s \equiv 2\sigma$ and $l_p \equiv \frac{\rho}{C}$ to obtain eq. (5.38).

5.9 Multi-state inference: the Conflict-Free Categorical (CFC)

BILD as presented in this chapter infers binary looping profiles: the system can be in a looped or an unlooped state. However, neither of these states is particularly special: at the end of the day, both are simply defined in terms of a specific connectivity matrix B for the Rouse model (section 5.4). As such, it should be quite straight-forward to generalize this inference method to work for an arbitrary collection of states of the associated Rouse model.

Let us consider an inference model with n states, labelled $\theta = 1, \ldots, n$. Possible transitions between these states are encoded in the boolean matrix S, where $S_{\theta\psi} = 1$ if and only if the transition $\theta \to \psi$ is allowed. This allows to enforce for example progressive iteration through the states, if so desired. Note that self-transitions are ill-defined (or lead to ill-defined parametrizations below), such that we always require $S_{\theta\theta} = 0 \, \forall \theta$.

While this setup does resemble a hidden Markov model, we point out that the model is manifestly not Markovian: memory about past states is stored in the polymer conformation. That this memory is indeed essential for accurate inference is demonstrated in section 5.1. Furthermore, we do not assume first order dynamics between the different states, though this is a minor point.

To implement this multi-state approach, we need to make only a few changes to the original (two-state) BILD method as laid out in section 5.3:

- Looping profiles now have to include n states. The space of all possible looping profiles thus becomes

$$\Theta \equiv \left\{ \theta : [0,T] \mapsto \{0,\ldots,n-1\} \;\middle|\; S_{\theta(t-)\theta(t+)} = 1 \,\forall \text{jump times } t \right\}, \tag{5.43}$$

  where we call the condition that state changes in the profile have to be allowed by the transition matrix S *conflict-free*.

- The parametrization in terms of k switches at positions $s_1,\ldots,s_k$ remains the same, except that now we have to explicitly specify the full *state trajectory* $(\theta_0,\ldots,\theta_k)$, i.e. the state assumed during each of the intervals. The full looping profile is then given by

$$\theta(t) = \theta_i \qquad \forall t \in [s_i, s_{i+1}), i = 0,\ldots,k, \tag{5.44}$$

  where $s_0 \equiv 0$ and $s_{k+1} \equiv T$ is the trajectory length.

  The conflict-free condition requires $S_{\theta_i\theta_{i+1}} = 1\,\forall i$; the number of possible state trajectories for k switches is then given by

$$N_{\text{states}} = \sum_{\theta_0=1}^{n} \cdots \sum_{\theta_k=1}^{n} S_{\theta_0\theta_1} S_{\theta_1\theta_2} \cdots S_{\theta_{k-1}\theta_k} \equiv \sum S^k, \tag{5.45}$$

  where the last sum should be understood to run over all entries of the matrix exponential S^k.

- The Rouse likelihood as described in section 2.2.2 natively works for $n > 2$ states; thus, no modification is needed.

- The uniform prior over profiles with k switches now becomes

$$\pi_k(\theta) = \frac{k!}{T^k N_{\text{states}}} = \frac{k!}{T^k \sum S^k}. \tag{5.46}$$

- We need a well-defined proposal distribution for state trajectories. Due to the conflict-free condition this turns out to require some thought and constitutes the rest of this section.

With these minor changes, BILD can be used as inference scheme for arbitrary numbers of states, as implemented in [99].

As noted above, the main challenge for multi-state inference is in the AMIS proposal distribution (5.11). While for two states the state trajectory is fully determined by the initial state θ_0 and we could thus simply use a Bernoulli distribution over θ_0, with $n > 2$ states this becomes more complicated. For one, the state space is now bigger; this is easily incorporated by generalizing the Bernoulli to a categorical distribution. Additionally, however, we have to make sure to respect the transition matrix S; we thus introduce the *Conflict-Free Categorical* distribution $\mathrm{CFC}_{p;S}(\boldsymbol{\theta})$.

The Conflict-Free Categorical (CFC) has to meet two objectives:

- it has to be flexible enough to be a useful proposal distribution in AMIS, and
- it has to respect the transition matrix S, i.e. be a distribution over conflict-free state trajectories.

In practice we need to be able to execute the following tasks:

- sample from the distribution,
- evaluate the distribution at a fixed point (i.e. calculate likelihoods), and
- estimate the parameters of the distribution from a given (weighted) sample.

We meet the flexibility criterion by starting from independent categorical distributions over each θ_i separately:

$$p(\theta_i = \psi) =: p_i^\psi \,, \qquad \sum_\psi p_i^\psi = 1 \,. \tag{5.47}$$

The p_i^ψ now define a distribution over state trajectories that does not necessarily respect the conflict-free condition. To achieve this second objective, we proceed by causal sampling:

- draw θ_0 according to the weights p_0;
- check which transitions are allowed from θ_0 and obtain their weights: $\tilde{p}_1^\psi = S_{\theta_0\psi} p_1^\psi$;
- sample θ_1 according to the renormalized weights $\frac{\tilde{p}_1}{\sum_\psi \tilde{p}_1^\psi}$;
- repeat until end of state trajectory is reached.

This procedure ensures the conflict-free property; obviously it also provides us with a sampling scheme, which was the first of three tasks we need to be able to execute.

The second task is likelihood evaluation, which is straight-forward by inspection of the sampling scheme:

$$p(\boldsymbol{\theta}) = p\left(\theta_k \mid \theta_{k-1}\right) \cdots p\left(\theta_1 \mid \theta_0\right) p\left(\theta_0\right) \tag{5.48}$$

$$= \frac{p_k^{\theta_k}}{\sum_\psi S_{\theta_{k-1}\psi} p_k^\psi} \cdots \frac{p_1^{\theta_1}}{\sum_\psi S_{\theta_0\psi} p_1^\psi} p_0^{\theta_0} \tag{5.49}$$

$$= \frac{p_k^{\theta_k} \cdots p_1^{\theta_1} p_0^{\theta_0}}{\sum_\psi S_{\theta_{k-1}\psi} p_k^\psi \cdots \sum_\psi S_{\theta_0\psi} p_1^\psi} , \tag{5.50}$$

where we assume that $\boldsymbol{\theta}$ is from the domain of the CFC and thus fulfills the conflict-free condition (otherwise we would just have to add appropriate factors of S).

So finally, we are left with estimation from a weighted sample. To that end, we adapt the commonly used method of moments into a "method of marginals": we define an estimator $\hat{p}_i^\psi$ for the parameters of the CFC by matching the marginals to the experimentally observed ones; the latter are straight-forward to compute from a weighted sample. Consider the CFC-marginals

$$f_i^\psi \equiv p\left(\theta_i = \psi\right) . \tag{5.51}$$

The causal sampling allows us to propagate marginals forward, such that one finds

$$f_0^\psi = p_0^\psi ; \qquad f_i^\psi = p_i^\psi \sum_\chi \frac{f_{i-1}^\chi S_{\chi\psi}}{\sum_\varphi S_{\chi\varphi} p_i^\varphi} \quad \forall i \geq 1 . \tag{5.52}$$

Given sample marginals f_i, we can then numerically find estimates $\hat{p}_i$ by fixed point iteration: starting from $\hat{p}_{i;(0)} = f_i$, update

$$\hat{p}_{i;(r+1)} = f_i \left(\sum_\chi \frac{f_{i-1}^\chi S_{\chi\psi}}{\sum_\varphi S_{\chi\varphi} \hat{p}_{i;(r)}^\varphi} \right)^{-1} . \tag{5.53}$$

This provides us with a method to estimate the parameters p_i^ψ from a weighted sample.

We have thus managed to define a suitable AMIS proposal distribution, which we can evaluate, sample from, and estimate. As outlined above, this allows us to generalize BILD to arbitrary number of states $n > 2$. Along the way we could even incorporate free choice of the admissible transitions between these states.

Chapter 6

Bayesian MSD fitting

The Mean Squared Displacement (MSD)[1] of a particle (fig. 6.1) is among the most central quantifications of any single particle tracking experiment. It is frequently reported to resemble a powerlaw, $\mathrm{MSD}(\Delta t) \sim \Delta t^{\alpha}$, which then renders the exponent α a central object of interest, since it characterizes the nature of the observed motion (fig. 6.1, **B**). If $\alpha = 1$ we talk about normal diffusion, i.e. a random walk: between any two time points, the particle simply picks a random direction to move in. In a cellular context one frequently observes $\alpha < 1$, indicating that the motion is somewhat recurrent, i.e. whenever the particle takes a step, it has a tendency to reverse direction on the next step. The extreme case $\alpha = 0$ is the white noise process, which just fluctuates about zero; this would be, for example, the behavior of a trapped particle (over long lag times). Similarly, for a particle on a polymer one would expect $\alpha \approx 0.5$ (eq. (2.87)), since whenever the particle moves somewhere, the rest of the chain tries to pull it back to its original position. $\alpha > 1$ indicates so-called *superdiffusion*, which is rarely observed in the cellular contexts of interest in this thesis, but occurs has been observed e.g. in the foraging behaviors of albatrosses [108] and spider monkeys [109]. Finally, the limiting case of superdiffusion is ballistic behavior with $\alpha = 2$. This indicates perfect correlations in the particle motion, such that it moves in a straight line.

As illustrated, the exponent α carries qualitative information about the type of motion and is therefore widely reported whenever a particle is tracked. Despite this widespread use, significant issues remain:

- Current SPT data does often not have the dynamic range to convincingly show powerlaw

[1] In this work we will always consider the time-averaged MSD, sometimes denoted TA-MSD.

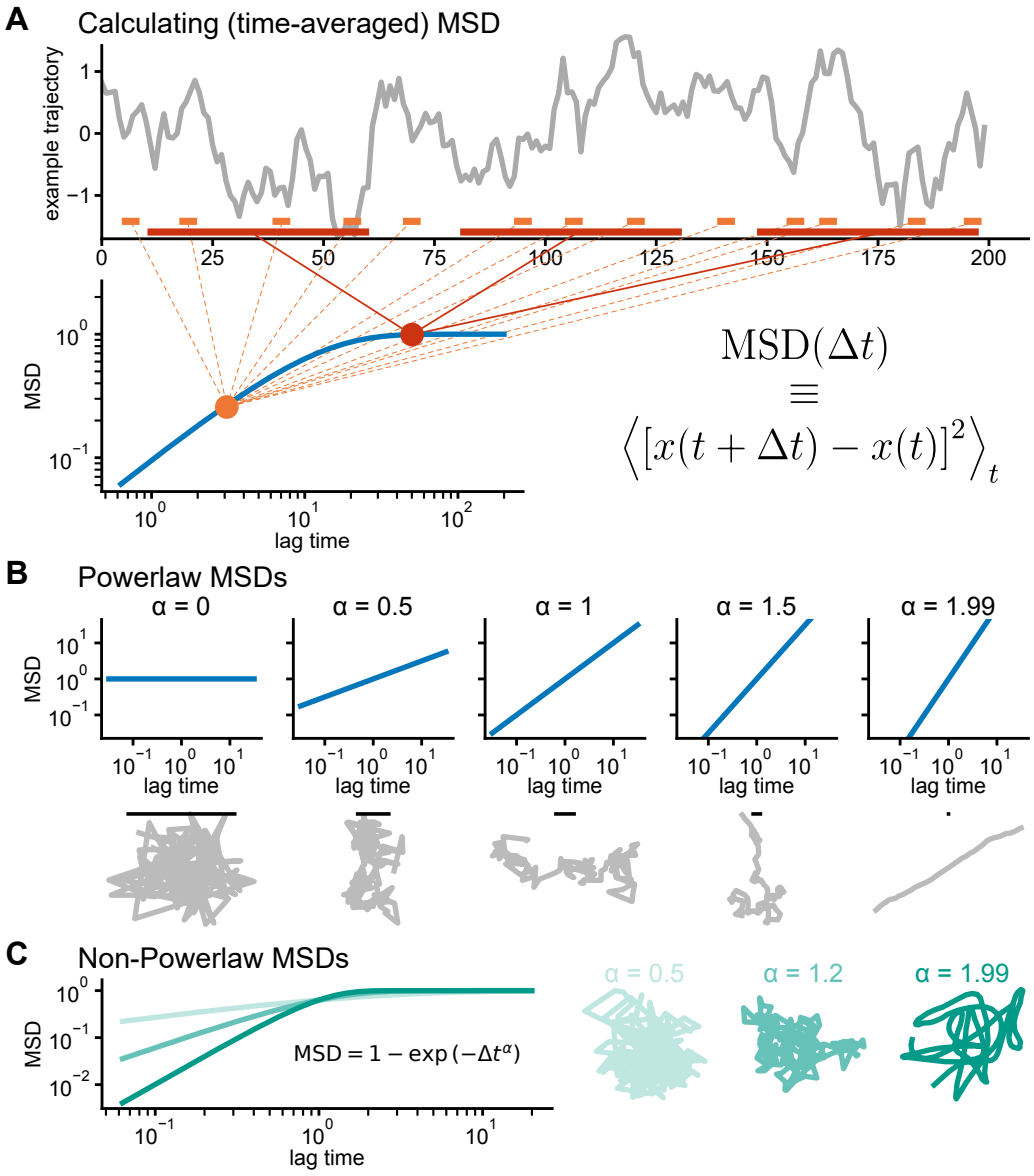


Figure 6.1: Mean Squared Displacement (MSD). (**A**) Definition and calculation of the (time-averaged) MSD. Top: a toy example trajectory (gray). Different points on the MSD curve (blue, bottom) are computed as squared average of the displacement over all possible windows with a given fixed lag time Δt. For MSD curves calculated from finite data, points at longer lag time (red) thus include markedly fewer samples than points at short times (orange). Note also that due to the time averaging any point on the MSD curve depends on *all* experimental data; we should thus expect highly correlated errors. (**B**) Example trajectories (gray, bottom) for powerlaw MSDs with different exponents (blue, top). The black scale bar in the trajectory plots has the same "physical" length for all trajectories, relative to the step size used in sampling. These trajectories are sampled from Gaussian processes with powerlaw MSD, which are known as *fractional Brownian motion*. (**C**) Examples of non-powerlaw MSDs and trajectories sampled from the associated Gaussian process.

scalings, rendering the central assumption that $\mathrm{MSD}(\Delta t) \sim \Delta t^{\alpha}$ questionable.

- Localization error and (potentially) motion blur cause systematic shifts in the MSD curve (section 6.5); they should be properly accounted for. In fact, it is often possible to determine the localization error from the observed data, reducing the reliance on orthogonal experiments for localization error determination (which often struggle to capture the exact acquisition conditions in production).

- Proper parameter inference from MSDs is statistically non-trivial, due to strong error correlations. To this end, note that because of the time averaging, *any* point of the MSD curve has a dependence on *all* the data (fig. 6.1, **A**). The field has therefore adopted heuristics like "only fit the first 10 data points of the MSD curve" [110]. Another instance of this correlation effect is described in [111]: for a diffusive process, the best estimator of the diffusion constant is exactly the first point of the MSD; fitting more of the curve makes the estimator *worse*.

- In line with the previous point, the statistical properties of MSD-based estimators are currently under so little control, that results are customarily reported without any error estimates. This makes it impossible to judge the significance of experimental discrepancies.

It thus seems desirable to get a better understanding of what we can and cannot learn from SPT data in terms of their MSD. This chapter lays out a rigorous approach to this question, based on the theory of Gaussian processes [112]. The basic idea is that for a suitably stationary Gaussian process (section 6.3), the MSD is a sufficient statistic (section 6.3.3). If the stationarity assumption is warranted, the MSD thus captures all relevant information about the data, up to second order; it should thus indeed be the central object of interest, e.g. for fitting parametric models to the data. Assuming that the data is well represented by a Gaussian process, we then lay out a Bayesian scheme for parameter inference (section 6.6). Since this approach fits the full correlation structure of the data, instead of just the plot of the MSD curve, it naturally takes into account the error correlations mentioned above. We can thus be confident about the statistical properties of the estimator; in fact, the Bayesian approach allows calculating credible intervals for all parameter estimates, which are quite notably lacking from the current literature.

Most of the treatment here does not rely on the Gaussianity assumption; that is needed "only" to define the likelihood for the actual inference in section 6.6 (which, of course, makes it

a central assumption in the inference part). To clearly distinguish where we do and do not need to assume Gaussianity, most of the chapter is formulated using the term "Gaussian-equivalent process" (definition 4). This should be thought of simply as a stochastic process that is defined only up to second order, i.e. we are agnostic about higher order moments.

A python implementation of the method developed in this chapter is provided in the package `bayesmsd` [113].

6.1 Overview

We are concerned with the inference problem for Gaussian-equivalent processes: given some trajectory data $\{x_i(t) \mid i = 1, \dots, N\}$, what are the best fit mean function $\mu(t)$ and covariance kernel $\Sigma(t, s)$? That parameter space—as characterized by theorem 1—has multiple functional degrees of freedom and is thus significantly larger than would be feasible to handle in practice. We therefore introduce additional assumptions to restrict the inference to a manageable set of parameters *a priori*.

In a first step we make a stationarity assumption, which can take one of two forms:

- we might want to assume the process $X(t)$ itself to be stationary (e.g. Ornstein-Uhlenbeck: particle in a harmonic potential; distance between two points on a fluctuating polymer; etc.), or
- assume only the *increments* of the process $X(t)$ to be stationary (e.g. freely diffusing tracer particle; individual chromosomal loci moving subdiffusively; etc.)

To clearly distinguish these two cases—while at the same time treating them under the same framework as far as possible—we refer to these as stationarity at level 0 and 1, respectively[2].

Either of these stationarity assumptions will reduce the parameter space to a single functional degree of freedom, which turns out to be the MSD. In a second step, we can then introduce parametric expressions for the MSD to reduce this functional degree of freedom to a finite number of real parameters.

[2]The idea here is to count differencing operations: level 0 means $X(t)$ is stationary, level 1 means $\Delta X(t)$ is stationary; in principle one might even consider higher level stationarity (e.g. stationary $\Delta\Delta X(t)$), but for those cases the MSD might not be the appropriate quantification anymore, since it relies on first differences. Since processes with stationarity at level > 1 also do not seem to have much (or any?) practical relevance, we will concern ourselves with levels 0 and 1 only.

Finally, we define a likelihood function over these remaining parameters by choosing for each Gaussian-equivalent process the corresponding Gaussian process as representative. This allows us to run standard Bayesian inference over the chosen parameter space.

Let us begin our theoretical treatment with the following central definition:

Definition 1 (MSD)**.** *The* mean square displacement (MSD) *of a stochastic process $X(t)$ is the function*

$$\psi(\Delta t) := \left\langle (X(t+\Delta t) - X(t))^2 \right\rangle . \tag{6.1}$$

Note that we implicitly assumed that the expectation value on the right hand side will be independent of absolute time t; this will be true (by construction) for level 0 and 1 stationary processes as considered below, but does not hold in general. This definition is thus of limited use until we introduce the rest of the framework below.

6.2 Gaussian-equivalent processes

Definition 2. *For a stochastic process $X(t)$, we define the*

$$\textit{mean function} \qquad \mu(t) \equiv \langle X(t) \rangle \qquad \textit{and} \tag{6.2}$$

$$\textit{covariance kernel} \qquad \Sigma(t,s) \equiv \langle X(t)X(s) \rangle_{\mathrm{c}} . \tag{6.3}$$

Definition 3. *We say two stochastic processes X and Y are* equivalent to second order *and write $X \sim Y$, if they have the same mean function and covariance kernel.*

Definition 4. *A* Gaussian-equivalent process *is an equivalence class of stochastic processes under the "equivalence to second order" of definition 3.*

By construction, any Gaussian-equivalent process is given by its mean function and covariance kernel. While essentially any function can be the mean of a Gaussian-equivalent process, the covariance kernel is more tightly constrained.

Definition 5 (Positive definite function; following [112])**.** *A function $f : \mathbb{R}^2 \to \mathbb{R}$ is said to be* positive (semi-)definite *if, for any finite set $\{t_i \mid i = 1, \ldots, n\}$, $n \in \mathbb{N}$, the matrix*

$$f_{ij} \equiv f(t_i, t_j) \tag{6.4}$$

is positive (semi-)definite.

Theorem 1. *The covariance kernel of a stochastic process $X(t)$ is a symmetric, positive definite function $\Sigma(t,s)$. Conversely, any pair $(\mu(t), \Sigma(t,s))$ with $\mu : \mathbb{R} \to \mathbb{R}$ arbitrary and $\Sigma : \mathbb{R}^2 \to \mathbb{R}$ symmetric positive definite uniquely defines a Gaussian-equivalent process.*

Proof. The proof relies on actual Gaussian processes as representatives of Gaussian-equivalent processes, cf. definition 11 and theorem 6. We will therefore only sketch it here.

- $\Sigma(t,s)$ being symmetric and positive definite for any $X(t)$ directly follows from the definition of $\Sigma(t,s)$.
- By definition of Gaussian-equivalent process, if the pair $(\mu(t), \Sigma(t,s))$ do define a Gaussian-equivalent process, then it is unique.
- What remains to be shown, then, is that indeed for any $(\mu(t), \Sigma(t,s))$ satisfying the constraints of the theorem, a Gaussian-equivalent process exists. This is where we take recourse to Gaussian processes as representatives of Gaussian-equivalent processes: quite unsurprisingly, each Gaussian-equivalent process has exactly one representative that is actually a Gaussian process (theorem 6). By construction, $\Sigma(t,s)$ being positive definite guarantees positive definiteness of the covariance matrix for any finite dimensional distribution, which is thus a well-defined Gaussian. Thus, the Gaussian process with $\mu(t)$ and $\Sigma(t,s)$ as in the theorem is well-defined and represents the corresponding Gaussian-equivalent process.

□

6.3 Stationarity assumptions

The key results of this section are theorems 2 and 5, which establish that stationary (at level 0 or 1) Gaussian-equivalent processes are fully specified by their MSD $\psi(\Delta t)$ (plus a constant and—for level 1 stationarity—a random variable; both of these will mostly be set identically zero such that we are not particularly concerned with them).

6.3.1 Level 0

Definition 6 (Level 0 stationarity)**.** *A stochastic process $X(t)$ is* stationary (at level 0) *if all its finite dimensional distributions are independent of absolute time. That is, for any finite*

collection $\{(t_i, x_i)\}_{i\in I}$, $|I| < \infty$ *and* $\Delta t \in \mathbb{R}$ *we have*

$$P\{X(t_i + \Delta t) = x_i \,\forall i \in I\} = P\{X(t_i) = x_i \,\forall i \in I\} \,. \tag{6.5}$$

Definition 7 (Stationarity of Gaussian-equivalent processes[3])**.** *A Gaussian-equivalent process is called stationary (at level 0) if its mean function and covariance kernel are independent of absolute time:*

$$\mu(t) = m \qquad \forall\, t\,, \tag{6.6}$$

$$\Sigma(t, s) = \gamma(|t - s|) \qquad \forall\, t, s\,. \tag{6.7}$$

In this work, we will always assume decaying correlations: $\lim_{\Delta t\to\infty} \gamma(\Delta t) = 0$. This allows us to avoid pathological cases like this

Example 1 (Non-ergodic, level 0 stationary Gaussian process)**.** *Consider the process* $X(t) := Y\,\forall t$*, with* $Y \sim \mathcal{N}(\mu, \sigma^2)$ *some normally distributed random variable. This is a level 0 stationary Gaussian process with mean* $m = \mu$ *and covariance* $\gamma(\Delta t) = \sigma^2$*.*

Theorem 2. *A level 0 stationary Gaussian-equivalent process with decaying correlations is fully specified by its mean* m *and MSD* $\psi(\Delta t)$*. Furthermore,* $\psi(\infty) < \infty$*.*

Proof. Writing out the expectation values shows that $\psi(\Delta t) = 2\gamma(0) - 2\gamma(\Delta t)$. Decaying correlations then allow us to identify $\psi(\infty) = 2\gamma(0) < \infty$, such that $\gamma(\Delta t) = \frac{1}{2}(\psi(\infty) - \psi(\Delta t))$. □

For future reference, we also note the following relation, which is similarly obtained by simple expansion of the terms in the expected value and application of $\gamma(\Delta t) = \frac{1}{2}(\psi(\infty) - \psi(\Delta t))$:

$$\begin{aligned} C^{\tau_1,\tau_2}(\Delta t) &\equiv \Big\langle \Big[X(t + \Delta t + \tau_1) - X(t + \Delta t)\Big]\Big[X(t + \tau_2) - X(t)\Big]\Big\rangle_{\mathsf{c}} && (6.8)\\ &= \; -\gamma(\Delta t + \tau_1) - \gamma(\Delta t - \tau_2) + \gamma(\Delta t) + \gamma(\Delta t + \tau_1 - \tau_2) \\ &= \frac{1}{2}\Big[\psi(\Delta t + \tau_1) + \psi(\Delta t - \tau_2) - \psi(\Delta t) - \psi(\Delta t + \tau_1 - \tau_2)\Big]\,. && (6.9) \end{aligned}$$

[3] This is often called *second-order stationarity*; second-order stationary processes (i.e. representatives of a stationary Gaussian-equivalent process) are not necessarily stationary under definition 6 ("strictly stationary"), since higher order moments might depend on time.

6.3.2 Level 1

Definition 8 (Increment process). *For a stochastic process $X(t)$ and time lag $\tau \in \mathbb{R}$ we define the* increment process $(\Delta^\tau X)(t)$ *as*

$$(\Delta^\tau X)(t) \equiv X(t+\tau) - X(t)\,. \tag{6.10}$$

Definition 9 (Increment stationarity). *A process $X(t)$ is called* increment stationary (stationary at level 1) *if all finite dimensional increment distributions are independent of absolute time. That is, for any set $\{(\tau_i, t_i, \Delta x_i) \mid i \in I\}$, $|I| < \infty$ and $\Delta t \in \mathbb{R}$ we have*

$$P\left\{(\Delta^{\tau_i} X)(t_i + \Delta t) = \Delta x_i \,\forall\, i \in I\right\} = P\left\{(\Delta^{\tau_i} X)(t_i) = \Delta x_i \,\forall\, i \in I\right\}\,. \tag{6.11}$$

Definition 10 (Gaussian-equivalent increment stationarity). *A Gaussian-equivalent process $X(t)$ is called increment stationary (stationary at level 1) if its mean increments*

$$v^\tau(t) \equiv \langle (\Delta^\tau X)(t) \rangle \tag{6.12}$$

and covariance structure

$$C^{\tau_1,\tau_2}(t,s) \equiv \langle (\Delta^{\tau_1} X)(t)\,(\Delta^{\tau_2} X)(s) \rangle_{\mathrm{c}} \tag{6.13}$$

are independent of absolute time t. In that case, we write

$$v^\tau(t) \equiv v^\tau\,, \tag{6.14}$$

$$C^{\tau_1,\tau_2}(t,s) \equiv C^{\tau_1,\tau_2}(|t-s|)\,. \tag{6.15}$$

Theorem 3. *The drift v^τ of an increment stationary process $X(t)$ is a linear function of the lag time τ.*

Proof. Consider

$$v^{\tau_1+\tau_2} = \langle X(t+\tau_1+\tau_2) - X(t+\tau_1) + X(t+\tau_1) - X(t) \rangle = v^{\tau_1} + v^{\tau_2} \qquad \forall\, \tau_1, \tau_2\,. \tag{6.16}$$

Thus, v^τ is linear in the argument τ. □

We can thus without loss of generality write $v^\tau \equiv v\tau$. A similarly strong constraint holds for the covariance structure (see also eq. (6.9)):

Theorem 4. *The covariance structure $C^{\tau_1,\tau_2}(\Delta t)$ of an increment stationary process $X(t)$ is equivalent to its MSD $\psi(\Delta t)$:*

$$\psi(\Delta t) = C^{\Delta t,\Delta t}(0)\,, \tag{6.17}$$

$$C^{\tau_1,\tau_2}(\Delta t) = \frac{1}{2}\Big[\psi(\Delta t+\tau_1)+\psi(\Delta t-\tau_2)-\psi(\Delta t)-\psi(\Delta t+\tau_1-\tau_2)\Big]\,. \tag{6.18}$$

Proof. Trivial by writing out the expectation values; we will show eq. (6.18) for reference. For notational simplicity, we introduce $t_i \equiv t+\Delta t$, $t_{i+1} \equiv t+\Delta t+\tau_1 = t_i+\tau_1$, $t_j \equiv t$, and $t_{j+1} \equiv t+\tau_2 = t_j+\tau_2$. Then

$$C^{\tau_1,\tau_2}(\Delta t) \equiv \Big\langle \Big[X(t_{i+1})-X(t_i)\Big]\Big[X(t_{j+1})-X(t_j)\Big]\Big\rangle_{\mathsf{c}} \tag{6.19}$$

$$= \Big\langle X(t_{i+1})X(t_{j+1}) - X(t_i)X(t_{j+1}) - X(t_{i+1})X(t_j) + X(t_i)X(t_j)\Big\rangle_{\mathsf{c}} \tag{6.20}$$

$$= \frac{1}{2}\Big\langle \begin{array}{lll} 2X(t_{i+1})X(t_{j+1}) & -X(t_{i+1})^2 & -X(t_{j+1})^2 \\ -2X(t_i)X(t_{j+1}) & +X(t_i)^2 & +X(t_{j+1})^2 \\ -2X(t_{i+1})X(t_j) & +X(t_{i+1})^2 & +X(t_j)^2 \\ +2X(t_i)X(t_j) & -X(t_i)^2 & -X(t_j)^2 \end{array}\Big\rangle_{\mathsf{c}} \tag{6.21}$$

$$= \frac{1}{2}\Big[-\psi(t_{i+1}-t_{j+1})+\psi(t_i-t_{j+1})+\psi(t_{i+1}-t_i)-\psi(t_i-t_i)\Big] \tag{6.22}$$

$$= \frac{1}{2}\Big[\psi(\Delta t+\tau_2)+\psi(\Delta t-\tau_1)-\psi(\Delta t)-\psi(\Delta t+(\tau_2-\tau_1))\Big]\,. \tag{6.23}$$

□

This finally brings us to our desired

Theorem 5. *A level 1 stationary Gaussian-equivalent process $X(t)$ is fully specified by its initial value $X(0)$, drift v, and MSD $\psi(\Delta t)$.*

Proof. $X(t)$ is given by its initial value and increment processes: $X(t) = X(0) + (\Delta^t X)(0)$. The increment processes in turn are fully specified by v and $\psi(\Delta t)$, according to theorems 3 and 4. □

Note that $X(0)$ in the above theorem is still a random variable in principle. For the purpose

of this work, however, we will just set $X(0) \equiv 0$. Similarly, we will mostly work with $v = 0$, but include it here for completeness.

6.3.3 Level 0 vs. Level 1

Summarizing theorems 2 and 5, we see that a Gaussian-equivalent process is mostly specified in terms of its MSD $\psi(\Delta t)$; we will set the remaining terms to zero for this section (for level 0: $m = 0$; level 1: $v = 0$ and $X(0) \equiv 0$). So given a specific MSD $\psi(\Delta t)$, can we judge whether we should think of this in terms of a level 0 or level 1 stationary process?

Essentially, yes:

- level 0 stationarity requires $\psi(\infty) < \infty$, so any MSD that grows indefinitely ($\psi(\infty) = \infty$) can only describe a level 1 stationary process.
- MSDs with $\psi(\infty) < \infty$ can describe level 0 or level 1 stationary processes. The difference then lies in the initial value, which for level 1 stationarity is independent of the process at other times (and we usually fix it to 0). This means we can think about a level 1 stationary process with $\psi(\infty) < \infty$ as sampling from the corresponding level 0 stationary process and then shifting the trajectory to match the specified value at $t = 0$. So while possible in principle, this seems quite unnatural.

So in summary, while not technically a strict statement, the intuition is that

$$\psi(\infty) < \infty \quad \Leftrightarrow \quad \text{level 0}\,, \tag{6.24}$$

$$\psi(\infty) = \infty \quad \Leftrightarrow \quad \text{level 1}\,; \tag{6.25}$$

technically accurate are only $\Leftarrow$ in the first and $\Rightarrow$ in the second line.

6.4 Parametrizations

In section 6.3 we established that stationary (level 0 or 1) Gaussian-equivalent processes are mostly specified by their MSD $\psi(\Delta t)$. This is still a functional degree of freedom, i.e. the parameter space is infinite dimensional[4]. We therefore customarily restrict this parameter space

[4]The parameter space is infinite dimensional, but not completely unconstrained: the MSD $\psi(\Delta t)$ still has to be such that the resulting covariance kernel $\Sigma(t, s)$ is positive definite. This is an opaque constraint, in that we are not aware of a more elementary formulation in general. Only for specific parametrizations this might translate

to a subset of MSD curves that can be described by a finite number of real parameters (e.g. powerlaws). Which parametrization to use is a modelling choice; `bayesmsd` therefore provides a few default parametrizations that might be useful in different scenarios, but also facilitates implementation of custom parametrizations depending on the problem at hand.

This section will describe the parametrizations included in the `bayesmsd` package [113].

6.4.1 `SplineFit`

This is a problem-agnostic (or "model-free") parametrization of MSD functions $\psi(\Delta t)$. We choose a number n of spline nodes, use their coordinates as parameter set, and define $\psi(\Delta t)$ by the cubic spline interpolation through these points.

Parameters. We define the spline interpolation (and the supporting nodes) in the transformed coordinates

$$x \equiv \begin{cases} \dfrac{4}{\pi} \arctan \dfrac{\log \Delta t - \log \Delta T_{\mathsf{min}}}{\log \Delta T_{\mathsf{max}} - \log \Delta T_{\mathsf{min}}} & \text{(level 0 stationarity)} \\ \dfrac{\log \Delta t - \log \Delta T_{\mathsf{min}}}{\log \Delta T_{\mathsf{max}} - \log \Delta T_{\mathsf{min}}} & \text{(level 1 stationarity)} \end{cases} , \quad y \equiv \log \psi(\Delta t) , \quad (6.26)$$

where ΔT_{min} is the minimum time between two frames (i.e. the inverse frame rate), while ΔT_{max} is the maximum time between two frames (i.e. the maximum trajectory length).

Note the following special values for the coordinate x as function of Δt:

$$x\left(\Delta T_{\mathsf{min}}\right) = 0 , \qquad x\left(\Delta T_{\mathsf{max}}\right) = 1 , \qquad x\left(\infty\right) = \begin{cases} 2 & \text{(level 0 stationarity)} \\ \infty & \text{(level 1 stationarity)} \end{cases} . \quad (6.27)$$

In principle, we need to describe $\psi(\Delta t)$ only for $\Delta t \in [\Delta T_{\mathsf{min}}, \Delta T_{\mathsf{max}}]$, i.e. $x \in [0, 1]$. For level 0 stationarity, however, $\psi(\Delta t \to \infty) < \infty$ is an important additional fit parameter, meaning that $\Delta t = \infty$ has to be accessible numerically. This motivates the parametrization in eq. (6.26), which ensures that $\Delta t = \infty \Leftrightarrow x = 2$ for level 0 stationarity.

Finally, the parameters for this fit are simply the coordinates $\{(x_i, y_i) \mid i = 1, \ldots, n\}$ for the

to bounds or constraints on the parameters: e.g. for powerlaw MSDs $\psi(\Delta t) = \Gamma\,|\Delta t|^{\alpha}$, positive definiteness constrains $\alpha \in [0, 2)$.

n spline nodes. To ensure coverage of the whole domain, we fix

$$x_1 = 0 \qquad\qquad x_n = \begin{cases} 2 & \text{(level 0 stationarity)} \\ 1 & \text{(level 1 stationarity)} \end{cases}, \tag{6.28}$$

such that ultimately we are left with $2n - 2$ independent fit parameters.

Boundary conditions. In addition to the nodes $\{(x_i, y_i) \mid i = 1, \ldots, n\}$, the spline interpolation needs two more constraints, which are usually written as boundary conditions; we require the following:

$$\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=0} = 0\,, \qquad\qquad \begin{cases} \left.\dfrac{\partial y}{\partial x}\right|_{x=2} = 0 & \text{(level 0 stationarity)} \\ \left.\dfrac{\partial^2 y}{\partial x^2}\right|_{x=1} = 0 & \text{(level 1 stationarity)} \end{cases}. \tag{6.29}$$

For level 1 stationarity, the boundary conditions in eq. (6.29) can be summarized as requiring a vanishing second derivative[5], which means that $y(x)$ is naturally extrapolated beyond the domain $x \in [0, 1]$ by linear functions on both ends. Since the coordinates x and y in this case are essentially the logarithms $\log \Delta t$ and $\log \psi$, this means that we choose the boundary conditions such that the MSD is naturally extrapolated beyond the existing data by powerlaws both for shorter and longer lag times.

For level 0 stationarity, the MSD has to plateau for $\Delta t \to \infty$, which in terms of our coordinates x and y means vanishing derivative of $y(x)$ at $x = 2$. On the short lag time end of the curve, we simply retain the natural boundary condition from level 1 stationarity.

Definition of $\psi(\Delta t)$. Given the spline points $\{(x_i, y_i) \mid i = 1, \ldots, n\}$ and boundary conditions (6.29), we can define a continuous function $y(x)$ by cubic spline interpolation [114, 115]. We then simply set

$$\psi(\Delta t) \equiv \exp y\big(x(\Delta t)\big)\,. \tag{6.30}$$

Note that the interpolation is performed in the coordinate space (x, y), i.e. the different sections of $\psi(\Delta t)$ will not be cubic polynomials themselves.

[5] often called "natural boundary condition" in the context of splines

Model selection. One major application of splines for MSD fitting is dimensionality reduction, i.e. understanding which features of an empirically calculated MSD curve are significant and which are just noise. To that end, one can perform model selection over the number of spline points n, thereby finding the "best" representation of the data, trading off model complexity (number of spline points) with goodness of fit (model likelihood). We commonly use the Akaike Information Criterion (AIC) for this purpose, though other methods could be considered.

6.4.2 `NPXFit`

"NPX" stands for "Noise + Powerlaw + arbitrary" and fundamentally constitutes a powerlaw MSD. The "arbitrary" part refers to a spline extension at long times, which allows representing MSDs that are assumed to match a powerlaw only at short times (e.g. if we want to fit a level 0 stationary process, where the MSD has to plateau at long times).

Noise. This refers to localization error and motion blur, and follows the generic treatment presented in section 6.5. With the fractional exposure time f fixed experimentally, this is described by the single parameter σ, the localization error.

Powerlaw. The central component of this parametrization is the powerlaw, i.e.

$$\psi(\Delta t) = \Gamma \, |\Delta t|^{\alpha} \ . \tag{6.31}$$

The two parameters for this part are the *prefactor* ("anomalous diffusion constant") $\Gamma \in \mathbb{R}_+$ and the *exponent* $\alpha \in [0, 2)$.

X: arbitrary. This parametrization allows to specify a number n of spline points that are be used to describe deviations from the powerlaw behavior (6.31) at long times. This reuses much of the `SplineFit` framework described in section 6.4.1, specifically the coordinates x and y defined in eq. (6.26).

Given the number n, we construct a spline with $n+1$ points, which are given by coordinates $\{(x_i, y_i) \mid i = 0, \ldots, n\}$. Since now the spline should cover only part of the time domain (at early times we have the powerlaw behavior of eq. (6.31)), x_0 is now free and determines the transition from powerlaw to spline.

At the transition point x_0, we can use the spline coordinate y_0 and the early time boundary condition to enforce continuity of the MSD and its first derivative. This gives

$$y_0 = \log \Gamma + \alpha \log \Delta t|_{x=x_0} \tag{6.32}$$

for continuity of the MSD and

$$\alpha \overset{!}{=} \left.\frac{\partial \log \psi}{\partial \log \Delta t}\right|_{x=x_0} = \left[\Delta t \frac{\partial y}{\partial x}\frac{\partial x}{\partial \Delta t}\right]_{x=x_0} \tag{6.33}$$

for continuity of the first derivative. Taking into account the different coordinate transformations $x(\Delta t)$ for level 0 and 1 stationary processes from eq. (6.26) we reformulate eq. (6.33) as proper boundary condition:

$$\left.\frac{\partial y}{\partial x}\right|_{x=x_0} = \begin{cases} \dfrac{\pi\alpha}{4}\left[1 + \left(\tan\dfrac{\pi x_0}{4}\right)^2\right] \log \dfrac{\Delta T_{\mathsf{max}}}{\Delta T_{\mathsf{min}}} & \text{(level 0 stationarity)} \\ \alpha \log \dfrac{\Delta T_{\mathsf{max}}}{\Delta T_{\mathsf{min}}} & \text{(level 1 stationarity)} \end{cases}. \tag{6.34}$$

Summary. Taking all three parts together, this parametrization has a total of $3 + 2n$ parameters, $n \in \mathbb{N}_0$. These are the localization error σ, the prefactor Γ, exponent α, and spline points $\{(x_i, y_i) \mid i = 0, \ldots, n\}$ with y_i and x_n fixed as outlined above.

6.4.3 `TwoLocusRouseFit`

This is an example of a fully model-driven MSD parametrization and in fact the origin of this Bayesian approach to MSDs. We used it in chapter 4 to fit a Rouse model to two-locus tracking data.

For a polymer with time dependent conformation $R(t, s)$ (s being the coordinate along the backbone of the polymer), we are interested in the dynamics of the two loci s_1 and s_2 relative to each other. Thus, the stochastic process of interest is $X(t) \equiv R(t, s_2) - R(t, s_1)$. Under the Rouse model of polymer dynamics (section 2.3), this is a Gaussian process and its MSD is given by eq. (2.88), reproduced here for convenience:

$$\psi(\Delta t) = 2\Gamma\sqrt{\Delta t}\left(1 - \mathrm{e}^{-\frac{t}{\tau}}\right) + 2J \operatorname{erfc}\sqrt{\frac{t}{\tau}}, \tag{6.35}$$

where Γ and J are free parameters, while $\tau \equiv \frac{1}{\pi}\left(\frac{J}{\Gamma}\right)^2$ is just notational convenience.

On top of this model we add imaging artifacts (localization error and motion blur) following section 6.5, thus bringing the total parameter count up to three: σ (localization error), Γ, and J.

6.5 Imaging artifacts

In general, when analyzing SPT data, we are looking at a convolution of the physical process of interest with imaging artifacts like localization error (due to finite photon count) and motion blur (due to finite exposure time). Both of these mean that the process we use to model the *observed* data should be slightly modified from the theoretical—"physical"—process we would usually want to model on the data. For a Gaussian-like process this means that the MSD will be modified from the shapes given in section 6.4; this section aims at understanding what these modifications should be.

Let us assume that we image the stochastic process $X(t)$ at discrete times $\{t_n\}$, thus obtaining the collection of random variables

$$Y_n \equiv \int_0^{\Delta T} \mathrm{d}\theta\, \zeta(\theta) X(t_n - \theta) + \sigma\xi_n\,, \tag{6.36}$$

where $\Delta T \equiv \min_{m,n|m\neq n} |t_m - t_n|$ is the minimum separation between time points, the *shutter function*[6] $\zeta : [0, \Delta T] \to \mathbb{R}_+$ satisfies $\int_0^{\Delta T} \mathrm{d}\theta\, \zeta(\theta) = 1$, σ is the standard deviation of localization error, and ξ_n are uncorrelated standard normal random variables.

Note that we did not assume even spacing of the $\{t_n\}$, although this is overwhelmingly the most common imaging modality. However, experimental data may have gaps (missing frames), which makes the spacing uneven (though still integral multiples of ΔT). For sake of generality we then simply assume fully unevenly spaced $\{t_n\}$. We then, however, have to take care of a few details; the following list should be taken as footnotes to the calculations below.

We define $\tau_{mn} \equiv t_m - t_n$ and note that

- $\tau_{km} + \tau_{mn} = \tau_{kn}$, since the τ are signed.
- We require $\tau_{mn} \neq 0\, \forall m \neq n$, i.e. differently indexed times should be different.

[6] Here we are assuming that light emission by the fluorophore is continuous in time. However, considering single photons arriving as inhomogeneous Poisson process with time dependent rate $\lambda(\tau)$ reproduces this treatment in expectation, with the shutter function $\zeta(\tau) \propto \lambda(\tau)$. We omit the discrete treatment here for brevity.

- For $n \neq m$, the integration domains for Y_m and Y_n in eq. (6.36) should not overlap, otherwise the localization error term would pick up correlations. This motivates the choice of $\Delta T = \min_{m,n|m\neq n} |\tau_{mn}|$.

- Make sure to distinguish continuous time variables and discrete indices. This will be especially important in the following study of increments, where clearly the "increment process" should be defined with a lag*time*, not a lag*index*; but this lagtime will only be allowed to take values from subsets of $\{\tau_{mn}\}$.

- Below we will write Kronecker δ's over the time indices; note that with the above requirement that $\tau_{mn} \neq 0 \, \forall m \neq n$ these can be expressed purely in terms of τ_{mn}, since $\delta_{mn} = 1 \Leftrightarrow \tau_{mn} = 0$.

6.5.1 Exact calculation of imaging artifacts for general MSD

Since eq. (6.36) is a linear transform, the increments of the process transform similarly:

$$\left(\Delta^{\tau_{mn}} Y\right)_n \equiv Y_m - Y_n = \int_0^{\Delta T} \mathrm{d}\theta \, \zeta(\theta) \Big[X(t_m - \theta) - X(t_n - \theta) \Big] + \sigma \left(\xi_m - \xi_n\right) . \tag{6.37}$$

If $X(t)$ is stationary (at level 0 or 1), its increment correlation takes the form $C_X^{\tau_1,\tau_2}(\Delta t)$. We can similarly calculate

$$\begin{aligned}
C_Y^{\tau_{kl},\tau_{mn}}(\tau_{ln}) &\equiv \Big\langle \left(\Delta^{\tau_{kl}} Y\right)_l \left(\Delta^{\tau_{mn}} Y\right)_n \Big\rangle_{\mathsf{c}} && (6.38)\\
&\equiv \Big\langle \left(Y_k - Y_l\right) \left(Y_m - Y_n\right) \Big\rangle_{\mathsf{c}} && (6.39)\\
&= \int_0^{\Delta T} \mathrm{d}\theta \, \mathrm{d}\theta' \, \zeta(\theta)\zeta(\theta') && (6.40)\\
&\qquad \times \Big\langle \Big[X(t_k - \theta) - X(t_l - \theta) \Big] \Big[X(t_m - \theta') - X(t_n - \theta') \Big] \Big\rangle_{\mathsf{c}} \\
&\qquad + \sigma^2 \Big\langle \left(\xi_k - \xi_l\right) \left(\xi_m - \xi_n\right) \Big\rangle_{\mathsf{c}} \\
&= \int_0^{\Delta T} \mathrm{d}\theta \, \mathrm{d}\theta' \, \zeta(\theta)\zeta(\theta') \, C_X^{\tau_{kl},\tau_{mn}}(\tau_{ln} + \theta' - \theta) && (6.41)\\
&\qquad + \sigma^2 \left(\delta_{km} - \delta_{lm} - \delta_{kn} + \delta_{ln}\right) .
\end{aligned}$$

We transform the integration variables as

$$\vartheta \equiv \theta' - \theta \in [-\Delta T, \Delta T] \,, \tag{6.42}$$

$$\bar{\theta} \equiv \frac{\theta' + \theta}{2} \in \left[\frac{|\vartheta|}{2}, \Delta T - \frac{|\vartheta|}{2}\right] \tag{6.43}$$

and use theorem 4 (for level 1 stationarity; but the same relations hold for level 0, cf. eq. (6.9)) to rewrite eq. (6.41) in terms of the MSDs ψ_X and ψ_Y, for $\Delta t > 0$:

$$\psi_Y(\Delta t) = C_Y^{\Delta t, \Delta t}(0) \tag{6.44}$$

$$= 2\sigma^2 + \int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) C_X^{\Delta t, \Delta t}(\vartheta) \tag{6.45}$$

$$= 2\sigma^2 + \frac{1}{2} \int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \Big[\psi_X(\vartheta + \Delta t) + \psi_X(\vartheta - \Delta t) - 2\psi_X(\vartheta)\Big] \,, \tag{6.46}$$

where

$$Z(\vartheta) \equiv \int_{\frac{|\vartheta|}{2}}^{\Delta T - \frac{|\vartheta|}{2}} \mathrm{d}\bar{\theta} \, \zeta\left(\bar{\theta} + \frac{\vartheta}{2}\right) \zeta\left(\bar{\theta} - \frac{\vartheta}{2}\right) . \tag{6.47}$$

Note that $Z(-\vartheta) = Z(\vartheta)$ and $\int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) = 1$ by the normalization of $\zeta(\theta)$. Furthermore, by definition $\psi(-\Delta t) = \psi(\Delta t)$, so we can exploit these parity symmetries and the symmetric ϑ-integration domain to write

$$\int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \psi(\vartheta - \Delta t) = \int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \psi(\Delta t - \vartheta) = \int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \psi(\Delta t + \vartheta) \,, \tag{6.48}$$

such that finally we find

$$\psi_Y(\Delta t) = 2\sigma^2 + \int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \psi_X(\Delta t + \vartheta) - \int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \psi_X(\vartheta) \,. \tag{6.49}$$

Below we will study the individual contributions to this expression in more detail. For intuition, note that

- The first term is the standard additive constant due to localization error
- The second term is a "washed-out" version of the original MSD ψ_X. Since $Z(\vartheta)$ is normalized, a first approximation for this term is $\int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta \, Z(\vartheta) \psi_X(\Delta t + \vartheta) = \psi_X(\Delta t)$. For common use cases, this approximation is actually accurate to within 2.5%; below we will improve upon this rough approximation.

- The third term does not depend on Δt and thus presents another—negative—additive constant. Note that it depends only on the MSD at $\vartheta \in [-\Delta T, \Delta T]$, i.e. lag times below a single frame.

6.5.2 Solution for powerlaw MSDs and "all-or-nothing" illumination

To investigate the behavior of eq. (6.49) in more detail, let us consider MSDs of the shape $\psi_X(\Delta t) = \Gamma\,|\Delta t|^\alpha$, $\alpha \in [0, 2)$ and shutter functions of the shape

$$\zeta(\tau) = \frac{1}{f\Delta T}\Theta\left(f\Delta T - \tau\right) \equiv \begin{cases} \frac{1}{f\Delta T} & \tau \in [0, f\Delta T]\ , \\ 0 & \text{else}\,, \end{cases} \tag{6.50}$$

with $f \in [0, 1]$ the *fractional exposure time*. Direct calculation then shows that

$$Z(\vartheta) = \frac{1}{f\Delta T}\left(1 - \frac{|\vartheta|}{f\Delta T}\right)\Theta\left(f\Delta T - |\vartheta|\right) \qquad \forall f > 0\,, \tag{6.51}$$

while for $f = 0$ (ideal stroboscopic illumination) we find $Z(\vartheta) = \delta(\vartheta)$.

Having found an expression for $Z(\vartheta)$ then allows us to calculate the integrals in eq. (6.49):

$$\int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta\, Z(\vartheta)\psi_X(\Delta t + \vartheta) = \Gamma \int_{-f\Delta T}^{f\Delta T} \frac{\mathrm{d}\vartheta}{f\Delta T}\left(1 - \frac{|\vartheta|}{f\Delta T}\right)|\Delta t + \vartheta|^\alpha \tag{6.52}$$

$$\text{let } \varphi \equiv \frac{f\Delta T}{\Delta t} \text{ and substitute } x \equiv \frac{\vartheta}{f\Delta T} \tag{6.53}$$

$$= \Gamma\,|\Delta t|^\alpha \int_{-1}^{1} \mathrm{d}x\ (1 - |x|)\,|1 + \varphi x|^\alpha \tag{6.54}$$

$$= \Gamma\,|\Delta t|^\alpha\, \frac{|1+\varphi|^{\alpha+2} + |1-\varphi|^{\alpha+2} - 2}{\varphi^2(\alpha+1)(\alpha+2)} \tag{6.55}$$

$$\equiv \psi_X(\Delta t) b(\Delta t, f, \alpha)\,. \tag{6.56}$$

Clearly the integral in the third term of eq. (6.49) is just the same expression, evaluated at $\Delta t = 0$. It can straightforwardly be calculated from eq. (6.52), or obtained as the $\varphi \to \infty$ limit

of eq. (6.55). Taking the latter route here gives

$$\int_{-\Delta T}^{\Delta T} \mathrm{d}\vartheta\, Z(\vartheta)\psi_X(\vartheta) = \Gamma\,|\Delta t|^{\alpha}\,\frac{2\,|\varphi|^{\alpha}}{(\alpha+1)(\alpha+2)} \tag{6.57}$$

$$= \frac{2\Gamma\,|f\Delta T|^{\alpha}}{(\alpha+1)(\alpha+2)} \tag{6.58}$$

$$\equiv 2B\,(\Gamma,\alpha,f)\ . \tag{6.59}$$

Finally, eq. (6.49)—for powerlaw MSDs—reads

$$\psi_Y(\Delta t) = \psi_X(\Delta t)b(\Delta t) - 2B + 2\sigma^2\,, \tag{6.60}$$

where we suppressed the dependence on Γ, α, f of the correction factors b and B.

6.5.3 Approximation for non-powerlaw MSD

In analyzing eq. (6.60), we note that the term $B \equiv \frac{1}{2}\int \mathrm{d}\vartheta\, Z(\vartheta)\psi_X(\vartheta)$ only depends on the "true" MSD ψ_X at very short time lags, namely time lags shorter than the exposure time $f\Delta T$ (since $Z(\vartheta) \neq 0$ only for $\vartheta \in (-f\Delta T, f\Delta T)$). In fitting experimental data, therefore, by construction we cannot give a good estimate for B, unless we make some assumption about how the MSD we are currently fitting extrapolates to lag times shorter than the exposure time. We will assume that on these sub-frame time scales, the "true" MSD ψ_X is well described by a powerlaw with an *effective short time scaling exponent* $\alpha_0 \equiv \lim_{\Delta t\to 0}\left(\Delta t \frac{\partial}{\partial \Delta t}\log\psi_X(\Delta t)\right)$. The correction term B in eq. (6.60) then in general reads

$$B = \frac{\psi_X(f\Delta T)}{(\alpha_0+1)(\alpha_0+2)}\,. \tag{6.61}$$

At first sight, this simple treatment in terms of an effective short time scaling exponent α_0 should not work for the $b(\Delta t)$ correction term in eq. (6.60), since this expression manifestly depends on the MSD at finite lag times. However, evaluating $b(\varphi,\alpha)$ numerically (fig. 6.2) shows that it is quite close to unity in general. Furthermore, relevant contributions essentially decay completely over the first 10 exposure times (note that this might be significantly shorter than 10 frames, if $f < 1$); so while in principle this term should be calculated from the local behavior of the MSD ψ_X at Δt, we can quite reasonably replace that with just the short time behavior, since the term is irrelevant elsewhere anyways. Thus, also for non-powerlaw MSDs,

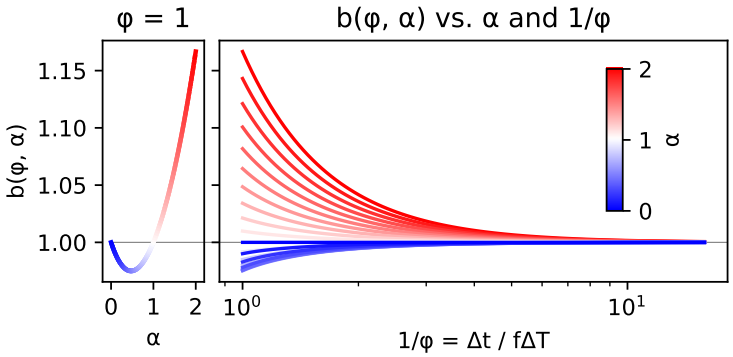


Figure 6.2: The correction term $b(\Delta t, f, \alpha)$. Equation (6.64) as function of $\varphi^{-1} \equiv \frac{\Delta t}{f\Delta T}$ for various α (right) and exponent α for fixed $\varphi = 1$ (left). Note how b is generally close to unity and in either case decays quickly over a few exposure times $\varphi^{-1} = \mathcal{O}(1)$.

we use the same correction term $b(\Delta t)$ as we would obtain for a powerlaw with exponent α_0:

$$b(\Delta t) \approx \frac{|1+\varphi|^{\alpha_0+2} + |1-\varphi|^{\alpha_0+2} - 2}{\varphi^2(\alpha_0+1)(\alpha_0+2)}, \qquad \text{with} \qquad \varphi \equiv \frac{f\Delta T}{\Delta t}. \tag{6.62}$$

6.5.4 Summary

If we observe particles whose dynamics are described by a stochastic process $X(t)$ (stationary at level 0 or 1) at discrete time points spaced by at least ΔT, the MSD of the observed process Y is given by eq. (6.49). Further assuming that illumination is constant over the exposure time $f\Delta T$ ($f \in [0,1]$) and making minor approximations for simplicity, this reduces to

$$\psi_Y(\Delta t) = b(\Delta t)\psi_X(\Delta t) - 2B + 2\sigma^2, \tag{6.63}$$

with

$$b(\Delta t) \equiv \frac{|1+\varphi|^{\alpha_0+2} + |1-\varphi|^{\alpha_0+2} - 2}{\varphi^2(\alpha_0+1)(\alpha_0+2)} \qquad \left(\varphi \equiv \frac{f\Delta T}{\Delta t}\right), \tag{6.64}$$

$$B \equiv \frac{\psi_X(f\Delta T)}{(\alpha_0+1)(\alpha_0+2)}, \tag{6.65}$$

$$\alpha_0 \equiv \left.\frac{\partial \log \psi_X}{\partial \log \Delta t}\right|_{\Delta t \to 0}. \tag{6.66}$$

Note that eq. (6.63) applies only for $\Delta t > 0$. By definition of MSD, $\psi_Y(0) = 0$, even though we might have $\lim_{\Delta t \to 0} \psi_Y(\Delta t) \neq 0$ according to eq. (6.63); the observed MSD $\psi_Y(\Delta t)$ does not have to (and usually will not) be continuous at $\Delta t = 0$.

6.6 Bayesian inference of parametrized MSDs

In section 6.3 we saw that stationary (level 0 or 1) Gaussian-equivalent processes are essentially characterized in terms of their MSD $\psi(\Delta t)$. We then introduced different parametrizations of such MSD functions in section 6.4, such that for the purpose of this section we can always assume a finite set of parameters θ; we will write $\psi = \Psi(\theta)$. Here we are now concerned with inferring the parameters θ from observed data D, i.e. a set of N trajectories $D \equiv \left\{ x^i(t^i_j) \,\middle|\, j = 1, \ldots, n_i \,,\, i = 1, \ldots, N \right\}$ (n_i denoting the number of frames for trajectory i).

The logic of this section is quite straight-forward: we first establish theorem 6, which states that any Gaussian-equivalent process has exactly one representative that is actually a Gaussian process. We then utilize the likelihood function derived from these Gaussian processes to perform Bayesian parameter inference.

Definition 11 (Gaussian Process)**.** *A stochastic process $X(t)$ is a* Gaussian process*, if all its finite dimensional distributions are Gaussian. That is, for any finite collection $\{(t_i, x_i)\}_{i \in I}$, $|I| < \infty$ we have*

$$P\left\{X(t_i) = x_i \,\forall i \in I\right\} = \mathcal{N}\left(x \,\middle|\, \mu^I, \Sigma^I\right) , \tag{6.67}$$

where $\mathcal{N}(x \mid \mu, \Sigma)$ is the normal distribution with

$$\text{mean } \mu^I_i \equiv \langle X(t_i) \rangle \equiv \mu(t_i) \quad \text{and} \tag{6.68}$$

$$\text{covariance } \Sigma^I_{ij} \equiv \langle X(t_i) X(t_j) \rangle_{\mathrm{c}} \equiv \Sigma(t_i, t_j) \,. \tag{6.69}$$

Theorem 6. *A Gaussian-equivalent process $X(t)$ has exactly one representative $X^G(t)$ that is a Gaussian process.*

Proof. By construction, mean function $\mu(t)$ and covariance kernel $\Sigma(t, s)$ of any Gaussian-equivalent process define a valid Gaussian process. In turn, assume there are two Gaussian process representatives $X^{G_1}(t)$ and $X^{G_2}(t)$ for the same Gaussian-equivalent process $X(t)$, i.e. $\mu^{G_1} \equiv \mu^{G_2}$ and $\Sigma^{G_1} \equiv \Sigma^{G_2}$. Since Gaussian processes are defined in terms of their finite dimensional distributions, which are in turn given by μ and Σ, these two processes are identical. □

Definition 12. *For a stationary (level 0 or 1) Gaussian-equivalent process with MSD $\psi(\Delta t)$,*

we denote its unique Gaussian representative by $X^G_\psi(t)$.

We can now directly write the likelihood of observing the data D, given MSD parameters θ:

$$P\left(D \mid \theta\right) = \prod_{i=1}^{N} P\left(x^i \;\middle|\; \theta\right) = \prod_{i=1}^{N} P\left(x^i \;\middle|\; X^G_{\Psi(\theta)}\right) , \tag{6.70}$$

where $P\left(x^i \;\middle|\; X^G_{\Psi(\theta)}\right)$ signifies evaluation of the finite dimensional distributions (6.67) pertaining to the process $X^G_{\Psi(\theta)}$ on the trajectories x^i in the dataset D.

Suppressing the dependence on data D, we finally introduce the log-likelihood function $\log \mathcal{L}(\theta)$ as

$$\log \mathcal{L}(\theta) \equiv \log P\left(D \mid \theta\right) = \sum_{i=1}^{N} \log P\left(x^i \;\middle|\; X^G_{\Psi(\theta)}\right) . \tag{6.71}$$

Prior. For given parametrization, we introduce a prior $\pi(\theta)$ over the parameters θ. By default we just use a flat prior; note, however, that this statement depends on the details of the parametrization. For example, in practice we parametrize the powerlaw part of `NPXFit` not by prefactor Γ and exponent α, but by $\log\Gamma$ and α. The flat prior over $\log\Gamma$ then becomes a log-flat prior over Γ, appropriate for a positive variable with unknown scale (since Γ carries units, its numerical value could be 10^{-5} just as well as 10^{5}).

Posterior. According to Bayes' rule, the posterior $p(\theta)$ over the parameters θ is now given by

$$p(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{\int \mathrm{d}\theta\, \mathcal{L}(\theta)\pi(\theta)} . \tag{6.72}$$

The integral in the denominator is a normalization constant and has no influence on the shape of the posterior. The *maximum posterior (MAP)* parameter estimate $\hat{\theta}$ can thus be found by numerical optimization of $\mathcal{L}(\theta)\pi(\theta)$:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta)\pi(\theta) . \tag{6.73}$$

Credible intervals. Beyond the point estimate (6.73), the Bayesian approach allows us to calculate credible intervals for the estimated parameters. This is especially useful when using an interpretable parametrization, such as a powerlaw or the Rouse model.

There are multiple ways to estimate credible intervals, with different use cases and computational feasibility:

- The theoretically most straight-forward approach is to evaluate the unnormalized posterior $\mathcal{L}(\theta)\pi(\theta)$ over the full parameter space, calculate the normalization factor in eq. (6.72) and delegate all further analysis to the true posterior distribution $p(\theta)$. This would allow for example to calculate exact credible intervals by finding the region in parameter space over which the posterior integrates to (say) 0.95. This approach is usually computationally not feasible, because it requires too many evaluations of the likelihood function $\mathcal{L}(\theta)$ (which is expensive)[7]

- An analytical approach to calculate asymptotically exact credible intervals employs Wilks' theorem, which asserts that the likelihood ratio statistic is asymptotically χ^2 distributed:

$$-2\log\Lambda = -2\log\frac{\mathcal{L}(\theta)\pi(\theta)}{\mathcal{L}(\hat{\theta})\pi(\hat{\theta})} \xrightarrow{D} \chi^2_d\,, \tag{6.74}$$

  where d is the dimensionality of the parameter space, i.e. the number of independent fit parameters. A $1-\alpha$ credible interval can then be found by moving each parameter independently and finding the boundaries where

$$\log\mathcal{L}(\theta)\pi(\theta) = \log\mathcal{L}(\hat{\theta})\pi(\theta) - \frac{1}{2}\chi^2_{d,\alpha}\,. \tag{6.75}$$

  Naïvely applying this prescription, one finds credible intervals for the parameter θ_i, *conditional on all other θ_j, $j \neq i$ taking the point estimate values*. Usually, the more interesting question is "which values of θ_i are consistent with the data, *when we adjust everything else properly*?" This question can be answered by using the profile likelihood approach instead: for parameter θ_i, define the *profile likelihood*

$$\mathcal{L}_i(\theta_i) \equiv \max_{\theta_j\,,\,j\neq i} \mathcal{L}(\theta)\pi(\theta)\,; \tag{6.76}$$

  The corresponding credible interval boundaries are then given by

$$\log\mathcal{L}_i(\theta_i) = \log\mathcal{L}_i(\hat{\theta}_i) - \frac{1}{2}\chi^2_{1,\alpha}\,. \tag{6.77}$$

  Note that now the χ^2-percentiles are evaluated for a single degree of freedom; also, by

[7] One could imagine getting this to work with some smart sampling scheme like AMIS [102]. We do not further pursue this avenue for now.

definition of the MAP estimate $\hat{\theta}$ we have $\mathcal{L}_i(\hat{\theta}_i) \equiv \mathcal{L}(\hat{\theta})\pi(\hat{\theta})$.

In practice, usually the profile likelihood credible intervals are the relevant ones.

- Finally, we can directly estimate the credible intervals by sampling. To that end we run (e.g.) a Markov chain Monte Carlo (MCMC) scheme to generate a posterior sample $\left\{\theta^k \;\middle|\; k = 1, \ldots, K\right\}$, from which we can estimate marginal posterior distributions over each individual parameter θ_i. We can then give credible intervals from these marginal distributions.

  A handy way to parametrize the MCMC is to set the step size in each parameter direction equal to the size of the profile likelihood credible interval. This means that each MCMC sample is close to independent from the previous one, thus essentially setting the burn-in time to zero.

Chapter 7

Scale-free models of chromosome structure, dynamics, and mechanics

This chapter was co-authored by Myself, Antoine Coulon, and Leonid Mirny.

It has been submitted for peer review; in the meantime, a pre-print is available online at the bioRxiv, doi: 10.1101/2023.04.14.536939

7.1 Abstract

Scale-free, or fractal, models are prevalent in the study of chromosome structure, dynamics, and mechanics. Recent experiments suggest the existence of scaling relationships; but currently there is no single model consistent with all observed exponents. We present a simple argument characterizing the space of scale-free models for chromosome structure, dynamics, and mechanics and discuss the implications for a consistent treatment. Our framework helps reconciling seemingly contradictory data and identifies specific experimental questions to be addressed in future work.

7.2 Main Text

The nucleus of a eukaryotic cell contains its genetic information in the form of chromatin—a composite polymer of DNA and associated proteins. The physical nature of this polymer, and specifically the local chromosomal context of a given locus, play crucial roles in determining how the information encoded on the DNA is processed [116]. So-called *enhancer* elements for example are thought to activate their target genes by "looping in" and physically contacting the target promoter to initiate transcription [117–119]. How this interaction is regulated between elements that can be separated by millions of base pairs remains an open question [120–122]; in fact, the structure and dynamics of even the "bare" chromatin polymer itself—without additional elements like enhancers and promoters—remain topics of active research [16, 123, 124].

Our understanding of the 3D structure of chromosomes has increased dramatically over the last decade, primarily due to experimental techniques like Hi-C [7, 125] (measuring pairwise contacts across the genome) and multiplexed FISH methods [15, 126] (visualizing chromosome conformations in 3D space). Both techniques show that the chromatin fiber adopts a *space-filling* conformation: two loci at a genomic separation s are on average separated in space by a distance $R(s) \sim s^{\frac{1}{3}}$ [123], corresponding to a confining volume $V(s) \sim R^3(s) \sim s$ [14]—thus the term "space-filling". The probability $P(s)$ of finding these two loci in contact is then given by the mean field approximation $P(s) \sim 1/V(s)$ [127]; $P(s) \sim s^{-1}$ was broadly observed in Hi-C experiments across vertebrate chromosome systems [7, 125]. Notably, this space-filling spatial organization is more compact than one would expect for an ideal chain in equilibrium, which should adopt a random walk conformation with $R(s) \sim s^{\frac{1}{2}}$, corresponding to $V(s) \sim s^{\frac{3}{2}}$ and $P(s) \sim s^{-\frac{3}{2}}$ [20].

The study of chromosome dynamics has not seen a breakthrough comparable to Hi-C yet and is therefore more heterogeneous. One main mode of investigation is fluorescent labelling and tracking of individual genomic loci in live cells, allowing for characterization of the particles' motion by the Mean Squared Displacement (MSD)

$$\mathrm{MSD}(\Delta t) := \left\langle (x(t+\Delta t) - x(t))^2 \right\rangle \sim (\Delta t)^{\mu} \ . \tag{7.1}$$

While a freely diffusive particle would exhibit a linear MSD curve ($\mu = 1$), a chromosomal locus (i.e. point on a long polymer) is expected to move subdiffusively ($\mu < 1$) due to the chain connectivity. Indeed, experiments show $\mu \approx 0.5 - 0.6$ in eukaryotic cells [13, 98]. Notably—

and in contrast to the spatial structure—this is consistent with an ideal chain ($R(s) \sim s^{\frac{1}{2}}$, $P(s) \sim s^{-\frac{3}{2}}$), for which the Rouse polymer model predicts $\mu = \frac{1}{2}$ [18, 19].

Taking an orthogonal angle on the question of chromatin dynamics, the present authors, together with others, recently developed an experimental system to measure the force response of a single genomic locus [128]. In response to a constant force switched on at $t = 0$ the locus displaced as $x(t; f) \sim t^{0.5}$, consistent with the same (Rouse) model for polymer dynamics that predicted the MSD scaling $\mu = 0.5$—but which is inconsistent with the structure $R(s) \sim s^{\frac{1}{3}}$ of real chromatin.

A model consistent with the space-filling structure of real chromatin is the *fractal globule*, which describes crumpling of the chain due to topological constraints [129]. This, however, predicts an MSD scaling of $\mu = \frac{2}{5}$ [17], markedly lower than the $\mu \approx 0.5 - 0.6$ observed in experiments. Within the context of commonly used polymer models for chromatin, we are thus left with two mutually contradictory observations: a fractal globule would reproduce the compact structure, but with slower dynamics[1]; the fast dynamics are consistent with the Rouse model, but that assumes an unrealistically open, equilibrium conformation. Does this point to some fundamental inconsistency in structural vs. dynamical observations, or is it simply that both models are wrong? If so, how can we reconcile all observations?

Chromosome structure and organization spans multiple orders of magnitude: from single nucleosomes (∼11 nm) to whole nuclei (∼2–20 µm). As such, *scale-free* models—such as Rouse or fractal globule—constitute useful null models for the description of chromosomes. Indeed, all the observables mentioned above—contact probability $P(s)$, spatial separation $R(s)$, $\mathrm{MSD}(\Delta t)$, force response $x(t; f)$—are expected to exhibit scaling behavior (read: are powerlaws), exactly because of this scale-free nature of the null model. The point we aim to highlight with this letter is that within the context of scale-free models, not only are these observables governed by powerlaws, but their exponents also have to satisfy hyperscaling relations that are a direct consequence of the scale-free assumption. These relations enable an informed discussion of the mismatch between structure, dynamics, and mechanics highlighted above; awareness of these relations is lacking in the literature [130].

Let us assume we have a scale-free model for chromatin structure, dynamics, and mechanics. Such a model should predict the behavior of the observables outlined in the introduction and because of the absence of finite length scales we expect to find powerlaws. Explicitly, we assume

[1] "slower" here technically meaning "more recurrent", i.e. a lower *exponent*.

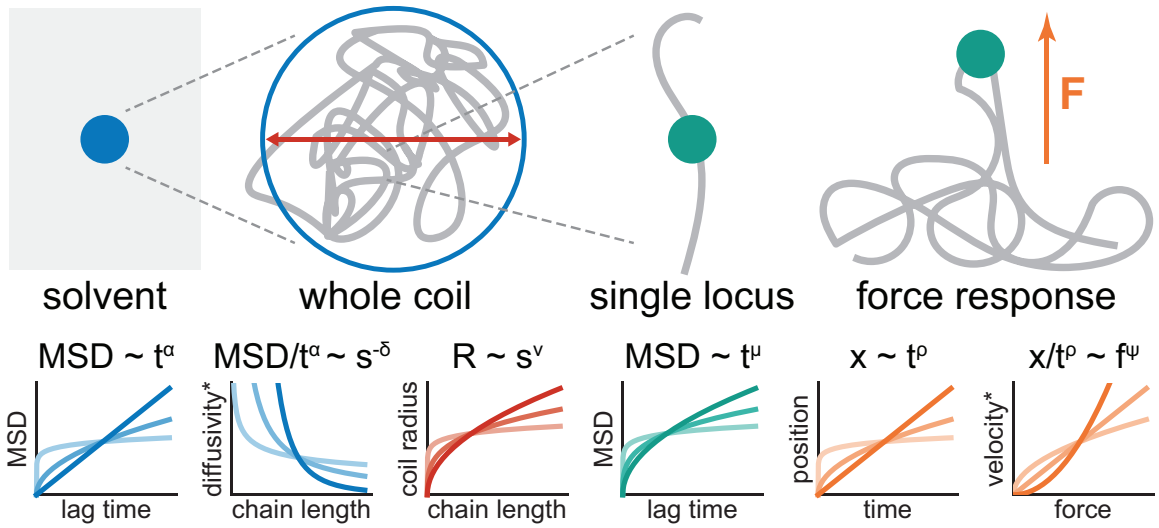


Figure 7.1: Summary of the exponents considered in the text and what part of the system they relate to. Left to right: α controls the viscoelasticity of the solvent, i.e. the MSD of a free tracer particle. Considering an isolated polymer coil as such a tracer particle, δ governs the dependence of its (anomalous) diffusivity on the chain length; ν gives the scaling of the physical radius of the coil. The motion of individual loci within the coil is characterized by μ. Upon application of an external force, such loci exhibit a powerlaw response with exponent ρ; the (fractional) velocity of this response is force dependent, as indicated by ψ. Colors indicate which constitutive relation an exponent is associated with: red for eq. (7.2), teal for eq. (7.3), orange for eq. (7.4), and blue for eq. (7.5).
*: "anomalous diffusivity" if $\alpha \neq 1$
"fractional velocity" if $\rho \neq 1$.

the forms

$$R(s) = Gs^{\nu} \tag{7.2}$$

for the spatial distance between two loci at a genomic separation s;

$$\mathrm{MSD}(\Delta t) = \Gamma\,(\Delta t)^{\mu} \tag{7.3}$$

for the MSD of a single genomic locus; and

$$x(t; f) = Af^{\psi}t^{\rho} \tag{7.4}$$

for displacement in response to a constant force f switched on at time $t = 0$ (red, teal, and orange in fig. 7.1). Note how the equations concerning dynamics and force response consider only a single locus, while the structural scaling refers to a finite stretch of chromatin. To bridge this gap and connect structure and dynamics, we consider the whole-coil diffusion of a finite and isolated stretch of chromatin. Over timescales longer than the internal relaxation time, we

expect this coil to diffuse like a free particle, quantified by an MSD of the form

$$\mathrm{MSD}_{\mathsf{coil}}(\Delta t; s) = D s^{-\delta} (\Delta t)^{\alpha} \tag{7.5}$$

(blue in fig. 7.1). Since we expect a free coil to undergo normal diffusion, $\alpha = 1$ seems like the most natural choice; however, allowing $\alpha < 1$ incorporates the possibility of a viscoelastic solvent, such that even a free tracer particle would undergo subdiffusion—which has been observed for the nucleoplasm, though estimates for α vary broadly ($\alpha \approx 0.5 - 1$) [131–133]. The exponent δ can be understood as incorporating long-range spatial interactions of different loci on the polymer. For a freely draining chain (such as the Rouse model), monomers are independent from each other; whole-coil diffusivity is thus simply inversely proportional to chain length, yielding $\delta = 1$. The Zimm model [19], in contrast, incorporates hydrodynamic interactions between the loci, which results in a hydrodynamic radius $R_{\mathsf{hydro}} \sim R(s)$. $\mathrm{MSD}_{\mathsf{coil}} \sim R_{\mathsf{hydro}}^{-1}$ then implies $\delta = \nu$.

The observables described by eqs. (7.2) to (7.5) all have units of length. But we assume that the underlying model itself does not have any finite length scale; this assumption would be inconsistent, if we could construct such a finite length scale from the constants in the model, i.e. from the prefactors in eqs. (7.2) to (7.5) and the thermal energy $k_{\mathsf{B}}T$. Their respective dimensions are

$$\begin{gathered} [k_{\mathsf{B}}T] = LF\,, \qquad [G] = LS^{-\nu}\,, \qquad [\Gamma] = L^2T^{-\mu}\,, \\ [A] = LF^{-\psi}T^{-\rho}\,, \qquad [D] = L^2S^{\delta}T^{-\alpha}\,, \end{gathered} \tag{7.6}$$

where we use the symbols L, F, S, T to denote length, force, genomic distance, and time, respectively.

A quantity

$$X := (k_{\mathsf{B}}T)^a\, G^b \Gamma^c A^d D^e \tag{7.7}$$

now has units

$$[X] = L^{a+b+2c+d+2e} F^{a-\psi d} S^{-\nu b+\delta e} T^{-\mu c-\rho d-\alpha e}\,; \tag{7.8}$$

setting $[X] = L$—attempting to construct a length scale—gives a system of four equations for the five variables a, b, c, d, e. Elementary substitutions reduce this system to one equation for two variables,

$$\left(1 + \frac{2\nu}{\delta} - \frac{2\alpha\nu}{\delta\mu}\right) b + \left(1 + \psi - \frac{2\rho}{\mu}\right) d = 1\,, \tag{7.9}$$

which has a one-parameter family of solutions (b, d)—unless both terms in brackets vanish. The scale-free assumption is thus only self-consistent if the exponents obey the two constraints

$$\frac{2\nu\alpha}{2\nu+\delta} = \mu = \frac{2\rho}{1+\psi}\,. \tag{7.10}$$

These two constraints ensure that both brackets in eq. (7.9) vanish, thus preventing the emergence of a finite length scale $[X] = L$. The first relation has been reported previously, in the special cases $\alpha = 1$, $\delta = 1$ [17, 134]; $\nu = \frac{1}{2}$, $\delta = 1$ [28]; and $\delta = 1$ [135]. The second relation connecting dynamics and force-response is satisfied explicitly by the Rouse model [128, 136], but has not been studied in generality.

Consider the force response experiments of [128], where we determined $\rho \approx 0.5$, $\psi \approx 1$, and $\mu \approx 0.5$, fully consistent with eq. (7.10). Notably, just the linear force response ($\psi = 1$) suffices to predict $\rho = \mu$; our measurement of the force response exponent $\rho \approx 0.5$ can thus be interpreted as an independent validation of earlier experiments finding $\mu \approx 0.5$ [137].

The first relation in eq. (7.10) connects the structural and dynamical scalings ν and μ, both of which have been investigated in various experimental systems (see table 7.1). While specifically yeast seems consistent with the Rouse expectations $\mu = 0.5$, $\nu = 0.5$, and $\alpha = 1$ [140], multicellular eukaryotes like fruit fly, mouse, or human, seem to behave differently. For the purpose of this discussion, let us consider the case $\mu = 0.5$, $\nu = 0.33$ (fig. 7.2); this seems consistent with best estimates, but is of course an idealization of the experimental situation. Importantly, eq. (7.10) holds true for any value of these exponents. As outlined in the introduction, $\mu = 0.5$ matches our expectations from the Rouse model, while $\nu = 0.33$ indicates a fractal globule; to the best of our knowledge, there is currently no consistent model reproducing both. Reformulating the first relation in eq. (7.10) as

$$\delta = 2\nu\left(\frac{\alpha}{\mu} - 1\right) \tag{7.11}$$

shows that we should expect a 1-parameter family of models with different α and δ that exhibit the desired scalings in μ and ν. We discuss a few of these options:

- In a freely draining chain (blue lines in fig. 7.2), individual monomers are independent, such that $\delta = 1$. This assumption is made in the Rouse model and in [17] for dynamics of the fractal globule. Equation (7.11) then becomes $\alpha = \frac{5}{4} > 1$, i.e. we would need

Organism	μ	ν	Ref.	Notes
H. sapiens				
HCT-116	-	0.3-0.4[1]	[87]	ΔRad21
	-	0.3-0.4[2]	[15]	
HeLa	0.5	-	[137]	Telomeric probes
U2OS	0.55	-	[137]	Telomeric probes
MF	0.7	-	[137]	Telomeric probes
M. musculus				
mESC	0.5[3]	-	[13]	WT and ΔRAD21
	0.6[3]	-	[98]	WT and ΔRAD21
	-	0.33-0.4[1]	[90]	ΔRAD21
	-	0.15-0.4[2]	[123]	
hepatocytes	-	0.4[1]	[138]	ΔNIPBL
3T3	0.4	-	[137]	Telomeric probes
D. melanogaster	0.58[3]	0.31[3]	[130]	μ and ν determined in same system
S. cerevisiae	-	0.5[1]	[139]	
	0.5	-	[62]	
E. coli	0.4	-	[27]	
Caulobacter	0.4	-	[27]	

[1] from Hi-C contact probability $P(s) \sim s^{-3\nu}$ [127]

[2] direct measurement from multiplexed FISH

[3] two-locus live-cell measurement

Table 7.1: Measured scalings for $\mathrm{MSD}(\Delta t) \sim (\Delta t)^{\mu}$ and $R(s) \sim s^{\nu}$. Chromosome structure is frequently not strictly fractal due to loop extrusion; therefore we here focus on experiments where loop extruding factors (Rad21, a component of the cohesin complex) or their loaders (Nipbl) were acutely degraded, where possible. This overview is not exhaustive.

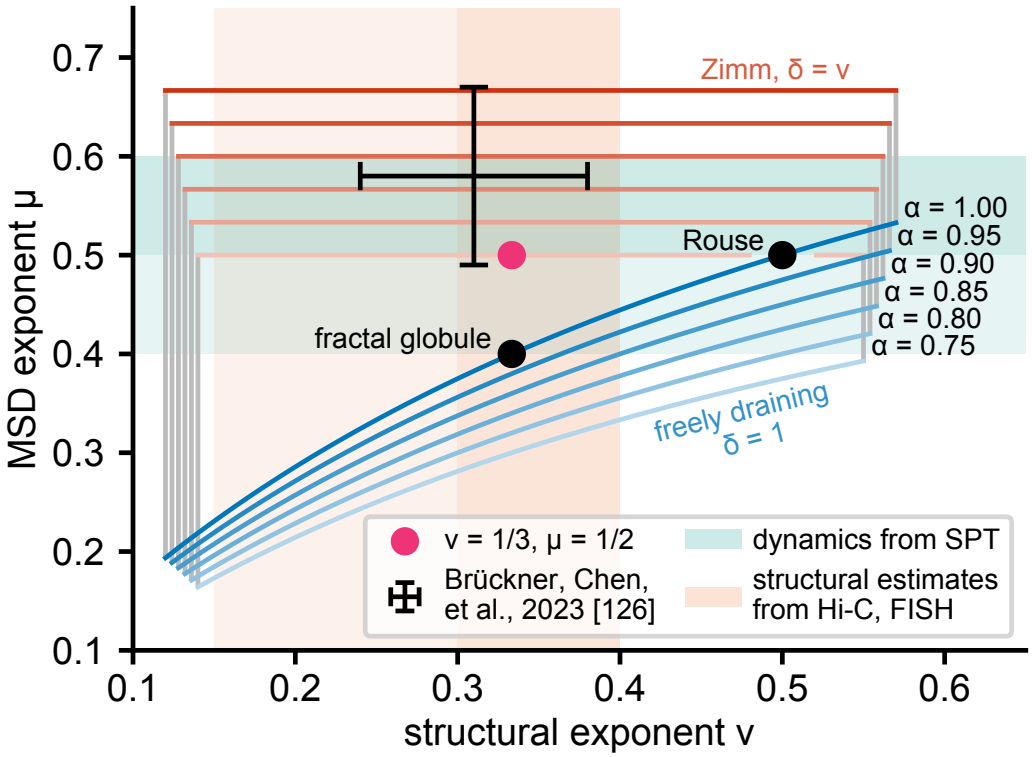


Figure 7.2: Experimental results in the context of eq. (7.10). Shaded regions are consistent with experimental determinations of structure ν (orange; directly from chromosome tracing or inferred from Hi-C, the latter densely shaded) or dynamics μ (green; from SPT; dense shade indicates eukaryotic estimate $\mu \approx 0.5 - 0.6$, light shade extends to bacterial estimate $\mu \approx 0.4$) respectively. Black error bars indicate estimate from [130]. Red circle marks $\nu = 0.33$, $\mu = 0.5$, which serves as example for discussion in the main text. Outlines show theoretically plausible regions (eq. (7.10)) for different α, as indicated. The top (red) edge of these regions is given by the Zimm condition $\delta = \nu$, while the bottom (blue) edge is given by the freely draining chain ($\delta = 1$); the points inbetween correspond to $\nu < \delta < 1$. Horizontal cutoffs (gray lines) are chosen for visual appeal. Common polymer models: Rouse chain and fractal globule are indicated as black circles; both are instances of a freely draining chain with $\alpha = 1$ (blue curve).

a medium in which free tracers undergo *superdiffusion*. This appears unrealistic for the nucleoplasm. While it might in principle be achieved by energy dependent processes like transcription or loop extrusion, we will not further pursue this point here.

- Including hydrodynamic interactions between different monomers amounts to $\delta = \nu$, such that $\mu = \frac{2}{3}\alpha$ independent of ν (red lines in fig. 7.2). This would allow matching $\mu \approx 0.5$ by tuning $\alpha \approx 0.75$. While this is within current estimates for nucleoplasm viscosity, these estimates scatter quite broadly ($\alpha \approx 0.5 - 1$), such that this consistency statement is rather weak. Furthermore, due to crowding we should expect hydrodynamic interactions to be screened in the nucleus [141, 142], such that $\delta = \nu$ appears questionable in the first place.

- Between the two canonical values of $\delta = 1$ (freely draining chain) and $\delta = \nu$ (hydrodynamic interactions), it is conceivable that chromatin loci in the nucleus do exhibit some (effective) long-range interaction with $\nu < \delta < 1$. In a purely viscous nucleoplasm ($\alpha = 1$), eq. (7.11) would then imply $\delta = 2\nu = \frac{2}{3}$, i.e. a whole-coil hydrodynamic radius scaling as $R_{\text{hydro}} \sim R^2(s)$ (cf. discussion below eq. (7.5)). While we are currently not aware of a physical model producing this behavior, this is an interesting possibility that certainly warrants further investigation.

Experimentally, there are two avenues to further narrow down the above 1-parameter family to a single, scale-free null model of chromatin in the nucleus: measuring α or δ. Observation of free particle diffusion in the nucleus has so far yielded conflicting results as to the viscoelastic properties of the nucleoplasm: while [132] report $\alpha = 0.5 - 0.6$, [143] measure $\alpha \approx 0.75$ in yeast; [25, 133] find normal diffusion $\alpha = 1$ on large scales (relative to multimeric GFP tracers), with intermediate behavior strongly probe dependent; most recently, [140] reported $\alpha \approx 0.9$ in yeast and $\alpha \approx 0.86$ in mammalian (hPNE) cells. Consensus about nucleoplasm viscosity is thus outstanding. Furthermore, the material whose viscoelasticity is probed here is a solution containing chromatin, which presumably contributes some (if not all) of the observed elastic response; it is thus even unclear what these results would imply for solvent viscoelasticity when modelling chromatin explicitly. We thus suggest that measuring δ (long-range interactions) instead of α (medium viscoelasticity) provides an orthogonal avenue towards a consensus model. Such measurements would require observing the diffusion of free chromatin chains (e.g. nucleosome arrays) of different length in the nucleus, which is feasible with current techniques. A major

challenge in such experiments would be to ensure that the probes still obey the same structural scaling ν as the rest of the genome. Early experimental work found $\delta \approx 0.72$ for naked DNA in aqueous solution [144]; we are not aware of similar measurements for chromatinized DNA inside the nucleus.

Our derivation of the exponent relations (7.10) strongly relied on the absence of finite length scales; of course, no real system is truly scale-free. So what would be implied by experimental data contradicting eq. (7.10)? First of all, it would simply mean that those data are not well approximated by a single, consistent, scale-free model. It then stands to reason that a more detailed model is required, which will most likely not predict powerlaws for the observables under study in the first place (in which case there are of course also no exponent relations to be satisfied). If any of the observables does indeed exhibit manifestly powerlaw scaling (a claim that is generally quite hard to justify rigorously [145]), the underlying reason might be quite interesting and should be investigated in detail. From a pragmatic point of view, eq. (7.10) might thus be interpreted simply as a check on the appropriateness of powerlaw fits to multiple observables.

To obtain the connection between unperturbed dynamics and response to an external force, we included the thermal energy $k_\mathrm{B}T$ in the set of model constants, because we assume the unperturbed dynamics to be driven by thermal fluctuations. This does not immediately imply an assumption about thermal equilibrium: $k_\mathrm{B}T$ in our treatment does not necessarily have to correspond to physical temperature, but should just be some energy scale of the fluctuations. However, many active processes act over a finite length scale, such that in presence of active fluctuations, the scale-free assumption might be questionable. In fact, assuming $T = 37\,^\circ\mathrm{C}$ (the physical temperature in the incubation chamber), we found good agreement between MSD and force response in our earlier work [128], suggesting that chromosome fluctuations are indeed largely driven by thermal noise.

We presented a streamlined version of the argument leading to eq. (7.10), tailored towards the application to polymer structure, dynamics, and mechanics; a more systematic approach to the dimensional analysis is given in chapter 8. Furthermore, the approach through dimensional analysis is of course nothing but a reformulation of the more classical approach based on the consideration of a finite subchain, as given e.g. in [17] for the freely draining chain; we provide that reformulation below.

7.3 Scaling of a finite subchain

The argument in section 7.2 is formulated in terms of dimensional analysis to emphasize that it is a necessary conclusion of the scale-free assumption. It is easily reformulated in a more physical language by considering a finite subchain of length $\mathfrak{s}$.

Equation (7.2) gives the physical size of this subchain as

$$\mathfrak{l} = R(\mathfrak{s}) = G\mathfrak{s}^{\nu} \,. \tag{7.12}$$

This allows for two independent definitions of a time scale: by setting $\mathrm{MSD}(\Delta t) = \mathfrak{l}^2$ we find

$$\mathfrak{t} = \left(\frac{G^2}{\Gamma}\right)^{\frac{1}{\mu}} \mathfrak{s}^{\frac{2\nu}{\mu}} \,; \tag{7.13}$$

letting instead $\mathrm{MSD}_{\mathsf{coil}}(\Delta t; s) = \mathfrak{l}^2$ yields

$$\tau = \left(\frac{G^2}{D}\right)^{\frac{1}{\alpha}} \mathfrak{s}^{\frac{2\nu+\delta}{\alpha}} \,. \tag{7.14}$$

Physically, both describe the relaxation time scale of the coil and should thus be equal (up to a numerical prefactor). This requires that the exponents on $\mathfrak{s}$ be the same, yielding the first relation in eq. (7.10).

Similarly, eqs. (7.3) and (7.4) and the thermal energy $k_{\mathsf{B}}T$ allow the construction of two orthogonal force scales associated with our subchain, both of which should exhibit the same scaling behavior with $\mathfrak{s}$ (or in this case $\mathfrak{l}$):

$$\mathfrak{f} \equiv \frac{k_{\mathsf{B}}T}{\mathfrak{l}} \overset{!}{\sim} \left(\frac{\mathfrak{l}}{A\mathfrak{t}^{\rho}}\right)^{\frac{1}{\psi}} = \left(\frac{\Gamma^{\frac{\rho}{\mu}}}{A}\right)^{\frac{1}{\psi}} \mathfrak{l}^{\frac{1}{\psi}-\frac{2\rho}{\psi\mu}} \,. \tag{7.15}$$

Again equating the exponents (on $\mathfrak{l}$) yields the second relation in eq. (7.10).

While the formulation in terms of a finite subchain can aid physical intuition, the core argument remains the same: if eq. (7.10) does not hold, a finite length scale emerges. To see this, consider:

$$q(\mathfrak{s}) := \frac{\tau(\mathfrak{s})}{\mathfrak{t}(\mathfrak{s})} = G^{\frac{2}{\alpha}-\frac{2}{\mu}} D^{-\frac{1}{\alpha}} \Gamma^{\frac{1}{\mu}} \mathfrak{s}^{\frac{2\nu+\delta}{\alpha}-\frac{2\nu}{\mu}} \,. \tag{7.16}$$

Since q is a dimensionless ratio, if the scaling with $\mathfrak{s}$ were non-trivial, the combination of

constants in front would have units of S to some power, translating to a length scale through eq. (7.2). Thus, within the framework of scale-free models, any finite scale (length, force, or otherwise) associated with the subchain $\mathfrak{s}$ has to be unique. This is ultimately what drives the scaling argument outlined here.

Chapter 8

Dimensional analysis in scale-free models of chromatin organization

In chapter 7 we developed a dimensional analysis argument to survey the space of scale-free models for chromatin organization. In order to put more focus on the conceptual implications there, we kept the dimensional analysis itself to a bare minimum and derived only the key relations (eq. (7.10)) in isolation. This chapter outlines a more systematic approach to the dimensional analysis that might also be useful in other contexts. We begin by developing some general machinery in section 8.1, apply these strategies to free particles (section 8.3), and subsequently polymers (section 8.4).

8.1 Machinery

Assume we have constants $A_1, \ldots, A_n$ with dimensions $[A_i] = D_1^{\gamma_1^1} \cdots D_m^{\gamma_m^1}$. Then we can build a constant X with dimension $[X] = D_1^{\xi_1} \cdots D_m^{\xi_m}$ by solving the linear system

$$Ca \equiv \begin{pmatrix} \gamma_1^1 & \cdots & \gamma_1^n \\ \vdots & \ddots & \vdots \\ \gamma_m^1 & \cdots & \gamma_m^n \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_m \end{pmatrix} \equiv \xi \,. \tag{8.1}$$

We ensure that this system is unique for given constants by requiring that C does not have any all-zero rows (which would correspond to unused dimensions). Then the image space of C has well-defined dimension m, such that a sufficient criterion for solvability of eq. (8.1) is

$\operatorname{rank} C = m$.

The key argument in the context of polymer scalings is then to set X to have units of length and ensure that eq. (8.1) is not solvable, such that no finite length scale emerges from the observed scaling laws. This means that we use the inverse of the previous statement: $\operatorname{rank} C < m$ is necessary to prevent solvability of eq. (8.1). This will usually also be sufficient, unless ξ just so happens to lie in the restricted image space; it is usually quite straight-forward to check whether this is the case.

So, at the end of the day, the main condition we need to ensure to prevent solvability of eq. (8.1) is $\operatorname{rank} C < m$. Since by construction C does not have any all-zero rows, this is equivalent to requiring that any matrix $\tilde{C}$ built from m columns of C has determinant zero.

The matrices we will consider are usually somewhat sparse; the following inductive reasoning might thus be useful: consider an N-by-N matrix where some row has only one non-zero entry; by Laplace expansion, the determinant of the matrix is zero if and only if the determinant of the $(N-1)$-by-$(N-1)$ submatrix obtained by removing the row and column corresponding to the non-zero entry is zero. This is visualized by the following example: let

$$C = \begin{pmatrix} a & b & c & d & e & f \\ g & h & & i & & \\ j & & k & l & & \\ m & & & & n & o \end{pmatrix}, \tag{8.2}$$

where zero entries are left empty to better visualize the structure. From this matrix, we would derive constraints as follows:

- choose columns 1, 3, 5, 6; row 2 then contains only one non-zero entry (g), so we can restrict to columns 3, 5, 6 with row 2 removed. Then the new row 2 has only one non-zero entry (k), so we can further restrict to columns 5 and 6, yielding the constraint

$$0 \stackrel{!}{=} \begin{vmatrix} e & f \\ n & o \end{vmatrix} = eo - fn\,. \tag{8.3}$$

- next, choose columns 1, 2, 3, 4, note that row 4 has three zeros; restricting to columns

2, 3, 4 then gives

$$0 \stackrel{!}{=} \begin{vmatrix} b & c & d \\ h & & i \\ & k & l \end{vmatrix} = dhk - bik - chl\,. \tag{8.4}$$

- finally, pick columns 1, 2, 3, 5, which does not reduce any further, so we calculate the determinant explicitly by Laplace expansion along the first row:

$$0 \stackrel{!}{=} \begin{vmatrix} a & b & c & e \\ g & h & & \\ j & & k & \\ m & & & n \end{vmatrix} = ahkn - bgkn - chjn - ehkm\,. \tag{8.5}$$

These three constraints ensure that at least three columns of C can be written as linear combination of the others, such that we are left with $\operatorname{rank} C \leq n - 3 = 3 < 4 = m$. We can then check explicitly that $\xi \notin \operatorname{im} C$ for whatever ξ we are aiming to prevent.

Note that each constraint obtained this way amounts to a linearly dependent set of column vectors; this translates to a dimensionless combination q of constants. So, for each exponent relation obtained from this construction we expect a corresponding relation between the constants A_i.

8.2 Notes

In the following, we will use the dimensions length (L), force (F), time (T), and size (S), the latter being a utility to describe polymeric systems (see below).

We generally require that no finite length scale emerges from the heuristic constants introduced, i.e. in the language of section 8.1 we set $[X] = L$.

All exponents are assumed to be strictly positive.

8.3 Free particle

Consider a (for now point-like) particle in a dissipative medium, at equilibrium at position $x = 0$, i.e. $x(t) = 0 \,\forall\, t \leq 0$. In response to a constant force f switched on at time $t = 0$, we expect

the particle to move as

$$x(t) = A f^{\varphi} t^{\sigma} \,, \qquad\qquad [A] = L F^{-\varphi} T^{-\sigma} \,. \tag{8.6}$$

Motivated by the fluctuation–dissipation theorem, in the absence of the force f we expect this particle to undergo fluctuations, characterized by

$$\mathrm{MSD}(t) \equiv \left\langle x^2(t) \right\rangle = D t^{\alpha} \,, \qquad\qquad [D] = L^2 T^{-\alpha} \,. \tag{8.7}$$

Note that so far the dimensions L, F, and T do not have much physical meaning, but just keep track of the exponents with which x, f, and t enter the equations, respectively. We now explicitly add some of that physical meaning by requiring—again motivated by the FDT—the existence of an energy scale, i.e. a constant with dimensions $(\text{length}) \cdot (\text{force})$, which we call *temperature*:

$$[k_\mathrm{B} T] = L F \,. \tag{I}$$

This constant provides a connection between the force in eq. (8.6) and the fluctuations in eq. (8.7). Accordingly, we now find a dimensional constraint:

$$0 \stackrel{!}{=} \begin{vmatrix} 1 & 2 & 1 \\ -\varphi & 0 & 1 \\ -\sigma & -\alpha & 0 \end{vmatrix} = -2\sigma + \varphi\alpha + \alpha \qquad \Leftrightarrow \qquad \alpha = \frac{2\sigma}{1+\varphi} \,, \tag{8.8}$$

along with the dimensionless quantity q_1 satisfying

$$(k_\mathrm{B} T)^{\varphi} = q_1 A^{-1} D^{\frac{1+\varphi}{2}} \,. \tag{8.9}$$

So far our test particle was a featureless point, i.e. all its properties and their interplay with the medium were lumped together in the constants A and D. In this framework, different particles would be characterized by the values of these constants; but so far we cannot make any statement about how the constants for different particles should be related to each other.

In our simple scaling description, test particles can differ by the material they are made of and how that interacts with the dissipative medium (which we will not describe in more detail), and by how much of that material there is; we call the latter the *size* s of the particle. Clearly

we expect the physical extent of the particle to change with its size, described by the relation

$$R(s) = Gs^{\nu}\,, \qquad\qquad [G] = LS^{-\nu}\,, \tag{II}$$

where $R(s)$ has the same units as $x(t)$ (length L) and the size s has an independent dimension S. Note that e.g. for a spherical particle we would usually identify its size and radius, such that $\nu = 1$ and eq. (II) is a simple conversion from the artificial unit S to physical length L. We introduce eq. (II) in this form mainly for later use, where we will characterize the "size" of a polymer coil by the length s of the chain contained, which is connected to the physical extent of the coil exactly through eq. (II), with the exponent ν indicative of the coil structure. Note, however, that this is also conceptually clearer: *a priori* it is not clear why particle size s and displacement x should have the same units, so we introduce them as independent and fix their relationship through eq. (II).

We now expect the constants A and D to depend on the just introduced particle size s. In a slight abuse of notation we will write $A \equiv A(s) \equiv As^{-\varepsilon}$ and $D \equiv D(s) \equiv Ds^{-\delta}$; note how eq. (8.9) now becomes

$$(k_{\mathrm{B}}T)^{\varphi} = q_1 A(s)^{-1} D(s)^{\frac{1+\varphi}{2}} = q_1 A^{-1} D^{\frac{1+\varphi}{2}} s^{\varepsilon - \frac{1+\varphi}{2}\delta}\,. \tag{8.10}$$

At this point we add another physical condition to our description: test-particles of different size should experience the same (effective) temperature, i.e. we set the above exponent on s to zero, leading to

$$\varepsilon = \frac{1+\varphi}{2}\delta\,. \tag{8.11}$$

We will rely on this relation for the remainder of this discussion and thus mostly omit ε from discussion.

Incorporating the size-dependence into eqs. (8.6) and (8.7), these equations become

$$x(t) = As^{-\varepsilon} f^{\varphi} t^{\sigma}\,, \qquad\qquad [A] = LS^{\varepsilon} F^{-\varphi} T^{-\sigma}\,, \tag{III}$$

$$\mathrm{MSD}(t) = Ds^{-\delta} t^{\alpha}\,, \qquad\qquad [D] = L^2 S^{\delta} T^{-\alpha}\,. \tag{IV}$$

The dimensional constraint resulting from eqs. (I) to (IV) then reads

$$0 \stackrel{!}{=} \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & -\nu & \varepsilon & \delta \\ 1 & 0 & -\varphi & 0 \\ 0 & 0 & -\sigma & -\alpha \end{vmatrix} = -\varphi\alpha\nu - \varepsilon\alpha + 2\nu\sigma + \sigma\delta - \nu\alpha \qquad \Leftrightarrow \qquad \alpha = \frac{2\nu + \delta}{(1+\varphi)\,\nu + \epsilon}\sigma\,, \tag{8.12}$$

which upon application of eq. (8.11) clearly reduces to eq. (8.8).

In summary, we can describe a test particle in dissipative medium by eqs. (I) to (IV), where the existence of a thermal energy scale (eq. (I)) and its independence of particle size (eq. (8.10)) are two physical assumptions about the system. The latter results in eq. (8.11), which together with the dimensional constraint eq. (8.12) can be summarized as the exponent relations

$$\alpha = \frac{2}{1+\varphi}\sigma\,, \qquad\qquad \epsilon = \frac{1+\varphi}{2}\delta\,. \tag{8.13}$$

The dimensional constraint also enforces the existence of a dimensionless constant (eq. (8.9)), given by

$$q_1 = \frac{D^{\frac{1+\varphi}{2}}}{A\,(k_\mathrm{B}T)^{\varphi}}\,. \tag{8.14}$$

For e.g. a Brownian particle (spherical particle in viscous solution) we have $\varphi = 1$, $\varepsilon = \delta = 1$, and $\alpha = \sigma = 1$, while the Einstein relation gives $q_1 = 2$ (note that the symbol D in our equations corresponds to *twice* the diffusion constant of the particle, since the MSD of a freely diffusing particle with diffusion constant $\tilde{D}$ is usually given by $\mathrm{MSD}(t) = 2\tilde{D}t$).

8.4 Polymer dynamics

Let us consider a quite specific type of test particle: a coil of polymer. A free (finite size) coil can be treated as a point particle and thus follows the treatment of the previous section, which accordingly is now interpreted as characterizing the interplay of the dissipative medium and the polymer. To access the internal dynamics of the polymer, we now connect infinitely many of these "test particle coils" together, such that we obtain an infinite polymer following the same structural scaling (II) as the individual particles. This last statement implies that the whole system is fractal, such that the scale of the initial coil is irrelevant and the following treatment applies on all scales (specifically, it also describes the internal dynamics of our original test

particle coil).

As mentioned before, we now take the size s of a finite (sub-)chain to be its length along the backbone, which relates to the physical extent R of the coil through eq. (II).

The internal dynamics of the coil are described by the behavior of a single locus on the chain. Similar to eqs. (8.6) and (8.7), this locus can respond to direct force application or fluctuations, resulting in the constitutive relations

$$x(t) = Bf^{\psi}t^{\rho}\,, \qquad [B] = LF^{-\psi}T^{-\rho}\,, \tag{V}$$

$$\mathrm{MSD}(t) = \Gamma t^{\mu}\,, \qquad [\Gamma] = L^2T^{-\mu}\,. \tag{VI}$$

Due to the algebraic similarity to eqs. (8.6) and (8.7) we expect that upon coupling these two equations by the constant $k_{\mathrm{B}}T$, we obtain the exponent relation

$$\mu = \frac{2}{1+\psi}\rho\,. \tag{8.15}$$

The full treatment of the complete system reads as follows. We assemble the matrix C (eq. (8.1)) from the constants in the constitutive relations (I) to (VI), giving

$$C = \begin{pmatrix} 1 & 1 & 1 & 2 & 1 & 2 \\ 0 & -\nu & \varepsilon & \delta & 0 & 0 \\ 1 & 0 & -\varphi & 0 & \psi & 0 \\ 0 & 0 & -\sigma & -\alpha & -\rho & -\mu \end{pmatrix}, \tag{8.16}$$

where the dimensions are sorted top to bottom as L, S, F, T. As outlined in section 8.1, the exponent relations then follow by setting all 4-by-4 sub-determinants to zero:

- Using columns 1 through 4 gives

$$\alpha = \frac{2\nu+\delta}{(1+\varphi)\nu+\varepsilon}\sigma\,, \tag{8.17}$$

  which splits into

$$\alpha = \frac{2}{1+\varphi}\sigma\,, \qquad \epsilon = \frac{1+\varphi}{2}\delta \tag{8.18}$$

  upon requiring a particle size independent temperature (eq. (8.10)). This implies the

existence of a dimensionless constant q_1, such that

$$D^{\frac{\sigma}{\alpha}} \left(G^{\frac{1}{\nu}}\right)^{\frac{\sigma}{\alpha}\delta-\varepsilon} = D^{\frac{1+\varphi}{2}} = q_1 A \left(k_{\mathrm{B}}T\right)^{\varphi} , \tag{8.19}$$

where the first expression corresponds to eq. (8.17), which simplifies to the second one upon applying eq. (8.18). This is a summary of section 8.3.

- Columns 1, 5, and 6—as expected—yield

$$0 \stackrel{!}{=} \begin{vmatrix} 1 & 1 & 2 \\ 1 & \psi & 0 \\ 0 & -\rho & -\mu \end{vmatrix} = -\psi\mu - 2\rho + \mu \qquad \Leftrightarrow \qquad \mu = \frac{2}{1+\psi}\rho \,. \tag{8.20}$$

This also implies the existence of a dimensionless constant q_2, satisfying

$$\Gamma^{\frac{1+\psi}{2}} = q_2 B \left(k_{\mathrm{B}}T\right)^{\psi} . \tag{8.21}$$

- Finally, we find a third constraint from columns 2, 4, and 6:

$$0 \stackrel{!}{=} \begin{vmatrix} 1 & 2 & 2 \\ -\nu & \delta & 0 \\ 0 & -\alpha & -\mu \end{vmatrix} = -\mu\delta + 2\nu\alpha - 2\nu\mu \qquad \Leftrightarrow \qquad \mu = \frac{2\nu\alpha}{2\nu+\delta} \,. \tag{8.22}$$

We define the corresponding dimensionless quantity q_3 such that

$$\Gamma^{\frac{1}{\mu}} = q_3 G^{\frac{\delta}{\nu\alpha}} D^{\frac{1}{\alpha}} . \tag{8.23}$$

From eqs. (8.19), (8.21) and (8.23) we see that we can express, respectively: A in terms of D and $k_{\mathrm{B}}T$; B in terms of Γ and $k_{\mathrm{B}}T$; and D in terms of G and Γ. In terms of the matrix C of eq. (8.16) this means that a maximal linearly independent subset of coulmns can be found

by eliminating the columns corresponding to A, B, and D, such that we are left with

$$c_1 \equiv \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \qquad c_2 \equiv \begin{pmatrix} 1 \\ -\nu \\ 0 \\ 0 \end{pmatrix} \qquad c_6 \equiv \begin{pmatrix} 2 \\ 0 \\ 0 \\ -\mu \end{pmatrix} . \tag{8.24}$$

Since $\nu \neq 0 \neq \mu$ by assumption, clearly $\hat{L} \equiv (1,0,0,0)^{\mathrm{T}} \notin \operatorname{span}\{c_1, c_2, c_6\} = \operatorname{im} C$, such that indeed the three constraints (8.18), (8.20) and (8.22) are not only necessary, but also sufficient to ensure that no length scale emerges.

8.5 Summary

We consider a system described by the constitutive relations

		$[k_{\mathrm{B}}T] = LF\,,$	(I)
	$R(s) = Gs^{\nu}$	$[G] = LS^{-\nu}\,,$	(II)
(free coil)	$x(t) = As^{-\varepsilon} f^{\varphi} t^{\sigma}$	$[A] = LS^{\varepsilon} F^{-\varphi} T^{-\sigma}\,,$	(III)
(free coil)	$\mathrm{MSD}(t) = Ds^{-\delta} t^{\alpha}$	$[D] = L^2 S^{\delta} T^{-\alpha}\,,$	(IV)
(locus)	$x(t) = Bf^{\psi} t^{\rho}$	$[B] = LF^{-\psi} T^{-\rho}\,,$	(V)
(locus)	$\mathrm{MSD}(t) = \Gamma t^{\mu}$	$[\Gamma] = L^2 T^{-\mu}\,.$	(VI)

The six constants $k_{\mathrm{B}}T$, G, A, D, B, and Γ can be combined into a length scale X—contradictory to our assumption of a scale-free model—*unless*

$$C = \begin{pmatrix} 1 & 1 & 1 & 2 & 1 & 2 \\ 0 & -\nu & \varepsilon & \delta & 0 & 0 \\ 1 & 0 & -\varphi & 0 & \psi & 0 \\ 0 & 0 & -\sigma & -\alpha & -\rho & -\mu \end{pmatrix} \tag{8.25}$$

has $\mathrm{rank} \leq 3$. Together with the physical requirement that temperature be independent of test particle size (eqs. (8.10) and (8.11)), this yields the exponent relations

$$\alpha = \frac{2}{1+\varphi}\sigma\,, \tag{8.26}$$

$$\varepsilon = \frac{1+\varphi}{2}\delta\,, \tag{8.27}$$

$$\mu = \frac{2}{1+\psi}\rho\,, \tag{8.28}$$

$$\mu = \frac{2\nu}{2\nu+\delta}\alpha\,. \tag{8.29}$$

Except for the physical constraint (8.27), these imply dimensionless quantities q_i, such that

$$D^{\frac{1+\varphi}{2}} = q_1 A\,(k_\mathrm{B}T)^{\varphi}\ , \tag{8.30}$$

$$\Gamma^{\frac{1+\psi}{2}} = q_2 B\,(k_\mathrm{B}T)^{\psi}\ , \tag{8.31}$$

$$\Gamma^{\frac{1}{\mu}} = q_3 G^{\frac{\delta}{\nu\alpha}} D^{\frac{1}{\alpha}}\,. \tag{8.32}$$

If we are interested predominantly in the internal polymer dynamics, we can summarize eqs. (8.28), (8.29), (8.31) and (8.32) as

$$\frac{2\nu\alpha}{2\nu+\delta} = \mu = \frac{2\rho}{1+\psi}\,, \tag{8.33}$$

$$\left(q_3 G^{\frac{\delta}{\nu\alpha}} D^{\frac{1}{\alpha}}\right)^{\mu} = \Gamma = \left(q_2 B\,(k_\mathrm{B}T)^{\psi}\right)^{\frac{2}{1+\psi}}\ . \tag{8.34}$$

8.6 Examples

The Rouse model of polymer dynamics (chapter 2) has

$$\text{(single locus dynamics)} \qquad \mu = \frac{1}{2}\,, \tag{8.35}$$

$$\text{(single locus force response)} \qquad \rho = \frac{1}{2}\,, \qquad \psi = 1\,, \tag{8.36}$$

$$\text{(whole coil diffusion)} \qquad \delta = 1\,, \qquad \alpha = 1\,, \tag{8.37}$$

$$\text{(free coil force response)} \qquad \epsilon = 1\,, \qquad \varphi = 1\,, \qquad \sigma = 1\,, \tag{8.38}$$

$$\text{(structure)} \qquad \nu = \frac{1}{2}\,; \tag{8.39}$$

$$q_1 = 2\,, \qquad q_2 = 2\,, \qquad q_3 = \frac{2}{\pi}\,. \tag{8.40}$$

Chapter 9

Conclusion and Outlook

Within the context of this thesis, chromosome structure can be subdivided into three aspects: structure, dynamics, and mechanics. While chromosome structure is by now a well-established area of research, studies of dynamics and mechanics are just growing out of their infancy. The present thesis presents my contributions to this program: we investigated the dynamics of chromatin loops with Anders Hansen, Christoph Zechner, and colleagues [13]; we probed the mechanical response of interphase chromatin with Antoine Coulon and colleagues [128]; and finally, we provided a scaling argument to connect structure, dynamics, and mechanics and devised a future path towards a unified model [146]. Multiple tools and techniques were developed along the way and made accessible to the community [21, 22, 99, 113, 147]. So, where do we go from here?

Starting from the end, I think one of the central goals to strive for over the next years is the development of a consistent null model of chromatin as a physical object—or the realization that no such thing exists in a useful way. Currently, depending on the use case and, frankly, what happens to fit the data best, one of a small set of scale-free models is often employed to "explain" experimental observations. The fact that the models used in the study of structure (fractal globule) and dynamics (Rouse) are mutually inconsistent is usually acknowledged, but otherwise ignored. As outlined in chapter 7, I believe that experiments in the not-too-distant future should be able to determine whether we can reconcile these models and build a single, consistent, scale-free null model of chromosome organization. Note that the counterfactual in this case is equally, if not even more exciting: if we find manifest violations of eq. (7.10), we might be forced to develop more detailed models of chromatin, moving the field beyond

its current obsession with powerlaw scalings. It then becomes an interesting question whether there even is such a thing as a useful null model of chromatin as a simple polymer, or whether it would be more fruitful to include more of the biological context in our descriptions.

On the applied side, both the locus pulling (chapter 3) as well as the loop quantification (chapter 4) projects provide initial forays into uncharted experimental territory and will lead to much exciting future work.

Our study of cohesin-mediated CTCF–CTCF-looping was the first to put a time scale to these loops, clearly pointing out their dynamic nature. Similar works have been published since then [98] or are in the works [148]; we expect these ongoing efforts to eventually lead to a comprehensive understanding of looping dynamics throughout the genome. Not least, of course, we also hope to further contribute to this field ourselves: the actual extrusion dynamics still evaded us in [13]. In fact, preliminary investigation with Henrik Pinholt showed that the data in that study seem to be symmetric under time reversal, indicating that loop extrusion as an *active* process might be mostly masked by experimental noise. Future improvements in the experimental system are expected to enable detection of this activity. More broadly, having established the statistical tools for looping inference (BILD; chapter 5), we now aim to move beyond studying purely structural features and investigate the interactions of enhancers and promoters. This should lead to major improvements in our mechanistic understanding of gene regulation.

Pulling on a genomic locus in interphase (chapter 3) has similarly not been achieved before, though here as well, exciting new developments are in the works [149]. Our study [128] was a proof of principle, opening the stage for more systematic exploration of biological questions. How is transcription impacted by these mechanical perturbations? Can we forcefully separate enhancer and promoter pairs, thus preventing transcription of target genes in the pulled part of the chromosome? This would provide an orthogonal—and quite unique—angle on enhancer–promoter interaction mentioned already above. In turn, how does transcription modify the mechanical properties of chromatin [66]? More mechanistically, how does strand passing (or the absence of it) affect the observed force response? We observed a response consistent with the Rouse model, which assumes a phantom chain (unhindered strand passing). Surprisingly enough, this process is actually possible in the nucleus, catalyzed by topoisomerase-II; inhibiting this enzyme would allow us to study to what extent the Rouse model is simply a useful effective description of the data, and to what extent it actually captures some of the underlying physics.

In a similar direction, what is the conformation of the pulled chain, and does it match the model predictions? In the existing data, we could label only the pulled locus itself; utilizing recent advances in imaging, we should be able to trace the conformation of the chromosome after the pulling, or potentially even track some parts of it during the pull. I believe that future studies of *in vivo* chromosome mechanics will lead to a host of valuable insights and therefore become firmly established as the third pillar in the study of chromosome organization, next to structure and dynamics.

Personally, I am excited to further explore these molecular, biological worlds. Much of what is so foundational to us as living beings is governed by physics that we have little natural intuition for, due to the vast mismatch of length scales—a human is about 10^5 times larger than a typical eukaryotic nucleus, and 10^8 times larger than a nucleosome, the smallest unit of genome structure (beyond the double helix itself). As such, studying any of these systems is usually indirect to some extent and requires rigorous data analysis and statistical treatment; often the fundamental task is not "go measure this, and see what it looks like", but rather "given these data, which part of it even contains any useful signal? What information is contained here and how do we access it?". In this regime, the region between what we can rigorously infer and what is fundamentally not knowable from a given data set is often just a thin ridge. Stringent method development then becomes imperative; this is where I see my comparative advantage as a theorist. Indeed, the biggest joy of working in an interdisciplinary field is that everyone has their comparative advantage: all my colleagues know things that I do not; and it is only by joining all our distinct expertise that major progress can be made.

Bibliography

[1] A. Piovesan, M. C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi, L. Vitale, On the length, weight and GC content of the human genome, *BMC Research Notes* **12**, 106 (2019), doi:10.1186/s13104-019-4137-z.

[2] J. D. Watson, F. H. C. Crick, Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid, *Nature* **171**, 737 (1953), doi:10.1038/171737a0.

[3] R. E. Franklin, R. G. Gosling, Molecular Configuration in Sodium Thymonucleate, *Nature* **171**, 740 (1953), doi:10.1038/171740a0.

[4] M. A. Ricci, C. Manzo, M. F. García-Parajo, M. Lakadamyali, M. P. Cosma, Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo, *Cell* **160**, 1145 (2015), doi:10.1016/j.cell.2015.01.054.

[5] F. Fatmaoui, P. Carrivain, D. Grewe, B. Jakob, J.-M. Victor, A. Leforestier, M. Eltsov, Cryo-electron tomography and deep learning denoising reveal native chromatin landscapes of interphase nuclei (2022). *bioRxiv* preprint, doi:10.1101/2022.08.16.502515.

[6] J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing Chromosome Conformation, *Science* **295**, 1306 (2002), doi:10.1126/science.1067799.

[7] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, *Science* **326**, 289 (2009), doi:10.1126/science.1181369.

[8] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, L. Mirny, Formation of Chromosomal Domains by Loop Extrusion, *Cell Reports* **15**, 2038 (2016), doi:10.1016/j.celrep.2016.04.085.

[9] G. Fudenberg, N. Abdennur, M. Imakaev, A. Goloborodko, L. Mirny, Emerging Evidence of Chromosome Folding by Loop Extrusion, *Cold Spring Harbor Symposia on Quantitative Biology* **82** (2018), doi:10.1101/sqb.2017.82.034710.

[10] M. Ganji, I. A. Shaltiel, S. Bisht, E. Kim, A. Kalichava, C. H. Haering, C. Dekker, Real-time imaging of DNA loop extrusion by condensin, *Science* **360**, 102 (2018), doi:10.1126/science.aar7831.

[11] S. Golfier, T. Quail, H. Kimura, J. Brugués, Cohesin and condensin extrude DNA loops in a cell cycle-dependent manner, *eLife* **9**, e53885 (2020), doi:10.7554/elife.53885.

[12] E. Kim, J. Kerssemakers, I. A. Shaltiel, C. H. Haering, C. Dekker, DNA-loop extruding condensin complexes can traverse one another, *Nature* **579**, 438 (2020), doi:10.1038/s41586-020-2067-5.

[13] M. Gabriele, H. B. Brandão, S. Grosse-Holz, A. Jha, G. M. Dailey, C. Cattoglio, T.-H. S. Hsieh, L. Mirny, C. Zechner, A. S. Hansen, Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging, *Science* **376**, 496 (2022), doi:10.1126/science.abn6583.

[14] A. N. Boettiger, B. Bintu, J. R. Moffitt, S. Wang, B. J. Beliveau, G. Fudenberg, M. Imakaev, L. A. Mirny, C.-t. Wu, X. Zhuang, Super-resolution imaging reveals distinct chromatin folding for different epigenetic states, *Nature* **529**, 418 (2016), doi:10.1038/nature16496.

[15] B. Bintu, L. J. Mateo, J.-H. Su, N. A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A. N. Boettiger, X. Zhuang, Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells, *Science* **362**, eaau1783 (2018), doi:10.1126/science.aau1783.

[16] A. Zidovska, D. A. Weitz, T. J. Mitchison, Micron-scale coherence in interphase chromatin dynamics, *Proceedings of the National Academy of Sciences* **110**, 15555 (2013), doi:10.1073/pnas.1220313110.

[17] M. Tamm, L. Nazarov, A. Gavrilov, A. Chertovich, Anomalous Diffusion in Fractal Globules, *Physical Review Letters* **114**, 178102 (2015), doi:10.1103/physrevlett.114.178102.

[18] P. E. Rouse, A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers, *The Journal of Chemical Physics* **21**, 1272 (1953), doi:10.1063/1.1699180.

[19] M. Doi, S. F. Edwards, *The theory of polymer dynamics*, International series of monographs on physics ; 73 (Clarendon Press, Oxford [Oxfordshire, 1988). Publication Title: The theory of polymer dynamics.

[20] M. Rubinstein, R. Colby, *Polymer Physics* (OUP Oxford, 2003).

[21] S. Grosse-Holz, `rouse`: An implementation of the rouse model of polymer dynamics, v0.1.0 (2022), pypi:rouse/0.1.0.

[22] S. Grosse-Holz, `rousepull`: Inference for locus pulling, based on the rouse model, v0.0.0 (2023), pypi:rousepull/0.0.0.

[23] R. E. Kalman, A New Approach to Linear Filtering and Prediction Problems, *Journal of Basic Engineering* **82**, 35 (1960), doi:10.1115/1.3662552.

[24] C. M. Bishop, *Pattern recognition and machine learning*, Information science and statistics (Springer, New York, 2006).

[25] F. Erdel, M. Baum, K. Rippe, The viscoelastic properties of chromatin and the nucleoplasm revealed by scale-dependent protein mobility, *Journal of Physics: Condensed Matter* **27**, 064115 (2015), doi:10.1088/0953-8984/27/6/064115.

[26] L. Liang, X. Wang, X. Da, T. Chen, W. R. Chen, Noninvasive determination of cell nucleoplasmic viscosity by fluorescence correlation spectroscopy, *Journal of Biomedical Optics* **14**, 024013 (2009), doi:10.1117/1.3088141.

[27] S. C. Weber, A. J. Spakowitz, J. A. Theriot, Bacterial Chromosomal Loci Move Subdiffusively through a Viscoelastic Cytoplasm, *Physical Review Letters* **104**, 238102 (2010), doi:10.1103/physrevlett.104.238102.

[28] S. C. Weber, J. A. Theriot, A. J. Spakowitz, Subdiffusive motion of a polymer composed of subdiffusive monomers, *Physical Review E* **82**, 011913 (2010), doi:10.1103/physreve.82.011913.

[29] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, S. Zhong, The 4D nucleome project, *Nature* **549**, 219 (2017), doi:10.1038/nature23884.

[30] D. Jost, P. Carrivain, G. Cavalli, C. Vaillant, Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains, *Nucleic Acids Research* **42**, 9553 (2014), doi:10.1093/nar/gku698.

[31] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, P. A. Sharp, A phase separation model for transcriptional control, *Cell* **169**, 13 (2017), doi:10.1016/j.cell.2017.02.007.

[32] L. A. Mirny, M. Imakaev, N. Abdennur, Two major mechanisms of chromosome organization, *Current Opinion in Cell Biology* **58**, 142 (2019), doi:10.1016/j.ceb.2019.05.001.

[33] E. J. Banigan, L. A. Mirny, Loop extrusion: theory meets single-molecule experiments, *Current Opinion in Cell Biology* **64**, 124 (2020), doi:10.1016/j.ceb.2020.04.011.

[34] J. D. Halverson, J. Smrek, K. Kremer, A. Y. Grosberg, From a melt of rings to chromosome territories: the role of topological constraints in genome folding, *Reports on Progress in Physics* **77** (2014), doi:10.1088/0034-4885/77/2/022601.

[35] C. Uhler, G. V. Shivashankar, Regulation of genome organization and gene expression by nuclear mechanotransduction, *Nature Reviews Molecular Cell Biology* **18**, 717 (2017), doi:10.1038/nrm.2017.101.

[36] A. A. Agbleke, A. Amitai, J. D. Buenrostro, A. Chakrabarti, L. Chu, A. S. Hansen, K. M. Koenig, A. S. Labade, S. Liu, T. Nozaki, S. Ovchinnikov, A. Seeber, H. A. Shaban, J.-H. Spille, A. D. Stephens, J.-H. Su, D. Wadduwage, Advances in chromatin and chromosome research: Perspectives from multiple fields, *Molecular Cell* **79**, 881 (2020), doi:10.1016/j.molcel.2020.07.003.

[37] T. Nozaki, R. Imai, M. Tanbo, R. Nagashima, S. Tamura, T. Tani, Y. Joti, M. Tomita, K. Hibino, M. T. Kanemaki, K. S. Wendt, Y. Okada, T. Nagai, K. Maeshima, Dynamic organization of chromatin domains revealed by super-resolution live-cell imaging, *Molecular Cell* **67**, 282 (2017), doi:10.1016/j.molcel.2017.06.018.

[38] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, B. Huang, Dynamic imaging of genomic loci in living human cells by an optimized crispr/cas system, *Cell* **155**, 1479 (2013), doi:10.1016/j.cell.2013.12.001.

[39] N. Khanna, Y. Zhang, J. S. Lucas, O. K. Dudko, C. Murre, Chromosome dynamics near the sol-gel phase transition dictate the timing of remote genomic interactions, *Nature Communications* **10**, 2771 (2019), doi:10.1038/s41467-019-10628-9.

[40] H. Strickfaden, T. O. Tolsma, A. Sharma, D. A. Underhill, J. C. Hansen, M. J. Hendzel, Condensed chromatin behaves like a solid on the mesoscale in vitro and in living cells, *Cell* **183**, 1772 (2020), doi:10.1016/j.cell.2020.11.027.

[41] A. D. Stephens, E. J. Banigan, S. A. Adam, R. D. Goldman, J. F. Marko, Chromatin and lamin a determine two different mechanical response regimes of the cell nucleus, *Molecular Biology of the Cell* **28**, 1984 (2017), doi:10.1091/mbc.e16-09-0653.

[42] Y. Shimamoto, S. Tamura, H. Masumoto, K. Maeshima, Nucleosome-nucleosome interactions via histone tails and linker dna regulate nuclear rigidity, *Molecular Biology of the Cell* **28**, 1580 (2017), doi:10.1091/mbc.e16-11-0783.

[43] K. Dahl, A. Engler, J. Pajerowski, D. Discher, Power-law rheology of isolated nuclei with deformation mapping of nuclear substructures, *Biophysical Journal* **89**, 2855 (2005), doi:10.1529/biophysj.105.062554.

[44] C. M. Hobson, M. Kern, E. T. O'Brien, A. D. Stephens, M. R. Falvo, R. Superfine, Correlating nuclear morphology and external force with combined atomic force microscopy and light sheet imaging separates roles of chromatin and lamin A/C in nuclear mechanics, *Molecular Biology of the Cell* **31**, 1788 (2020), doi:10.1091/mbc.e20-01-0073.

[45] A. Tajik, Y. Zhang, F. Wei, J. Sun, Q. Jia, W. Zhou, R. Singh, N. Khanna, A. S. Belmont, N. Wang, Transcription upregulation via force-induced direct stretching of chromatin, *Nature Materials* **15**, 1287 (2016), doi:10.1038/nmat4729.

[46] A. H. B. de Vries, B. E. Krenn, R. van Driel, V. Subramaniam, J. S. Kanger, Direct observation of nanomechanical properties of chromatin in living cells, *Nano Letters* **7**, 1424 (2007), doi:10.1021/nl070603+.

[47] Y. Shin, Y.-C. Chang, D. S. W. Lee, J. Berry, D. W. Sanders, P. Ronceray, N. S. Wingreen, M. Haataja, C. P. Brangwynne, Liquid nuclear condensates mechanically sense and restructure the genome, *Cell* **175**, 1481 (2018), doi:10.1016/j.cell.2018.10.057.

[48] C. M. Caragine, S. C. Haley, A. Zidovska, Nucleolar dynamics and interactions with nucleoplasm in living cells, *eLife* **8** (2019), doi:10.7554/elife.47533.

[49] M. S. Syrchina, A. M. Shakhov, A. V. Aybush, V. A. Nadtochenko, Optical trapping of nucleolus reveals viscoelastic properties of nucleoplasm inside mouse germinal vesicle oocytes (2020). *bioRxiv* preprint, doi:10.1101/2020.03.19.999342.

[50] M. Mittasch, A. W. Fritsch, M. Nestler, J. M. Iglesias-Artola, K. Subramanian, H. Petzold, M. Kar, A. Voigt, M. Kreysing, Active gelation breaks time-reversal-symmetry of mitotic chromosome mechanics (2018). *bioRxiv* preprint, doi:10.1101/296566.

[51] B. Seelbinder, M. Jain, E. Erben, S. Klykov, I. D. Stoev, M. Kreysing, Non-invasive Chromatin Deformation and Measurement of Differential Mechanical Properties in the Nucleus (2021). *bioRxiv* preprint, doi:10.1101/2021.12.15.472786.

[52] S. Janicki, T. Tsukamoto, S. Salghetti, W. Tansey, R. Sachidanandam, K. Prasanth, T. Ried, Y. Shav-Tal, E. Bertrand, R. Singer, D. Spector, From silencing to gene expression: Real-time analysis in single cells, *Cell* **116**, 683 (2004), doi:10.1016/s0092-8674(04)00171-0.

[53] C. Monzel, C. Vicario, J. Piehler, M. Coppey, M. Dahan, Magnetic control of cellular processes using biofunctional nanoparticles, *Chemical Science* **8**, 7330 (2017), doi:10.1039/c7sc01462g.

[54] D. Lisse, C. Monzel, C. Vicario, J. Manzi, I. Maurin, M. Coppey, J. Piehler, M. Dahan, Engineered ferritin for magnetogenetic manipulation of proteins and organelles inside living cells, *Advanced Materials* **29** (2017), doi:10.1002/adma.201700189.

[55] T. Tsukamoto, N. Hashiguchi, S. Janicki, T. Tumbar, A. Belmont, D. Spector, Visualization of gene activity in living cells, *Nature Cell Biology* **2**, 871 (2000), doi:10.1038/35046510.

[56] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, R. H. Singer, In vivo dynamics of rna polymerase ii transcription, *Nature Structural & Molecular Biology* **14**, 796 (2007), doi:10.1038/nsmb1280.

[57] L. Toraille, K. Aizel, E. Balloul, C. Vicario, C. Monzel, M. Coppey, E. Secret, J.-M. Siaugue, J. Sampaio, S. Rohart, N. Vernier, L. Bonnemay, T. Debuisschert, L. Rondin, J.-F. Roch, M. Dahan, Optical magnetometry of single biocompatible micromagnets for quantitative magnetogenetic and magnetomechanical assays, *Nano Letters* **18**, 7635 (2018), doi:10.1021/acs.nanolett.8b03222.

[58] M. Bongaerts, K. Aizel, E. Secret, A. Jan, T. Nahar, F. Raudzus, S. Neumann, N. Telling, R. Heumann, J.-M. Siaugue, C. Ménager, J. Fresnais, C. Villard, A. E. Haj, J. Piehler, M. A. Gates, M. Coppey, Parallelized Manipulation of Adherent Living Cells by Magnetic Nanoparticles-Mediated Forces, *International Journal of Molecular Sciences* **21**, 6560 (2020), doi:10.3390/ijms21186560.

[59] E. A. Galburt, S. W. Grill, C. Bustamante, Single molecule transcription elongation, *Methods* **48**, 323 (2009), doi:10.1016/j.ymeth.2009.04.021.

[60] B. Gu, T. Swigut, A. Spencley, M. R. Bauer, M. Chung, T. Meyer, J. Wysocka, Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements, *Science* **359**, 1050 (2018), doi:10.1126/science.aao3136.

[61] J. Li, A. Dong, K. Saydaminova, H. Chang, G. Wang, H. Ochiai, T. Yamamoto, A. Pertsinidis, Single-molecule nanoscopy elucidates rna polymerase ii transcription at single genes in live cells, *Cell* **178**, 491 (2019), doi:10.1016/j.cell.2019.05.029.

[62] H. Hajjoul, J. Mathon, H. Ranchon, I. Goiffon, J. Mozziconacci, B. Albert, P. Carrivain, J.-M. Victor, O. Gadal, K. Bystricky, A. Bancaud, High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome, *Genome Research* **23**, 1829 (2013), doi:10.1101/gr.157008.113.

[63] I. Solovei, A. S. Wang, K. Thanisch, C. S. Schmidt, S. Krebs, M. Zwerger, T. V. Cohen, D. Devys, R. Foisner, L. Peichl, H. Herrmann, H. Blum, D. Engelkamp, C. L. Stewart, H. Leonhardt, B. Joffe, Lbr and lamin a/c sequentially tether peripheral heterochromatin and inversely regulate differentiation, *Cell* **152**, 584 (2013), doi:10.1016/j.cell.2013.01.009.

[64] C. Chuang, A. Carpenter, B. Fuchsova, T. Johnson, P. de Lanerolle, A. Belmont, Long-range directional movement of an interphase chromosome site, *Current Biology* **16**, 825 (2006), doi:10.1016/j.cub.2006.03.059.

[65] N. Khanna, Y. Hu, A. S. Belmont, Hsp70 transgene directed motion to nuclear speckles facilitates heat shock activation, *Current Biology* **24**, 1138 (2014), doi:10.1016/j.cub.2014.03.053.

[66] S. Leidescher, J. Ribisel, S. Ullrich, Y. Feodorova, E. Hildebrand, A. Galitsyna, S. Bultmann, S. Link, K. Thanisch, C. Mulholland, J. Dekker, H. Leonhardt, L. Mirny, I. Solovei, Spatial organization of transcribed eukaryotic genes, *Nature Cell Biology* **24**, 327 (2022), doi:10.1038/s41556-022-00847-6.

[67] C. P. Caridi, C. D'Agostino, T. Ryu, G. Zapotoczny, L. Delabaere, X. Li, V. Y. Khodaverdian, N. Amaral, E. Lin, A. R. Rau, I. Chiolo, Nuclear f-actin and myosins drive relocalization of heterochromatic breaks, *Nature* **559**, 54 (2018), doi:10.1038/s41586-018-0242-8.

[68] J. A. Beagan, J. E. Phillips-Cremins, On the existence and functionality of topologically associating domains, *Nature Genetics* (2020), doi:10.1038/s41588-019-0561-1.

[69] Y. Kim, Z. Shi, H. Zhang, I. J. Finkelstein, H. Yu, Human cohesin compacts DNA by loop extrusion, *Science* **366**, 1345 LP (2019), doi:10.1126/science.aaz4475.

[70] I. F. Davidson, B. Bauer, D. Goetz, W. Tang, G. Wutz, J.-M. Peters, DNA loop extrusion by human cohesin, *Science* **366**, 1338 LP (2019), doi:10.1126/science.aaz3418.

[71] S. Golfier, T. Quail, H. Kimura, J. Brugués, Cohesin and condensin extrude DNA loops in a cell cycle-dependent manner, *eLife* **9**, e53885 (2020), doi:10.7554/elife.53885.

[72] O. Symmons, V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller, F. Spitz, Functional and topological characteristics of mammalian regulatory domains, *Genome Research* (2014), doi:10.1101/gr.163519.113.

[73] M. V. Arrastia, J. W. Jachowicz, N. Ollikainen, M. S. Curtis, C. Lai, S. A. Quinodoz, D. A. Selck, R. F. Ismagilov, M. Guttman, Single-cell measurement of higher-order 3D genome organization with scSPRITE, *Nature Biotechnology* (2021), doi:10.1038/s41587-021-00998-1.

[74] Q. Szabo, D. Jost, J.-M. Chang, D. I. Cattoni, G. L. Papadopoulos, B. Bonev, T. Sexton, J. Gurgo, C. Jacquier, M. Nollmann, F. Bantignies, G. Cavalli, TADs are 3D structural units of higher-order chromosome organization in Drosophila, *Science Advances* **4** (2018), doi:10.1126/sciadv.aar8082.

[75] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman, E. D. Laue, 3D structures of individual mammalian genomes studied by single-cell Hi-C, *Nature* **544**, 59 (2017), doi:10.1038/nature21429.

[76] E. H. Finn, G. Pegoraro, H. B. Brandão, A.-L. Valton, M. E. Oomen, J. Dekker, L. Mirny, T. Misteli, Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization, *Cell* **176**, 1502 (2019), doi:10.1016/j.cell.2019.01.020.

[77] D. I. Cattoni, A. M. Cardozo Gizzi, M. Georgieva, M. Di Stefano, A. Valeri, D. Chamousset, C. Houbron, S. Déjardin, J.-B. Fiche, I. González, J.-M. Chang, T. Sexton, M. A.

Marti-Renom, F. Bantignies, G. Cavalli, M. Nollmann, Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions, *Nature Communications* **8**, 1753 (2017), doi:10.1038/s41467-017-01962-x.

[78] H. B. Brandão, M. Gabriele, A. S. Hansen, Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging, *Current Opinion in Cell Biology* **70**, 18 (2021), doi:10.1016/j.ceb.2020.11.002.

[79] H. Chen, M. Levo, L. Barinov, M. Fujioka, J. B. Jaynes, T. Gregor, Dynamic interplay between enhancer–promoter topology and gene activity, *Nature Genetics* **50**, 1296 (2018), doi:10.1038/s41588-018-0175-z.

[80] J. M. Alexander, J. Guan, B. Li, L. Maliskova, M. Song, Y. Shen, B. Huang, S. Lomvardas, O. D. Weiner, Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity, *eLife* **8**, e41769 (2019), doi:10.7554/elife.41769.

[81] Z. Hensel, X. Weng, A. C. Lagda, J. Xiao, Transcription-factor-mediated dna looping probed by high-resolution, single-molecule imaging in live e. coli cells, *PLOS Biology* **11**, 1 (2013), doi:10.1371/journal.pbio.1001591.

[82] E. de Wit, E. S. Vos, S. J. Holwerda, C. Valdes-Quezada, M. J. Verstegen, H. Teunissen, E. Splinter, P. J. Wijchers, P. H. Krijger, W. de Laat, CTCF Binding Polarity Determines Chromatin Looping, *Molecular Cell* **60**, 676 (2015), doi:10.1016/j.molcel.2015.09.023.

[83] T. Germier, S. Kocanova, N. Walther, A. Bancaud, H. A. Shaban, H. Sellou, A. Z. Politi, J. Ellenberg, F. Gallardo, K. Bystricky, Real-Time Imaging of a Single Gene Reveals Transcription-Initiated Local Confinement, *Biophysical Journal* **113**, 1383 (2017), doi:10.1016/j.bpj.2017.08.014.

[84] T.-H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, X. Darzacq, Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding, *Molecular Cell* (2020), doi:10.1016/j.molcel.2020.03.002.

[85] A. S. Hansen, T.-H. S. Hsieh, C. Cattoglio, I. Pustova, R. Saldaña-Meyer, D. Reinberg, X. Darzacq, R. Tjian, Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF, *Molecular Cell* (2019), doi:10.1016/j.molcel.2019.07.039.

[86] T. Natsume, T. Kiyomitsu, Y. Saga, M. T. Kanemaki, Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors, *Cell Reports* **15**, 210 (2016), doi:10.1016/j.celrep.2016.03.001.

[87] S. S. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, E. L. Aiden, Cohesin Loss Eliminates All Loop Domains, *Cell* **171**, 305 (2017), doi:10.1016/j.cell.2017.09.026.

[88] E. P. Nora, A. Goloborodko, A. L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, B. G. Bruneau, Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization, *Cell* (2017), doi:10.1016/j.cell.2017.05.004.

[89] G. Wutz, C. Várnai, K. Nagasaka, D. A. Cisneros, R. R. Stocsits, W. Tang, S. Schoenfelder, G. Jessberger, M. Muhar, M. J. Hossain, N. Walther, B. Koch, M. Kueblbeck, J. Ellenberg, J. Zuber, P. Fraser, J.-M. Peters, Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins, *The EMBO Journal* (2017), doi:10.15252/embj.201798004.

[90] T.-H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, X. Darzacq, R. Tjian, Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1, *Nature Genetics* **54**, 1919 (2022), doi:10.1038/s41588-022-01223-8.

[91] A. Tedeschi, G. Wutz, S. Huet, M. Jaritz, A. Wuensche, E. Schirghuber, I. F. Davidson, W. Tang, D. A. Cisneros, V. Bhaskara, T. Nishiyama, A. Vaziri, A. Wutz, J. Ellenberg, J.-M. Peters, Wapl is an essential regulator of chromatin structure and chromosome segregation, *Nature* **501**, 564 (2013), doi:10.1038/nature12471.

[92] J. H. I. Haarhuis, R. H. van der Weide, V. A. Blomen, J. O. Yáñez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit, B. D. Rowland, The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension, *Cell* **169**, 693 (2017), doi:10.1016/j.cell.2017.04.013.

[93] Y. Li, J. H. I. Haarhuis, Á. Sedeño Cacciatore, R. Oldenkamp, M. S. van Ruiten, L. Willems, H. Teunissen, K. W. Muir, E. de Wit, B. D. Rowland, D. Panne, The structural basis for cohesin–CTCF-anchored loops, *Nature* **578**, 472 (2020), doi:10.1038/s41586-019-1910-z.

[94] G. Wutz, R. Ladurner, B. G. St Hilaire, R. R. Stocsits, K. Nagasaka, B. Pignard, A. Sanborn, W. Tang, C. Várnai, M. P. Ivanov, S. Schoenfelder, P. van der Lelij, X. Huang, G. Dürnberger, E. Roitinger, K. Mechtler, I. F. Davidson, P. Fraser, E. Lieberman-Aiden, J.-M. Peters, ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesinSTAG1 from WAPL, *eLife* **9**, e52091 (2020), doi:10.7554/elife.52091.

[95] C. Cattoglio, I. Pustova, N. Walther, J. J. Ho, M. Hantsche-Grininger, C. J. Inouye, M. J. Hossain, G. M. Dailey, J. Ellenberg, X. Darzacq, R. Tjian, A. S. Hansen, Determining cellular CTCF and cohesin abundances to constrain 3D genome models, *eLife* **8**, e40164 (2019), doi:10.7554/elife.40164.

[96] A. S. Hansen, I. Pustova, C. Cattoglio, R. Tjian, X. Darzacq, CTCF and cohesin regulate chromatin loop stability with distinct dynamics, *eLife* **6**, e25776 (2017), doi:10.7554/elife.25776.

[97] H. B. Brandão, Z. Ren, X. Karaboja, L. A. Mirny, X. Wang, DNA-loop-extruding SMC complexes can traverse one another in vivo, *Nature Structural & Molecular Biology* **28**, 642 (2021), doi:10.1038/s41594-021-00626-1.

[98] P. Mach, P. I. Kos, Y. Zhan, J. Cramard, S. Gaudin, J. Tünnermann, E. Marchi, J. Eglinger, J. Zuin, M. Kryzhanovska, S. Smallwood, L. Gelman, G. Roth, E. P. Nora, G. Tiana, L. Giorgetti, Cohesin and CTCF control the dynamics of chromosome folding, *Nature Genetics* **54**, 1907 (2022), doi:10.1038/s41588-022-01232-7.

[99] S. Grosse-Holz, `bild`: Bayesian inference of looping dynamics, v0.0.4 (2022), pypi:bild/0.0.4.

[100] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118 of *Lecture Notes in Statistics* (Springer New York, New York, NY, 1996).

[101] D. MacKay, J. MacKay, D. Kay, C. U. Press, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).

[102] J.-M. Cornuet, J.-M. Marin, A. Mira, C. P. Robert, Adaptive Multiple Importance Sampling, *Scandinavian Journal of Statistics* **39**, 798 (2012), doi:10.1111/j.1467-9469.2011.00756.x.

[103] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, J. Gottschlich, *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett, eds. (Curran Associates, Inc., 2018), vol. 31.

[104] E. L. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association* **53**, 457 (1958), doi:10.1080/01621459.1958.10501452.

[105] J. Kalbfleisch, R. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics (Wiley, 2002).

[106] D. W. Hosmer, S. Lemeshow, S. May, *Applied survival analysis: regression modeling of time-to-event data*, Wiley series in probability and statistics (Wiley-Interscience, Hoboken, N.J, 2008), second edn. OCLC: ocn154798742.

[107] J.-M. Arbona, S. Herbert, E. Fabre, C. Zimmer, Inferring the physical properties of yeast chromatin through Bayesian analysis of whole nucleus simulations, *Genome Biology* **18**, 81 (2017), doi:10.1186/s13059-017-1199-x.

[108] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, H. E. Stanley, Lévy flight search patterns of wandering albatrosses, *Nature* **381**, 413 (1996), doi:10.1038/381413a0.

[109] G. Ramos-Fernández, J. L. Mateos, O. Miramontes, G. Cocho, H. Larralde, B. Ayala-Orozco, Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi), *Behavioral Ecology and Sociobiology* **55**, 223 (2004), doi:10.1007/s00265-003-0700-6.

[110] E. Kepten, A. Weron, G. Sikora, K. Burnecki, Y. Garini, Guidelines for the Fitting of Anomalous Diffusion Mean Square Displacement Graphs from Single Particle Tracking Experiments, *PLOS ONE* **10**, e0117722 (2015), doi:10.1371/journal.pone.0117722.

[111] C. L. Vestergaard, P. C. Blainey, H. Flyvbjerg, Optimal estimation of diffusion coefficients from single-particle trajectories, *Physical Review E* **89** (2014), doi:10.1103/physreve.89.022726.

[112] M. B. Marcus, *Markov processes, Gaussian processes, and local times*, no. 100 in Cambridge studies in advanced mathematics (Cambridge University Press, Cambridge ;, 2006).

[113] S. Grosse-Holz, `bayesmsd`: Bayesian MSD fitting, v0.1.2 (2023), pypi:bayesmsd/0.1.2.

[114] C. de Boor, *A Practical Guide to Splines* (Springer, 1978).

[115] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nature Methods* **17**, 261 (2020), doi:10.1038/s41592-019-0686-2.

[116] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell* (Garland Science, Taylor & Francis Group, LLC, New York, 2015), 6th edn.

[117] B. Doyle, G. Fudenberg, M. Imakaev, L. A. Mirny, Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions, *PLOS Computational Biology* **10**, e1003867 (2014), doi:10.1371/journal.pcbi.1003867.

[118] J. Zuin, G. Roth, Y. Zhan, J. Cramard, J. Redolfi, E. Piskadlo, P. Mach, M. Kryzhanovska, G. Tihanyi, H. Kohler, M. Eder, C. Leemans, B. van Steensel, P. Meister, S. Smallwood, L. Giorgetti, Nonlinear control of transcription through enhancer–promoter interactions, *Nature* **604**, 571 (2022), doi:10.1038/s41586-022-04570-y.

[119] M. A. Zabidi, A. Stark, Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors, *Trends in Genetics* **32**, 801 (2016), doi:10.1016/j.tig.2016.10.003.

[120] B. Lim, M. S. Levine, Enhancer-promoter communication: hubs or loops?, *Current Opinion in Genetics & Development* **67**, 5 (2021), doi:10.1016/j.gde.2020.10.001.

[121] O. Kyrchanova, P. Georgiev, Mechanisms of Enhancer-Promoter Interactions in Higher Eukaryotes, *International Journal of Molecular Sciences* **22**, 671 (2021), doi:10.3390/ijms22020671.

[122] C. C. Galouzis, E. E. M. Furlong, Regulating specificity in enhancer–promoter communication, *Current Opinion in Cell Biology* **75**, 102065 (2022), doi:10.1016/j.ceb.2022.01.010.

[123] Y. Takei, J. Yun, S. Zheng, N. Ollikainen, N. Pierson, J. White, S. Shah, J. Thomassie, S. Suo, C.-H. L. Eng, M. Guttman, G.-C. Yuan, L. Cai, Integrated spatial genomics reveals global architecture of single nuclei, *Nature* **590**, 344 (2021), doi:10.1038/s41586-020-03126-2.

[124] S. Ide, S. Tamura, K. Maeshima, Chromatin behavior in living cells: Lessons from single-nucleosome imaging and tracking, *BioEssays* **44**, 2200043 (2022), doi:10.1002/bies.202200043.

[125] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping, *Cell* **159**, 1665 (2014), doi:10.1016/j.cell.2014.11.021.

[126] S. Wang, J.-H. Su, B. J. Beliveau, B. Bintu, J. R. Moffitt, C.-t. Wu, X. Zhuang, Spatial organization of chromatin domains and compartments in single chromosomes, *Science* **353**, 598 (2016), doi:10.1126/science.aaf8084.

[127] J. D. Halverson, W. B. Lee, G. S. Grest, A. Y. Grosberg, K. Kremer, Molecular dynamics simulation study of nonconcatenated ring polymers in a melt. I. Statics, *The Journal of Chemical Physics* **134**, 204904 (2011), doi:10.1063/1.3587137.

[128] V. I. P. Keizer, S. Grosse-Holz, M. Woringer, L. Zambon, K. Aizel, M. Bongaerts, F. Delille, L. Kolar-Znika, V. F. Scolari, S. Hoffmann, E. J. Banigan, L. A. Mirny, M. Dahan, D. Fachinetti, A. Coulon, Live-cell micromanipulation of a genomic locus reveals interphase chromatin mechanics, *Science* **377**, 489 (2022), doi:10.1126/science.abi9810.

[129] A. Grosberg, Y. Rabin, S. Havlin, A. Neer, Crumpled Globule Model of the Three-Dimensional Structure of DNA, *Europhysics Letters* **23**, 373 (1993), doi:10.1209/0295-5075/23/5/012.

[130] D. B. Brückner, H. Chen, L. Barinov, B. Zoller, T. Gregor, Stochastic motion and transcriptional dynamics of distal enhancer–promoter pairs on a compacted chromosome (2023). *bioRxiv* preprint, doi:10.1101/2023.01.18.524527.

[131] I. Golding, E. C. Cox, Physical Nature of Bacterial Cytoplasm, *Physical Review Letters* **96**, 098102 (2006), doi:10.1103/physrevlett.96.098102.

[132] M. Weiss, Probing the Interior of Living Cells with Fluorescence Correlation Spectroscopy, *Annals of the New York Academy of Sciences* **1130**, 21 (2008), doi:10.1196/annals.1430.002.

[133] M. Baum, F. Erdel, M. Wachsmuth, K. Rippe, Retrieving the intracellular topology from multi-scale protein mobility mapping in living cells, *Nature Communications* **5**, 4494 (2014), doi:10.1038/ncomms5494.

[134] P. G. De Gennes, Dynamics of Entangled Polymer Solutions. I. The Rouse Model, *Macromolecules* **9**, 587 (1976), doi:10.1021/ma60052a011.

[135] K. Polovnikov, M. Gherardi, M. Cosentino-Lagomarsino, M. Tamm, Fractal Folding and Medium Viscoelasticity Contribute Jointly to Chromosome Dynamics, *Physical Review Letters* **120**, 088101 (2018), doi:10.1103/physrevlett.120.088101.

[136] H. Schiessel, G. Oshanin, A. Blumen, Dynamics and conformational properties of polyampholytes in external electrical fields, *The Journal of Chemical Physics* **103**, 5070 (1995), doi:10.1063/1.470593.

[137] I. Bronshtein, I. Kanter, E. Kepten, M. Lindner, S. Berezin, Y. Shav-Tal, Y. Garini, Exploring chromatin organization mechanisms through its dynamic properties, *Nucleus* **7**, 27 (2016), doi:10.1080/19491034.2016.1139272.

[138] W. Schwarzer, N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N. A. Fonseca, W. Huber, C. H. Haering, L. Mirny, F. Spitz, Two independent modes of chromatin organization revealed by cohesin removal, *Nature* **551**, 51 (2017), doi:10.1038/nature24281.

[139] S. Kim, I. Liachko, D. G. Brickner, K. Cook, W. S. Noble, J. H. Brickner, J. Shendure, M. J. Dunham, The dynamic three-dimensional organization of the diploid yeast genome, *eLife* **6**, e23623 (2017), doi:10.7554/elife.23623.

[140] T. Shu, T. Szórádi, G. R. Kidiyoor, Y. Xie, N. L. Herzog, A. Bazley, M. Bonucci, S. Keegan, S. Saxena, F. Ettefa, G. Brittingham, J. Lemiere, D. Fenyö, F. Chang, M. Delarue, L. J. Holt, nucGEMs probe the biophysical properties of the nucleoplasm (2021). *bioRxiv* preprint, doi:10.1101/2021.11.18.469159.

[141] P. G. De Gennes, Dynamics of Entangled Polymer Solutions. II. Inclusion of Hydrodynamic Interactions, *Macromolecules* **9**, 594 (1976), doi:10.1021/ma60052a012.

[142] J. Skolnick, Perspective: On the importance of hydrodynamic interactions in the subcellular dynamics of macromolecules, *The Journal of Chemical Physics* **145**, 100901 (2016), doi:10.1063/1.4962258.

[143] I. M. Tolić-Nørrelykke, E.-L. Munteanu, G. Thon, L. Oddershede, K. Berg-Sørensen, Anomalous Diffusion in Living Yeast Cells, *Physical Review Letters* **93**, 078102 (2004), doi:10.1103/physrevlett.93.078102.

[144] G. L. Lukacs, P. Haggie, O. Seksek, D. Lechardeur, N. Freedman, A. S. Verkman, Size-dependent DNA Mobility in Cytoplasm and Nucleus, *Journal of Biological Chemistry* **275**, 1625 (2000), doi:10.1074/jbc.275.3.1625.

[145] M. P. H. Stumpf, M. A. Porter, Critical Truths About Power Laws, *Science* **335**, 665 (2012), doi:10.1126/science.1216142.

[146] S. Grosse-Holz, A. Coulon, L. Mirny, Scale-free models of chromosome structure, dynamics, and mechanics (2023). *bioRxiv* preprint, doi:10.1101/2023.04.14.536939.

[147] S. Grosse-Holz, `noctiluca`: SPT trajectory analysis, v0.1.1 (2023), pypi:noctiluca/0.1.1.

[148] T. Sabaté, B. Lelandais, E. Bertrand, C. Zimmer, Polymer simulations guide the detection and quantification of chromatin loop extrusion by imaging, *Nucleic Acids Research* **51**, 2614 (2023), doi:10.1093/nar/gkad034.

[149] A. R. Strom, Y. Kim, H. Zhao, N. Orlovsky, Y.-C. Chang, A. Košmrlj, C. Storm, C. P. Brangwynne, Condensate-driven interfacial forces reposition DNA loci and measure chromatin viscoelasticity (2023). *bioRxiv* preprint, doi:10.1101/2023.02.27.530281.