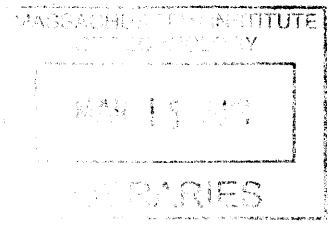
Computational Tools for Modeling and Measuring Chromosome Structure



by

Brian Christopher Ross

B.S., University of Maryland (1998)

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

September 2012

Author .....
Department of Physics
June 11, 2012

Certified by .....
Alexander van Oudenaarden
Professor
Thesis Supervisor

Certified by .....
Paul Wiggins
Assistant Professor
Thesis Supervisor

Accepted by .....
Krishna Rajagopal
Associate Department Head for Education

Computational Tools for Modeling and Measuring Chromosome Structure

by

Brian Christopher Ross

Submitted to the Department of Physics
on June 11, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Physics

Abstract

DNA conformation within cells has many important biological implications, but there are challenges both in modeling DNA due to the need for specialized techniques, and experimentally since tracing out *in vivo* conformations is currently impossible. This thesis contributes two computational projects to these efforts. The first project is a set of online and offline calculators of conformational statistics using a variety of published and unpublished methods, addressing the current lack of DNA model-building tools intended for general use. The second project is a reconstructive analysis that could enable *in vivo* mapping of DNA conformation at high resolution with current experimental technology.

Thesis Supervisor: Alexander van Oudenaarden
Title: Professor

Thesis Supervisor: Paul Wiggins
Title: Assistant Professor

Acknowledgments

I want to acknowledge at the outset that much of the work presented here as my own originated from conversations of my advisor, coworkers, thesis committee and even friends and family. In particular I am grateful to Hyun Jin Lee for help with microscopy, David Johnson for advice on the Traveling Salesman Problem, Henrik Vestermark for the use of his complex root solver in the Wormulator, and numerous coworkers for their helpful comments on both the oral and written parts of my thesis. Most of all I want to thank my research mentor Paul Wiggins for his patient guidance and consistent support of me throughout my graduate career. Paul gave me more credit than I deserved for many of his ideas, and trusted me with my own ideas more than I had any right to expect, for which I am enduringly grateful.

Finally, this thesis would have been much more difficult without the considerable patience and support from my friends and family. Longtime friends Pak and John helped me find a research group; many other friends gave critical support and advice during this time. My lovely girlfriend Yanghee endured many late evenings at the lab while I was working on my 3d-alignment method. A number of friends and family (from Maryland!) gave considerable moral support from the back rows at my thesis defense. My heartfelt thanks go out to everyone who supported and encouraged me throughout this long process.

Contents

List of Figures

Chapter 1

Introduction

The goal of structural biology is to build an engineering blueprint of an organism that catalogs all of its molecular parts, where they reside and how they fit together to form a cell. Of all these molecules, none is more famous or (arguably) more important than the double strands of DNA called chromosomes whose base-pair sequences control the cell. Owing to the central role of DNA and the accumulating lines of evidence that DNA structure plays a large factor in many biological processes, a major research effort is now devoted to understanding how chromosomes are arranged inside of the cell. The importance of this effort was captured in the closing paragraph of a recent commentary on epigenomics[7], which christened the task of uncovering chromosome structure the 'Fourth Era' in the genomics revolution.

The sheer size of a chromosome gives it a structure unlike that of any other molecule. A single chromosome can sprawl from one end of its cell to the other a thousand times, forming an unfathomable tangle that is apparently organized in biologically important ways. Modeling this enormous and complex structure is complicated by the fact that the small-scale mechanics are poorly understood, and that a wide range of length and energy scales are important for various biological processes. Experimentally photographing the structure of a chromosome in a single cell is quite impossible with current technology, although coarse-grained and cell-averaged structural data are now becoming available.

This thesis contributes two computational tools to the current effort to understand

chromosome structure. The first of these, called the 'Wormulator', bundles a suite of methods for calculating conformational statistics into a usable modeling tool for the scientific community. Chapter 2 describes the Wormulator and demonstrates how it can be applied to refine models of DNA bending. In Chapter 3 we propose an experiment that could obtain extended high-resolution, single-cell conformations of DNA. The proposed experiment requires a '3d-alignment' data analysis which we demonstrate with our second tool.

The remainder of Chapter 1 briefly reviews the current state of knowledge of DNA mechanics, the *in vivo* conformation of chromosomes, and relevant experimental techniques with an emphasis on super-resolution fluorescence imaging.

1.1 DNA mechanics

A general goal of DNA structure models is to determine the free energy required to impose certain conformational constraints on the DNA molecule. For example, the cell may wish to bring together a gene and an enhancer, or wrap 150 base pairs around a nucleosome, or supercoil a topological domain. These activities are typical of those involved in chromosome packaging and segregation, gene regulation and other essential cellular activities. The energetics of these processes help determine when and to what extent they happen in a live cell, and those energetics are in turn determined by basic mechanical properties of DNA.

1.1.1 DNA mechanical parameters

The starting point for any model of large-scale chromosome structure is a fine-scale DNA model which specifies mechanical properties such as stiffness. Measurements of these properties have typically fallen into two classes: those related to the consecutive base pairs that form a dinucleotide step, and bulk properties such as the persistence length that average over many bases.

A standard reference frame has been developed[31, 101] for specifying the positions and orientations of single DNA base pairs. Within this frame, the x axis points

along the major groove, the y axis runs along the sequence strand, and z is defined by $x \times y$. Translations along x, y and z are called shift, slide and rise respectively; rotations about those axes are tilt, roll and twist respectively. In accordance with these definitions, both the mean values and variances of each of these six variables have been determined between each possible pair of base sequences, based on X-ray crystallographic structures of DNA-protein complexes[102]. The mean values determine the unstressed conformation of a DNA contour, while the variances determine the stiffnesses in various degrees of freedom. An alternative convention is to define the base steps relative to a helical axis rather than a body-fixed axis. A full description of a base pair involves the relative position and orientation of one of the two bases with respect to the other: these translations and rotations are called shear, stretch, stagger, buckle, propeller, and opening.

Over the span of many subunits, any relaxed polymer of identical subunits takes the shape of a helix (or, in limiting cases, a line or a circle). Coarse-grained descriptions of DNA describe the mechanics of DNA in terms of the rate of bending of, and twisting about, the helical axis, rather than the translations and rotations of individual subunits. The mean bend angle of a stretch of DNA about any perpendicular axis is zero because the helicity makes the bending isotropic (i.e. it is as likely to bend left as right, forwards or backwards). However, the mean twist is nonzero: relaxed DNA twists in the right-handed sense about its helical axis on average once per every 10.5 base pairs. Over- and under-twisting are referred to as positive and negative supercoiling respectively.

In a thermal environment the bending and twisting profile of a polymer will differ from that of the unstressed conformation, depending on the polymer stiffness and temperature of the thermal bath. The expected deviations in bending and twisting are parametrized by the bending and twisting persistence lengths. The mean dot product of the helical axis vectors between two loci decays exponentially with the interlocus spacing, and the length scale for this decay is the (bending) persistence length. Likewise the twist persistence length determines the mean decay of correlations in the rotation angle about the tangent, after having adjusted for the mean

twist. By convention the term 'persistence length' refers to the bending persistence length.

The bending persistence length of double-stranded DNA has been experimentally measured using many methods, with the consensus being that it is at about 50 nm *in vitro* (reviewed in [52]). An early technique for measuring the bending persistence length relied on the absorbance of a DNA solution where the DNA molecules had been aligned by an electric field, then allowed to relax diffusively for a short time. The degree of relaxation was then measured by measuring the opacity of the solution to light polarized along the electric field. The relaxation rate in turn depends on the stiffness of the molecule[147], yielding a persistence length[111]. Translational diffusion can also be used to measure the persistence length. The persistence length of single-stranded DNA at electrophoresis conditions was measured to be about 4 nm[146] using fluorescence recovery after photobleaching (FRAP), in which a solution of fluorescently-labeled DNA is bleached by overexposure within a small region of the solution, and the subsequent rate of recovery of fluorescence in that region is measured to determine the diffusion rate of the DNA and indirectly the persistence length.

A technique for measuring DNA persistence length that is very sensitive to highly-bent conformations is to measure the cyclization rate of short ($\sim$ 100 base pair) oligonucleotides[131]. In this assay a dilute mixture of DNA is ligated, and the ratio of intermolecular ligations to cyclizations gives the effective concentration of one end relative to the other, which is called the J factor. From the J factor one can determine the persistence length[161, 129]. The single-stranded DNA persistence length has been measured using the proximity of the two ends, rather than cyclization which requires strict matching of the two ends' orientations[92]. The single-stranded measurement used Förster resonance energy transfer (FRET), a fluorescence method whereby an excited donor fluorophore on one end of the DNA transfers some of its energy to an acceptor fluorophore on the other end, causing the latter to emit a photon. The persistence length of single-stranded DNA was measured by FRET to be only 1.25 - 3 nm, depending on NaCl concentration, which indicates that the mutual support of the complementary strands contributes considerably to the stiffness of double-stranded

DNA.

The stiffness of DNA has been probed more directly by pulling on it with optical tweezers[6, 154], in a setup in which one end of a DNA oligo is fixed to a surface while the other is attached to a transparent bead in an optical trap. The optical trap is a converging beam of light refracts through, and thereby imparts momentum to, the bead thus creating a force on the bead towards the focus of the beam. The displacement of the bead from the center of the beam indicates a net external force which can be determined from the magnitude of the displacement. From the force required to pull the two ends of a given length of DNA to a given separation distance, one can obtain the persistence length[100, 85]. There are non-optical ways of stretching DNA as well. One experiment fixed one end of a DNA oligo, pulled the other end with an atomic force microscope, and inferred the pulling force from the deflection of the tip[117]. Finally, bending-force measurements were made by pulling DNA oligos between a bead caught in a pipette and the tip of a perpendicular optical fiber; the displacement of the fiber was imprinted on an optical beam passing through it and recorded by a photodetector[25].

Electron microscopy has also been used to measure the DNA persistence length by directly imaging DNA contours in two[40] and three[8] dimensions. By measuring the tangents along the contour at various locations on a DNA sample, and tracing out the contour distances between those points, one can measure the decay constant in tangent correlations that defines the bending persistence length. A correction has to be made for dimensionality: DNA prepared on electron microscope slides that has equilibrated in two dimensions will have a larger persistence length by a factor of 2 than a three-dimensional contour, although the extent to which the samples are truly two-dimensional on a microscope slide can be unclear.

The twist persistence length l_t has been estimated in a number of ways. One type of measurement exploits the fact that fluorophores have nonisotropic polarization cross sections for excitation and emission; exciting fluorescently labeled DNA with polarized light and measuring the rate at which the polarization of emission decays to isotropic can be used to find the twist persistence length[65] (this method can

also measure bending stiffness). Another estimate[90] of the twist persistence length comes from fixing one end of a stretch of DNA, attaching a magnetic bead to the other end, rotating the bead by an external magnet and measuring the fluctuations of the bead at different torsional angles[143]: this measurement yielded $l_t = 120$ nm. Other measurement methods compare the cyclization efficiency of DNA oligos differing slightly in length[130], or the efficiency of cyclizing at various linking numbers[29]. The various estimates of l_t are generally close to 100 nm.

1.1.2 Modeling DNA Conformation

The most accurate computational models of molecules are those that compute the trajectories of each bonded atom in the molecule from first principles, a technique which is called molecular dynamics (MD)[12]. Using MD to model DNA requires simulation not just of the DNA oligonucleotide but also of the surrounding solution consisting of water molecules and ions, since the solvent interacts strongly with the negatively-charged DNA backbone. In order to initialize a MD simulation the system is first brought to a minimum-energy state close to a known structural conformation of the system, then evolved in time until the initial state is forgotten. MD can resolve dynamical processes such as the structural transition from A-form DNA to B-form[19], equilibrium conformational properties such as mean bending between base pairs[165], and mechanical properties that depend on thermal fluctuations such as the bending and twisting stiffnesses[77, 99, 108]. Simulations by MD usually require a lot of computation time.

Coarse-grained models of DNA omit the details of the individual atoms, and instead resolve the contour of either the helix or the helical axis. These models fall into two classes: discrete and continuous[160]. In a discrete model the contour is represented as a series of rigid line segments connected by joints (the segments may or may not correspond to individual base pairs). In a continuous model the DNA contour is represented as a mathematically smooth curve in space. Within these categories the various models differ in how rotations and twists occur between adjacent segments along the contour. Some models treat only bending, by parametrizing each segment

with a tangent vector pointing along the contour; other models use both a tangent vector and a normal vector which establishes a twist sense. Different models treat polymers as inextensible by fixing each segment length, or extensible in which the segment length varies according to some distribution. A final distinction is between models that consider non-adjacent interactions between segments, such as excluded volume, and those that ignore them which makes most analyses much more tractable. Excluded volume is often ignorable if either: the interactions between segments are either very rare, or (paradoxically) if they are so common that very long range, isotropic interactions dominate[33]. Models that ignore excluded volume are called phantom-chain models.

Common discrete polymer models are the freely-jointed chain model, the freely-rotating chain model, and the rotational isomeric state model[160], in order of increasing complexity. In the freely-jointed chain model each segment is free to rotate arbitrarily with no energy penalty with respect to its neighbors. In the freely-rotating chain model the bend angle θ of each joint is fixed but the direction of bending ϕ may freely take any value from 0 to 2π. The rotational isomeric state model fixes not only θ but also ϕ to one or several values. Both of these models deal with inextensible polymers. Unlike the first two models the rotational isomeric state model cannot ignore twisting, since the twist angle defines the allowed bending directions between each segment.

A popular continuous polymer model is the wormlike chain model (WLC)[75], in which the bending energy is given by the function $E = (l_p/2)\int(du/dl)^2 dl$. The WLC can be understood as the limit of a discrete polymer divided into infinitesimal segments connected by infinitely stiff springs, where the limits are taken so that the overall bending scale is the persistence length l_p. The most common wormlike chain models are isotropic in bending and may or may not deal with twist. An extended model called the helical wormlike chain[160] includes twist and allows for a mean bending angle relative to the twist vector, so that the minimum energy configuration is a helix rather than a straight line.

Linear models such as the wormlike chain work well by definition for small de-

viations of the DNA contour from the minimum-energy configuration, but are not necessarily correct for sharply bent or twisted contours. Indeed, sharp bends or kinks may form in DNA at a much higher rate than a linear model would predict[24], although this is controversial[35]. That kinks would form more easily than linear theory predicts is perhaps not surprising given that DNA is often very sharply bent *in vivo*. One way to incorporate kinking into polymer models is to assume a smooth wormlike chain contour with a quadratic bending energy that is punctuated by a number of kinks, where each kink has an energetic cost that is independent of, or weakly dependent on, bending angle[162, 110, 115, 157, 156].

Over many persistence lengths a polymer effectively performs a random walk in steps of the Kuhn length $a_K = 2l_p$, resulting in a near-Gaussian probability distribution between the two ends. This fact inspires the Gaussian chain model, which imagines replacing a length-L polymer having persistence length l_p with a polymer having length L/a and persistence length $l_p a$ while taking the limit $a \to 0$, such that the end-to-end distribution becomes exactly Gaussian whose decay constant converges to a finite value. The Gaussian chain model predicts only the relative displacements of the two ends, not the intermediate contour. The relative twist between the two ends is completely uniform (because the length is effectively infinite), and the probability distribution for the end-to-end displacement $\mathbf{R}$ is determined only be the separation distance R by the formula $p(\mathbf{R}) = (\alpha/\pi)^{3/2} \exp(-\alpha R^2)$, where $\alpha = 3/lL$ for the freely jointed chain (l is the segment length) and $\alpha = 3/a_K L$ for a wormlike chain.

One structural phenomenon of DNA *in vivo* for which excluded volume cannot be neglected is supercoiling. A supercoiled polymer is subjected to torsion, which is relieved by a mixture of twist and writhe (coiled bending). Analytically, it is possible to impose some small amount of twist and writhe within the framework of phantom chain models[15] as long the configuration is not destabilized much from the unstressed conformation. However, strong supercoiling causes forward-and-back excursions of the contour called plectonemes, in which the forward and backwards halves coil around and support one another. A proper model of a plectoneme thus requires modeling of the support interactions between nonadjacent parts of the polymer; generally this is

done numerically (for example see [17]).

1.1.3 Solution methods for conformational statistics

Starting from a given model of the local mechanical properties of a polymer, a variety of analytical and computational techniques can generate statistics relating to large-scale conformation. These statistics relate to biologically-important quantities such as the rates of DNA looping and cyclization (end-to-end joining of an oligo with matching tangents and twists), and the free energies of supercoiling, compacting and packaging DNA. Each calculational technique is generally tailored to a certain DNA model or class of models, either discrete or continuous, and works best in a certain regime of DNA length (relative to the persistence length) and the degree of bending and twisting (relative to thermal fluctuations).

The conformational statistic we will be most concerned with is the end-to-end probability distribution of a DNA segment. This is the probability density for finding the two ends of a length-L polymer segment in a given relative position and/or orientation. For example, suppose we would like to know the free-energy cost paid by a DNA-bridging protein when it binds two locations on the DNA and holds them together. If, for simplicity, we ignore crowding effects between different parts of the polymer (as we will do consistently here), then the free-energy penalty is entirely due to the fact that the protein restricts the allowed configurations of the *intervening* segment between the two binding sites; the remainder of the polymer does not affect the free energy and can thus be ignored. Specifically, we can find the free energy penalty by integrating the length-L end-to-end distribution over the positions and orientations that satisfy the boundary conditions of the protein's binding sites. The particular distribution of interest depends on the boundary conditions. For a rigid bridging protein, the two bound spots are essentially fixed relative to one another, and we need to work with a probability density that accounts for both relative position and orientation. If a sufficiently flexible linker in the protein connects its two binding sites, then the relative orientation may be considered free, so it is simpler to integrate, up to the maximum spacing of the linker length, a reduced distribution

that is only a function of the separation distance of the two ends.

A very general method for obtaining statistics of discrete conformations is Monte Carlo sampling. A Monte Carlo algorithm draws representative values for the free parameters of a contour (for example the bending and twisting angle between each pair of consecutive segments) based on some prior distribution that comes from the DNA model being used, then generates conformations using those parameters and uses those conformations to obtain the statistics of interest. Monte Carlo can accommodate very general polymer models, which may or may not involve extension of the contour, nonharmonic energy functions, sequence dependence along the contour, etc.. Since the common form of Monte Carlo samples conformations in proportion to their occurrence in a random thermal environment, and because computational sampling is usually much slower than the thermal sampling performed by a physical system, it can be difficult to sample very rare conformations, implying that Monte Carlo generally works best in the low-energy regime.

A variant of Monte Carlo, called the Metropolis method[89], constructs conformations by a series of iterative perturbations on a starting conformation. The conformation at each iterative step n is subjected to some permutation in an attempt to generate the next conformation at step $n+1$. The new conformation is accepted if: the new conformation satisfies any user-imposed constraints; and if the ratio of statistical weights p_{n+1}/p_n is either greater than one or greater than a randomly-generated number on the uniform interval $[0,1]$, a rule which ensures the proper weighting of samples. If any of these tests fails then the $n+1$th conformation must be resampled using a different permutation of the conformation at step n. Only a well-separated subset of conformations form the sample set, in order to ensure that the samples are relatively uncorrelated. Metropolis sampling is good for enforcing constraints that would rarely be satisfied by the basic Monte Carlo method.

Several techniques have been developed to deal with conformations of discrete polymers that are rare because they are sharply bent or twisted relative to the size of thermal fluctuations. These methods take advantage of the fact that such conformations tend to be sharply clustered around the minimum-energy conformation

that satisfies a set of constraints. The general method is to find this minimum-energy conformation and calculate the end statistics by perturbation. The original method by Zhang and Crothers[166] took a quadratic expansion in the energy about the minimum energy configuration, and integrated them exactly along with the Fourier-transformed constraint functions; this method can accommodate sequence-dependent models since every joint is accounted for separately. A later technique by Wilson and coworkers[158] calculates the normal modes of the polymer about the minimum-energy conformation, subject to endpoint constraints, and integrates over the amplitudes of these modes. The normal-mode analysis is similar to an earlier approximate analytical treatment of cyclized DNA[129]. The methods of Zhang and Crothers disagreed somewhat with those of Wilson et. al., and although the latter claim to be in better agreement with earlier work[129] it is still unclear which is the more accurate. Both techniques fail outside of the high-bending regime.

A transfer matrix technique exists[163, 164, 162] that can calculate end statistics in both high- and low-energy regimes, using a formalism that discretizes not only the chain contour but also the orientation of each segment. Each element of a transfer matrix $T_i^{(n;k)j}$ contains the statistical weight for evolving segment n at orientation i to segment $n+1$ at orientation j; each matrix also absorbs the Fourier-transformed positional end constraint as denoted by the wavenumber k. Any bending and twisting energy functions may be used, and by interposing different transfer matrices sequence-dependent models can be accommodated.

The end statistics of continuous model polymers can be calculated using formal methods that were originally developed for quantum mechanics. The partition function that integrates over the bending and twisting angles along the DNA length is formally equivalent (up to an imaginary factor) with a path integral over the rotations of a quantum spinor in time. Converting the path integral of the orientation-only partition function into a Schrödinger equation allows one to borrow the solution from quantum mechanics, which is an eigenbasis of Wigner functions[160]. The perturbative method of Spakowitz[139, 140, 138, 88] extends the method to include a relative position constraint, by adding off-diagonal terms to the orientation-only Hamiltonian,

which a perturbative analysis reduces to a solution in terms of continued fractions weighting the eigenstates. This work has been extended beyond the wormlike chain by Wiggins et. al., who developed the continuous kinkable wormlike chain model[157] in which each kink pays a fixed energy penalty regardless of the discontinuous bend angle it imposes, and the subelastic chain model[156] which incorporates an energy that is first-order in the bend angle.

1.2 DNA conformation *in vivo*

In all organisms the physical length of genetic material ($\sim 0.3 - 300$ mm) greatly exceeds the dimensions of the cell ($\sim 1 - 100\mu$m). Compacting and arranging the genetic material is a necessary chore that every cell must perform. Furthermore, the genetic material must be faithfully untangled and segregated with each cell division cycle. Maintaining such a lengthy genome is a necessary hassle, but many cells (mostly eukaryotes) exploit the rich palette of possible conformational states to aid various cellular processes.

Cells compact their DNA using a variety of DNA-condensing proteins[83]. In bacteria, H-NS and Lrp are proteins containing both DNA-binding domains and dimerization (H-NS) or multimerization (Lrp) domains; these associate *in vivo* to form protein complexes with multiple DNA-binding sites that may help compact the DNA by bringing distal regions together. Lrp forms either octomeric or hexadecameric complexes depending on the leucine concentration, indicating that this amino acid sensor may transduce a chemical signal into a physical rearrangement of the chromosome. Various structural maintenance of chromosome (SMC) proteins such as MukB in *E. coli* are believed to form dimers that loop around DNA, tying multiple DNA strands together either within a single dimer or through the association of dimers into multimeric 'rosettes'. The proteins IHF, HA and Fis introduce bends into the DNA, leading to an effective shortening of the persistence length and compaction of the chromosome. Fis may also help compact the genome by multimerizing.

Nucleoid-associated proteins cause higher-order structures in bacterial chromo-

somes at a variety of scales. Electron microscopy of extracted bacterial chromosomes reveals them as supercoiled loops emanating from a dense central core[72, 112] likely composed of RNA and protein, although the exact picture depends strongly on the preparation. The most recent estimate[112] gives the loop lengths roughly an exponential distribution with a mean of 10 kb. Experiments measuring the level of supercoiling in different parts of the genome suggest that the chromosome is divided into topological domains, where the level of supercoiling equilibrates throughout each domain but cannot propagate beyond the two flanking domain walls. The size of the topological domains in *E. coli* was originally estimated at 100 kb [72], but newer studies have argued for smaller domains of 10 kb [112], consistent with the sizes of the DNA loops. On larger scales, recombination experiments[149] and FISH[96] suggest that the *E. coli* genome may organize into 2-4 larger 'macrodomain' structures.

On the global scale, many prokaryotic genomes seem to be arranged in an orderly way down the length of the cell. For example, the *E. coli* genome is packaged linearly along the long axis of the cell, with the origin located at mid-cell and the two arms spreading towards opposite poles[155]. Because the *E. coli* genome seems to be a closed loop, as are most bacterial chromosomes (though there is some controversy about this—see [9]), it seems that the terminus region of the genome must stretch tightly from one end of the cell to the other in order to connect the two arms, although this has not yet been seen directly. The arrangement of the *Caulobacter* genome is also linear at the global scale[152], except that the terminus is at one cell pole, the origin of replication is at the other, and the two arms between origin and terminus are both stretched in parallel along the full length of the cell. Less is known about radial positioning; one very recent study found that loci encoding membrane proteins in the *E. coli* genome are apparently pulled to the cell periphery when those genes are expressed[81], affecting the positioning of loci up to 100 kb away.

In both *E. coli* and *Caulobacter*, newly-replicated genomic loci are segregated rapidly to their appropriate locations within the two daughter cells[152]. There are two replisomes (replication machines) per parent chromosome in *E. coli*. Both replisomes begin their work at the middle of the parent cell where the origin of replication

is located, but soon thereafter follow the two replication forks out towards the opposite poles, then come back again to the center of the parent cell as the replication forks meet again at the terminus[116].

The linear arrangement of DNA in *E. coli*[155] and *Caulobacter*[152] suggests that in these organisms, and perhaps in the majority of prokaryotes, genetic material is simply arranged for convenient packaging and not organized for higher-order control over processes such as transcription. However, nature does exploit the fact that gene copy number and therefore transcription levels correlate with *genomic* position in dividing cells[137], since origin-proximal regions are replicated first and are therefore present in higher numbers during DNA replication. Genomic position in turn correlates with physical position along the long axis of the cell in many species[155, 152].

Genomic compaction in eukaryotes relies largely on histone proteins which form octameric spool-like complexes called nucleosomes[74]. Each nucleosome tightly wraps 146 base pairs of supercoiled DNA around its outside, corresponding to about $1\frac{3}{4}$ complete turns of the DNA around the nucleosome particle. Nucleosomes are separated by 'linker' stretches of DNA of random lengths averaging around 50 base pairs; due to their tight packing the mass of nucleosomes approximately equals the mass of DNA. Protruding tails from the individual histone proteins interact with the tails of other nucleosome particles, causing them to aggregate into higher-order structures and leading to compaction of the DNA.

Nucleosome-bound DNA, called chromatin, has different mechanical properties from bare DNA. It seems that short lengths of chromatin are more flexible than bare DNA[118]. At longer lengths it is believed that nucleosomes bind one another and arrange the DNA into a thicker fiber whose structure is still controversial. The traditional view is that nucleosomes package DNA into a regular '30-nm fiber'[26], although newer measurements suggest that the fiber may under some conditions become about 50% thicker[120], and others indicate that such a fiber is nonexistent[37] which would imply a disordered arrangement of eukaryotic DNA.

Nucleosome-bound DNA exists in one of two forms: transcriptionally-active euchromatin, and repressed heterochromatin. While both forms of chromatin may fold

into a 30-nm fiber, heterochromatin is further compacted on larger scales[45] causing steric exclusion of DNA-binding factors required for transcription and thus preventing expression of genes within that region. Chemical modifications to the nucleosomes nucleate heterochromatic regions, which then spread along open DNA through the action of proteins such those encoded by the silent information regulator (Sir) genes in yeast. The boundaries between euchromatic and heterochromatic regions are demarcated by 'boundary elements' that in some or all cases are associated with CCTF and cohesin[51]. Cohesin is a member of the SMC protein family, and like MukB is believed to dimerize to form a ring that can ensnare several strands of DNA at a time, thereby bringing them together[64]. It seems likely that CCTF recruits cohesin to boundary elements, and that cohesin in turn somehow prevents the spread of heterochromatin.

Cohesin also plays a role in activating the transcription of eukaryotic genes by altering the conformation of the chromosome. Transcription initiation requires the assembly of a number of protein factors, some of which bind at the gene promoter and some of which bind to distal enhancer elements which can be up to hundreds of kilobases away. A protein complex containing both cohesin and the mediator complex forms contacts between enhancer elements and gene promoters[97, 71] in order to activate transcription. Other elements termed insulators block the action of enhancers. Insulators are associated with CTCF and cohesin, and it is thought that the cohesin complex tethers DNA in a way that blocks the enhancers from looping over to their target promoters[1]. Cohesin thus upregulates or downregulates gene expression, depending on which protein partners it associates with.

Transcribed genomic loci frequently associate in the cell in DNA-protein complexes called 'transcription factories'[68, 103]. Actively dividing HeLa cells are estimated to have several thousand transcription factories at any given time. It is not clear whether the colocalization of genes has a purpose or is just a byproduct of active transcription, but a recent experiment[79] discovered numerous interacting promoters in human cells whose gene expression levels were both correlated and affected by the presence of the nearby promoters. The authors suggested that the physical coupling of genes during

transcription is involved in the combinatorial regulation of gene expression.

In some cases the positioning of genomic loci relative to the nuclear envelope coordinates gene expression and DNA maintenance. Repressed genes are often associated with the nuclear envelope in yeast[1] or the nuclear lamina in mammalian cells[46]. In yeast, association with the nuclear pores can, paradoxically, increase the expression levels of certain genes[1], and damaged DNA has been shown to be recruited to nuclear pores for repair[93].

Recombination, both intentional and accidental, is influenced by the mutual accessibility of the recombining genomic regions. Examples of intentional recombination in mammals are the generation of unique immunoglobulins (which include antibodies) in mammalian B cells, and of T cell receptors in T cells, in which recombination involves the looping of the recombinant regions[70]. Unintentional, mutagenic recombination usually occurs between genomic regions that are nearby in physical space[87], a fact which likely implicates DNA conformation in a cell's predisposition to certain types of cancer.

Eukaryotic chromosomes are organized into megabase-scale domains bounded by CTCF-enriched boundary elements[32, 98]. A given locus associates much more often with chromatin within its own domain than loci residing outside the domain, and gene expression level is strongly correlated within each domain. On the largest scales, individual interphase chromosomes of higher eukaryotes tend to arrange in distinct, non-overlapping chromosomal territories[87], although the territory occupied by each chromosome depends partly on cell type and state and varies considerably between cells. During mitosis and meiosis the chromosomes condense with the aid of condensins, which are members of the SMC protein and operate similarly to cohesin. Kinetochore motors walking along microtubule spindles partition daughter chromosomes between dividing daughter cells. The large-scale organization of interphase chromosomes after mitosis is somewhat consistent between cells, but the mechanisms behind, and significance of, this organization are still unknown.

1.3 Experimental techniques for measuring DNA conformation

A number of experiments have probed conformation in an indirect way, without resolving explicit genomic interactions or positions. Information about the compactness of DNA comes from studying the permeability of different cellular regions; high permeability presumably corresponds to low DNA packing density and vice versa. The permeability can be mapped out directly on the micrometer scale by measuring the mobility of diffusing fluorescent molecules using pair-correlation analysis[63]. Another method of probing cellular permeability is to measure the frequency with which transposons jump or copy to various genomic loci; this maps density as a function of genomic position rather than density as a function of spatial position. A number of transposon systems have been used for this purpose: recombination between the $\gamma\delta$ transposon[62], self-inhibition of transposition by the Tn7 transposon[27], and recombination between the phage λ *attL/R* sites[43, 149], and there have been a number of incidental reports on transposition frequencies from studies of IS1-flanked transposons, Tn10, and bacteriophage Mu (referenced in the Discussion of [27]).

1.3.1 Chromosome conformation capture

One of the biologically important effects of DNA conformation is that physical contact tend to happen between pairs of nearby genomic loci that are close in space. The standard technique for measuring the frequencies of these interactions between genomic regions is called Chromosome Conformation Capture (CCC or 3C)[28]. The first step in the 3C protocol is to fix cells with formaldehyde, which binds DNA to protein and hence can indirectly couple proximal DNA segments. Then the DNA is extracted from the cell and digested with a restriction enzyme, yielding many single oligonucleotides of DNA along with pairs of oligos bound together by protein. This DNA-protein mix is then diluted and self-ligated using a DNA ligase. Upon ligation most single oligonucleotides will simply circularize; however pairs A and B of oligonu-

cleotides that are bound together by protein will sometimes join end-to-end to form a single circular strand of DNA with a hybrid sequence $A-B$. In the final steps of the 3C protocol, the crosslinking is reversed, and qPCR is performed with one primer for each pair of loci that one is interested in. The interaction frequency between any two loci should be proportional to the strength of the PCR signal for that pair of primers. Each qPCR reaction is compared to a control reaction containing all $A-B$ hybrid sequences in equal amounts, produced by random ligation of digested chromosomal DNA, in order to normalize the PCR signals. Both reactions can be conducted over a range of template concentrations to determine the linear range over which product is proportional to template; over this linear range the relative fold enrichment over the control can be accurately measured.

A disadvantage of the original 3C technique was the need to run a separate qPCR reaction for each *pair* of loci whose interactions were to be measured. Thus to completely map out the interactions between N loci required $\mathcal{O}(N^2)$ separate PCR reactions. A sequence of improvements to the protocol has reduced the number of separate PCR reactions that need to be performed, allowing large contact maps to be produced with much higher throughput.

The original improved protocols, made independently by two research teams which were each called 4C (Circular CCC[167] or CCC-on-Chip[136]), allow the interaction frequencies of all genomic regions with a given target locus A to be measured in a single experimental step. The 4C protocols crosslink and digest the chromosome as in 3C, and then ligate separate crosslinked DNA strands into a circular DNA loop. The methods differ from each other in the way the circular loop is produced: in one case the digested chromosome is ligated extensively to directly circularize the long fragments[167], and in the other the fragmented DNA is further digested using a 4-cutter (DpnI) to speed the ligation[136]. The final steps in both methods are to isolate any fragment B that has bound to target locus A (and thus incorporated into the circular DNA loop) by quantitative inverse PCR where both outward-facing primers reside within locus A, and then to identify those PCR fragments using a microarray.

The next improvement, called 5C (3C-Carbon Copy)[34], allowed all pairwise

interactions between a set of defined loci to be measured in a single experimental step. The 5C protocol follows the 3C protocol up through the construction of the 3C library (the mix of religated oligos whose crosslinks have been removed). At this point the 5C library is constructed from the 3C library using multiplex ligation-mediated amplification (MLMA)[124]. The first step in MLMA is to anneal a mixture of oligonucleotides to the 3C library, where one end of each oligo is complementary to one of the target sequences, and the other end is an adaptor for PCR. Any pair of crosslinked loci in the 3C library will allow two oligos to bind, directly abutting one other without a single base pair gap. Ligation is then performed to tie abutting oligos into a single strand of DNA containing both target sequences, and these ligated oligo pairs are amplified by PCR using the flanking adaptor sequences. The target sequences in the amplified 5C library are assayed by microarray or sequencing.

Whereas in 3C each pair of potentially interacting loci must be separately queried, 4C finds all interacting partners B of the single query locus A, and 5C obtains all interactions within a set of query loci. In contrast, Hi-C[82] enables query-free interrogation of the entire genome. In High-C, the ends of the DNA are labeled with biotin prior to (circular) ligation, generating a library identical to the 3C library except that each circularized DNA fragment has biotin tags at the ligated junctions. The arms of each circularized fragment containing the interacting sequences are generally too long to sequence with high-throughput methods, so their length is reduced by shearing the 5C library and purifying the junction regions using streptavidin. Adaptors are ligated to the ends of these sheared and purified fragments, and those fragments are PCR-amplified and sequenced with a high-throughput machine (Illumina). This allows a comprehensive interaction map of an entire genome to be made in a single experimental step.

A method called Chromatin Interaction Analysis by Paired-End Tag sequencing method (Chia-PET)[41, 79] is a cousin to the 3C techniques that measures the frequency of interaction between genetic loci and a target protein of interest. The Chia-PET protocol initially follows that of 3C: the cells are crosslinked, and their DNA is fragmented. The DNA fragments bound to the protein of interest are en-

riched by immunoprecipitation, in which an antibody specific to that protein is used to isolate the protein from the cell lysate. At this point a dilute mixture of the protein-DNA complexes are ligated to adaptors which are then ligated to each other. The adaptors have inverted restriction sites for an enzyme (MmeI) that cuts 18 base pairs away from the restriction sequence; digestion with MmeI therefore releases the restriction site along with enough of either flanking sequence that they can generally be uniquely mapped to the genome using high-throughput sequencing. Each piece of DNA is either a single sequence interrupted by a restriction site, indicating a simple DNA-protein binding event, or else two genomic loci separated by the MmeI site, indicating that the protein must have bound, or at least been very close to, two or more genomic regions simultaneously. The sequences thus reveal all single and pairwise DNA interactions at the locus of the target protein or protein assembly, in a single experimental step.

1.3.2 Electron microscopy

Electron microscopy has been used to study chromosomes *in vivo* and *ex vivo* at high resolution[72, 112]. The *ex vivo* preparations involve lysing the cell, transferring the chromosome onto a flat grid and imaging it in two dimensions using a transmission electron microscope (TEM). Using this approach researchers have traced short stretches of the *E. coli* chromosome, enabling estimates of the sizes of the loops coming out of the central core. There are a number of problems with the use of this method to map out full conformations. First, although it is easy to trace regions of the chromosomal contour that are well separated from other DNA in TEM images, a large part of the nucleoid exists in a dense mass that makes tracing the contour very difficult. A second problem is that labeling techniques for EM are primitive, so it is extremely difficult to identify individual genetic loci. Third, the *ex vivo* preparation collapses the native 3-dimensional conformation onto a 2-D surface, which hugely perturbs the structure and erases out-of-plane conformational information. Finally, positioning information relative to other cellular structures is lost.

TEM has also been performed *in vivo*[38], by vitrification (ice-free cooling) of a

live sample, followed by sectioning and imaging. This method can resolve general questions about the ordering of DNA in the nucleoid. Unfortunately, due to the thickness of the sections most images show overlapping DNA contours that in general cannot be separated, limiting the usefulness for following individual DNA contours. Furthermore, in order to follow a contour over long distances one would have to capture most or all sections and computationally align their images, while accounting for distortions of each section. As with the *ex vivo* preparations, genomic regions are difficult to label, and the monochromatic images preclude multichannel labeling.

1.3.3 Fluorescence microscopy

Optical microscopy is routinely used to image chromosomes with the aid of DNA-binding fluorescent labels or dyes. Fluorescent dyes absorb light of one wavelength to become excited, then radiate part of that excitation energy as photons of a longer wavelength. The wavelength difference is critical for microscopy as it allows the emission to be cleanly distinguished from scattered light by means of highly-selective color filters. One popular fluorescent molecule is the UV-to-blue fluorophore called DAPI (4',6-diamidino-2-phenylindole) which stains all of the genetic material of a cell. DAPI passes freely through cellular membranes and nonspecifically binds DNA. Other nonspecific DNA stains are ethidium bromide (usually used *in vitro*) and the various Hoescht stains. The DNA inside cells that have been fluorescently labeled using these nonspecific dyes generally shows up in an image as a single mass taking up the volume of the chromosome, due to the high cellular density of genetic material. These images are useful for measuring the bulk shape of genetic material in the cell, but cannot be used to resolve the DNA contour.

DNA labeling methods

More sophisticated fluorescent labeling schemes target only certain genetic loci as opposed to the entire chromosome. All these methods involve a fluorescent molecule or molecular domain which is fused to a molecule or domain that binds a certain genomic

region, or set of genomic regions. There are two main techniques for accomplishing targeted fluorescent labeling of genetic loci *in vivo*: fluorescent repressor-operator systems (FROS) and fluorescent *in situ* hybridization (FISH). These methods differ in the molecule used to target the genetic loci, in the need for fixation and permeabilization (which kills the cell) and in their ability to target endogenous versus exogenous DNA sequences.

Using the FROS method, a protein fusion is engineered that connects a fluorescent protein such as GFP to a DNA-binding protein that binds some defined sequence in live cells. To create a cell line that can be used for FROS imaging, one first clones an array of repeats of the binding sequence (so that enough fluorescent proteins will bind for easy imaging) into a defined genomic locus in the cell of interest. Then, the fluorescent-fusion protein is introduced into the cell, typically by introducing a plasmid bearing the gene for the fusion protein which can be inducibly expressed at high levels. Upon induction, the cell produces the fluorescent fusion proteins which then bind to the exogenous binding array. To date, FROS has been performed with both the lacI-*lacO* and tetR-*tetO* combinations of binding protein and DNA sequence. The related parB-parS system[80] involves the spontaneous polymerization of a (fluorophore-conjugated) ParB protein from a single 286-bp *parS* locus. These two techniques have been used to image genetic loci in live bacteria[152, 155], yeast[142] and mammals[119], often for visualizing dynamics (chromosome and plasmid segregation) in the live cells. For example, FROS studies have revealed that the mechanism by which *E. coli* positions the F-plasmid is different than its mechanism for positioning its centromere, because the centromere initially gets pushed to the poles of the dividing cell whereas the F-plasmid localizes to their quarter-points (the midpoints of the future daughters)[44]. An advantage of FROS is that it doesn't involve any physical perturbation to the cells once the cell lines have been produced; the main disadvantage is that it requires a genetic perturbation (insertion of an operator sequence). Common protein fluorophores include GFP[18], CFP[56], YFP[94] and mCherry[126] (green, cyan, yellow and red respectively).

The FISH technique uses fluorescent dyes conjugated to oligonucleotides that are

complementary to the genomic sequences of interest. Cells are fixed prior to visualization, their membranes are permeabilized, the chromosomes are denatured (i.e. their strands made to separate), and the probes (oligos) are allowed to enter the cell and anneal to the target DNA region. Competitor DNA is also introduced and subsequently removed, in order to remove most of the nonspecifically-bound probes and reduce background. FISH has been used to systematically map a number of genomic loci in *E. coli*[96] and in *Caulobacter*[152]; in both organisms the chromosome was found to be packed linearly along the axis of the cell. FISH has also been used to locate plasmids within the cell[95]. A major advantage of FISH is that it can target endogenous loci in virtually any type of cell. The disadvantages are that the fixation kills the cell, precluding measurements of dynamic processes, and that the denaturation may perturb the native structure of the DNA. FISH is compatible with many small-molecule organic fluorophores; commonly-used fluorophores are the cyanine and Alexa[105] series of dyes, and fluorescein and rhodamine.

There are ways to convert the nucleic acids themeselves into fluorophores, with possible applications to DNA labeling. One approach[104] uses a single-stranded nucleic acid (RNA) as a scaffold that binds the central amino acid ring in GFP and maintains it in a fluorescent state, by preventing non-fluorescent decay channels. Another technique[47] replaces DNA nucleotide bases with organic fluorophores, which interact via FRET to collectively fluoresce with unique excitation and emission spectra that depend on the arrangement of individual fluorophores.

Superresolution fluorescence microscopy

A major limitation of optical microscopy is fact that, for realistic microscopes, an infinitesimal emitter in the sample will produce an image of finite size, called a point spread function (PSF). The intensity profile of the PSF is nearly Gaussian, centered about the true position of the point source, and for optical wavelengths and standard microscopes the width of the Gaussian is on the order of 100 nm. A single emitter can be localized much more accurately than the width of the PSF by numerically fitting the profile of the PSF (usually approximated by a Gaussian) to the observed PSF and

finding its center; in this way the position of the emitter can theoretically be found to arbitrary accuracy with unlimited photons. However, there is a problem when two or more emission sources have PSFs that overlap heavily, as the two sources become nearly indistinguishable and are very hard to localize. This phenomenon has been termed the diffraction limit of light microscopy, as the size of the PSF has historically been the resolution limit that can be attained in most practical situations where the emitters are closely spaced.

A number of recent techniques have broken the diffraction limit of fluorescence microscopy[58]. These 'superresolution' ('subdiffraction') methods fall into three categories. The first category is of methods that reduce the size of the PSF. The second class is composed of methods that control the position of the emitters very precisely by suppressing fluorescent emission outside of a small subdiffraction target region. Methods falling in the third category build up a dense fluorescence image (containing many emitters within a given PSF volume) from many sparse images with well-separated fluorophores that can be localized precisely, by turning most of the emitters off each imaging cycle. Methods falling into the second and third categories apply only to fluorescence microscopy.

There are several ways to reduce the size of the PSF, which break one or another of the assumptions that mathematically imply a wavelength-sized PSF. One of these assumptions is that the detector is located in the optical far field of the emitter. A conceivable way to reduce the size of the PSF is therefore to position the detectors very close to the sample ($\ll$ 1 wavelength) and scan it. In practice, it is not the detector but rather a pinhole aperture that is placed near the sample and scanned over the sample surface; at every step in the scan the mechanically-controlled location of this element is used to assign a location to the corresponding point in the image[78]. This method, which is a type of near-field microscopy, can in principle reduce the PSF arbitrarily but the proximity requirement is severely limiting: for example, one cannot perform near-field imaging far from the sample surface.

The size of the PSF also depends upon the range of solid angles that emitted photons are collected from. If light could be collected from all 4π steradians then

the PSF would shrink to zero size even when imaging in the far field. Two methods, called 4Pi microscopy[49] and I^5M microscopy[50], collect light using lenses both in front of and behind the sample, thereby doubling the collecting area and increasing resolution by a factor of about 3 to 7. The disadvantages are that these methods requires a more complex experimental setup, and that in many cases deconvolution is required after image acquisition.

The size of the PSF is limiting only because emitters are typically localized by finding their location in an image. If the emitters have been localized prior to imaging then the diffraction limit no longer applies. One way to do this is with a second type of near-field microscopy, similar to the method we discussed earlier except that the excitation light is scanned in the near field over the sample rather than an aperture in front of the detector[10]. As before, the resolution is now determined by the precision with which the excitation light can be delivered, which depends on both the mechanical accuracy with which the excitation fiber is positioned as well as the size of the collecting area as determined by the proximity of the excitation beam to the sample.

An alternative to activating certain emitters within a small region is to activate the emitters within a large (PSF-sized) region and then turn all of them off again except for a small group in a very small (sub-PSF) sub-region. This class of methods is generically termed RESOLFT[57], although there are many specific implementations. The original RESOLFT method is called stimulated emission depletion (STED) microscopy[60, 73]. STED constrains the excitation volume to far below the wavelength limit ($\sim$ 20 nm) by surrounding the excitation laser beam by an inactivating beam in the shape of a doughnut. Sub-wavelength resolution is possible because the point-spread function becomes relatively steeper in the fringes; thus by turning the inactivating laser to high power the edge of the doughnut becomes very sharp and the activating volume can be squeezed very tight. In STED, the depletion beam causes fluorophores in the excited state to fall into a near-ground state, which rapidly decays into the true ground state so that the depletion beam cannot repopulate the excited state. A related method called ground-state-depletion (GSD) microscopy[59]

makes use of a third intermediate metastable state, and has the advantage that the depletion can be done using a low-power continuous laser rather than a high-power pulsed picosecond laser.

A variation on the RESOLFT concept is saturated structured illumination microscopy (SSIM)[48], in which stripes of excitation cover the image and the dark unexcited regions in between are localized to subdiffraction resolution. Superresolution is obtained in the direction perpendicular to the stripes. Acquiring an image involves rotating the stripes in many directions, and at each rotation step translating the stripes so that the dark regions cover the entire field of view. Image processing is required to generate a SSIM image from the fluorescence data.

The final way to beat the diffraction limit is to image only a few fluorophores at a time, so that each fluorophore's PSF is unlikely to overlap that of another. Every fluorophore can then be localized to high (subdiffraction) accuracy by fitting a PSF to the image of that spot and measuring its centroid. After the first image is produced, the fluorophores are turned off, and an independent set of fluorophores is turned on; then another image is taken. By repeating this process of localizing fluorophores, a superresolution image is gradually built up. The crucial technology for this method is fluorescent molecules whose on-off state can be controlled by the experimenter.

Three groups independently developed photoswitching systems and applied them to superresolution imaging, under the names PALM[11], FPALM[61] and STORM[123]. The three implementations are very similar, the main difference being that STORM uses photoswitchable pairs of inorganic fluorescent dyes[4] whereas PALM and FPALM use photoswitchable fluorescent proteins. (A variant of STORM, called direct STORM or dSTORM[55], switches inorganic dyes directly without the need for a proximal activator dye.) The techniques are complementary: inorganic dyes are bright but must be delivered exogenously, whereas photoswitchable proteins can be produced *in vivo* and fused to other proteins but they are bulky and relatively dim. Typical resolution with the PALM/STORM methods has been on the order of 20 nm, but by careful calibration, in particular correcting for the variable efficiency of pixels in the CCD array, a resolution of ~ 1 nm seems within reach[109].

The original PALM, FPALM and STORM methods imaged only one type of fluorophore, but improvements soon enabled 'multicolor' imaging in which each color in the image represents one type of fluorophore. Multicolor imaging is complicated by the fact that fluorophores are typically photoswitched by one wavelength of light, activated in the fluorescent state by another wavelength and emit in a third wavelength; due to the width of the emission and absorption spectra each fluorophore is thus involved with a significant part of the optical spectrum and separating the channels is difficult. Nevertheless PALM imaging was soon demonstrated in 2 colors[132] and STORM imaging in three[5]. The different colors are typically imaged sequentially, by selectively activating, exciting and/or imaging one particular type of fluorophore very exclusively using appropriate lasers and filters. However, it is possible to discriminate between several types of fluorescent proteins in single image[14] by collecting the fluorophore emission in two or more different color channels and comparing the amounts of light collected in each channel. This method can relax the separation requirement for the various emission, activation and excitation spectra, which is a serious impediment to multicolor superresolution imaging.

Since publication of the original methods, there has been considerable effort on methods that enable axial localization (perpendicular to the imaging plane), as well as lateral (in-plane) localization. The STORM group accomplished three-dimensional imaging[67] by placing a cylindrical lens in the beam path which caused the PSF to elongate horizontally when the fluorophore is above the imaging plane, and the vertically when the fluorophore is below the imaging plane. The axial resolution in 3D STORM is about half the resolution in the lateral directions ($\sim 50 - 60$ nm), although it should be possible to improve this by a better-shaped PSF[114]. Similar axial resolution has been obtained by scanning over optical sections, using a two-photon process for fluorophore activation where the photon density is highest in a single imaging plane[150]. An advantage of the sectioning technique is the large depth of field (~ 10 μm). A third method, applied to a PALM microscope, is to insert a 45-degree mirror that reflects a side view of the sample onto the imaging plane along with the face-on image[145]; this results in an axial resolution that is

essentially the same as the lateral resolution ($\sim$ 20 nm). A fourth technique splits the PSF into two spots, where axial position is encoded as the relative rotation of these spots[107]. The axial resolution using this double-helical PSF technique is as good as or better than the lateral resolution ($\sim 10 - 20$ nm). A final approach is to use interferometry to axially localize emitters. Interferometric PALM (iPALM)[135] has yielded an axial resolution about twice that in the lateral directions ($\sim$ 10 nm) with high photon-collection efficiency, although the depth of field is limited.

Other methods exist for stochastically switching sparse subsets of fluorophores for high-precision localization. Ground state depletion microscopy followed by individual molecule return (GSDIM)[39] exploits a metastable dark state present in nearly all fluorescent molecules to temporarily inactivate most fluorphores in the image. The advantage of GSDIM is that it allows nearly any conventional fluorophore to be used[141]. A method called points accumulation for imaging in nanoscale topography (PAINT)[128] images fluorophores that are bound to a target of interest in a bath of freely-diffusing fluorescent molecules. The unbound fluorophores do not produce a sharp PSF in the image and so are not visible, so there is no need for photoactivation and conventional fluorophores can be used.

Several methods perform superresolution imaging by exploiting the blinking, bleaching, or photoswitching of individual fluorophores that may have overlapping PSFs. Super-resolution optical fluctuation imaging (SOFI)[30] computes higher powers of the PSF to reduce the size of the PSF and thereby localize individual fluorescent molecules whose PSFs overlap. Disadvantages of SOFI are that the contrast is enhanced, so the dynamic range in brightness is quite limited, and it can be computationally intensive. Another technique, called bleaching/blinking assisted localization microscopy (BaLM)[16], simply subtracts successive images from a movie, looking for single blinking or bleaching events which can be localized using the subtracted images. Both SOFI and BaLM require no photoactivation and thus no special fluorophores. The resolution of both of these methods is $\sim$ 50 nm. Finally, there exist nonlinear deconvolution techniques[66, 91] that estimate centroids of overlapping PSFs while introducing less noise than linear deconvolution.

Most superresolution methods that rely on photoactivation require special photoswitchable fluorophores[53], most of which are either organic dyes as used in STORM or photoswitchable protein fluorophores as used in PALM. Most of these molecules convert from a dark state to an active state when exposed to light of a particular range of activation energies; fluoresce for a time when exposed to excitation light; then stochastically enter a dark state (bleach). A reversible fluorophore is one for which the dark state can be converted back into the active state again by the activation laser. For some fluorophores the activation laser causes a photoconversion between two fluorescent states that are distinguished by their excitation and/or emission spectra, rather than a transition from a strictly dark state to an active state.

The original STORM method[123] involved activator/emitter pairs of organic cyanine dyes held together either by linker DNA or by an antibody. The fluorophores interact via a FRET-like mechanism[4] to enable activation of the emitter dye by light of a wavelength outside of its normal excitation band. Common organic fluorophores are Cy3, Cy5, Cy7, and Alexa647 and Alexa680. The dSTORM technique subsequently showed that some single cyanine dyes, namely Cy5 and Alexa647, can photoswitch without the need for a proximal activator dye[54, 55]. Advantages of these organic dyes are that they are small and can therefore position very close to their targets, and that they emit many photons before bleaching which allows high-precision localization.

A number of monomeric protein fluorophores[84] have been engineered to be photoswitchable: PA-GFP[106], PA-CFP[23] and PA-mRFP1[151] having green (G), cyan (C) or red (R) fluorescence; PAmCherry (red)[144]; and Dronpa-3[2], mEos2[86], and bsDronpa and Padron[3] which all fluoresce green. A number of other proteins, such as Kaede[76] and KFP1 (kindling fluorescent protein)[22], form tetramers. Dronpa and KFP1 are both reversible fluorophores. The unmodified EYFP fluorophore frequently used for standard-resolution yellow-fluorescence imaging has been used as an accidentally-photoswitchable fluorophore for superresolution[13].

Chapter 2

End Statistics Calculator

Here we present our modeling tool, named the Wormulator from a contraction of "wormlike-chain calculator" (although the discrete-chain calculations can work with any DNA model that ignores excluded volume, not just the wormlike chain model). This program computes end-to-end distributions of DNA, under the phantom-chain approximation whereby excluded-volume interactions are ignored. These statistics govern many important biological processes, such as DNA-looping free energies for transcription factors, and the constraints they impose on genomic sequences([125]), etc. Our goal in writing the Wormulator was to provide a comprehensive and user-friendly tool for computing end-to-end statistics of DNA, addressing the current lack of a such a tool written for the scientific public. Although the Wormulator was designed with DNA in mind, the program can equally be applied to any polymer whatsoever, as long as the phantom chain approximation can be used.

Our calculator can use three complementary methods. The eigenfunction-based method, due to Spakowitz and Wang[140][138], is best-suited for cases in which the end-to-end separation is much greater than the bending scale of the DNA. The numerical Monte Carlo method efficiently computes statistics of shorter DNA contours. Finally, the method of Zhang and Crothers[166] handles polymers that are sharply deformed due to positional and/or orientational constraints. The program can be either downloaded and run on a personal computer, or accessed via the web. The online calculator has a straightforward and intuitive interface but is somewhat restricted in

what it can do. The offline program uses a command-prompt interface, but has the full range of capabilities and can be used for intensive calculations.

2.1 End-to-end distribution

The mathematical form of the full end-to-end distribution can be formally written as[160]:

$$p(\mathbf{R}_2, \mathbf{\Omega}_2 | \mathbf{R}_1, \mathbf{\Omega}_1; L).$$

$\mathbf{R}$ denotes the position of one end, $\mathbf{\Omega}$ denotes its orientation, and L is the length of intervening DNA. The meaning of this expression is a probability density, per unit of space and orientation, for a length-L polymer whose first end lies at position $\mathbf{R}_1$ and orientation Ω_1 to have the second end at position $\mathbf{R}_2$ and orientation Ω_2, assuming that the configuration was randomly sampled in a thermal environment. Each orientation $\mathbf{\Omega}$ consists of a tangent vector $\mathbf{u}$ which points in some positive sense parallel to the contour, and a twist angle. Various reduced distributions can be obtained by integrating (averaging) the full distribution over variables before (after) the conditional bar. For example, $p(\mathbf{R}_2, \mathbf{u}_2 | \mathbf{R}_1, \mathbf{u}_1; L)$ ignores the relative end twists, and $p(R_2 | R_1; L)$ considers only distances between the ends.

If we ignore excluded volume, then the end-to-end distribution completely determines all of the conformational properties of a polymer. The only forces acting on each monomer are those of its immediate neighbors, so the influence of some segment of a polymer on the conformation of the rest of the polymer is mediated by the positions and orientations of its two endpoints, which are determined by the end-to-end distribution. Compounding this logic, one can reconstruct the probability distributions of the positions and orientations of any number of points along the polymer, using only the pairwise end-to-end statistics of adjacent segments.

Neglecting excluded volume makes calculation of end-to-end statistics much more

tractable, and this approximation is justifiable in several situations. One such situation occurs when the density of DNA in the surrounding medium is low, in which case contacts between distal segments are rare and contribute little to the distributions. Short and highly-stressed polymer configurations are usually quite stiff and may not form distal contacts at an appreciable frequency. Surprisingly, in the opposite case of a very high DNA density, excluded volume is also ignorable[33], because the effective external pressure is isotropic and equal over all regions of the polymer. One notable phenomenon for which excluded volume is *not* possible to ignore is supercoiling: the two individual strands of a supercoiled plectoneme cannot be resolved separately under the phantom-chain approximation, because their conformations are supported by contact with their neighbors. However, if one is willing to treat the two strands of a plectoneme together as a single effective polymer with altered material properties, then the phantom chain model can be applied. Supercoiled DNA tends to branch, although we do not treat branched polymers in our work.

2.2 Method

2.2.1 Gaussian chain

The intrinsic length scale for DNA bending is its (bending) persistence length l_p[160], which is roughly 50 nm for naked DNA[52]. For this reason we will call segments of DNA 'long' or 'short' if the ratio of the contour length to the persistence length is significantly greater or less than one. An analogous quantity, called the twist persistence length and which we shall denote l_t, sets the length scale for computing statistics of DNA twist. Various measurements of the twist persistence length of DNA give roughly 100-120 nm[52].

In the limit of very long contours, the DNA essentially performs a random walk in which the step taken over any individual persistence length is much smaller than the total distance traversed. In this situation the end-to-end distribution approaches the Gaussian chain distribution[160]:

$$p(\Delta\mathbf{R}) \longrightarrow \left(\frac{3}{4\pi l_p L}\right)^{3/2} \cdot e^{-3(\Delta R)^2/4l_p L} \tag{2.1}$$

where the orientational distributions are uniform. This long-chain limit is trivial to calculate and is included in our program.

2.2.2 Eigenfunction method

A more accurate long-chain model than the simple Gaussian distribution was obtained by Spakowitz and Wang[140, 138]. They computed the complete distribution as a perturbation series about the Gaussian, summed into the following expression:

$$p(\mathbf{R}_2, \mathbf{\Omega}_2 | \mathbf{R}_1, \mathbf{\Omega}_1; L) = \sum_{l_0, l_f, m, j} \mathcal{F}^{-1}\left[\mathcal{L}^{-1}\left\{f^{mj}_{l_0 l_f}(\mathbf{\Omega}_1, \mathbf{\Omega}_2, \hat{\mathbf{k}}) \cdot \mathcal{G}^{mj}_{l_0 l_f}(k, p)\right\}\right]. \tag{2.2}$$

Here $\mathcal{F}^{-1}$ is the inverse Fourier operator that converts the variable $\mathbf{k}$ (having magnitude k and direction $\hat{\mathbf{k}}$) into $\mathbf{R}_2 - \mathbf{R}_1$; and $\mathcal{L}^{-1}$ is the inverse Laplace operator which converts p to L. The functions f (a product of Wigner functions[160]) and $\mathcal{G}$ (a product of continued fractions in the dummy variable 'l') are given explicitly in [140, 138]. The variables l_0 and l_f technically range from 0 to infinity in both the sums and the continued fractions, but because the contributions of the higher terms tend towards zero, in practice we drop all terms above some cutoff l_{max} in order to perform a real calculation. The shorter the contour (relative to a persistence length), the higher l_{max} must be to achieve a given accuracy.

One difficulty with this perturbative expression is that it explicitly involves eight nested iterations: the four sums over l_0, l_f, m and j; the three inverse Fourier integrals; and the inverse Laplace transform. In order to speed up the calculation, our implementation pre-computes and stores the roots of the continued-fraction polynomials that contribute to the residues of the inverse Laplace transform. Effectively, we compute the following:

1. $g^{mj}_{l_0 l_f}(k,L) = \mathcal{L}^{-1}\left\{\mathcal{G}^{mj}_{l_0 l_f}(k,p)\right\}$

2. $p(\mathbf{R}_2,\mathbf{\Omega}_2|\mathbf{R}_1,\mathbf{\Omega}_1;L) = \sum_{l_0,l_f,m,j} \mathcal{F}^{-1}\left[f^{mj}_{l_0 l_f}(\mathbf{\Omega}_1,\mathbf{\Omega}_2,\hat{\mathbf{k}})\cdot g^{mj}_{l_0 l_f}(k,L)\right].$

The limiting step 2 now involves only seven nested sums, which is still intensive but much faster than the original expression. The memory required to store the results of step 1 can be significant, but overall we have found this tradeoff to be worthwhile.

We obtain further boosts in speed by exploiting symmetries of the perturbation series that arise in various distributions. For the full distribution (2.2) we can take advantage of the fact that the continued fractions are nearly symmetric with respect to m and j. The only exception is a term α that can be removed by changing to the Laplace variable $\tilde{p} = p + \alpha$ in step 1 above, and factoring it back in at step 2 after the transform. Thus we only have to compute terms for $j \leq m$.

To obtain the reduced distribution $p(\mathbf{R}_2,\mathbf{u}_2|\mathbf{R}_1,\mathbf{u}_1;L)$ that ignores the relative twist of the two ends, one integrates the full distribution over the twist component of $\mathbf{\Omega}_2$. Since the dependence on $f^{mj}_{l_0 l_f}$ on $\mathbf{\Omega}_2$ is of the form $e^{ij\psi}$, where i is the imaginary constant and ϕ is the relative twist, all terms vanish except for those where $j = 0$; thus we can avoid the sum over j and just set it to zero. To obtain the distribution $p(\mathbf{R}_2|\mathbf{R}_1;L)$ which completely ignores the orientations of the two ends, we integrate over both the relative tangents and twists, which eliminates both the m and j sums and we simply set $m = j = 0$. These reduced distributions are much faster to evaluate than the full distribution.

To obtain the orientation-only distribution $p(\mathbf{\Omega}_2|\mathbf{\Omega}_1;L)$ one integrates over all $\mathbf{R}_2$. Equivalently, we replace the inverse-Fourier operation with a simple evaluation at $\mathbf{k} = 0$ in the program, thereby removing the three nested Fourier integrations from the evaluation and (due to the form of g) restricting $l_0 = l_f$. This calculation involves only three nested iterations and is very fast. One can also obtain the distribution involving only the separation distance between the two ends of the polymer, $p(R_2,\mathbf{\Omega}_2|R_1,\mathbf{\Omega}_1;L)$, by integrating out the relative direction of the two ends $\hat{\mathbf{R}}$ from the full distribution.

These two integrals can be done analytically, as they act only on the inverse Fourier factor, so the distance distribution requires the same calculational cost as the full distribution in Eq. (2.2).

One final interesting special case is that of cyclization, where the contour loops back on itself so that $\mathbf{R}_2 = \mathbf{R}_1$ and $\mathbf{\Omega}_2 = \mathbf{\Omega}_1$. In this case the two angular Fourier integrals (in a spherical basis) act only on the Wigner functions in $f_{l_0 l_f}^{mj}$, which, due to the properties of the Wigner functions[160], restricts $l_0 = l_f$. Thus we can both ignore the angular integrals and replace the two sums over l_0 and l_f with a single sum over l. We then exploit the fact that g is symmetric with respect to m and j, and with m and $-m$ modulo a complex conjugation, to restrict the ranges of the other two sums to $0 \leq m \leq l$ and $0 \leq j \leq m$ while multiplying by prefactors of 2 and 4 and manually effecting the complex conjugations.

Our inverse-Laplace solver uses (with permission) a C++ implementation of the complex-valued Jenkins-Traub root-finding algorithm written by Henrik Vestermark (`http://www.hvks.com/Numerical/ports.html`). A root-polisher using Newton's method ensures that the roots are at machine precision.

2.2.3 Monte Carlo

Our calculator also employs the sampling-based method called Monte Carlo for calculating the end-to-end distribution function. The Monte Carlo prescription is to computationally generate a large number of representative random polymer configurations that all begin from $(\mathbf{R}_1, \mathbf{\Omega}_1)$, and to derive end-to-end statistics just by counting the number of chains whose other end lies within some finite window of the desired $(\mathbf{R}_2, \mathbf{\Omega}_2)$. The distribution function evaluated at $p(\mathbf{R}_2, \mathbf{\Omega}_2)$ is then approximately the fraction of chains falling within the window divided by the window size. In order to generate representative conformations, the algorithm must be given the end-to-end distribution for a single segment; in other words, Monte Carlo constructs $p(\mathbf{R}_2, \mathbf{\Omega}_1 | \mathbf{R}_1, \mathbf{\Omega}_2; L)$ from $p(\mathbf{R}_2, \mathbf{\Omega}_2 | \mathbf{R}_1, \mathbf{\Omega}_1; l_s)$ where l_s is the segment length. A typical ensemble of chains generated by our calculator is shown in Figure 2-1. This Monte Carlo method works best at short contour lengths, because short polymers are fast

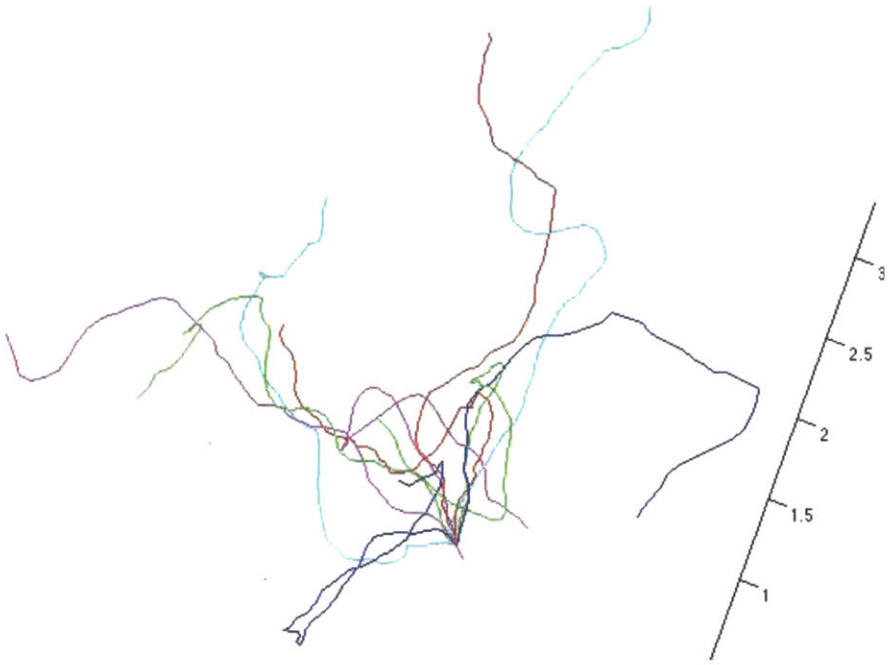


Figure 2-1: **Monte Carlo-generated conformations.** Simulated chains have a segment length of $l_s = 0.1l_p$ and a total length of $L = 5l_p$.

to evolve, and have close end-to-end spacing which allows high-resolution sampling of the distribution. It is therefore complementary to the eigenfunction technique of Spakowitz, which works best at long contour lengths.

The first step in Monte Carlo is to define the single-step end-to-end distribution $p(\mathbf{R}_2, \mathbf{\Omega}_2|\mathbf{R}_1, \mathbf{\Omega}_1; l_s)$ which defines the polymer model. Using translational and rotational invariance we recast this quantity as $p(\mathbf{x}; l_s)$ where $\mathbf{x} = \{\Delta r_\parallel, \Delta r_\perp, \Delta r_\times, \theta, \phi, \psi\}$ describes translations and rotations using axes affixed to the polymer. Here θ is the bending angle, ϕ the azimuthal angle of the bending axis, and ψ is the twist angle. In our Monte Carlo implementation $p(\mathbf{x}; l_s)$ can be specified arbitrarily: the user provides an energy function $E(\mathbf{x})$ for bending/twisting/translating assuming a contour length l_s, and a discretized probability function is computed using the Boltzmann factor $p \propto J(\mathbf{x})e^{-E(\mathbf{x})}$ where $J(\mathbf{x})$ is the volume factor appropriate to the system. Note that this is still considerably more general than the single-parameter wormlike-chain model, as it allows for nonharmonic energy functions, extensible polymers, and coupling between translations, bending and/or torsion.

In order to accommodate arbitrary single-step distributions, our program precomputes $p(\mathbf{x}; l_s)$ at discrete values of each x_i when the model is defined, then upon chain-generation approximates the function at arbitrary $\mathbf{x}$ by linearly interpolating between those values. One difficulty is that the memory and the computation time required to generate an interpolation table both increase exponentially with the dimensionality of that table. For this reason, we will want to factorize the tables as much as possible. The wormlike-chain model factorizes as $p(\mathbf{x}) = p(r_1)p(r_2)p(r_3)p(\theta)p(\phi)p(\psi)$, which is convenient because six one-dimensional tables are computationally far less expensive than a single six-dimensional table. Couplings between different degrees of freedom must be described with multi-dimensional single-segment distributions; for example, accounting for twist/roll coupling involves the three-dimensional function $p(\theta, \phi, \psi)$. Our Monte Carlo method supports practically any way of factorizing the single-joint distribution, and so can accommodate all of these models.

The arrangement of dinucleotides has traditionally been described using local axes attached to the DNA labeled $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$[31], where $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$ and $\hat{\mathbf{z}}$ is the direction of base pair stacking. Unfortunately, $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ are also generally used for fixed axes, which we need in order to measure the absolute position and orientation of the base pairs. In our program we reserve $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ for the fixed axes, and instead use the triad $(\hat{\mathbf{n}}, \hat{\mathbf{b}}, \hat{\mathbf{u}})$ to denote the local DNA-attached axes, where $\hat{\mathbf{n}} \times \hat{\mathbf{b}} = \hat{\mathbf{u}}$ and $\hat{\mathbf{u}}$ points along base pair stacking. In wormlike-chain terminology[160] these latter axes are respectively termed the normal, binormal and tangent unit vectors. We shall however use traditional dinucleotide terminology[36] in speaking about translations and rotations: translations in $\hat{\mathbf{n}}$, $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ are respectively termed 'shift', 'slide' and 'rise'; and rotations about those axes are 'tilt' $(\theta \sin\phi)$, 'roll' $(\theta \cos\phi)$ and 'twist' (ψ) in the same order, using the definitions of [36].

Once a set of interpolation tables describing the single-joint distribution have been generated, they can be repeatedly sampled to specify representative trajectories of displacements, bends and twists between subunits that determines a polymer conformation. To facilitate sampling, our program normalizes and then integrates each distribution, so that the interpolation tables contain (normalized) cumulative distri-

bution functions. If the table is one-dimensional then the sampling procedure is to draw a uniform random number on the interval $0 \leq r \leq 1$, then solve $C(x_i) = r$ for x_i using our numerical table $C(x_i) = \int_{-\infty}^{x_i} p(y_i)dy_i$. For general n-dimensional tables we store a cumulative distribution function for each variable i to be sampled, where each function consists of $i-1$ one-dimensional tables (one for each discretized choice of prior variables); and integrates over the $n-i$ subsequent variables. For example, to sample $p(x_i, x_j)$ we would first solve $C_i(x_i) = r_1$ for x_i, then linearly interpolate x_j between the solutions of $C_j^{a_i}(x_j) = r_2$ and $C_j^{a_i+1}(x_j) = r_2$ using the bracketing tables at $a_i \leq x_i \leq a_i + 1$.

The procedure for evolving the polymer given its trajectory follows the convention outlined in Ref. [36], which we will summarize briefly. Define Ω as the twist angle about $\hat{\mathbf{u}}$; Γ as the total bend angle (root-sum-of-squares of roll and tilt angles); and let ϕ be the direction of the 'total bending' axis as an angle from $\hat{\mathbf{b}}$ towards the direction of $\hat{\mathbf{n}}$ (note that this is opposite the direction of positive twist). Given the orientation of polymer segment n and a set of bend/twist angles, we compute the orientation of the following segment $n+1$ by rotating the local axes at step n as follows: 1) rotation by $\Omega/2 - \phi$ about $\hat{\mathbf{u}}$; 2) rotation by Γ about the new $\hat{\mathbf{b}}$ from step (1); 3) rotation by $\Omega/2 + \phi$ about the new $\hat{\mathbf{u}}$ from step (2). Translations are performed in the directions of either the initial or final axes, or the 'middle' axes; the last of which are obtained from the initial axes by: a) rotation by $\Omega/2 - \phi$ about $\hat{\mathbf{u}}$; b) rotation by $\Gamma/2$ about the new $\hat{\mathbf{b}}$ from step (a); c) rotation by ϕ about the new $\hat{\mathbf{u}}$ from step (b).

Most of the computational expense of Monte Carlo is in generating the simulated polymers, so runtime is proportional to the number of simulated chains N and inversely proportional to the discretization length l. Since the sampling the statistics is much faster than generating the polymers, it makes sense to decouple these two processes in some cases. Our implementation therefore first generates the chains, selecting (based on user criteria) certain chains for which the positions and the orientations of select mid/endpoints are stored; then, in subsequent steps, mine the table of selected chains to recover various statistics such as the end-to-end distributions. The tables require some memory, but they allow one to rapidly draw multiple statistics

from one set of simulated data without having to specify the criteria a-priori.

The three sources of Monte Carlo error are: the discretization of the polymer chain, $1/\sqrt{N}$ sampling error, and the averaging of the distribution over a finite sampling window. The discretization error is due to the fact that we typically discount internal structure of each segment which is often impractical to resolve in the single-segment distributions the user provides. Discretization error can be brought to tolerable levels by taking l to a few times smaller than a persistence length. The remainder of the error implies a tradeoff between the resolution one can obtain from the distribution, and the density of sampling and hence cost of the simulation. If we sample at a given $(\mathbf{R}_2, \mathbf{\Omega}_2)$, then the counting error in terms of the distribution density p, the number of samples N and the window volume V is $\Delta_c \approx \sqrt{p/NV}$. Capping Δ_c at a tolerable maximum value thus implies that the best resolution we can get at that point in the distribution is $V = p/(N\Delta_c^2)$. Of course this can be improved by increasing the number of samples N. If we are sampling the full distribution in position and orientation, V is a volume window in both $\mathbf{R}_2$ and $\mathbf{\Omega}_2$, so one can trade positional resolution for angular resolution and vice versa.

For short polymers, many endpoint configurations are sampled sparsely or not at all because they require the polymer to be sharply bent and energetically disfavored. Since thermal sampling on molecular scales happens so fast, some of these sparse regions may be relevant to biology despite being hard to access numerically. To improve the results, our Monte Carlo code has the ability to bias its sampling towards certain conformations by sampling from a different single-segment distribution which is somewhat different from the distribution that defines the model. The bias is then corrected for by post-weighting. This generic name for this technique is 'importance sampling'.

In order to see how the post-weighting procedure works, we collect all the variables that parameterize the conformation (bend and twist angles, extensions of the contour, etc. of every polymer segment) into a vector $\mathbf{y}$. Each scalar parameter y_i will be drawn from a Boltzmann distribution of an unconstrained polymer $p_i(y_i) = \frac{1}{Z_i} e^{-E(y_i)}$, giving us $p(\mathbf{y}) = \prod p_i(y_i)$ which is the overall multiplier for the entire conformation. This

distribution must be reflected somehow in the Monte Carlo method. For example, one could draw all y_i uniformly, and post-weight each sample n by $\prod p_i(y_i)$.

$$\begin{aligned} p &= \int p(\mathbf{y})\delta(f(\mathbf{y}))d\mathbf{y} \\ &\approx \frac{1}{NV}\sum_i p(\mathbf{y_n})\sigma(f(\mathbf{y_n}) < \epsilon) \\ &= \frac{1}{V}\left\langle p(\mathbf{y})\sigma(f(\mathbf{y}) < \epsilon)\right\rangle \end{aligned}$$

Traditional unbiased Monte Carlo *samples* y_i in proportion to $p(y_i)$, rather than explicitly multiplying by that factor.

$$p_{\mathrm{MC}} \approx \frac{1}{V}\left\langle \sigma(f(\mathbf{y}) < \epsilon)\right\rangle_{\mathbf{y}\in p}$$

Our biased sampling method splits the difference between these two alternatives. We factor $p_i(y_i) = s_i(y_i)w_i(y_i)$, where $s()$ is also normalized; use $s_i(y_i)$ as the sampling bias and retain $w(\mathbf{y}) = \prod_i w_i(y_i)$ as the explicit weight.

$$p_{\mathrm{bMC}} \approx \frac{1}{V}\left\langle w(\mathbf{y})\sigma(f(\mathbf{y}) < \epsilon)\right\rangle_{\mathbf{y}\in s}$$

Our Monte Carlo method estimates the sampling error of a general biased sample by binning the weighting factors and estimating the counting error in each bin b:

$$N^2V^2\left\langle \delta p^2\right\rangle = \sum_b w_b^2\left(\left\langle n_b^2\right\rangle - \left\langle n_b\right\rangle^2\right)$$

If we take these bins to be very small so that $\langle n_b\rangle \ll 1$, then n_b is almost certain to be zero or one, in which case $n_b^2 = n_b$. Then

$$N^2V^2\left\langle \delta p^2 \right\rangle \approx \sum_b w_b^2 \left(\langle n_b \rangle - \langle n_b \rangle^2 \right)$$
$$\approx \sum_b w_b^2 \langle n_b \rangle$$

We can estimate the error using the expression $\delta p \approx \sqrt{\sum_i w_i^2}/NV$ using the sample set w_i. In the special case of an unweighted sample this reduces to $\delta p \approx p/\sqrt{n_{hits}}$, but in this case we explicitly use $\delta p \approx p/\sqrt{n_{hits}-1}$ to remove the bias of having estimated the mean from the same sample set.

In addition to the various end-to-end distributions, we include routines for measuring the various moments of the distribution: the mean end-to-end *distance* function $\langle R^{2n} \rangle$, and the mean of $\langle (\mathbf{R} \cdot \mathbf{u}_0)^n \rangle$ for any n, where $\mathbf{R}$ is the end-to-end displacement and $\mathbf{u}_0$ is the initial tangent vector. These functions complement analytical results of these same quantities[160], as those can be difficult to evaluate. To estimate the error in the moments, the program divides the set of N conformations into m disjoint subsets, computes the moment separately using each subset, and then estimates the error based on the variance in the moments of the subsets.

2.2.4 Harmonic approximation method

The eigenfunction and Monte Carlo methods described above are most accurate when there are low-energy polymer conformations that satisfy the end-to-end constraints. To complement these, we have also implemented the 'harmonic approximation' (HA) method of Zhang and Crothers[166] which works best in the regime of high-energy, sharply-bent conformations. The HA method estimates the probability function by integrating about the minimum-energy configuration of the polymer that satisfies the given constraints. When the polymer is sharply bent, the energy trough tends to be steep, fluctuations are small and approximations made in the perturbative integral become ignorable. As in the case of Monte Carlo, the HA method can incorporate

any number of positional and/or orientational constraints along the length of the polymer.

Our HA calculator is essentially an extension of our Monte Carlo calculator, so it can be applied to the same variety of DNA models as Monte Carlo deals with. Sequence-dependence, extensibility, coupled degrees of freedom and non-harmonic energy functions can all be accounted for using HA. The HA method itself is described in detail in [166], although because the details of our implementation differ slightly we briefly rederive the result below.

The objective is to approximate the constrained partition function

$$Z_c = \int dx_1 dx_2 \ldots dx_N \left(e^{-E/k_B T} \delta(f^{(1)}) \delta(f^{(2)}) \ldots \delta(f^{(m)}) \right) \tag{2.3}$$

by integrating over a 2nd-order expansion of the argument of the exponential in the degrees of freedom and the Lagrange multipliers that arise from enforcing the constraints. The degrees of freedom x_i are the translations and rotations at each segment of the polymer, given that the initial position and orientation of the first segment are fixed. The calculation proceeds in two steps: first, a minimum-energy configuration is found that satisfies the constraints along with a set of Lagrange multipliers; secondly, the derivatives of the energy function about the minimum-energy configuration allow us to approximate Z_c. By integrating over the interpolation tables we find the unconstrained Z, which allows us to calculate the probability function $p = Z_c/Z$.

We find the minimum-energy conformation of the polymer by simultaneously satisfying the constraints and a set of force-balance equations, with Lagrange multipliers transmitting the constraint forces. Borrowing the terminology of Zhang and Crothers, E is the configuration energy, $f^{(i)}$ denotes the ith scalar constraint function, m is the number of constraints, and N is the total number of degrees of freedom. At equilibrium these variables satisfy:

$$\mathbf{y} = \begin{bmatrix} \frac{dE}{dx_i} + \sum_{j=1}^{m} \lambda_j \frac{df^{(j)}}{dx_i} \\ f^{(j)} \end{bmatrix} = 0. \tag{2.4}$$

To find the x_{0i} and λ_j that give the minimum-energy configuration $\mathbf{y} = 0$, we use a general-purpose multi-dimensional root finder from the GNU Scientific Library (GSL)[42].

If a minimum-energy configuration is found, the probability density function p, which is the ratio of the constrained to full partition functions Z_c/Z, is found by the technique which is explained carefully in [166]. The constrained partition function evaluates to $Z_c = e^{-E_s}\sqrt{\pi^{N-m}/|\mathbf{A}'||\mathbf{F}'|}$, where E_s is the energy of the configuration found in the first step, and the matrices $\mathbf{A}'$ and $\mathbf{F}'$ are defined by:

$$\mathbf{A}'_{ik} = \frac{1}{2}\frac{d^2E}{dx_i^2}\delta_{ik} + \frac{1}{2}\sum_{j=1}^{m}\lambda_j \frac{d^2f^{(j)}}{dx_i dx_k}$$
$$\mathbf{F}'_{jl} = \sum_{i=1}^{N}\sum_{k=1}^{N} A'^{-1}_{ik}\frac{df^{(j)}}{dx_i}\frac{df^{(l)}}{dx_k}.$$

Meanwhile, we calculate the unconstrained partition function by summing the numerical interpolation tables. Taking the ratio of the two partition functions gives:

$$p = \frac{Z_c}{Z} = \frac{e^{-E_s}}{Z}\sqrt{\frac{\pi^{N-m}}{|\mathbf{A}'||\mathbf{F}'|}}.$$

2.2.5 Finite-width delta function

For some purposes we would like to know the normal modes of a constrained polymer. Unfortunately, real-valued normal modes do not come directly out of the Zhang and Crothers analysis, because the delta-function constraints are made tractable for integration by Fourier-transforming them into complex space. However, if we replace the singular delta-functions with a narrow Gaussian, then both the energy and the

constraint appear straightforwardly within an exponential, which we can expand to second order and straightforwardly convert to real-valued normal modes.

As before, let N be the number of degrees of freedom and m the number of scalar-valued constraints. The constrained partition function with Gaussian constraints is:

$$Z_c = \int dx_1 dx_2 \dots dx_N \left(e^{-E} e^{-\frac{k_1}{2} f_1^2} e^{-\frac{k_2}{2} f_2^2} \dots e^{-\frac{k_m}{2} f_m^2} \right) \times \sqrt{\frac{k_1}{2\pi}} \sqrt{\frac{k_2}{2\pi}} \dots \tag{2.5}$$

$$= \sqrt{\frac{\prod_i k_i}{(2\pi)^m}} \int d^N \mathbf{x} \exp \left[-E - \sum_{i=1}^{m} \frac{k_i}{2} f_i^2 \right] \tag{2.6}$$

$$= \sqrt{\frac{|\mathbf{K}_c|}{(2\pi)^m}} \int d^N \mathbf{x} \exp \left[-E - \frac{1}{2} \mathbf{f}^T \mathbf{K}_c \mathbf{f} \right] \tag{2.7}$$

where the third line introduced a general constraint stiffness matrix $\mathbf{K}_c$. Expanding the energy E and constraint functions f_i to second order in $\mathbf{x}$ about the minimum-energy configuration gives:

$$\begin{aligned} Z_c \approx \sqrt{\frac{|\mathbf{K}_c|}{(2\pi)^m}} \int d^N \Delta\mathbf{x} \exp \Big[& -E_0 - (\nabla_\mathbf{x} E) \cdot (\Delta\mathbf{x}) - \frac{1}{2} (\Delta\mathbf{x})^T \cdot (\nabla_\mathbf{x} \nabla_\mathbf{x} E) \cdot (\Delta\mathbf{x}) \\ & - \frac{1}{2} \mathbf{f}_0^T \mathbf{K}_c \mathbf{f}_0 - \mathbf{f}_0^T \mathbf{K}_c \cdot (\nabla_\mathbf{x} \mathbf{f} \cdot \Delta\mathbf{x}) \\ & - \frac{1}{2} (\nabla_\mathbf{x} \mathbf{f} \cdot \Delta\mathbf{x})^T \cdot \mathbf{K}_c \cdot (\nabla_\mathbf{x} \mathbf{f} \cdot \Delta\mathbf{x}) \\ & - \frac{1}{4} \mathbf{f}_0^T \mathbf{K}_c \cdot \big((\Delta\mathbf{x})^T \cdot (\nabla_\mathbf{x} \nabla_\mathbf{x} \mathbf{f}) \cdot (\Delta\mathbf{x}) \big) \\ & - \frac{1}{4} \big((\Delta\mathbf{x})^T \cdot (\nabla_\mathbf{x} \nabla_\mathbf{x} \mathbf{f}) \cdot (\Delta\mathbf{x}) \big)^T \cdot \mathbf{K}_c \mathbf{f}_0 \Big] . \end{aligned} \tag{2.8}$$

The first-order terms cancel:

$$(\nabla_\mathbf{x} E) \cdot (\Delta\mathbf{x}) + \mathbf{f}_0^T \mathbf{K}_c \cdot (\nabla_\mathbf{x} \mathbf{f} \cdot \Delta\mathbf{x}) = 0 \tag{2.9}$$

$$\longrightarrow \mathbf{f}_0^T \mathbf{K}_c (\nabla_\mathbf{x} \mathbf{f}) = -\nabla_\mathbf{x} E \tag{2.10}$$

and Z_c reduces to:

$$Z_c \approx \sqrt{\frac{|\mathbf{K}_c|}{(2\pi)^m}} e^{-E_0'} \int d^N \Delta\mathbf{x} \exp\left[-\frac{1}{2}(\Delta\mathbf{x})^T \cdot \mathbf{M} \cdot (\Delta\mathbf{x})\right] \tag{2.11}$$

$$= (2\pi)^{(N-m)/2} \sqrt{\frac{|\mathbf{K}_c|}{|\mathbf{M}|}} e^{-E_0'} \tag{2.12}$$

where

$$E_0' = E_0 + \frac{1}{2}\mathbf{f}_0^T \mathbf{K}_c \mathbf{f}_0 \tag{2.13}$$

and

$$M_{ij} = \partial_{x_i}\partial_{x_j} E + \left(\partial_{x_i}\mathbf{f}\right)^T \mathbf{K}_c \left(\partial_{x_j}\mathbf{f}\right) + \frac{1}{2}\left[\left(\partial_{x_i}\partial_{x_j}\mathbf{f}\right)^T \cdot \mathbf{K}_c\mathbf{f}_0 + \mathbf{f}_0^T\mathbf{K}_c \cdot \left(\partial_{x_i}\partial_{x_j}\mathbf{f}\right)\right] \tag{2.14}$$

$$= \partial_{x_i}\partial_{x_j}\left(E + \lambda \cdot \mathbf{f}\right) + \left(\partial_{x_i}\mathbf{f}\right)^T \mathbf{K}_c \left(\partial_{x_j}\mathbf{f}\right). \tag{2.15}$$

The last step follows from the identification $\lambda = \mathbf{K}_c\mathbf{f}_0$ from the Zhang-Crothers analysis. Force balance implies that $\mathbf{f} \sim |\mathbf{K}|^{-1}$, so in the delta-function limit where $|\mathbf{K}| \to \infty$ the energy reduces to $E_0' \to E_0$.

The eigenmodes of the constrained polymer are useful because they can form a basis for biased Monte Carlo sampling. For good sampling statistics, we want to choose $\mathbf{K}_c$ so that the variances in the f_i will be of the same order as the respective sampling window sizes. Specifically, we will dial in $\mathbf{K}_c$ such that the projection of $\mathbf{M}$ into $\mathbf{f}$-space is diagonal with entries $(\mathbf{PMP}^T)_{ii} = T_{ii} = 1/\sigma_i^2$, where σ_i is the window size of constraint i. Projections between $\Delta\mathbf{x}$ and $\mathbf{f}$ are effected by $\mathbf{f} = \boldsymbol{\beta}\Delta\mathbf{x}$ and $\Delta\mathbf{x} = \mathbf{Pf}$, where $\beta_{ij} = \partial_{x_i} f_j$ and $\mathbf{P} = \left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right)\boldsymbol{\beta}^T$. Writing $\mathbf{M}$ in the form $\mathbf{M} = \mathbf{M}_0 + \boldsymbol{\beta}\mathbf{K}_c\boldsymbol{\beta}^T$, we find our desired constraint stiffness matrix to be $\mathbf{K}_c = \mathbf{T} - (\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\mathbf{M}_0\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}$.

2.3 Implementation

The core program described here is a text-based tool, in which the user types commands and obtains results on the screen or in output files. This program can be downloaded and run on a desktop computer. For convenience, we have also provided an online calculator with a web interface, which has much of the capability of the command-line tool. The URL containing both the online calculator and the download for the command-line tool is:

`http://mtshasta.phys.washington.edu/wormulator/`

Web-based documentation supports the online calculator, while the command-line tool comes with a help file that explains its use and provides examples.

2.3.1 Web interface

The online calculator allows the user to measure the full end-to-end distribution $p(\mathbf{R}_2, \mathbf{\Omega}_2 | \mathbf{R}_1, \mathbf{\Omega}_1; L)$ for single values of $\{\Delta\mathbf{R}, \mathbf{\Omega}_1, \mathbf{\Omega}_2, L\}$ by providing a polymer length, endpoint tangents and relative displacement and twist. The wormlike chain model is used for all calculations. By default the material parameters (bending/twist persistence lengths and intrinsic twist) are those of DNA, although one can change these to model other polymers. Checkbox options to sum over $\mathbf{R}$, tangents and/or twists allow the various reduced distributions to be computed. One convenience, not present in the command-line tool, is that the program supports several length units (nanometers, persistence lengths, base pairs of DNA, etc.) which may differ between the various input fields and the program's output. Online computations may be performed using the eigenfunction and Monte Carlo methods.

The output of the basic computation outlined above is a single number: a probability density (probability per unit volume and/or unit angular volume) for the polymer's second end to be in the given position and/or orientation relative to the first end. Frequently, the user would like to map this distribution over a range of values

in some parameter – for example, to predict the efficiency of cyclization over a range of polymer lengths. It would be tedious to do these multiple evaluations manually, so the online calculator incorporates a 'counter' for accomplishing this automatically: given a range of values of the counter variable, the program will evaluate multiple times, once for each value of the counter. To set up unique conditions for every run the user writes the counter variable 'c' into the input fields: for example the twist field might read `sin((pi/8)*c)`, or the length may be `e^-c`. If the counter is used, then the output of the calculator will be a table, where the values of the counter are displayed alongside the outputs of each evaluation of the distribution.

It is easy for the user to request a calculation that will either require too much memory or run practically forever, especially when using the eigenfunction method which involves many nested loops. In order to avoid overtaxing the server, the online calculator restricts the permitted ranges of those parameters that affect memory usage and computation time: the maximum l-value and the number of integration steps in the eigenfunction calculation, the number of samples and discrete segments that Monte Carlo generates, and the range of the counter variable. Because of these restrictions, intensive calculations can only be done using the command-line tool.

2.3.2 Command-line tool

To perform a calculation using the command-line tool, the user may enter commands directly into the interactive prompt, or else place those commands into a file and have the program execute them all at once. A basic calculation takes two or three commands; the help file includes examples demonstrating how to perform each type of calculation. Additional commands allow the user to generate and save tables, inspect intermediate stages of the calculation, control the random sequences, and measure computation time and memory usage.

The Monte Carlo component of the command-line tool has several capabilities that are not available from the web site. One is the ability to use very general 2D or 3D polymer models, including those with non-harmonic energy functions, coupled degrees of freedom, sequence dependence and extensible segments, along with the trick

of biased sampling. Additionally, only the command-line tool has the perturbative method of Zhang and Crothers. The command-line tool can perform very lengthy calculations that are forbidden online. Finally, the ability to export tables is useful for storing results, making plots, and troubleshooting.

2.4 Results and Discussion

2.4.1 Validation

In order to validate our program, we performed a number of calculations that could be checked either explicitly or against a different method. For the perturbative calculation, we compared selected computations of Euler angles and Wigner functions with hand-derived results, verified that the distribution asymptotically approaches the expected Gaussian for long chains, and reproduced the cyclization plot given in ref. [138]. We also compared probability densities given by our implementation of the perturbative method with equivalent calculations performed using a symbolic calculator (Mathematica[159]), drawing test cases from the full distribution and from the orientation-only and cyclization distributions. Tests of the Monte Carlo method included explicit checks on the propagation and rotation of individual segments, on the sampling of the bending/twisting energy functions (see Figure 2-2 for wormlike chain, biased distribution where $E_{bias} = E/4$), and evaluations of $\langle \mathbf{R} \cdot \mathbf{u}_0 \rangle$ which as expected approach $l_p \left(1 - e^{-L/l_p}\right)$. In all cases the results agreed with the predictions within numerical precision, as long as the parameters controlling accuracy (l_{max}, K_{max}, etc. for the perturbative method; segment length and number of runs for Monte Carlo) were made stringent enough.

When a polymer's length is on the order of a couple of persistence lengths, both the perturbative calculation and Monte Carlo can give good answers with reasonable computational cost, and of course their results should agree. We generated end-to-end distributions of a 3-persistence-length stretch of DNA using these two methods, which are compared in Figure (2-3). Each plot shows a 2-D slice of the distribution

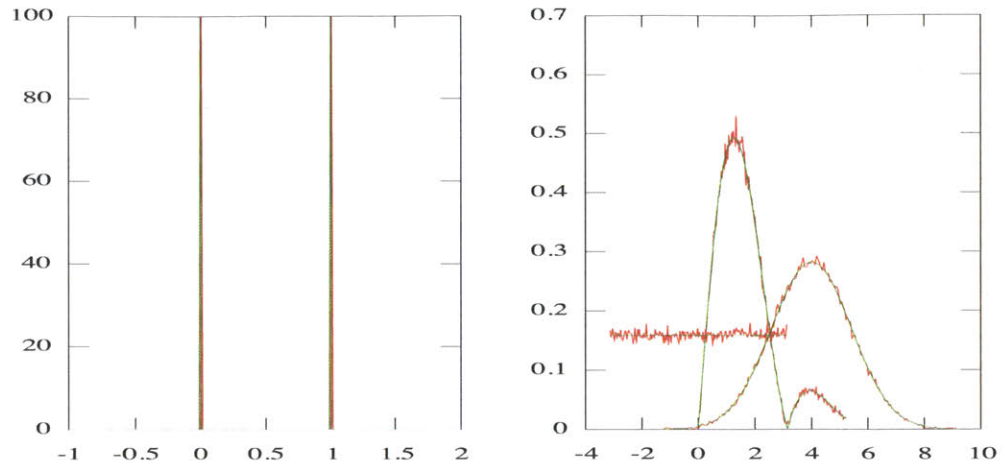


Figure 2-2: **Monte Carlo sampling of single-segment distributions.** Sampled segment evolution parameters (red) are overlaid with target sampling function (green), validating our sampling method. Left panel: shift, slide, rise distributions; right panel: bend, azimuth, and twist.

over the spatial separation of the ends, for the given fixed end-orientations.

As our final test, we compared the probability density for cyclization of wormlike-chain DNA between the three main calculational techniques: Spakowitz's eigenfunction method, Monte Carlo, and the Zhang-Crothers perturbative method, using a DNA chain whose length ranged over $L = 100 - 400$ bp (Figure 2-4). Monte Carlo results were computed using both an unbiased and biased sampling; in the latter case the bias was to lower the energy by a factor of 1.5. The perturbative results were obtained by discretizing each chain with 500 segments; by contrast the Monte Carlo results used a much lower discretization of 20 segments in order to obtain good statistics when using the biased sampling. A perturbative result of Yamakawa[160] is included to compare against that of Zhang and Crothers. By convention all probability densities are expressed as J-factors[69] defined by $J = 8\pi p$. The J-factor accounts for overall rotations of the entire system, and is related to the looping free energy by $G = -k_B T \ln J$. The eigenfunction result for DNA, using somewhat lower tolerances and computed over the range $L = 200 - 600$ bp, is also used as an example for the online calculator.

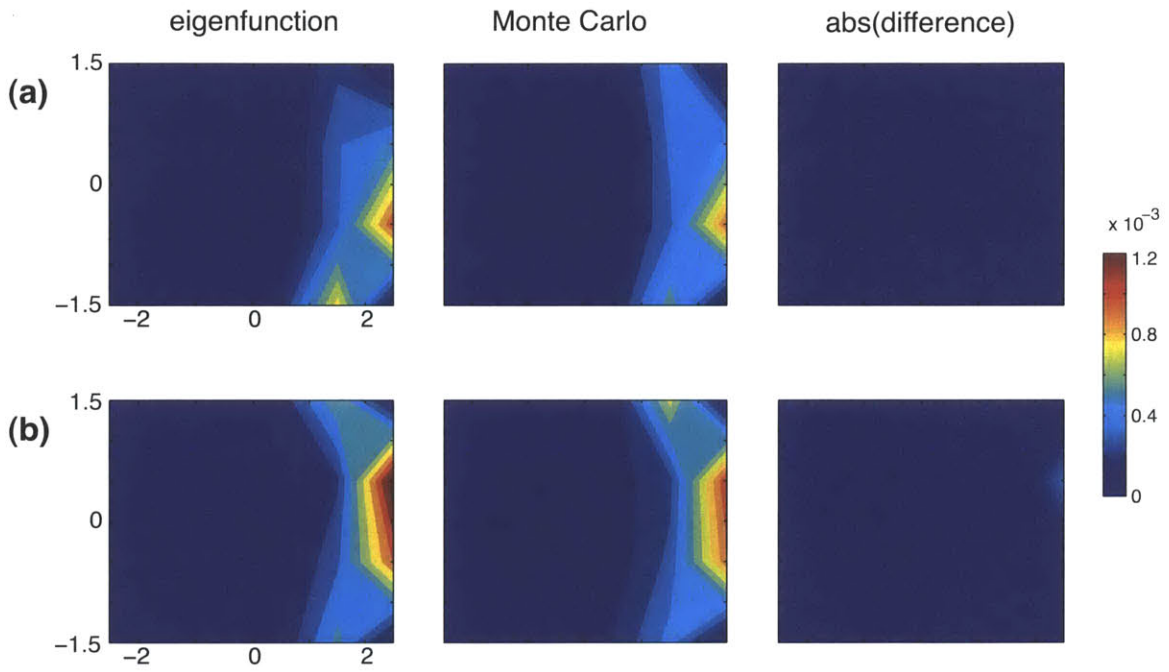


Figure 2-3: **Comparison of distributions calculated by eigenfunction and Monte Carlo methods.** Two slices (top and bottom) through the three-dimensional end-to-end position distributions calculated by the eigenfunction method (left panels) and Monte Carlo (middle panels), and their difference (right panels). For this run, $L = 3$, $l_p = 1$, $l_t = 2.08$, twist $= 1.5$, $\hat{\mathbf{u}}_0 = \hat{\mathbf{z}}$, $\hat{\mathbf{n}}_0 = \hat{\mathbf{x}}$, $\hat{\mathbf{u}}_f = (0, -.6, .8)$, $\hat{\mathbf{n}}_f = (-1, 0, 0)$.

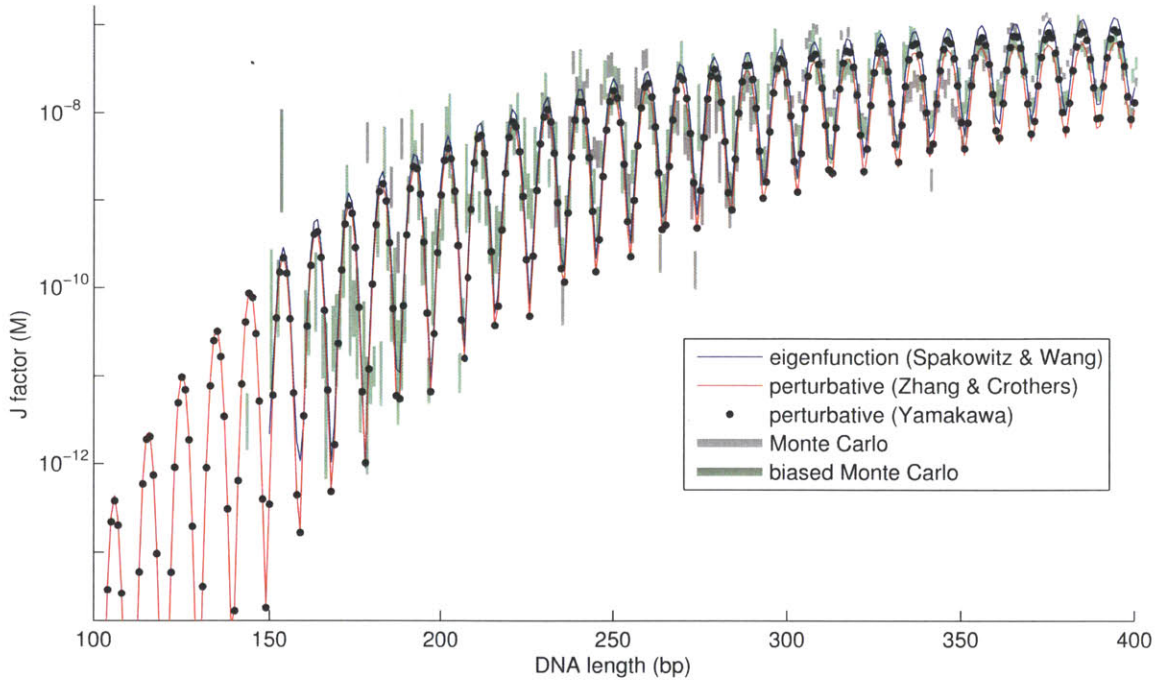


Figure 2-4: **Comparison of methods for calculating DNA cyclization rates.** DNA cyclization frequency $J(L)$ in nanomolars was calculated using the eigenfunction, perturbative and Monte Carlo methods (with and without biasing).

2.4.2 Example: going beyond wormlike chain using cyclization data

We have shown that our calculator correctly constructs the multi-segment distribution $p()$ given a single-segment polymer model. We will now apply our methods to the inverse problem: using end-to-end statistics of an extended DNA chain to learn something about the poorly-characterized single-segment distribution. Experimental evidence indicates that DNA bends sharply much more often than the wormlike chain model predicts[24]. Thus the energy penalty for at least some large bending angles is probably less than $E_{wormlike} = (l_p/2)\dot{\theta}^2$, although the amount of softening and the range of bend angles over which this occurs is still unknown.

Here we will compare three candidate models for high-energy DNA bending, and show using simulations how an experiment may be performed that can discriminate between these models. Model I is the familiar wormlike chain model. Model II differs from the wormlike chain in that its energy function remains constant beyond

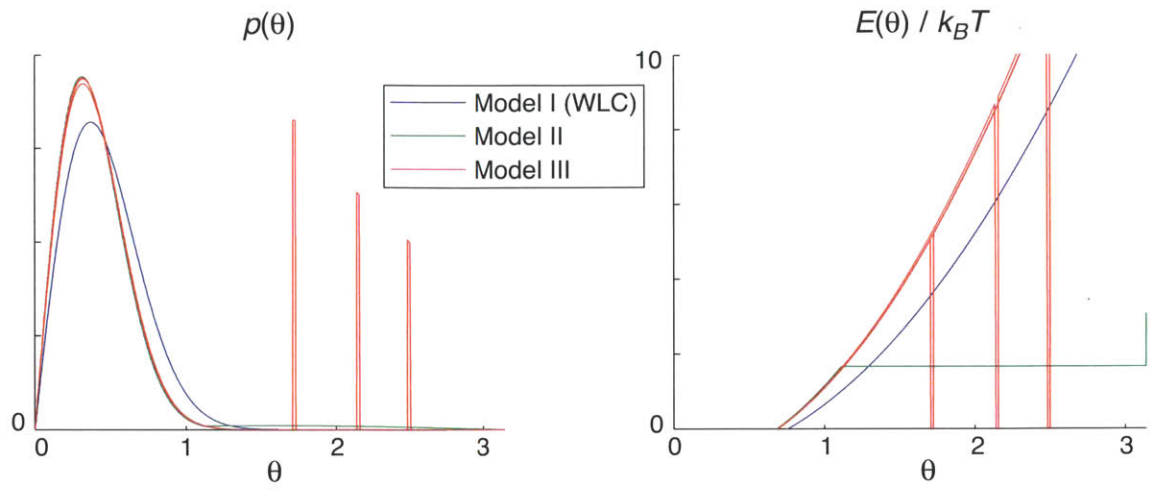


Figure 2-5: **Cyclization of nicked DNA: energy functions of different models.** Models I, II, and III are defined by $p(\theta) = \exp\left(-E(\theta)/k_BT\right)$; note that three variants of Model III are superimposed on these plots. Part of the green line for Model II is hidden behind the red lines of Model III.

a certain bending angle: $E\left(\theta > \theta_{max}\right) = E\left(\theta_{max}\right)$. Model III has a wormlike chain energy function for all bending angles except near a certain kinking angle where we set the energy penalty to zero: $E\left(|\theta - \theta_{kink}| < \epsilon\right) = 0$. We note that similar models have been studied analytically[157] and computationally in other contexts[20][168]. The low-energy stiffnesses of the various kinking models are set slightly greater than higher than in the wormlike chain model so that the long-chain distributions of all three of these models converge; thus the three models would become indistinguishable by experiments that measure end statistics over long contour lengths. The energy functions and single-segment distributions of each model are shown in figure 2-5.

The experiment we consider is a cyclization experiment, in which linear double-stranded DNA fragments are converted to closed circles in the presence of DNA ligase at a rate proportional to the probability density for end-to-end looping with matching tangents and twists: $p\left(\mathbf{R}_1 = \mathbf{R}_2, \Omega_2 = \Omega_1\right)$. In our proposed experiment, the double-stranded DNA contour to be cyclized bears two nicks, so that the cyclized loop will consist of two double-stranded regions termed the long and short arms of the loop separated by flexible nicked hinges. Similar DNA configurations have been used in prior experiments that probed the stretching behavior of the short arm[21, 153, 148,

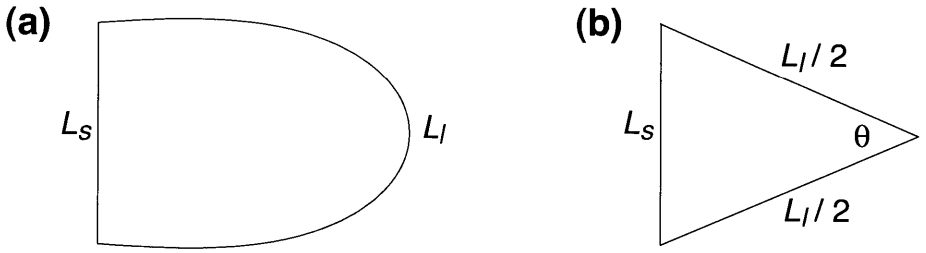


Figure 2-6: **Cyclization of nicked DNA: energy-minimized kinked and unkinked contours.** (a) Energy-minimized conformation of wormlike-chain polymer with 2 free joints. (b) Relaxed conformation with a kink angle θ absorbing the bending in the long arm.

113] or the distance between the ends of the short arm[133, 134]. Single-stranded nicks are quite flexible in bending so we will assume that they are free joints. The result of having these two joints is that the short arm will be stretched nearly taut between the two ends of the long arm in the cyclized conformation, as shown in Figure 2-6a. We note that single-stranded nicks are also completely free to rotate in twist, eliminating the energy penalty for relative twisting along the contour and allowing us to apply tangent-only orientation conditions on the ends in our calculations.

In the cyclized conformation, the sharpest bending tends to be concentrated near the midpoint of the long arm, so any kinks in the DNA will likely form there. In one idealized scenario, a central kink absorbs all of the bending of the long arm, forming a triangle in which the kink angle θ_k is opposite the fully-extended short arm as shown in figure 2-6b. Assuming that cyclized conformations would be variations on this idealized triangle, we expected Model III to show an enhanced cyclization rate when θ_k is close to the model's preferred kinking angle. Model II should also shown an enhanced cyclization rate for sharp bending relative to the wormlike chain prediction, but over a wide range of θ_k.

We used Monte Carlo to compare the predicted end distribution $p\left(\Delta\mathbf{R}0, \Omega_2 = \Omega_1\right)$ between the three models, including three variants of Model III with different preferred kink angles. We measured the cyclization probability over a range of lengths of both the long and short arms (L_l and L_s), where the length of each arm was always a multiple of 21 bp to minimize helical asymmetry. Figure 2-7 shows the results of our

sampling, plotted by $\theta = 2\arcsin(L_s/L_l)$ on the horizontal axis and the total length $L_l + L_s$ on the vertical. The free energy gain in moving subunits from the long arm to the short arm is consistently below 1 k_BT per base pair, so we do not anticipate melting at the double-stranded ends to skew our results. Not only are the three models easily discriminated, but any preferred kink angle is clearly visible. Since twisting does not enter into these distributions, we believe that this experiment could cleanly extract the bending component of the energy function at high bend angles.

2.5 Appendix A: Derivatives of the constraint functions

The harmonic approximation of Zhang and Crothers[166] requires efficient and accurate calculation of the first and second derivatives of the constraints with respect to each of the coordinates x_i. Each vector constraint is enforced by a set of scalar $f^{(j)}$ factors, each of which is the vector constraint on one of the unit vectors $\mathbf{e}_c^r$ projected along some axis $\mathbf{e}_j$. Let a denote the segment immediately after step x_i, and c denote the segment at which the constraint is applied; necessarily $c \geq a$ or else the derivative vanishes. We will ignore the constraints on the position and orientation of the initial segment, as our calculator does not allow those to vary. For all other scalar constraints we calculate a first-order derivative by

$$\frac{\partial f^{(j)}}{\partial x_i} = \frac{\partial f^{(j)}}{\partial \mathbf{R}_a} \cdot \frac{\partial \mathbf{R}_a}{\partial x_i} + \sum_{p=1}^{3} \left(\frac{\partial f^{(j)}}{\partial \mathbf{e}_a^p} \cdot \frac{\partial \mathbf{e}_a^p}{\partial x_i} \right) \tag{2.16}$$

$$\tag{2.17}$$

where each right-hand term is a dot product of two 3-vectors. The $\mathbf{e}_a^p$ are the three orientation vectors at segment a. For simplicity the derivatives of $\mathbf{R}_a$ and $\mathbf{e}_a^p$ with respect to x_i are evaluated numerically by two symmetric perturbations about x_{0i}. Perturbing a translational degree of freedom affects only $\mathbf{R}_a$, whereas a general per-

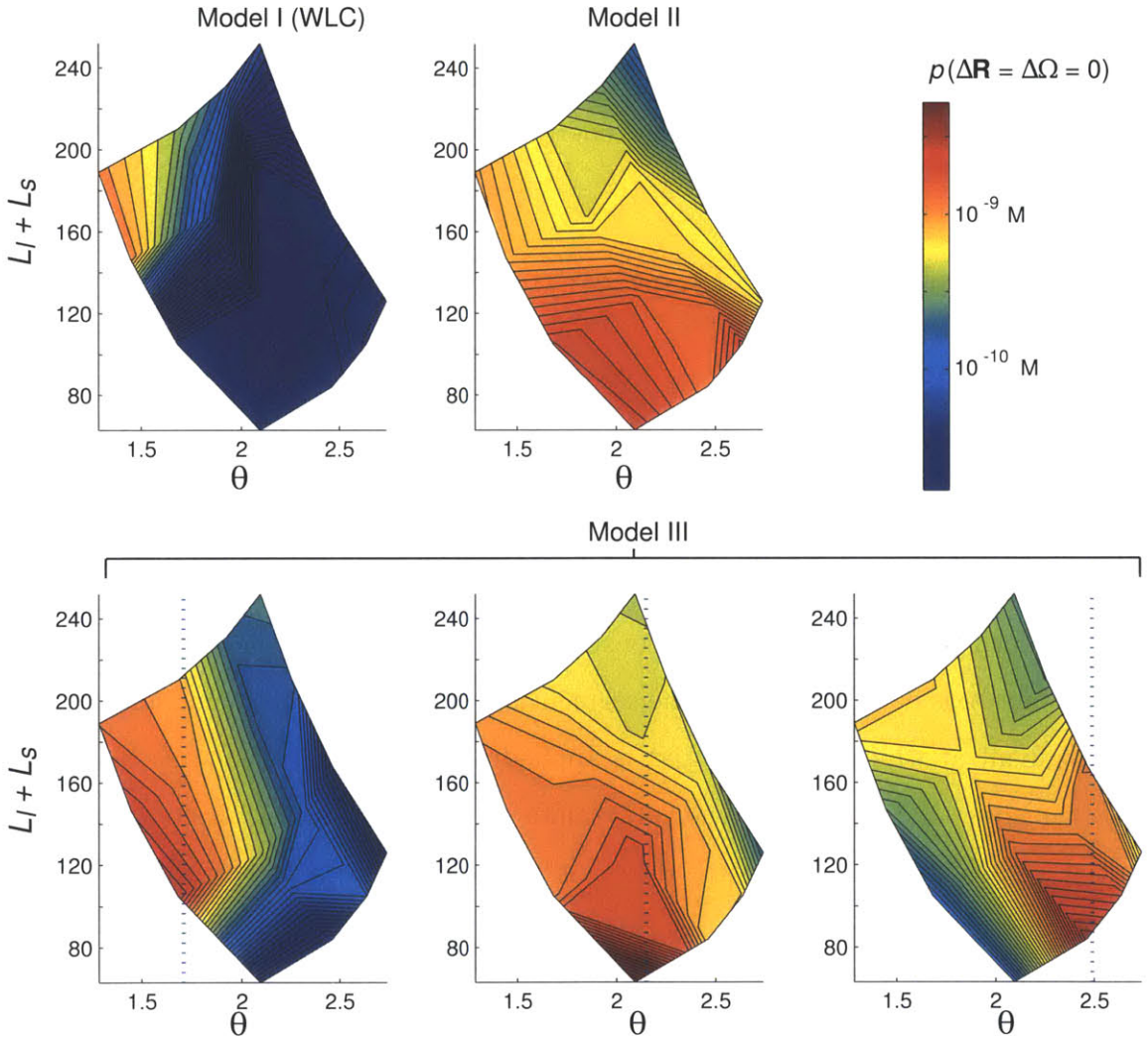


Figure 2-7: **Cyclization of nicked DNA: cyclization frequencies.** Cyclization frequencies were estimated for each model (including three variants of Model III having different kink angles θ_k, indicated by a dotted line) using Monte Carlo. Horizontal axis is the kink angle θ needed to relax the long arm; vertical axis is the total length of the polymer $L_l + L_s$.

turbation in a rotational degree of freedom affects $\mathbf{R}_a$ as well as the three orientational unit vectors $\mathbf{e}_a^p$.

The partial derivatives of $f^{(j)}$ can be calculated geometrically. Positional constraints where $a > a_0$ have derivatives from both translation of $\mathbf{R}_a$ and rotation of the intervening length of polymer between a and c. In this case we have

$$\frac{\partial f^{(j)}}{\partial \mathbf{R}_a} = \frac{\partial f^{(j)}}{\partial \mathbf{R}_c} = \mathbf{e}_j \tag{2.18}$$

$$\frac{\partial f^{(j)}}{\partial \mathbf{e}_a^p} = \frac{\partial f^{(j)}}{\partial \mathbf{R}_c}\frac{\mathbf{R}_c}{\partial \mathbf{e}_a^p} \tag{2.19}$$

$$= \mathbf{e}_j \cdot \frac{\partial}{\partial \mathbf{e}_a^p} \sum_{p'} \left(\mathbf{e}_a^{p'} \cdot (\mathbf{R}_c - \mathbf{R}_a) \right) \mathbf{e}_a^{p'} \tag{2.20}$$

The vector $\mathbf{R}_c - \mathbf{R}_a$ maintains a constant projection upon the unit vectors $\mathbf{e}_a$, so the derivative operator $\partial_{\mathbf{e}_a^p}$ acts only upon the second unit vector.

$$\frac{\partial f^{(j)}}{\partial \mathbf{e}_a^p} = \mathbf{e}_j \cdot (\mathbf{e}_a^p \cdot (\mathbf{R}_c - \mathbf{R}_a))\, \mathbf{I} \tag{2.21}$$

$$= [(\mathbf{e}_a^p \cdot \mathbf{R}_c - \mathbf{R}_a)]\, \mathbf{e}_j. \tag{2.22}$$

Orientational constraints upon the vector $\mathbf{e}_c^r$ are only affected by rotations in the $\mathbf{e}_a$. Because the projections $\mathbf{e}_a^p \cdot \mathbf{e}_c^r$ are constant, we ignore the action of the derivative $\partial_{\mathbf{e}_a^p}$ on the component of $\mathbf{e}_c^r$ in the $\mathbf{e}_a$ basis.

$$\frac{\partial f^{(j)}}{\partial \mathbf{R}_c} = 0 \tag{2.23}$$

$$\frac{\partial f^{(j)}}{\partial \mathbf{e}_a^p} = \frac{\partial f^{(j)}}{\partial \mathbf{e}_c} \cdot \frac{\partial \mathbf{e}_c}{\partial \mathbf{e}_a^p} \tag{2.24}$$

$$= \mathbf{e}_j \cdot \frac{\partial}{\partial \mathbf{e}_a^p} \sum_{p'} \left(\mathbf{e}_a^{p'} \cdot \mathbf{e}_c \right) \mathbf{e}_a^{p'} \tag{2.25}$$

$$= (\mathbf{e}_a^p \cdot \mathbf{e}_c^r)\, \mathbf{e}_j. \tag{2.26}$$

We calculate second derivatives of the constraint function $\partial^2 f^{(j)}/\partial x_i \partial x_k$ in two different ways. Let a and b correspond to the segments at which x_i and x_k act, where we take $a \leq b$. When $a = b$ there are complicated couplings between x_i and x_k, so we numerically perturb both coordinates symmetrically by $\epsilon/2$ and calculate the difference $(f^{(j)}_{++} - f^{(j)}_{+-} - f^{(j)}_{-+} + f^{(j)}_{--})/\epsilon^2$. When $a < b$ it is much more efficient to evaluate second derivatives using the first derivatives already calculated, noting that the only nonzero second derivatives when $a \neq b$ involve two orientational degrees of freedom. We first calculate the derivatives $\partial \mathbf{f}^r_c/\partial x_k$ for each constrained vector $\mathbf{e}^r_c$ along all three axes $\mathbf{e}_b$, regardless of how many components of the vector have been constrained. We then take a further derivative in x_i by noting that for both positional and orientational constraints the constraint derivative vector co-rotates with the unit vectors $\mathbf{e}_a$.

$$\frac{\partial^2 f^{(j)}}{\partial x_i \partial x_k} = \frac{\partial}{\partial x_i}\frac{\partial f^{(j)}}{\partial x_k} \tag{2.27}$$

$$= \left(\frac{\partial}{\partial x_i}\frac{\partial \mathbf{f}^r_c}{\partial x_k}\right) \cdot \mathbf{e}_j \tag{2.28}$$

$$= \sum_{pq} (\mathbf{e}^p_a \cdot \mathbf{e}^q_b)\left(\mathbf{e}^q_b \cdot \frac{\partial \mathbf{f}^r_c}{\partial x_k}\right)\left(\frac{\partial \mathbf{e}^p_a}{\partial x_i} \cdot \mathbf{e}_j\right) \tag{2.29}$$

$$\tag{2.30}$$

2.6 Appendix B: Converting p into a density in angular coordinates

Due to the way the angular constraints are represented, they give rise to a probability density per unit volume in Cartesian space (for example, $du_x du_y dn_x$) of various components of the angular unit vectors. We want to output densities in angular space ($\sin\theta \cdot d\theta d\phi d\psi$), so we have to multiply our computed density by the Jacobian factor between the two spaces (e.g. $d\hat{\mathbf{u}}/d\Omega$) in terms of the known parameters ($\hat{\mathbf{u}}$).

The simplest case is of that of a single scalar constraint. There is no fundamental

difference between $u/n/b$ constraints, so we can parametrize the problem in terms of angles any way we like as long as we express our final answer in terms of the constrained vector components. Without loss of generality we will take our constrained vector to be $\hat{\mathbf{u}}$ and choose the simple parametrization $u_z = \cos\theta$.

$$u_z = \cos\theta$$
$$\left|\frac{du_z}{d\theta}\right| = |\sin\theta| = \sqrt{1-u_z^2}$$

In the case of a constraint with two scalar components, both components u_x and u_z generally apply to the same angular vector. Then we have:

$$u_x = \sin\theta\cos\phi$$
$$u_z = \cos\theta$$
$$\frac{1}{\sin\theta}\frac{d(u_x u_z)}{d(\theta\phi)} = \frac{1}{\sin\theta}\begin{vmatrix} \cos\theta\cos\phi & -\sin\theta\sin\phi \\ -\sin\theta & 0 \end{vmatrix} = \sin\theta\sin\phi$$
$$= \sqrt{1-u_x^2-u_z^2}.$$

For the general three-component case we can take $\hat{\mathbf{u}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ as before. Bending is effected by a rotation vector $\hat{\omega} = \hat{\mathbf{z}}\times\hat{\mathbf{u}}/\sin\theta = (-\sin\phi, \cos\phi, 0)$. In this frame two vectors orthogonal to $\hat{\mathbf{u}}$ are $\hat{\mathbf{n}}_0 = \hat{\omega}\times\hat{\mathbf{u}} = (\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)$ and $\hat{\mathbf{b}}_0 = \hat{\mathbf{u}}\times\hat{\mathbf{n}} = (-\sin\phi, \cos\phi, 0)$, a general linear combination of which gives us $\hat{\mathbf{n}} = (\cos\theta\cos\phi\cos\psi - \sin\phi\sin\psi, \cos\theta\sin\phi\cos\psi + \cos\phi\sin\psi, -\sin\theta\cos\psi)$. We have chosen a convention in which $\hat{\mathbf{u}}$ by angle θ rotates into $\hat{\mathbf{n}}$ when there is no twist ($\psi = 0$).

There can be only two independent constraints on a given unit vector because normalization constrains the third component. We will assume that two of the constraint components apply to unit vector A and the remaining component applies to another unit vector B. There are two possibilities: 1) either the constrained compo-

nent of B is in the same direction as one of the constrained components of A, or else 2) the constraint components are all along different axes. Without losing generality we will assume $\hat{\mathbf{u}}$ corresponds to A and $\hat{\mathbf{n}}$ to B; other combinations will have the same result with appropriate permutations of the variable names. We compute case (1) for constraints on u_x, u_z and n_z.

$$\begin{aligned}
\frac{1}{\sin\theta}\frac{d(u_x u_z n_x)}{d(\theta\phi\psi)} &= \frac{1}{\sin\theta}\begin{vmatrix} \cos\theta\cos\phi & -\sin\theta\sin\phi & 0 \\ -\sin\theta & 0 & 0 \\ -\cos\theta\cos\psi & 0 & \sin\theta\sin\psi \end{vmatrix} \\
&= |\sin\theta\sin\phi|\,|\sin\theta\sin\psi| \\
&= \sqrt{1-u_x^2-u_z^2}\sqrt{1-u_z^2-n_z^2}
\end{aligned}$$

For case (2) we constrain u_x, u_y and n_z.

$$\begin{aligned}
\frac{1}{\sin\theta}\frac{d(u_x u_z n_x)}{d(\theta\phi\psi)} &= \frac{1}{\sin\theta}\begin{vmatrix} \cos\theta\cos\phi & -\sin\theta\sin\phi & 0 \\ \cos\theta\sin\phi & \sin\theta\cos\phi & 0 \\ -\cos\theta\cos\psi & 0 & \sin\theta\sin\psi \end{vmatrix} \\
&= |\cos\theta|\,|\sin\theta\sin\psi| \\
&= \sqrt{1-u_x^2-u_y^2}\sqrt{u_x^2+u_y^2-n_z^2}
\end{aligned}$$

Chapter 3

Measuring Chromosome Conformation *In Vivo*

In this chapter we turn away from the prediction of DNA configurations based on thermodynamic models, and concentrate on how to directly measure the conformations of chromosomes in live or intact cells. Various methods might be used for making this measurement, which we will categorize as being either continuous or discrete. A continuous rendering of a conformation would be some sort of traced contour, for example as seen in some high-resolution EM images[112], coupled with information allowing each genomic locus to be mapped to its respective region on the contour. A discrete conformation maps a selected set of genomic loci to their locations inside the three-dimensional space of the cell, leaving one to interpolate the contour between these loci.

I have chosen to focus on discrete measurements of conformation, out of the belief that they are easier to parametrize and analyze than continuous conformations, and also because of the experimental difficulty in tracing *in vivo* DNA trajectories and the relative ease of labeling discrete loci. One disadvantage is that discrete conformations are inherently coarse-grained relative to continuous ones: bends of the DNA in between adjacent loci cannot be resolved. This problem is essentially solved if the locus spacing is comparable to or below the bending scale of DNA inside the cell, because in that case interpolation accurately reproduces the intervening contour. A

discrete measurement made at persistence-length resolution shows practically all major bends of the DNA; there is little further conformational information to be gleaned from spacing the loci closer together.

A serious difficulty in making a discrete measurement of a complex conformation is that a large number of loci need to be not only localized but also *individually identified.* Typically, the number of loci needed to discretize the contour far exceeds the number of loci that can be uniquely distinguished ('colors' in our terminology). To get an idea of the discrepancy, note that full chromosomes range in length from about $\sim 10^4$ to $\sim 10^7$ persistence lengths, so to resolve even 1% of a chromosome at high resolution will involve hundreds or thousands of loci. For comparison, fluorescence microscopy that distinguishes three colors is considered state-of-the-art. This problem must be addressed in order to measure extensive conformations, either experimentally by drastically increasing the number of available labeling colors into the thousands, or computationally through techniques for inferring the identity of each locus despite a multitude of look-alikes. In this chapter[121] we will take the latter route.

3.1 Proposed experiment

We propose to reconstruct the *in vivo* conformation of some long stretch of a chromosome in a single-cell and at high resolution, using a three step procedure. The first step is to target a large number of predefined genetic loci using labels of only a few distinguishable colors. In the second step those labels are 'imaged' or otherwise localized *in vivo* (whether by fluorescence microscopy or some other means is not important). The final step is an analysis, described in the next section, that maps each imaged label to its respective locus by computationally looking for low-energy conformations connecting the imaged labels, given their known spacing and color-ordering along the DNA contour. The original contribution in this thesis is the analysis. We will start by outlining the sorts of experiments that could be employed for the analysis, while emphasizing that no experiment was actually performed for this thesis.

We begin by assuming that the localization is done by fluorescence microscopy,

because this method is routine, high-resolution (if super-resolution techniques are employed), and because the marking of genetic loci with fluorophores is a highly-developed art. Super- (subdiffraction) resolution is important because, as mentioned, it allows for a close locus spacing which aids the subsequent analysis. The most promising super-resolution methods for our purposes photoswitch and centroid-fit discrete fluorophores (PALM[11], FPALM[61] and STORM[123]); these have already demonstrated the ability to image in three dimensions[67][135] and in multiple color channels[5][132]. The resolution from these techniques ($\sim$ 30 nm) implies a locus spacing on the order of an *in vitro* persistence length, which is around the optimal spacing for building a discrete conformation. One disadvantage of these imaging methods is that they are slow, so fixation is required to prevent movement at small scales.

There are a number of ways to fluorescently label the genome *in vivo*. A standard method for targeting loci in fixed cells is fluorescence in-situ hybridization (FISH). A FISH labeling experiment would probably use end-labeled probes, in order to sharply define the locus and because the aforementioned super-resolution techniques are inherently single-molecule. An alternative to FISH is FROS (fluorescent repressor-operator system)[119], which would require the cloning of operator binding sites into the region of the chromosome to be mapped. Finally, if we are willing to consider *ex vivo* experiments such as the imaging of flat DNA spreads, one could target restriction sites using fluorescently-fused DNA-binding proteins such as restriction enzymes; unbound, freely-diffusing enzymes are hard to detect by PALM/STORM. If the probes can be replaced during the experiment then different 'colors' can be imaged sequentially, obviating the need for multiple fluorescence channels.

It is important to use an irregular label spacing and color ordering, so that every region of DNA can be uniquely identified. For short DNA contours it is easy to engineer each individual probe or binding site to satisfy this requirement. For longer contours, various barcoding techniques might be used that exploit the random distribution of restriction sites to heterogenously label the DNA in a single step. One advantage of an *ex vivo* experiment using labeled restriction enzymes is that

barcoding is automatic without any extra experimental steps.

If the FISH labeling method is employed, one can automatically generate barcoded probe templates where each restriction site corresponds to a spot of a unique color in the labeling scheme. The probe templates are generated by serial genomic digests with several different restriction enzymes. Both ends of each probe are then ligated to adaptors, with a unique adaptor sequence corresponding to each unique restriction overhang. The final probes are produced by PCR using end-labeled primers complementary to the adaptor sequences. This protocol guarantees that each fluorophore color is conjugated to only one primer sequence, which matches one type of adaptor, which ligates to one restriction site, thus mapping each restriction site to a unique color. Note that this will method double-label every restriction site in a perfect experiment, since each restriction site will contact two probes at the 5' end and two probes at the 3' end, and each probe is (singly) labeled at the 5' end. Double-labeling would have to be accounted for in a preprocessing step (or else one can accept a high false-positive rate); however, it could considerably suppress the false negative rate which we have found to be a more harmful source of error for the final analysis.

Automated barcoding using a FROS method involves generating a genomic array of different operator binding sites, in random order and separated by random-length spacer sequences. This can be done automatically using digested DNA for the spacers using a three-step process. First, digested DNA of random lengths is ligated to the operator sequences to form a library of operator-spacer fusions. In the second step, operator-spacer fusions are serially stitched together to form progressively longer arrays; this step is repeated until a suitably long array is produced. The third step is to sequence the final array and insert it onto the genome. The length of the binding array will be limited by cloning capacity (probably about 10 kb).

3.2 Analysis: the 3D-alignment method

The computational method we will present in this section infers probabilistic conformations from a) the genomic position and label color of each locus, and b) the physical

position and color of each imaged label. The constraint that potentially allows such conformations to be identified is that labels close together along the genomic contour must also be close in physical space; the closer the label spacing the tighter the constraints. False positives (nonspecifically-bound labels) and false negatives (undetected labels) are both taken into account by our method, as is localization error. Our method makes an approximation that lowers the accuracy but greatly speeds up the analysis, so that reconstructions involving $\sim$ 1000 labeled loci ($\sim$ 100 kb at persistence-length resolution) are completely feasible.

Our algorithm outputs a table of probabilities for mapping genomic loci to imaged spots. In order to properly construct these probabilities one should consider only conformations in which no two genomic loci map to the same spot in the image; unfortunately, enforcing this rule exactly makes the problem intractable when more than a handful of loci are involved. Our method ignores this 'no-overlap rule' except between adjacent mappings (i.e. two loci that are consecutive or separated only by false negatives), thereby making the solution much more tractable but introducing considerable error into the probability table. In order to recover some of this error we associate a binding energy with every spot in the image, and iteratively recompute the probabilities while adjusting the binding energies in order to minimize a cost function C. The cost function establishes two constraints on the probability table: 1) the expected number of mapped loci (excluding false negatives) should match the estimated number of spots in the image that are not false positives; and 2) the normalization condition $\sum_i p_{i;\alpha} \leq 1$ must hold for each spot in the image. Details of the algorithm are given in the Analysis section.

In order to evaluate the quality of a locus-to-spot mapping, we measure the Shannon information[127] per locus needed to specify the unique conformation, given the partial information already contained in the mapping probabilities. Shannon information is a positive number that is inversely related to the probability assigned to the *correct* mapping; a final information score of zero implies a perfect mapping ($p = 1$ for the correct conformation). A probabilistic mapping *lacks* this amount of information relative to the perfect mapping, so the objective of our algorithm is to recover

as much of this missing information as possible and thereby minimize the information measure. This information score can be measured in simulations for which the correct mapping is known, but unfortunately not in a real experiment. To estimate the mapping quality in a real experiment, we average the information metric over all possible mappings weighted by the mapping probabilities, obtaining a score we term 'entropy'. Entropy is therefore an estimate of the information score that does not require knowledge of the true mapping; the entropy and information scores should be nearly equal if the mapping probabilities accurately reflect their likelihoods of being correct (see Figure 3-1). Note that accurate mapping probabilities are a necessary, but not sufficient, condition for a *good* mapping: a state of uniform probabilities is accurate but uninformative, and consequently scores high on both information and entropy.

3.2.1 The partition function Z

A particular conformation is determined by a set of mapping variables $\{\alpha_1, \alpha_2, ..., \alpha_N\}$, where each α_i either indexes the spot in the image corresponding to the ith locus along the contour, or else $\alpha_i = \varnothing$ in the case of a false negative. For example, '$\{3, \varnothing, 2\}$' denotes the conformation where the first contour locus maps to spot 3 in the image, the third contour locus maps to spot 2 in the image, the second contour locus was not imaged at all, and the presumed spot 1 in the image must have been a false positive.

The statistical weight of a given mapping between two loci a and b is a simple product of terms connecting each pair of loci along the contour, where each term is the statistical weight required to stretch the length of intervening DNA between the two loci by the separation distance of the two spots in the image. We sum over all conformations between loci a and b, and multiply each conformation by 'no-overlap' terms which remove conformations that reuse one or more spots:

$$e^{-F(\alpha)} = \sqrt{f_{a;\alpha_a} f_{b;\alpha_b}} \prod_{i=a}^{b-n_f} \sqrt{f_{i;\alpha_i}} e^{-F(l_{i+n_i} - l_i, \mathbf{R}_{\alpha_i}, \mathbf{R}_{\alpha_{i+n_i}})} \sqrt{f_{i+n_i;\alpha_{i+n}}} w^{n_i - 1} \prod_{j<i} \epsilon_{i;\alpha_i}^{j;\alpha_j} \quad (3.1)$$

$$\equiv \sqrt{f_{a;\alpha_a} f_{b;\alpha_b}} \prod_{i=a}^{b-n_f} \sigma_{i;\alpha_i}^{i+n_i;\alpha_{i+n_i}} \prod_{j<i} \epsilon_{i;\alpha_i}^{j;\alpha_j} \quad (3.2)$$

Here l_i is the distance of locus i from one end of the DNA, measured along the contour; $\mathbf{R}_{\alpha_i}$ is the location in three-dimensional space of imaged spot α_i; and $\sigma_{i;\alpha_i}^{i+n;\alpha_{i+n}}$ is the statistical weight corresponding to the stretching of two loci i and $i+n$ between the locations of spots α_i and α_{i+n} in the image. We have associated an unphysical free parameter $f_{i;\alpha_i}$ with each mapping $i \to \alpha_i \neq \varnothing$, and an unbinding penalty w with each false-negative $j \to \alpha_j = \varnothing$. The summation index i ranges only over mappings where $\alpha_i \neq \varnothing$, so each mapped locus i is separated from the next mapped locus by $n_i - 1$ intervening false negatives. We enforce the no-overlap condition by the $\epsilon_{i;\alpha_i}^{j;\alpha_j}$ factors which are 1 if $\alpha_i \neq \alpha_j$ and zero if $\alpha_i = \alpha_j$.

If we ignore correlations between non-adjacent α_i, the only overlap terms are $\epsilon_{i;\alpha_i}^{i+n_i;\alpha_{i+n_i}}$ which can be absorbed into the $\sigma_{i;\alpha_i}^{i+n;\alpha_{i+n}}$ by setting those to zero when $\alpha_i = \alpha_{n+i}$. In that case the partition function simplifies to:

$$e^{-F(\alpha)} = \sqrt{f_{a;\alpha_a} f_{b;\alpha_b}} \prod_{i=a}^{b-n_f} \sqrt{f_{i;\alpha_i}} e^{-F(l_{i+n_i} - l_i, \mathbf{R}_{\alpha_i}, \mathbf{R}_{\alpha_{i+n_i}})} \sqrt{f_{i+n_i;\alpha_{i+n}}} w^{n_i - 1} \quad (3.3)$$

$$\equiv \sqrt{f_{a;\alpha_a} f_{b;\alpha_b}} \prod_{i=a}^{b-n_f} \sigma_{i;\alpha_i}^{i+n_i;\alpha_{i+n_i}} \quad (3.4)$$

We will use the approximate partition function (3.4) in our 3d-alignment algorithm which we describe below. After that we will consider how to add some of the non-adjacent ϵ_i^j into the calculation to improve the accuracy.

We have opted to model the statistical weighting factor $\sigma(R; L)$ using the Gaussian chain distribution[160], where $\sigma(R; L) = (2\pi s^2)^{-3/2} \exp(-R^2/2s^2)$ and $s^2 = 2l_p L/3$. (The Gaussian chain fails definitively at short contour lengths, so for $L < 2l_p$ we

set $s^2 = L^2/3$.) There are several reasons to use a Gaussian model. 1) The true form of $\sigma()$ is unfortunately poorly-constrained *in vivo*, and almost certainly differs between and within organisms. Since the Gaussian chain model is a smooth diffusive distribution, it should not fail sharply if the *in vivo* polymer bends in unexpected ways. 2) There is only one free parameter: the persistence length l_p, which is the bending length scale of the polymer (approximately 50 nm $\approx$ 150 bp for *in vitro* DNA). 3) Convolving the localization error of the two loci (usually assumed Gaussian, having variances s_1^2 and s_2^2) against a Gaussian $\sigma()$ simply results in a new Gaussian $\sigma'()$ having $s'^2 = s^2 + s_1^2 + s_2^2$.

It turns out that nearly every quantity that depends upon a contour locus i also depends on the corresponding mapping variable α_i. For notational convenience, we will henceforth omit the explicit dependences of most quantities on the α_i variables. As a general rule, *every* superscripted or subscripted index attached to a quantity implies that it is also a function of the respective mapping variable 'α' corresponding to that index: f_a is shorthand $f_{a;\alpha_a}$, σ_i^{i+n} is shorthand for $\sigma_{i;\alpha_i}^{i+n;\alpha_{i+n}}$, ϵ_i^j stands for $\epsilon_{i;\alpha_i}^{j;\alpha_j}$, etc.

3.2.2 Step 1: calculate Z

Consider the quantity Z_a^b (shorthand for $Z_{a;\alpha_a}^{b;\alpha_b}$) that sums all possible mappings of a stretch of consecutive loci bounded by $a \leq i \leq b$, where the endpoints are fixed by the implicit α_a and α_b but each intervening α_i is summed over all imaged spots of the appropriate color.

$$Z_a^b = \sum_{\alpha_{a+1} \ldots \alpha_{b-1}} \prod_{i=a}^{b-n_f} \sigma_i^{i+n_i} \tag{3.5}$$

By setting a to the first mapped locus on the entire chain and summing over α_a, we obtain the 'half-partition function' Z^b which accounts for all mappings that end at (b, α_b). Because we have relaxed the no-overlap condition, Z^b can be efficiently calculated using a recursive rule.

$$Z^b \equiv \sum_{a<b} \sum_{\alpha_a} w^{a-1} \sqrt{f_a} Z_a^b \tag{3.6}$$

$$= \sum_{m=1}^{b-1} \sum_{\alpha_{b-m}} \sigma_{b-m}^b Z^{b-m} + w^{b-1} \sqrt{f_b}. \tag{3.7}$$

The boundary terms can be absorbed by imagining that all chains begin from some outside locus 0, having $Z^{0;0} = 1$ and $\sigma_0^b = w^{b-1}\sqrt{f_b}$.

$$Z^b = \sum_{m=1}^{b} \sum_{\alpha_{b-m}} \sigma_{b-m}^b Z^{b-m} \tag{3.8}$$

In like manner we obtain the second half-partition function Z_a, which accounts for all mappings beginning at (a, α_a). We account for the boundary terms by considering all chains to end at an imaginary $N+1$ locus having the properties $Z_{N+1} = 1$, $\sigma_a^{N+1} = w^{N-a}\sqrt{f_a}$ and $\sigma_0^{N+1} = w^N$.

$$Z_a \equiv \sum_{b>a} \sum_{\alpha_b} Z_a^b \sqrt{f_b} w^{N-b} \tag{3.9}$$

$$= \sum_{n=1}^{N-a+1} \sum_{\alpha_{a+n}} \sigma_a^{a+n} Z_{a+n} \tag{3.10}$$

Finally, the full partition function Z, which contains all possible mappings of the entire chain, is the summation of Z_a^b over both endpoints. In practice Z is best calculated by choosing a locus x, summing over all possible mappings of that locus and adding terms arising from $\alpha_x = \varnothing$.

$$Z \equiv \sum_{a=1}^{N} \sum_{\alpha_a} \sum_{b=a}^{N} \sum_{\alpha_b} w^{a-1} \sqrt{f_a} Z_a^b \sqrt{f_b} w^{N-b} \tag{3.11}$$

$$= \sum_{\alpha_x} Z^x Z_x + \sum_{m=1}^{x} \sum_{n=1}^{N-x+1} \sum_{\alpha_{x-m}} \sum_{\alpha_{x+n}} Z^{x-m} \sigma_{x-m}^{x+n} Z_{x+n} \tag{3.12}$$

The various partition functions immediately give us the mapping probabilities $p(x \to \alpha_x)$:

$$p_x = \frac{Z^x Z_x}{Z}. \tag{3.13}$$

The amount of unrecovered information associated with this probability matrix is $I = -\sum_i \log p_{i;\alpha_i'}$ where the vector α' contains the true mapping; this score is used to evaluate the algorithm on simulated inputs for which α' is known. The entropy of the probability matrix, defined by $S = -\sum_i \sum_{\alpha_i} p_i \log p_i$, gives an estimate of the unrecovered information when the true mapping is not known.

3.2.3 Step 2: adjust weighting factors

To enforce proper normalization of the probability array and tune the false-negative rate, we perform gradient descent on the following cost function with respect to the weighting parameters $\mathbf{w} = \{f_i, w\}$:

$$C = \frac{1}{2} \sum_{\alpha} \left(\max \left\{ 0, \sum_i p_{i;\alpha} - 1 \right\} \right)^2 + \frac{K}{2} \left(\left(\sum_i \sum_{\alpha_i} p_{i;\alpha_i} \right) - N_s \right)^2 \tag{3.14}$$

where N_s is the estimated number of imaged spots that are not false positives. We break up the cost-function gradient in the following way:

$$\frac{dC}{dw_k} = \sum_i \sum_{\alpha_i} \frac{dC}{dp_i} \left[\sum_j \sum_{\alpha_j} \left(\frac{\partial p_i}{\partial Z^j} \frac{dZ^j}{dw_k} + \frac{\partial p_i}{\partial Z_j} \frac{dZ_j}{dw_k} \right) + \frac{\partial p_i}{\partial w_k} \right] \tag{3.15}$$

$$= \sum_i \sum_{\alpha_i} \left(\frac{\partial C}{\partial Z^i} \frac{dZ^i}{dw_k} + \frac{\partial C}{\partial Z_i} \frac{dZ_i}{dw_k} + \frac{dC}{dp_i} \frac{\partial p_i}{\partial w_k} \right). \tag{3.16}$$

Four of the six terms are straightforward:

$$\frac{dC}{dp_i} = \max \left\{ 0, \sum_j p_{j;\alpha_i} - 1 \right\} + K \left(\left(\sum_j \sum_{\alpha_j} p_j \right) - N_s \right) \tag{3.17}$$

$$\frac{\partial C}{\partial Z^i} = \frac{dC}{dp_i} \frac{Z_i}{Z} + \frac{dC}{dZ|_i} Z_i - \sum_{n=1}^{N-i+1} \left(\sum_{j=i+1}^{i+n-1} \sum_{\alpha_j} \frac{dC}{dp_j} \frac{p_j}{Z} \right) \sum_{\alpha_{i+n}} \sigma_i^{i+n} Z_{i+n} \tag{3.18}$$

$$\frac{\partial C}{\partial Z_i} = \frac{dC}{dp_i} \frac{Z^i}{Z} + \frac{dC}{dZ|_i} Z^i - \sum_{m=1}^{i} \left(\sum_{j=i-m+1}^{i-1} \sum_{\alpha_j} \frac{dC}{dp_j} \frac{p_j}{Z} \right) \sum_{\alpha_{i-m}} Z^{i-m} \sigma_{i-m}^i \tag{3.19}$$

$$\frac{\partial p_i}{\partial w_k} = \frac{dC}{dZ|_i} \sum_{m=1}^{i} \sum_{n=1}^{N-i+1} \sum_{\alpha_{i-m}} \sum_{\alpha_{i+n}} Z^{i-m} \frac{d\sigma_{i-m}^{i+n}}{dw_k} Z_{i+n} \tag{3.20}$$

where

$$\frac{dC}{dZ|_i} \equiv -\sum_{\alpha_i} \frac{dC}{dp_i} \frac{p_i}{Z} \tag{3.21}$$

$$\frac{d\sigma_{i-m}^{i+n}}{df_j} = \sigma_{i-m}^{i+n} \cdot \frac{\delta_{i-m,j}\delta_{\alpha_{i-m},\alpha_j} + \delta_{i+n,j}\delta_{\alpha_{i+n},\alpha_j}}{2f_j} \tag{3.22}$$

$$\frac{d\sigma_{i-m}^{i+n}}{dw} = \sigma_{i-m}^{i+n} \cdot \frac{m+n-1}{w}. \tag{3.23}$$

The remaining step is to calculate dZ^i/dw_k and dZ_i/dw_k, each of which involves chains of derivatives owing to the recursive structure of the half-partition functions. To calculate these derivatives efficiently, we note an analogy between our computation of Z^i and Z_i and the propagation of signals through neural networks, and borrow the famous backpropagation algorithm[122] in order to compute our derivatives at

the same order as the rest of our calculation. Consider a linear feedforward neural network with connectivity between all layers, where the output of neuron α_i in layer i is x_i (shorthand for $x_{i;\alpha_i}$), and the weight connecting neurons between two layers as W_i^j. To compute the x_i we use the recursive rule that $x_i = \sum_{m\geq 1}\sum_{\alpha_{i-m}} W_{i-m}^i x_{i-m}$; this is equivalent to our calculation of Z^i or Z_i. The gradient $s_i \equiv dC/dx_i$, called a sensitivity, can be propagated backwards in a second recursive step: $s_i = \partial C/\partial x_i + \sum_{m\geq 1}\sum_{\alpha_{i+m}} W_i^{i+m} s_{i+m}$. The gradient in the weights is then: $dC/dW_{i-m}^i = x_{i-m}s_i$. The correspondence with our situation is that Z^i or Z_i is analogous to the output x_i of a given neuron, and σ_a^b plays the role of the weight matrix W_a^b. Using this analogy, we backpropagate the sensitivities using the following rule:

$$s^i = \frac{\partial C}{\partial Z^i} + \sum_{m=1}^{N-i+1}\sum_{\alpha_{i+m}} \sigma_i^{i+m} s^{i+m} \tag{3.24}$$

$$s_i = \frac{\partial C}{\partial Z_i} + \sum_{n=1}^{i}\sum_{\alpha_{i-n}} \sigma_{i-n}^i s_{i-n}. \tag{3.25}$$

The component of $\partial C/\partial \sigma_a^b$ that comes through the dependence of C on the half-partition functions is:

$$\left.\frac{\partial C}{\partial \sigma_a^b}\right|_{Z^i, Z_i} = s^b Z^a + s_a Z_b \tag{3.26}$$

so the component of $\partial C/\partial f_i$ and $\partial C/\partial w$ coming through the half-partition functions is:

$$\sum_j \sum_{\alpha_j} \left(\frac{\partial C}{\partial Z^j} \frac{dZ^j}{df_i} + \frac{\partial C}{\partial Z_j} \frac{dZ_j}{df_i} \right) = \frac{1}{2f_i} \left[\sum_{m=1}^{i} \sum_{\alpha_{i-m}} \left(\sigma_{i-m}^{i} s^i Z^{i-m} + \sigma_{i-m}^{i} s_{i-m} Z_i \right) + \sum_{n=1}^{N-i+1} \sum_{\alpha_{i+n}} \left(\sigma_i^{i+n} s^{i+n} Z^i + \sigma_i^{i+n} s_i Z_{i+n} \right) \right] \tag{3.27}$$

$$= \frac{1}{f_i} \left[Z^i s^i + Z_i s_i - \frac{1}{2} \left(Z^i \frac{\partial C}{\partial Z^i} + Z_i \frac{\partial C}{\partial Z_i} \right) \right] \tag{3.28}$$

$$\sum_j \sum_{\alpha_j} \left(\frac{\partial C}{\partial Z^j} \frac{dZ^j}{dw} + \frac{\partial C}{\partial Z_j} \frac{dZ_j}{dw} \right) = \sum_{a=0}^{N} \sum_{b=a+1}^{N+1} \sum_{\alpha_a} \sum_{\alpha_b} \frac{b-a-1}{w} \left(Z^a \sigma_a^b s^b + s_a \sigma_a^b Z_b \right). \tag{3.29}$$

Our derivation assumes a separate weighting factor f_i for each mapping $i \to \alpha_i$. In practice much better results are obtained when there is just one weighting factor per spot in the image, which all mappings to that spot use: $f_{i;\alpha} = f_{j;\alpha} = f_{;\alpha}$ for all i which implies that $dC/df_{;\alpha} = \sum_i dC/f_{i;\alpha}$. The gradient optimizer actually works with the binding energies $\mathbf{E} = -\log \mathbf{w}$ rather than the weighting factors themselves, in order to keep all $\mathbf{w} = \{f_{;\alpha}, w\}$ positive. The cost function derivatives in the weighting energies are: $dC/dE_i = -w_i(dC/dw_i)$.

There are roughly N^2 ($N_c \times N_f$) elements of each array (Z^x, Z_x, s^x, s_x), each of which couples to all $\sim N$ spots of all $\sim N$ preceding layers (due to false negatives) in both the forward and backward steps of the calculation. We greatly shorten both halves of the calculation by truncating the latter two sums: we ignore long runs of consecutive false negatives, and we ignore all σ_a^b factors between distant a and b, subject to some probability threshold in either case. As a result each iteration of our computation is roughly of order $N^2 \cdot \rho$, where ρ is the density (mean number of neighboring spots).

3.2.4 Connection to the Traveling Salesman Problem

Our spot-mapping problem is remarkably similar to the famous traveling salesman problem (TSP), which is the problem of finding the unique shortest path connecting a set of cities. Cities are analogous to spots in the image, and the contour connecting the spots is the path of the traveling salesman. There are some differences between the two problems, some trivial and some not.

- The 'distance' metric between two 'cities' (spots) in our problem is not Euclidean: $\sigma(\Delta\mathbf{R}, L) \neq f(\Delta R_x, \Delta R_y, \Delta R_z)$. This is only a superficial difference because it does not affect our solution method, and non-Euclidean TSP problems are well-studied in the literature.

- The cities (spots) have 'colors', and paths are constrained in the order of colors that they connect. This is again a superficial difference because one interpretation of the color constraint is that $\sigma() = \infty$ between any two cities where one or both cities have the wrong color.

- We do not know exactly where the cities (spots) are due to localization error. Again, this simply affects the metric $\sigma()$ so does not introduce any additional complexity into the problem.

- The distance metric $\sigma(\Delta\mathbf{R}, L)$ between two cities/spots depends on the stop along the tour, owing to the dependence on L which varies according to the nonuniform spot spacing. This changes the problem significantly: we are actually working with a generalization of the TSP called the time-dependent TSP (TDTSP).

- A realistic algorithm for an experiment will have to deal with both false positives (cities that are not supposed to be on the tour) and false negatives (cities that have mysteriously vanished). I have not seen (TD)TSP literature that deals with these sources of error.

The 3D-alignment problem reduces to traditional (non-metric) TSP in the special case of: equally spaced spots of one color, and no false positives or false negatives.

3.2.5 General comments

There are differences between our problem and those of studied TSP-type problems, but perhaps the more significant difference is in our approach to solving the problem. Traditionally, the (TD)TSP problems are solved for the exact unique optimum tour (conformation). This contrasts with our 3d-alignment algorithm which sums over ensembles of solutions, and does so most inexactly owing to the the no-overlap approximation. Obviously, having to make no approximation is a major plus for the traditional TSP solver. Whether the true optimal conformation is more significant than the sum over conformations depends on whether the optimal conformation or something similar to it is overwhelmingly likely to be correct. If the energy landscape contains many separated basins that are each somewhat probable, then a single unique conformation may not be very useful. One justification for obtaining a solution in terms of mapping probabilities that consider all conformations is that the p-values are frank about their uncertainty in the final answer.

The second justification for using the partition-function method is that it is much less computationally expensive than finding an optimal solution, and therefore more easily scaled up to mapping long conformations. The exact solution to a TSP is unfortunately believed to be an exponential problem, although heuristic algorithms can solve typical data sets in much less time. The TDTSP is necessarily as or more complicated compared to the TSP, and has been studied much less: solutions for the TDTSP have been limited to very small problems (tens of cities/spots). The obvious way of calculating the TDTSP partition function exactly—by brute-force enumeration—unfortunately scales exponentially in the number of loci N. Using our approximate algorithm, memory requirements and computation time (per iteration) are both roughly proportional to N^4 for the full calculation, brought down to N^2 times the spot density by truncating the inner two sums.

One peculiarity of our approach is that our solution is not unique. The N_f independent free parameters (the f_i; w is not independent) are solved using N_f inequality constraints (on the normalization of each field spot α_i) and one equality constraint

(on the overall false-negative rate), so if the false positive rate is greater than zero the field-normalization constraints describe a N_f-dimensional region rather than a point. The steepest descent algorithm returns the point on the boundary of that region that it hits when starting from its initial state. Although we always set the initial state to the same state $f_i = 0$ and so reproducibly obtain the same answer with every run, the fact that the output depends on this particular choice of initial state as well as on the vagaries of the subsequent trajectory means the final answer contains a degree of arbitrariness that increases with the false positive rate.

3.3 Performance of 3D-alignment algorithm on simulated data

Several checks on our numerics give us confidence in our results. The part of the program that generates random conformations is a reuse of the Monte Carlo method in the Wormulator, which was tested as explained in Chapter 2. The feed-forward step in our program was tested alone by comparing the half and full partition functions with those computed by direct enumeration of all contours (which is quite doable for a few loci), in the special case where all same-color pairs of loci are adjacent. This works because the no-overlap condition is enforced for neighboring loci. Finally, we verified that the numerical gradient of the cost function dC/df_i, where C is computed by the feedforward step of the algorithm, equals (to numerical precision) the analytical gradient computed by the backpropagation step.

For our numerical experiments we generated random conformations having the *in vitro* 50 nm persistence length (the true *in vivo* figure almost certainly depends on the organism, locus position, cell state, etc.). Our initial simulations generated short 10-kb contours that were decorated at 30 random loci using 3 label colors (10 loci per color). Both the false negative and false positive rates were assumed to be 10%. False positives were randomly interspersed within the box described by the minimum and maximum (x, y, z) of imaged spots along the contour. Localization

error, superimposed on the true position of each spot, was assumed Gaussian, having a standard deviation $10+(2/15)|z|$ nm in the xy-plane and $22+(1/15)|z|$ in the z-plane, where $|z|$ represents the distance to the focal plane (depending on the superresolution method out-of-focus spots are usually harder to localize).

In order to evaluate the quality of the p-values and the associated entropy measure, we generated 1000 10-kb conformations, having the parameters just described, iterated our algorithm, and stored the mapping probabilities for each run that converged within 100 iterations. Figure 3-1 compares the p-values generated from simulated conformations to the 'hit rates' corresponding to these p-values. Ideally these would be equal: for example a mapping variable $p = 0.4$ should have a roughly 40% likelihood of being correct. The S-shape in this graph indicates that our algorithm is in truth somewhat under-confident in driving mapping probabilities from the starting mean: low probabilities are not low enough, and high probabilities are not high enough. There are three possible causes of this systematic bias: 1) our use of the Gaussian chain model for the alignment which is different from the wormlike chain model for generating the contours; 2) the no-overlap approximation; and 3) the nonuniqueness of our alignment solution. One ought to be able to rescale the p-values to eliminate the bias, but doing so would introduce inconsistencies into the probability array and so we have not used rescaling in any of the results we will present here.

Since the program will output some probability table regardless of the quality of the data, it would be helpful for there to be some flag that an experimenter can use to know whether the data is indeed reliable or not. To this end we generated a set of controls in which the contour, the genomic position of every locus and color ordering along the contour were unchanged but in which the locus colors in the control images were scrambled. Since the expected color ordering of the spots along the contour no longer matched the ordering used in generating the images, we expected three things to happen in the control mappings. 1) The convergence rate of the cost function should drop, since easy mappings converge quickly. 2) The entropy should increase. A higher entropy corresponds to lower expected information recovery. (Ideally one would measure information recovery which indeed reliably dropped or went negative,

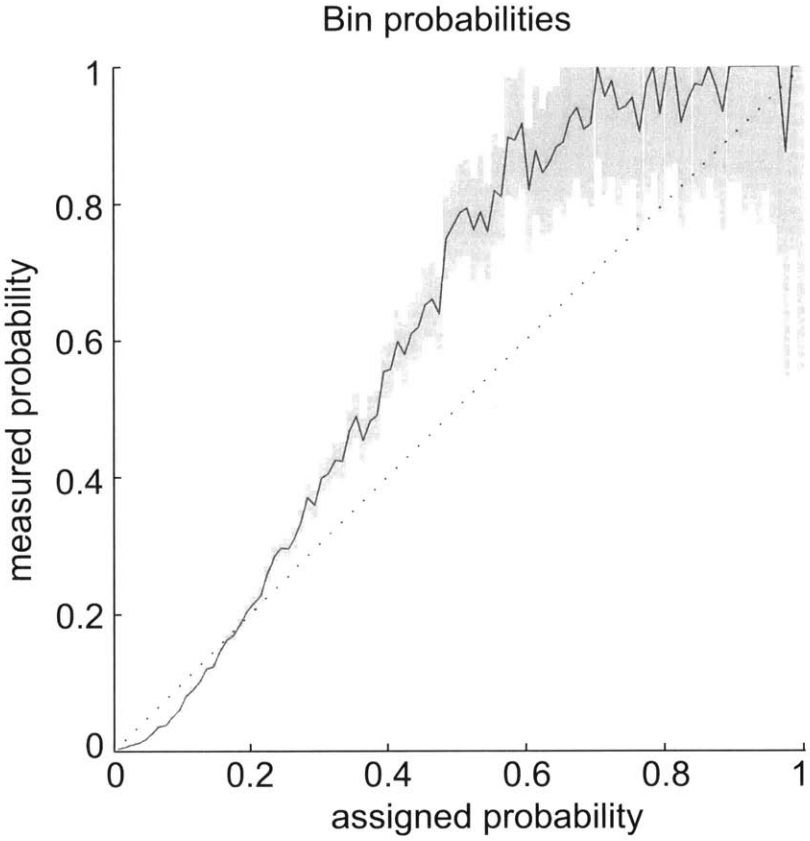


Figure 3-1: **Quality of mapping probabilities.** Mapping probabilities from converged mappings (x-axis) versus their empirical likelihood of being a correct mapping (y-axis). Shaded region indicates the region of uncertainty. 1000 10-kb conformations were attempted (3 colors, 10 spots per label); the mapping probabilities used here are taken from the attempted mappings which converged within 100 iterations of the algorithm. Localization error was Gaussian, of standard deviation $\sigma_x = \sigma_y = 10 + (2/15)|z|$ nm and $\sigma_z = 22 + (1/15)|z|$, where $|z|$ represents the distance to the focal plane. False-positive rate and false-negative rate were both 10%.

but that is not experimentally knowable). 3) The (logarithm of the) partition function should decrease. The reason is that the DNA should have to be overstretched at a high energetic cost in order to fit the more awkward arrangement of dots in the control mappings.

The comparison of normal to control mappings was performed using 5 data sets; each data set contained 100 conformations with the same experimental parameters used to bin the p-values (10 kb, 3 colors, 30 spots, experimental error as before). The five data sets tested were: a low-error set, a high microscope (MS) error set, a set with a high false negative (FN) rate, a high false positive (FP) rate set, and a set with a mixed FN/FP/MS error rate. For each normal mapping a set of 5 color-scrambled control mappings were performed. The performance of the true mappings relative to the controls is shown in Figure 3-2. The fact that we only plotted the data points for controls which converged shows immediately that (1) convergence is indeed markedly poorer on the controls. For example, when microscope error was high 3 out of 4 of the normal runs converged in 100 iterations, whereas only 11/100 runs had at least 2/5 controls converge. In fact, for two of the data sets either 0 or 1 of the conformations had two converged controls, so those were not even plotted. Secondly, the entropy (2) was generally lower in the real mappings than in the controls, although the presence of false negatives makes this metric less reliable. The least reliable metric was the logarithm of the partition function (3): $\log Z$ was indeed larger in the true mapping than the control mappings when the false negative rate was low, but for high false negatives the reverse was true. The reasons for this are unknown but may have been biased by the fact that we did not plot the poorly-converged controls.

We next used the short 30-spot problem for visualizing the mapping probabilities. Conformations was randomly generated as before ($l_p = 50$ nm, 3 colors, error parameters as in Figure 3-1), mapped using our algorithm, and the entropies before and after each mapping were recorded. We repeated this procedure until we found a mapping that would look good to an experimenter, where the entropy drop exceeded one bit per labeled locus, and that had a significant number of false negatives for demonstration purposes. This mapping, which came from the fifth overall conforma-

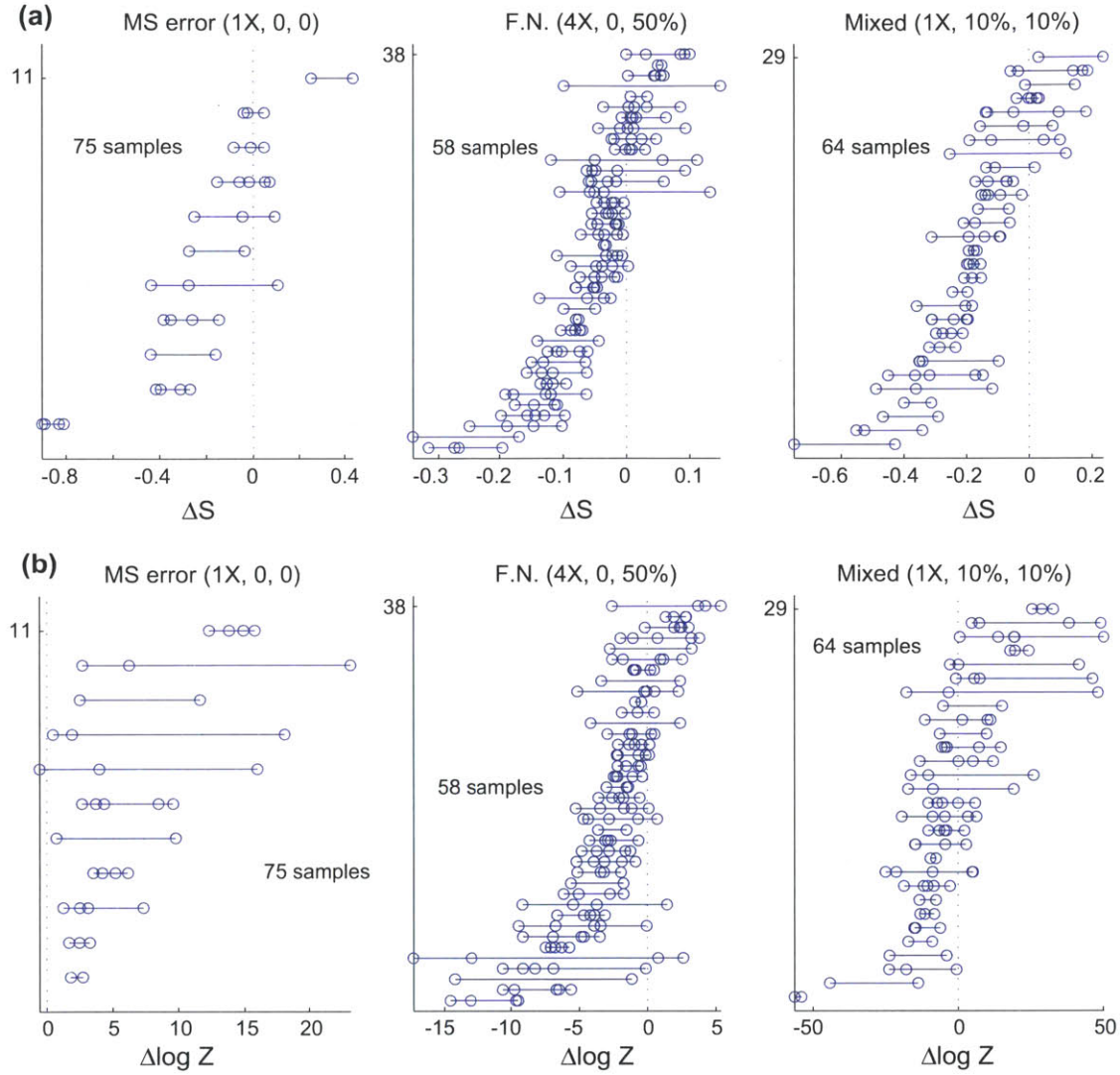


Figure 3-2: **Comparison with control mappings.** (a) The entropy of simulated contour mappings minus the entropy of color-permuted control mappings. 100 conformations were generated for each given set of experimental conditions; those conformations for which the cost function of the real mapping converged within 100 iterations formed the sample set of that experiment. For each sample, mappings were attempted for 5 different color-scrambled controls. The sample was plotted if at least 2 of the controls converged. The title gives the experimental conditions (localization error reduction factor, false positive rate, false negative rate). Between the two data sets having both low localization error and a low false negative rate, only a single sample had two controls converge, so those data sets have been omitted. (b) The logarithm of the partition function of the simulated contours minus that of the control mappings, using the same data sets as in (A). The $\log Z$ measure is seen to be unreliable unless the false negative rate is very low.

tion examined, is plotted in Figure 3-3 along with the associated DNA contour and colored labels.

To demonstrate how such an array of mapping probabilities might be used to construct a conformation, we imagined a DNA contour in which each locus is mapped to the imaged spot having the highest p-value for that locus; if the false negative p-value was the largest then that spot was skipped. This conformation is shown in Figure 3-4c. The error in this conformation is due to a combination of discretization error, localization error in the microscopy, and mapping error, as the progression in Figure 3-4 shows. This method of generating a conformation is rather crude, as it allows different loci to map to the same imaged spot.

The quality of a mapping depends partly on the three experimental error parameters: the false positive rate, false negative rate and localization error. We generated 100 conformations of 10-kb DNA and labeled them randomly in three colors as before, except that the error parameters were varied. Figure 3-5 shows the information recovery, entropy change and error parameters for each of the 89 runs for which the cost function C converged near to zero, along with the conformation of Figure 3-3 shown in green. For comparison, we also generated and analyzed ensembles of conformations for two-dimensional 10-kb contours having the same persistence length, along with 100 kb three-dimensional contours with the same labeling density (Figure 3-6). Encouragingly, we find that positive information is almost always recovered in cases of low experimental error, and that entropy is generally a reasonable proxy for information, although entropy change tends to underestimate information recovery for good mappings.

Mean computation time was about one second for each 10-kb mapping, and roughly 35 minutes for each 100-kb mapping. Only a few mappings converged to our tolerances before reaching the maximum number of iterations that we set (100 iterations for the 10-kb mappings, 400 iterations for the 100-kb mappings). The average computation time *per iteration* was therefore about 0.01 seconds for the 10-kb mapping and 5 seconds for the 100-kb mapping. The 5-fold difference between the ratio $\tau_{100}/\tau_{10} = 500$ and the ratio $N_{100}^2/N_{10}^2 = 100$ one might expect from the scaling

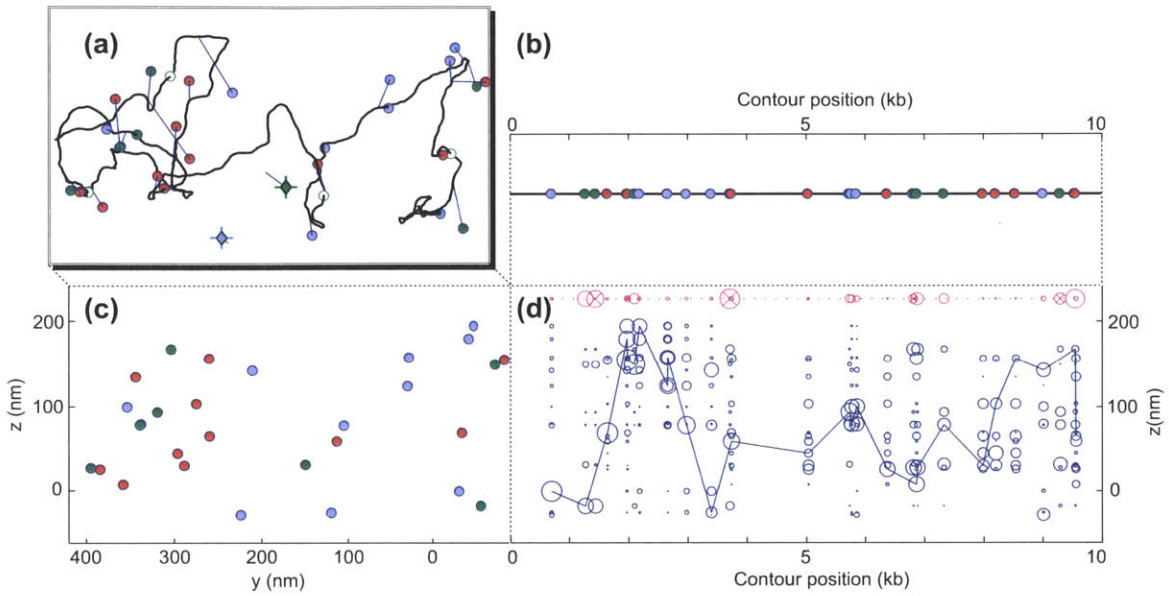


Figure 3-3: **Mapping of simulated 10 kb conformation.** (a) Simulated 3-dimensional 10-kb DNA contour (l_p = 50 nm) decorated with colored fluorophores using the error parameters described in the text. Localization error is indicated by the lines connecting the recorded positions of the spots to their true locations on the contour. False positives have a four-pointed star shape; open circles indicate false negatives. (b-c) The inputs to the mapping algorithm are the genomic (b) and imaged (c) positions of a set of labels along with their colors. (d) A graphical representation of the mapping probabilities output by the algorithm. Each element of the probability array, which maps the genomic locus at l base pairs to the spot imaged at (x, y, z), is represented by a circle centered at (l, z) and having an area proportional to the probability. False negative probabilities are given by the line of pink circles along the top. The solid line connects the correct mappings, skipping false negatives which are instead marked with X's inside their respective pink circles. The fact that the true conformation generally passes through the largest circle of each locus located between 0 and 8 kb indicates a good mapping within this region.

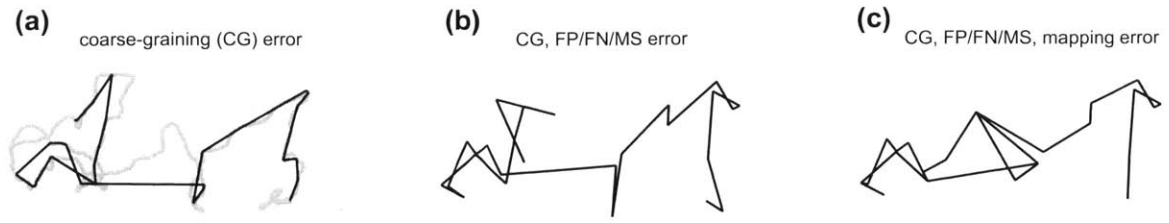


Figure 3-4: **Discrete conformations** (a) The coarse-grained approximation to the contour that connects labels at their true positions, superimposed upon the full conformation (shaded line). Mismatch regions can occur where consecutive labels are widely spaced (due the random label spacing), making it impossible to resolve the intervening contour. (b) The coarse-grained contour connecting the imaged spots in their proper order, taking experimental error into account. False negative errors cause some spots to be missed, and localization error offsets the remaining spots. (c) The coarse-grained contour that connects the imaged spots (with experimental error) based on the maximum computed mapping probability for each locus (note that this heuristic does not strictly enforce no-overlap).

rule $\tau \propto N^2\rho$ is partly due to the fact that the mean density ρ of neighboring spots increases with increasing contour length, since distal regions can sometimes overlap.

All calculations were performed using a custom program written in C and Yazoo, and compiled on a Macbook Pro. The gradient-optimization routines and Gaussian random number generator used the GNU Scientific Library version 1.14[42]. Source code and a compiled binary can be downloaded from the following URL:

`www.phys.washington.edu/~pwiggins/align3d/`.

3.4 Conclusions and Outlook

Any future DNA-mapping experiment that is both large-scale and high-resolution will involve hundreds or thousands of labels, far more than the number of labeling colors currently available. Using the algorithm we have demonstrated, one can reconstruct much of the conformational information from images of labeled DNA even when the number of labels far exceeds the number of labeling colors, if the spacing and color ordering of labels along the DNA strand is known to the experimenter. Our simulations demonstrated 10-kb and 100-kb mappings using three distinguishable

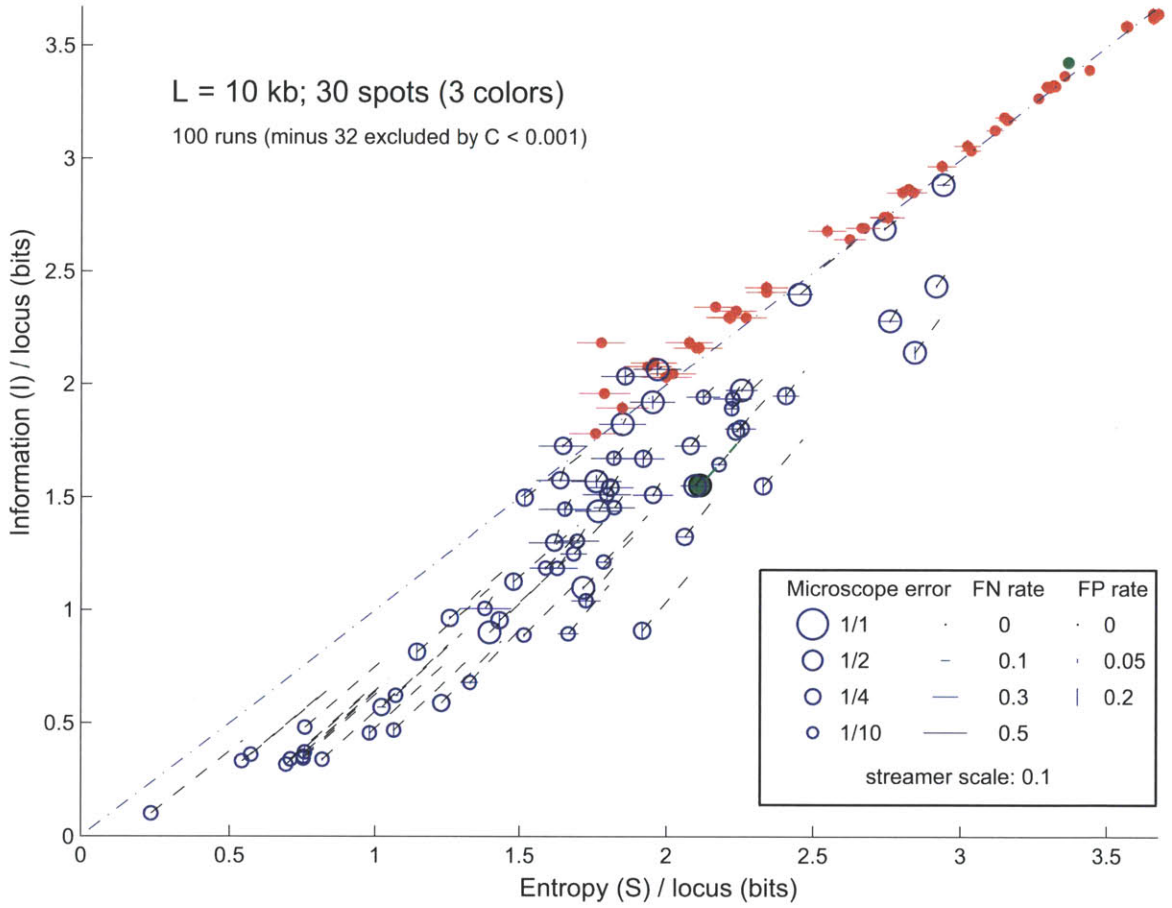


Figure 3-5: **Information recovery and experimental error of 3D 10 kb conformations.** A more detailed view of the 3D 10-kb mappings shown in Figure 3-6. Each dot is replaced by a symbol denoting the error parameters, and a 'streamer' points from each final state one tenth of the way back to its respective initial state. The microscope error factor m is a m-fold reduction in the localization error relative to the error $\sigma_x = \sigma_y = 10 + (2/15)|z|$ nm and $\sigma_z = 22 + (1/15)|z|$ used elsewhere in the text. The green dots give the initial and final state of the run from Figure 3-3.

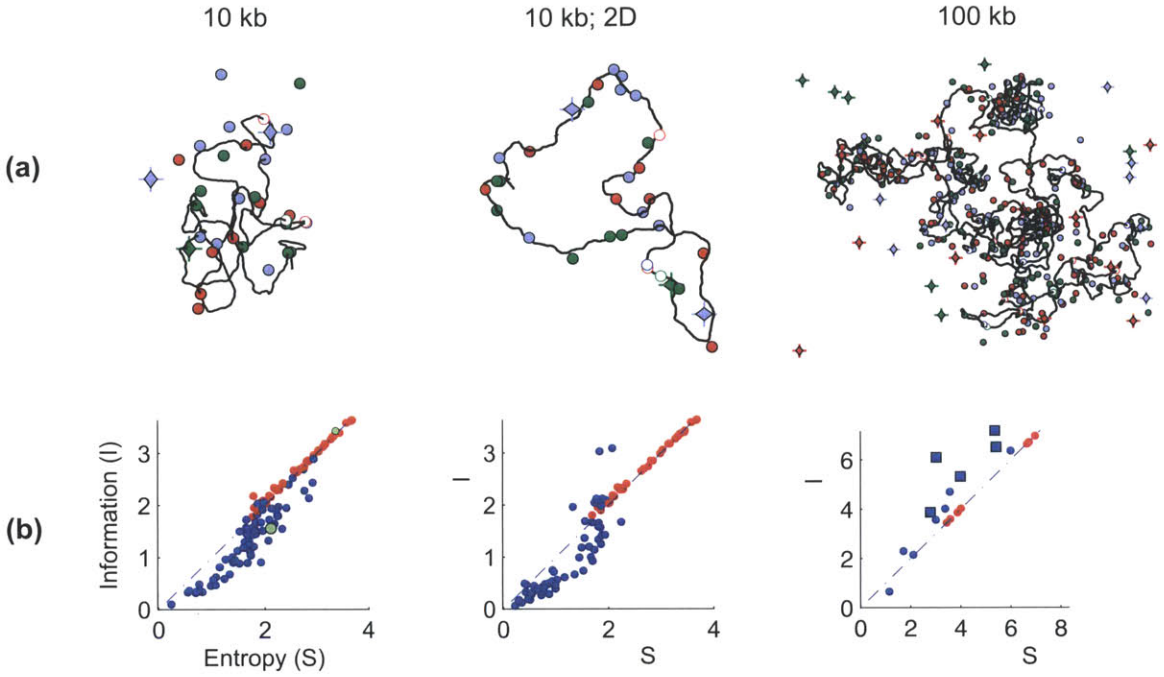


Figure 3-6: **Information recovery from simulated conformations.** Three sets of simulated DNA conformations were generated: 3D and 2D 10-kb contours, and 3D 100-kb contours. Three labeling colors were used, at a labeling density of approximately one label per 1000 base pairs per color; however the experimental error parameters were varied randomly with each conformation. (a) Sample contours and labelings for each set. False positives, false negatives and microscope error are all shown. (b) Change in entropy (S) and information scores (I) over the course of the mapping procedure, from an initial state of uniform probabilities that enforce the false negative rate (red) to the final state (blue). There is one red and blue dot for each mapping in a set that converged after 100 (400) iterations for the 10 kb (100 kb) mappings. 100 conformations were generated per set. The 100 kb set converged considerably slower, so the tolerances were relaxed; squares have looser tolerances than circles and show poorer information recovery. The upper/lower green dots in the 3D 10-kb plot respectively give the initial/final states of the run from Figure 3-3.

colors, which is the current state-of-the-art in multicolor superresolution microscopy.

We wish to point out that our use of the word 'color' refers only to distinguishability, and that there are ways of distinguishing groups of labels that do not require the use of spectrally distinct fluorophores. For example, a FISH experiment might introduce and image several groups of probes separately from one another, and if the experimenter knows which group each probe is in then probes in different groups effectively have different colors even if they use the same fluorophore. An increase in the number of effective colors not only improves the mapping quality but also reduces the memory requirements and speeds the analysis, by truncating the various color-restricted sums over pairs of neighboring spots, and (we expect) by speeding the convergence.

It should be possible to lower the experimental tolerances by improving the mapping algorithm. One obvious potential improvement is to enforce the no-overlap constraint between certain pairs of non-adjacent loci. By targeting additional no-overlap constraints to those pairs of loci that cause the most confusion in the probability matrix, it may be possible to improve the information recovery significantly without undue computational overhead. One might also be able to improve the mappings by adjusting the labeling densities of each color.

One final difficulty is in the interpretation of the probability matrix, which is not a conformation and from which there is not even an obvious way to get an optimal conformation. Indeed, our approach was guided by a belief that a conformation by itself would not be a very helpful output of the analysis, since one should also like to know about the uncertainty and the range of different conformations. There is thus a need to develop reductions of the probability matrix that are more directly interpretable than, but nearly as informative as, the probability matrix itself. Such a reduced output might be helpful in comparing the error due to the experimental steps to that from the mapping procedure (see Figure 3-4). For example, if the reduced output is a weighted ensemble of conformations, then one can compare the (expected) RMS deviations of the contour with and without discretization error, experimental error and mapping error. The error metric might even be used to provide further

constraints on the non-unique solution space of our algorithm. We expect that these post-analysis tools will be easier to develop and much less computationally intensive than the 3d-alignment algorithm described here, and we hope that they will make our conformational analysis accessible to the general scientific public.

Bibliography

[1] A. Akhtar and S.M. Gasser. The nuclear envelope and transcriptional control. *Nature Reviews Genetics*, 8(7):507–517, 2007.

[2] R. Ando, C. Flors, H. Mizuno, J. Hofkens, and A. Miyawaki. Highlighted generation of fluorescence signals using simultaneous two-color irradiation on dronpa mutants. *Biophysical journal*, 92(12):L97–L99, 2007.

[3] M. Andresen, A.C. Stiel, J. Fölling, D. Wenzel, A. Schönle, A. Egner, C. Eggeling, S.W. Hell, and S. Jakobs. Photoswitchable fluorescent proteins enable monochromatic multilabel imaging and dual color fluorescence nanoscopy. *Nature biotechnology*, 26(9):1035–1040, 2008.

[4] M. Bates, T.R. Blosser, and X. Zhuang. Short-range spectroscopic ruler based on a single-molecule optical switch. *Physical review letters*, 94(10):108101, 2005.

[5] M. Bates, B. Huang, G.T. Dempsey, and X. Zhuang. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science*, 317(5845):1749, 2007.

[6] C.G. Baumann, S.B. Smith, V.A. Bloomfield, and C. Bustamante. Ionic effects on the elasticity of single dna molecules. *Proceedings of the National Academy of Sciences*, 94(12):6185, 1997.

[7] S.B. Baylin and K.E. Schuebel. Genomic biology: the epigenomic era opens. *Nature*, 448(7153):548–549, 2007.

[8] J. Bednar, P. Furrer, V. Katritch, A. Stasiak, J. Dubochet, and A. Stasiak. Determination of dna persistence length by cryo-electron microscopy. separation of the static and dynamic contributions to the apparent persistence length of dna. *Journal of molecular biology*, 254(4):579–594, 1995.

[9] A.J. Bendich. The form of chromosomal dna molecules in bacterial cells. *Biochimie*, 83(2):177–186, 2001.

[10] E. Betzig, RJ Chichester, F. Lanni, and DL Taylor. Near-field fluorescence imaging of cytoskeletal actin. *Bioimaging*, 1(3):129–135, 1993.

[11] E. Betzig, G.H. Patterson, R. Sougrat, O.W. Lindwasser, S. Olenych, J.S. Bonifacino, M.W. Davidson, J. Lippincott-Schwartz, and H.F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642, 2006.

[12] D.L. Beveridge. Molecular dynamics: Dna. *Encyclopedia of Computational Chemistry*, 2002.

[13] J.S. Biteen, M.A. Thompson, N.K. Tselentis, G.R. Bowman, L. Shapiro, and WE Moerner. Super-resolution imaging in live caulobacter crescentus cells using photoswitchable eyfp. *Nature methods*, 5(11):947–949, 2008.

[14] M. Bossi, J. Fölling, V.N. Belov, V.P. Boyarskiy, R. Medda, A. Egner, C. Eggeling, A. Schönle, and S.W. Hell. Multicolor far-field fluorescence nanoscopy through isolated detection of distinct molecular species. *Nano letters*, 8(8):2463–2468, 2008.

[15] C. Bouchiat and M. Mezard. Elastic rod model of a supercoiled dna molecule. *The European Physical Journal E: Soft Matter and Biological Physics*, 2(4):377–402, 2000.

[16] D.T. Burnette, P. Sengupta, Y. Dai, J. Lippincott-Schwartz, and B. Kachar. Bleaching/blinking assisted localization microscopy for superresolution imaging using standard fluorescent molecules. *Proceedings of the National Academy of Sciences*, 108(52):21081–21086, 2011.

[17] Y. Burnier, J. Dorier, and A. Stasiak. Dna supercoiling inhibits dna knotting. *Nucleic acids research*, 36(15):4956–4963, 2008.

[18] M. Chalfie, Y. Tu, G. Euskirchen, W.W. Ward, and D.C. Prasher. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802, 1994.

[19] TE Cheatham III and PA Kollman. Observation of the a-dna to b-dna transition during unrestrained molecular dynamics in aqueous solution. *Journal of molecular biology*, 259(3):434–444, 1996.

[20] H. Chen and J. Yan. Effects of kink and flexible hinge defects on mechanical responses of short double-stranded dna molecules. *Physical Review E*, 77(4):041907, 2008.

[21] B. Choi, G. Zocchi, Y. Wu, S. Chan, and L. Jeanne Perry. Allosteric control through mechanical tension. *Physical review letters*, 95(7):78102, 2005.

[22] D.M. Chudakov, V.V. Belousov, A.G. Zaraisky, V.V. Novoselov, D.B. Staroverov, D.B. Zorov, S. Lukyanov, and K.A. Lukyanov. Kindling fluorescent proteins for precise in vivo photolabeling. *Nature biotechnology*, 21(2):191–194, 2003.

[23] D.M. Chudakov, V.V. Verkhusha, D.B. Staroverov, E.A. Souslova, S. Lukyanov, and K.A. Lukyanov. Photoswitchable cyan fluorescent protein for protein tracking. *Nature biotechnology*, 22(11):1435–1439, 2004.

[24] T.E. Cloutier and J. Widom. Spontaneous sharp bending of double-stranded dna. *Molecular cell*, 14(3):355–362, 2004.

[25] P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J.L. Viovy, D. Chatenay, and F. Caron. Dna: an extensible molecule. *Science*, 271(5250):792, 1996.

[26] HG Davies, JV Small, et al. Structural units in chromatin and their orientation on membranes. *Nature*, 217(5134):1122, 1968.

[27] R.T. DeBoy and N.L. Craig. Tn7 transposition as a probe of cis interactions between widely separated (190 kilobases apart) dna sites in the escherichia coli chromosome. *Journal of bacteriology*, 178(21):6184, 1996.

[28] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306, 2002.

[29] DE Depew and J.C. Wang. Conformational fluctuations of dna helix. *Proceedings of the National Academy of Sciences*, 72(11):4275, 1975.

[30] T. Dertinger, R. Colyer, G. Iyer, S. Weiss, and J. Enderlein. Fast, background-free, 3d super-resolution optical fluctuation imaging (sofi). *Proceedings of the National Academy of Sciences*, 106(52):22287, 2009.

[31] RE Dickerson, M. Bansal, and C.R. Calladine. Definitions and nomenclature of nucleic acid structure parameters. *The EMBO Journal*, 8:1–4, 1989.

[32] J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 2012.

[33] M. Doi and S.F. Edwards. *The theory of polymer dynamics*, volume 73. Oxford University Press, USA, 1988.

[34] J. Dostie, T.A. Richmond, R.A. Arnaout, R.R. Selzer, W.L. Lee, T.A. Honan, E.D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.

[35] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii. Cyclization of short dna fragments and bending fluctuations of the double helix. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5397, 2005.

[36] MA El Hassan, CR Calladine, et al. The assessment of the geometry of dinucleotide steps in double-helical dna; a new local calculation scheme. *Journal of molecular biology*, 251(5):648–664, 1995.

[37] M. Eltsov, K.M. MacLellan, K. Maeshima, A.S. Frangakis, and J. Dubochet. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proceedings of the National Academy of Sciences*, 105(50):19732, 2008.

[38] M. Eltsov and B. Zuber. Transmission electron microscopy of the bacterial nucleoid. *Journal of structural biology*, 156(2):246–254, 2006.

[39] J. Fölling, M. Bossi, H. Bock, R. Medda, C.A. Wurm, B. Hein, S. Jakobs, C. Eggeling, and S.W. Hell. Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nature methods*, 5(11):943–945, 2008.

[40] C. Frontali, E. Dore, A. Ferrauto, E. Gratton, A. Bettini, MR Pozzan, and E. Valdevit. An absolute method for the determination of the persistence length of native dna from electron micrographs. *Biopolymers*, 18(6):1353–1373, 1979.

[41] M.J. Fullwood, M.H. Liu, Y.F. Pan, J. Liu, H. Xu, Y.B. Mohamed, Y.L. Orlov, S. Velkov, A. Ho, P.H. Mei, et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.

[42] Mark Galassi et al. *GNU Scientific Library Reference Manual.* Network Theory Ltd, 3rd edition, January 2009.

[43] N. Garcia-Russell, T.G. Harmon, T.Q. Le, N.H. Amaladas, R.D. Mathewson, and A.M. Segall. Unequal access of chromosomal regions to each other in salmonella: probing chromosome structure with phage λ integrase-mediated long-range rearrangements. *Molecular microbiology*, 52(2):329–344, 2004.

[44] G.S. Gordon, D. Sitnikov, C.D. Webb, A. Teleman, A. Straight, R. Losick, A.W. Murray, and A. Wright. Chromosome and low copy plasmid segregation in e. coli: visual evidence for distinct mechanisms. *Cell*, 90(6):1113–1121, 1997.

[45] S.I.S. Grewal and D. Moazed. Heterochromatin and epigenetic control of gene expression. *Science*, 301(5634):798, 2003.

[46] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M.B. Faza, W. Talhout, B.H. Eussen, A. de Klein, L. Wessels, W. de Laat, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, 2008.

[47] J. Guo, S. Wang, N. Dai, Y.N. Teo, and E.T. Kool. Multispectral labeling of antibodies with polyfluorophores on a dna backbone and application in cellular imaging. *Proceedings of the National Academy of Sciences*, 108(9):3493, 2011.

[48] M.G.L. Gustafsson. Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13081, 2005.

[49] M.G.L. Gustafsson, D.A. Agard, and J.W. Sedat. Sevenfold improvement of axial resolution in 3d wide-field microscopy using two objective lenses. In *Proceedings of SPIE*, volume 2412, page 147, 1995.

[50] MGL Gustafsson, DA Agard, JW Sedat, et al. I5m: 3d widefield light microscopy with better than 100nm axial resolution. *Journal of microscopy*, 195(1):10–16, 1999.

[51] S. Hadjur, L.M. Williams, N.K. Ryan, B.S. Cobb, T. Sexton, P. Fraser, A.G. Fisher, and M. Merkenschlager. Cohesins form chromosomal cis-interactions at the developmentally regulated ifng locus. *Nature*, 460(7253):410–413, 2009.

[52] P.J. Hagerman. Flexibility of dna. *Annual review of biophysics and biophysical chemistry*, 17(1):265–286, 1988.

[53] M. Heilemann, P. Dedecker, J. Hofkens, and M. Sauer. Photoswitches: Key molecules for subdiffraction-resolution fluorescence imaging and molecular quantification. *Laser & Photonics Reviews*, 3(1-2):180–202, 2009.

[54] M. Heilemann, E. Margeat, R. Kasper, M. Sauer, and P. Tinnefeld. Carbocyanine dyes as efficient reversible single-molecule optical switch. *Journal of the American Chemical Society*, 127(11):3801–3806, 2005.

[55] M. Heilemann, S. van de Linde, M. Schüttpelz, R. Kasper, B. Seefeldt, A. Mukherjee, P. Tinnefeld, and M. Sauer. Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angewandte Chemie International Edition*, 47(33):6172–6176, 2008.

[56] R. Heim, D.C. Prasher, and R.Y. Tsien. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proceedings of the National Academy of Sciences*, 91(26):12501, 1994.

[57] S.W. Hell. Strategy for far-field optical imaging and writing without diffraction limit. *Physics Letters A*, 326(1-2):140–145, 2004.

[58] S.W. Hell. Microscopy and its focal switch. *Nature Methods*, 6(1):24–32, 2008.

[59] SW Hell and M. Kroug. Ground-state-depletion fluorscence microscopy: A concept for breaking the diffraction resolution limit. *Applied Physics B: Lasers and Optics*, 60(5):495–497, 1995.

[60] S.W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics letters*, 19(11):780–782, 1994.

[61] S.T. Hess, T.P.K. Girirajan, and M.D. Mason. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical journal*, 91(11):4258–4272, 2006.

[62] N.P. Higgins, X. Yang, Q. Fu, and J.R. Roth. Surveying a supercoil domain by using the gamma delta resolution system in salmonella typhimurium. *Journal of bacteriology*, 178(10):2825, 1996.

[63] E. Hinde, F. Cardarelli, M.A. Digman, and E. Gratton. Changes in chromatin compaction during the cell cycle revealed by micrometer-scale measurement of molecular flow in the nucleus. *Biophysical Journal*, 102(3):691–697, 2012.

[64] T. Hirano. At the heart of the chromosome: Smc proteins in action. *Nature Reviews Molecular Cell Biology*, 7(5):311–322, 2006.

[65] M. Hogan, J. LeGrange, and B. Austin. Dependence of dna helix flexibility on base composition. *Nature*, 1983.

[66] S.J. Holden, S. Uphoff, and A.N. Kapanidis. Daostorm: an algorithm for high-density super-resolution microscopy. *nature methods*, 8(4):279–280, 2011.

[67] B. Huang, W. Wang, M. Bates, and X. Zhuang. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science*, 319(5864):810, 2008.

[68] F.J. Iborra, A. Pombo, D.A. Jackson, and P.R. Cook. Active rna polymerases are localized within discrete transcription "factories' in human nuclei. *Journal of cell science*, 109(6):1427–1436, 1996.

[69] H. Jacobson and W.H. Stockmayer. Intramolecular reaction in polycondensations. i. the theory of linear systems. *The Journal of Chemical Physics*, 18:1600, 1950.

[70] S. Jhunjhunwala, M.C. van Zelm, M.M. Peak, and C. Murre. Chromatin architecture and the generation of antigen receptor diversity. *Cell*, 138(3):435–448, 2009.

[71] M.H. Kagey, J.J. Newman, S. Bilodeau, Y. Zhan, D.A. Orlando, N.L. van Berkum, C.C. Ebmeier, J. Goossens, P.B. Rahl, S.S. Levine, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.

[72] R. Kavenoff and B.C. Bowen. Electron microscopy of membrane-free folded chromosomes from escherichia coli. *Chromosoma*, 59(2):89–101, 1976.

[73] T.A. Klar, S. Jakobs, M. Dyba, A. Egner, and S.W. Hell. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proceedings of the National Academy of Sciences*, 97(15):8206, 2000.

[74] R.D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, review fundamental particle of the eukaryote chromosome. *Cell*, 98:285–294, 1999.

[75] O. Kratky and G. Porod. Röntgenuntersuchung gelöster fadenmoleküle. *Recueil des Travaux Chimiques des Pays-Bas*, 68(12):1106–1122, 1949.

[76] Y.A. Labas, NG Gurskaya, Y.G. Yanushevich, AF Fradkov, KA Lukyanov, SA Lukyanov, and MV Matz. Diversity and evolution of the green fluorescent protein family. *Proceedings of the National Academy of Sciences*, 99(7):4256, 2002.

[77] F. Lankas, J. Sponer, J. Langowski, and T.E. Cheatham III. Dna basepair step deformability inferred from molecular dynamics simulations. *Biophysical journal*, 85(5):2872–2883, 2003.

[78] A. Lewis, M. Isaacson, A. Harootunian, and A. Muray. Development of a 500 spatial resolution light microscope:: I. light is efficiently transmitted through [lambda]/16 diameter apertures. *Ultramicroscopy*, 13(3):227–231, 1984.

[79] G. Li, X. Ruan, R.K. Auerbach, K.S. Sandhu, M. Zheng, P. Wang, H.M. Poh, Y. Goh, J. Lim, J. Zhang, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98, 2012.

[80] Y. Li, K. Sergueev, and S. Austin. The segregation of the escherichia coli origin and terminus of replication. *Molecular microbiology*, 46(4):985–996, 2002.

[81] E.A. Libby, M. Roggiani, and M. Goulian. Membrane protein expression triggers chromosomal locus repositioning in bacteria. *Proceedings of the National Academy of Sciences*, 109(19):7445–7450, 2012.

[82] 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, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289, 2009.

[83] M.S. Luijsterburg, M.C. Noom, G.J.L. Wuite, and R.T. Dame. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *Journal of structural biology*, 156(2):262–272, 2006.

[84] K.A. Lukyanov, D.M. Chudakov, S. Lukyanov, and V.V. Verkhusha. Photoactivatable fluorescent proteins. *Nature Reviews Molecular Cell Biology*, 6(11):885–890, 2005.

[85] J.F. Marko and E.D. Siggia. Stretching dna. *Macromolecules*, 28(26):8759–8770, 1995.

[86] S.A. McKinney, C.S. Murphy, K.L. Hazelwood, M.W. Davidson, and L.L. Looger. A bright and photostable photoconvertible fluorescent protein. *Nature methods*, 6(2):131–133, 2009.

[87] K.J. Meaburn and T. Misteli. Cell biology: chromosome territories. *Nature*, 445(7126):379–381, 2007.

[88] S. Mehraeen, B. Sudhanshu, E.F. Koslover, and A.J. Spakowitz. End-to-end distribution for a wormlike chain in arbitrary dimensions. *Physical Review E*, 77(6):061803, 2008.

[89] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087, 1953.

[90] J.D. Moroz and P. Nelson. Torsional directed walks, entropic elasticity, and dna twist stiffness. *Proceedings of the National Academy of Sciences*, 94(26):14418, 1997.

[91] E.A. Mukamel, H. Babcock, and X. Zhuang. Statistical deconvolution for superresolution fluorescence microscopy. *Biophysical Journal*, 102(10):2391–2400, 2012.

[92] MC Murphy, I. Rasnik, W. Cheng, T.M. Lohman, and T. Ha. Probing single-stranded dna conformational flexibility using fluorescence spectroscopy. *Biophysical journal*, 86(4):2530–2537, 2004.

[93] S. Nagai, K. Dubrana, M. Tsai-Pflugfelder, M.B. Davidson, T.M. Roberts, G.W. Brown, E. Varela, F. Hediger, S.M. Gasser, and N.J. Krogan. Functional targeting of dna damage to a nuclear pore-associated sumo-dependent ubiquitin ligase. *Science*, 322(5901):597, 2008.

[94] T. Nagai, K. Ibata, E.S. Park, M. Kubota, K. Mikoshiba, and A. Miyawaki. A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature biotechnology*, 20(1):87–90, 2002.

[95] H. Niki and S. Hiraga. Subcellular distribution of actively partitioning f plasmid during the cell division cycle in e. coli. *Cell*, 90(5):951–957, 1997.

[96] H. Niki, Y. Yamaichi, and S. Hiraga. Dynamic organization of chromosomal dna in escherichia coli. *Genes & development*, 14(2):212, 2000.

[97] I.K. Nolis, D.J. McKay, E. Mantouvalou, S. Lomvardas, M. Merika, and D. Thanos. Transcription factors mediate long-range enhancer–promoter interactions. *Proceedings of the National Academy of Sciences*, 106(48):20222, 2009.

[98] E.P. Nora, B.R. Lajoie, E.G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N.L. van Berkum, J. Meisig, J. Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 2012.

[99] A. Noy, A. Perez, F. Lankas, F. Javier Luque, and M. Orozco. Relative flexibility of dna and rna: a molecular dynamics study. *Journal of molecular biology*, 343(3):627–638, 2004.

[100] T. Odijk. Stiff chains and filaments under tension. *Macromolecules*, 28(20):7016–7018, 1995.

[101] W.K. Olson, M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey, U. Heinemann, X.J. Lu, S. Neidle, Z. Shakked, et al. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of molecular biology*, 313(1):229–237, 2001.

[102] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin. Dna sequence-dependent deformability deduced from protein–dna crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19):11163, 1998.

[103] C.S. Osborne, L. Chakalova, K.E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J.A. Mitchell, S. Lopes, W. Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065–1071, 2004.

[104] J.S. Paige, K.Y. Wu, and S.R. Jaffrey. Rna mimics of green fluorescent protein. *Science*, 333(6042):642, 2011.

[105] N. Panchuk-Voloshina, R.P. Haugland, J. Bishop-Stewart, M.K. Bhalgat, P.J. Millard, F. Mao, W.Y. Leung, and R.P. Haugland. Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *Journal of Histochemistry & Cytochemistry*, 47(9):1179, 1999.

[106] G.H. Patterson and J. Lippincott-Schwartz. A photoactivatable gfp for selective photolabeling of proteins and cells. *Science*, 297(5588):1873, 2002.

[107] S.R.P. Pavani, M.A. Thompson, J.S. Biteen, S.J. Lord, N. Liu, R.J. Twieg, R. Piestun, and WE Moerner. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences*, 106(9):2995, 2009.

[108] A. Pérez, F. Lankas, F.J. Luque, and M. Orozco. Towards a molecular dynamics consensus view of b-dna flexibility. *Nucleic acids research*, 36(7):2379–2394, 2008.

[109] A. Pertsinidis, Y. Zhang, and S. Chu. Subnanometre single-molecule localization, registration and distance measurements. *Nature*, 466(7306):647–651, 2010.

[110] Y.O. Popov and A.V. Tkachenko. Effects of kinks on dna elasticity. *Physical Review E*, 71(5):051905, 2005.

[111] D. Porschke. Persistence length and bending dynamics of dna from electrooptical measurements at high salt concentrations. *Biophysical chemistry*, 40(2):169–179, 1991.

[112] L. Postow, C.D. Hardy, J. Arsuaga, and N.R. Cozzarelli. Topological domain structure of the escherichia coli chromosome. *Genes & development*, 18(14):1766, 2004.

[113] H. Qu, C.Y. Tseng, Y. Wang, A.J. Levine, and G. Zocchi. The elastic energy of sharply bent nicked dna. *EPL (Europhysics Letters)*, 90:18003, 2010.

[114] S. Quirin, S.R.P. Pavani, and R. Piestun. Optimal 3d single-molecule localization for superresolution microscopy with aberrations and engineered point spread functions. *Proceedings of the National Academy of Sciences*, 109(3):675–679, 2012.

[115] P. Ranjith, P.B.S. Kumar, and G.I. Menon. Distribution functions, loop formation probabilities, and force-extension relations in a model for short double-stranded dna molecules. *Physical review letters*, 94(13):138102, 2005.

[116] R. Reyes-Lamothe, C. Possoz, O. Danilova, and D.J. Sherratt. Independent positioning and action of *Escherichia coli* replisomes in live cells. *Cell*, 133(1):90–102, 2008.

[117] M. Rief, H. Clausen-Schaumann, H.E. Gaub, et al. Sequence-dependent mechanics of single dna molecules. *nature structural biology*, 6:346–350, 1999.

[118] L. Ringrose, S. Chabanis, P.O. Angrand, C. Woodroofe, and A.F. Stewart. Quantitative comparison of dna looping in vitro and in vivo: chromatin increases effective dna flexibility at short distances. *The EMBO journal*, 18(23):6630–6641, 1999.

[119] C.C. Robinett, A. Straight, G. Li, C. Willhelm, G. Sudlow, A. Murray, and A.S. Belmont. In vivo localization of dna sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *The Journal of cell biology*, 135(6):1685, 1996.

[120] P.J.J. Robinson, L. Fairall, V.A.T. Huynh, and D. Rhodes. Em measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *PNAS*, 103(17):6506–6511, 2006.

[121] B. Ross and P. Wiggins. Measuring chromosome conformation with degenerate labels (submitted). *Physical Review E*, 2012.

[122] D.E. Rumelhart, G.E. Hintont, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[123] M.J. Rust, M. Bates, and X. Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature methods*, 3(10):793–796, 2006.

[124] J.P. Schouten, C.J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic acids research*, 30(12):e57–e57, 2002.

[125] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A.C. Thåström, Y. Field, I.K. Moore, J.P.Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[126] N.C. Shaner, R.E. Campbell, P.A. Steinbach, B.N.G. Giepmans, A.E. Palmer, and R.Y. Tsien. Improved monomeric red, orange and yellow fluorescent proteins derived from discosoma sp. red fluorescent protein. *Nature biotechnology*, 22(12):1567–1572, 2004.

[127] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[128] A. Sharonov and R.M. Hochstrasser. Wide-field subdiffraction imaging by accumulated binding of diffusing probes. *Proceedings of the National Academy of Sciences*, 103(50):18911, 2006.

[129] J. Shimada and H. Yamakawa. Ring-closure probabilities for twisted wormlike chains. application to dna. *Macromolecules*, 17(4):689–698, 1984.

[130] D. Shore and R.L. Baldwin. Energetics of dna twisting*:: I. relation between twist and cyclization probability. *Journal of molecular biology*, 170(4):957–981, 1983.

[131] D. Shore, J. Langowski, and R.L. Baldwin. Dna flexibility studied by covalent closure of short fragments into circles. *Proceedings of the National Academy of Sciences*, 78(8):4833, 1981.

[132] H. Shroff, C.G. Galbraith, J.A. Galbraith, H. White, J. Gillette, S. Olenych, M.W. Davidson, and E. Betzig. Dual-color superresolution imaging of genetically expressed probes within individual adhesion complexes. *Proceedings of the National Academy of Sciences*, 104(51):20308, 2007.

[133] H. Shroff, B.M. Reinhard, M. Siu, H. Agarwal, A. Spakowitz, and J. Liphardt. Biocompatible force sensor with optical readout and dimensions of 6 nm3. *Nano letters*, 5(7):1509–1514, 2005.

[134] H. Shroff, D. Sivak, J.J. Siegel, AL McEvoy, M. Siu, A. Spakowitz, P.L. Geissler, and J. Liphardt. Optical measurement of mechanical forces inside short dna loops. *Biophysical journal*, 94(6):2179–2186, 2008.

[135] G. Shtengel, J.A. Galbraith, C.G. Galbraith, J. Lippincott-Schwartz, J.M. Gillette, S. Manley, R. Sougrat, C.M. Waterman, P. Kanchanawong, M.W. Davidson, et al. Interferometric fluorescent super-resolution microscopy resolves 3d cellular ultrastructure. *Proceedings of the National Academy of Sciences*, 106(9):3125, 2009.

[136] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. De Wit, B. Van Steensel, and W. De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348–1354, 2006.

[137] C. Sousa, V. de Lorenzo, and A. Cebolla. Modulation of gene expression through chromosomal positioning in escherichia coli. *Microbiology*, 143(6):2071, 1997.

[138] AJ Spakowitz. Wormlike chain statistics with twist and fixed ends. *EPL (Europhysics Letters)*, 73:684, 2006.

[139] A.J. Spakowitz and Z.G. Wang. Exact results for a semiflexible polymer chain in an aligning field. *Macromolecules*, 37(15):5814–5823, 2004.

[140] A.J. Spakowitz and Z.G. Wang. End-to-end distance vector distribution with fixed end orientations for the wormlike chain model. *Physical Review E*, 72(4):041802, 2005.

[141] C. Steinhauer, C. Forthmann, J. Vogelsang, and P. Tinnefeld. Superresolution microscopy on the basis of engineered dark states. *Journal of the American Chemical Society*, 130(50):16840–16841, 2008.

[142] A.F. Straight, A.S. Belmont, C.C. Robinett, and A.W. Murray. Gfp tagging of budding yeast chromosomes reveals that protein-protein interactions can mediate sister chromatid cohesion. *Current Biology*, 6(12):1599–1608, 1996.

[143] TR Strick, J.F. Allemand, D. Bensimon, A. Bensimon, and V. Croquette. The elasticity of a single supercoiled dna molecule. *Science*, 271(5257):1835, 1996.

[144] F.V. Subach, G.H. Patterson, S. Manley, J.M. Gillette, J. Lippincott-Schwartz, and V.V. Verkhusha. Photoactivatable mcherry for high-resolution two-color fluorescence microscopy. *Nature methods*, 6(2):153–159, 2009.

[145] J. Tang, J. Akerboom, A. Vaziri, L.L. Looger, and C.V. Shank. Near-isotropic 3d optical nanoscopy with photon-limited chromophores. *Proceedings of the National Academy of Sciences*, 107(22):10068, 2010.

[146] B. Tinland, A. Pluen, J. Sturm, and G. Weill. Persistence length of single-stranded dna. *Macromolecules*, 30(19):5763–5765, 1997.

[147] M.M. Tirado and J.G. de La Torre. Rotational dynamics of rigid, symmetric top macromolecules. application to circular cylinders. *The Journal of Chemical Physics*, 73:1986, 1980.

[148] C.Y. Tseng, A. Wang, G. Zocchi, B. Rolih, and A.J. Levine. Elastic energy of protein-dna chimeras. *Physical Review E*, 80(6):061912, 2009.

[149] M. Valens, S. Penaud, M. Rossignol, F. Cornet, and F. Boccard. Macrodomain organization of the escherichia coli chromosome. *The EMBO journal*, 23(21):4330–4341, 2004.

[150] A. Vaziri, J. Tang, H. Shroff, and C.V. Shank. Multilayer three-dimensional super resolution imaging of thick biological samples. *Proceedings of the National Academy of Sciences*, 105(51):20221, 2008.

[151] V.V. Verkhusha and A. Sorkin. Conversion of the monomeric red fluorescent protein into a photoactivatable probe. *Chemistry & biology*, 12(3):279–285, 2005.

[152] P.H. Viollier, M. Thanbichler, P.T. McGrath, L. West, M. Meewan, H.H. McAdams, and L. Shapiro. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial dna replication. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9257, 2004.

[153] A. Wang and G. Zocchi. Elastic energy driven polymerization. *Biophysical journal*, 96(6):2344–2352, 2009.

[154] J.R. Wenner, M.C. Williams, I. Rouzina, and V.A. Bloomfield. Salt dependence of the elasticity and overstretching transition of single dna molecules. *Biophysical journal*, 82(6):3160–3169, 2002.

[155] P.A. Wiggins, K.C. Cheveralls, J.S. Martin, R. Lintner, and J. Kondev. Strong intranucleoid interactions organize the escherichia coli chromosome into a nucleoid filament. *Proceedings of the National Academy of Sciences*, 107(11):4991–4995, 2010.

[156] P.A. Wiggins and P.C. Nelson. Generalized theory of semiflexible polymers. *Physical Review E*, 73(3):031906, 2006.

[157] P.A. Wiggins, R. Phillips, and P.C. Nelson. Exact theory of kinkable elastic polymers. *Physical Review E*, 71(2):021909, 2005.

[158] D.P. Wilson, A.V. Tkachenko, and J.C. Meiners. A generalized theory of dna looping and cyclization. *EPL (Europhysics Letters)*, 89:58005, 2010.

[159] Inc. Wolfram Research. *Mathematica*. Wolfram Research, Inc., Champaign, Illinois, version 5.1 edition, 2004.

[160] H. Yamakawa. *Helical wormlike chains in polymer solutions*. Springer Berlin, 1997.

[161] H. Yamakawa and W. H. Stockmayer. Statistical mechanics of wormlike chains. ii. excluded volume effects. *Journal of chemical physics*, 57(7):2843–1254, 1972.

[162] J. Yan, R. Kawamura, and J.F. Marko. Statistics of loop formation along double helix dnas. *Physical Review E*, 71(6):061905, 2005.

[163] J. Yan and J.F. Marko. Effects of dna-distorting proteins on dna elastic response. *Physical Review E*, 68(1):011905, 2003.

[164] J. Yan and J.F. Marko. Localized single-stranded bubble mechanism for cyclization of short double helix dna. *Physical review letters*, 93(10):108108, 2004.

[165] MA Young and DL Beveridge. Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn1. *Journal of molecular biology*, 281(4):675–687, 1998.

[166] Y. Zhang and D.M. Crothers. Statistical mechanics of sequence-dependent circular dna and its application for dna cyclization. *Biophysical journal*, 84(1):136–153, 2003.

[167] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K.S. Sandhu, U. Singh, et al. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, 38(11):1341–1347, 2006.

[168] X. Zheng and A. Vologodskii. Theoretical analysis of disruptions in dna minicircles. *Biophysical journal*, 96(4):1341–1349, 2009.