INTRODUCTION TO SINGLE-DNA MICROMECHANICS

(AUGUST 9, 2004)

John F. Marko

Department of Physics, University of Illinois at Chicago ,
Chicago, IL, USA

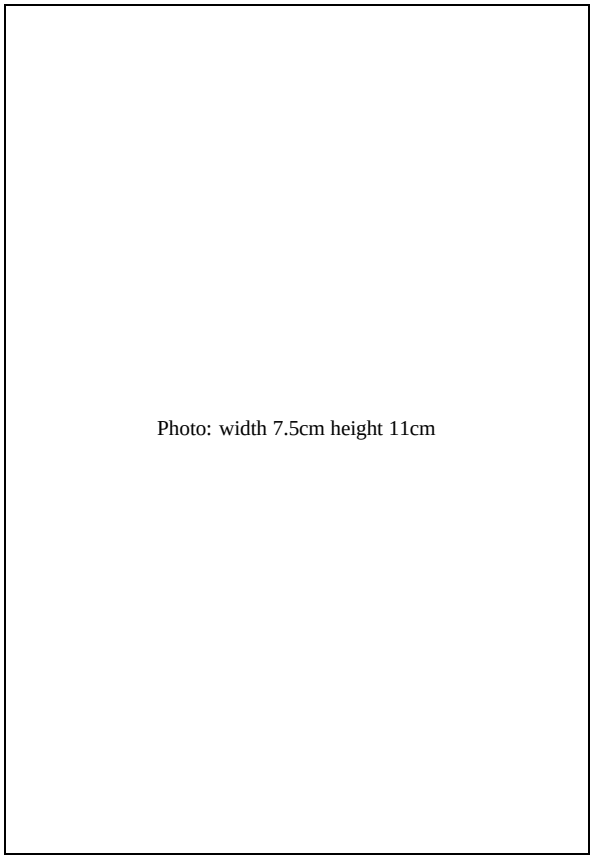
Photo: width 7.5cm height 11cm

Contents

1. Introduction 5
2. The Double Helix is a Semiflexible Polymer 7
   - 2.1. Structure 7
   - 2.2. DNA Bending 9
     - 2.2.1. Discrete-segment model of a semiflexible polymer 10
     - 2.2.2. Bending elasticity and the persistence length 12
     - 2.2.3. End-to-end distance 13
     - 2.2.4. DNA loop bending energies 14
     - 2.2.5. Site-juxtaposition probabilities 15
     - 2.2.6. Permanent sequence-driven bends 15
   - 2.3. Stretching out the double helix 16
     - 2.3.1. Small forces ($< k_BT/A = 0.08$ pN) 17
     - 2.3.2. Larger forces ($> k_BT/A = 0.08$ pN) 18
     - 2.3.3. Free energy of the semiflexible polymer 20
     - 2.3.4. Really large forces (> 10 pN) 20
3. Strand Separation 22
   - 3.1. Free-energy models of strand separation 22
     - 3.1.1. Sequence-dependent models 23
     - 3.1.2. Free energy of internal 'bubbles' 24
     - 3.1.3. Small internal bubbles may facilitate sharp bending 25
   - 3.2. Stretching single-stranded nucleic acids 26
   - 3.3. Unzipping the double helix 28
     - 3.3.1. Effect of torque on dsDNA end 30
     - 3.3.2. Fixed extension versus fixed force for unzipping 30
4. DNA Topology 32
   - 4.1. DNA supercoiling 32
     - 4.1.1. Twist rigidity of the double helix 32
     - 4.1.2. Writhing of the double helix 33
     - 4.1.3. Simple model of plectonemic supercoiling 34
   - 4.2. Twisted DNA under tension 36
   - 4.3. High forces and torques cause structural changes in the double helix 40
   - 4.4. DNA knotting 41
     - 4.4.1. Cells contain active machinery for removal of knots and other entanglements of DNA 41
     - 4.4.2. Knotting a molecule is surprisingly unlikely 41
     - 4.4.3. Condensation-resolution mechanism for disentangling long molecules 42
5. DNA-Protein Interactions 43
   - 5.1. How do sequence-specific DNA-binding proteins find their targets? 44
     - 5.1.1. Three-dimensional diffusion to the target 44
     - 5.1.2. Nonspecific interactions can accelerate targeting 45
   - 5.2. Single-molecule study of DNA-binding proteins 46
     - 5.2.1. DNA-looping protein: equilibrium 'length-loss' model 46

5.2.2. Loop formation kinetics 47
5.2.3. DNA-bending proteins 47
References 48

1. Introduction

The past 10 years has seen the development of new experimental techniques to look at DNA and the machines that process it. These experimental techniques are often called 'DNA micromanipulation', 'single-DNA' or slightly more generally 'single-molecule' experiments. These lectures focus on mechanical properties of DNA, which are crucial to the design and interpretation of single-DNA experiments and to the understanding of how DNA is processed, and therefore functions, inside the cell.

A seminal example of a single-DNA experiment is the experiment of Jeff Gelles, Steve Block and co-workers to measure the force exerted by RNA polymerase. Gene sequences in DNA are 'read' by RNApol, which synthesizes an RNA copy of a DNA sequence. To give one example of their utility, single-molecule experiments have revealed that as a RNApol moves along a DNA, it is able to pull with up to 30×10^{-12} Newtons, or about 30 piconewtons (pN) of force. In the single-molecule world, this is a hefty force: the motor proteins which generate your muscle contractions, called *myosin* generate only about 5 pN.

RNApol is an example of a *processive enzyme* which works rather like a macroscopic engine, using stored chemical energy to catalyze not only the synthesis of mRNA, but also converting some of that energy to mechanical work. This mechanical work is absolutely necessary for RNApol's function: it must move 'processively' along the DNA double helix in order to make a faithful copy of DNA. Another important DNA-processing enzyme is *DNA polymerase* which is able to synthesize a copy of a DNA strand; this is important in cell division, since in order to make a copy of itself, a cell must faithfully copy its chromosomal DNAs. Proper understanding of this kind of DNA-processing enzyme machinery requires us to first understand the mechanical properties of DNA itself.

DNA has extremely interesting and unique polymer properties. In double helix form it is a water-soluble, semiflexible polymer which can be obtained in gigantic lengths. A human genome contains 3×10^9 bases divided into 23 chromosomes. Each chromosome therefore contains roughly 10^8 bases. Since each chromosome is a single linear DNA molecule, chromosomal DNAs are the longest polymers known. Furthermore, the base-paired complementary-strand

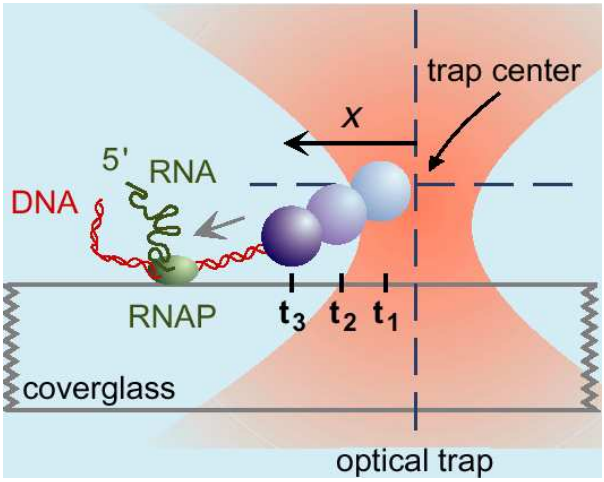

Fig. 1. Sketch of single-DNA experiment of Wang et al to measure force generated by RNA polymerase (reproduced from Ref. [1]). The polymerase is attached to the glass, and the DNA is pulled through it. A bead at the end of the DNA is held in a laser trap; deflection of the bead in the trap indicates the applied force.

structure of the double helix offers up completely new polymer physics problems. Exploration of the basic polymer physics of the double helix, with an emphasis on applications to the theory behind single-DNA experiments, is the main objective of these lectures.

Just a note on physical scales relevant to these lectures. The fundamental length scale of molecular biology is the nanometer (nm); this is a distance a few atoms long, the size of a single nucleic acid base, or a single amino acid. Cells maintain their organizational structure at room temperature: this requires that the components be acted on by forces of roughly
$1\ k_BT/\text{nm} = 4 \times 10^{-21}\ \text{J}/10^{-9}\ \text{m} = 4 \times 10^{-12}\ \text{N} = 4\ \text{pN}$.
We can expect the forces generated by single mechanoenzymes to be on the pN scale. If RNApol generated smaller forces than this, it would get pushed around by thermal forces, and would be unable to read DNA sequence in a processive manner.

Problem: Consider a molecule localized by a harmonic force $f = -kx$. What force constant is necessary to have 5A $\langle x^2 \rangle = 1\ \text{nm}^2$? What is the average force that is applied to the molecule in this case? Repeat this calculation if the localization is done to 1 Å (atomic) accuracy.

Problem: Consider a nanowire made of some elastic material, with circular

cross-section of diameter d. In any cross-section of the wire, what will be the typical elongational stress (force per area)? What does this suggest about the Young modulus of the material that you might try to use to make a nanowire?

Problem: Consider a linear random sequence of DNA bases which is 48502 bases long. How many times do you expect the sequences AATT, ACTAGT and GGCCGGCC to occur?

2. The Double Helix is a Semiflexible Polymer

The double helix (sometimes called the 'B-form') is taken by DNA most of the time in the cell. In this form, it has a regular helix structure with remarkably uniform mechanical properties. This section will focus on the bending flexibility of the double helix, which gives rise to polymer elasticity effects of biological importance, and accessible in biophysical experiments.

2.1. Structure

The double helix is made of two DNA polymer molecules. Each DNA polymer is a string of four interchangable types of 'monomers', which can be strung together in any sequence. The monomers each carry a *sugar-phosphate backbone* element: these are covalently bound together in the polymer. However, each monomer also carries, attached to the sugar (which is deoxyribose), one of four possible 'bases': either adenine (A), thymine (T), guanine (G) or cytosine (C).

The length of each backbone unit is about 0.7 nm when extended. The bases are each about 1 nm wide, and 0.3 nm thick.

The structure of each polymer gives it a definite 'polarity'. It is conventional to report DNA sequence along each strand in the direction read by RNA polymerase, from 5' to 3' (the number refer to carbon atoms in the deoxyriboses). Often people just omit the leading 5': in this case it is almost always in 5' to 3' order.

The bases have shapes and hydrogen-bonding sites which make A-T and G-C bonds favorable, under the condition that the two strands are anti-aligned (see sketch). Such *complementary strands* will bind together, making inter-strand hydrogen bonds, and intra-strand *stacking interactions*. The latter interactions, driven mainly by the hydrophobicity of the flat bases, and twist the two strands around one another. The result is an approximately regular helical structure, since each base is only about 0.3 nm thick while the backbones are roughly 0.7 nm long per base.

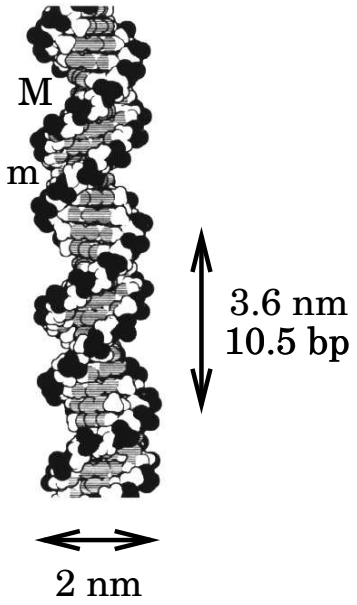

Fig. 2. DNA double helix structure. The two complementary-sequence strands noncovalently bind together, and coil around one another to form a regular helix. The two strands can be seen to have directed chemical structures, and are oppositely directed. Note the different sizes of the major (M) and minor (m) grooves. The helix repeat is 3.6 nm, and the DNA cross-sectional diameter is 2 nm.

We can roughly estimate the helix parameters of the double helix, assuming that the backbones end up tracing out a helical path on the surface of a radius cylinder (the bases are 1 nm wide). Since each base is 0.34 nm thick, and traces a helix contour length of 0.7 nm, the circumference occupied by each base is $\sqrt{0.7^2 - 0.34^2}$ nm = 0.61 nm. Dividing the total circumference (6.3 nm) by this indicates that the double helix contains about 10.3 base pairs (bp) per helical turn. This is very close to the number usually quoted of 10.5 bp/turn; the double helix therefore makes one turn for every 10.5×0.34 = 3.6 nm.

Don't forget that the B-form double helix is *right-handed*. Also, note that the opposite directions taken by the two backbones mean that there are two types of 'grooves' between the backbones: these are in fact rather different in size in B-DNA, and are called the 'major groove' and the 'minor groove'.

Also don't forget the important conversion factor for the double helix, length: 1 bp = 0.34 nm; thus each micron (1000 nm) worth of DNA contains about 3000 bp = 3 kilobp (kb); one whole human genome is thus close to 10^9 nm = 1 m in length.

Problem: Consider a hypothetical form of double helix formed of two *parallel-orientation* strands. Describe the grooves between the backbones.

Problem: A student proposes that for two complementary-sequence biological DNA strands, there must be an equivalent form of double helix, of free energy equal to B-DNA, which is instead left-handed. Explain under what circumstances of symmetry of the monomers this conjecture can be expected to be true. Based on textbook pictures of the base and backbone chemical structures, what is your conclusion?

Problem: Do you expect the average helix repeat (base-pairs per turn) of the double helix to increase or decrease with increased temperature?

Problem: Estimate the 'Young modulus' of the double helix, using the assumption that the single-base helix parameters described above apply to room-temperature structure of DNA to roughly 1 Å precision.

2.2. DNA Bending

Although the structure of DNA is often presented in books as if it is static, at room temperature and in solution the double helix undergoes continual thermally excited changes in shape. Per base pair, these amount to only small motions (a few degrees of bend, 0.03 nm average separations of the bases) but over long stretches of double helix, these build up to signifi cant, thermally excited random bends.

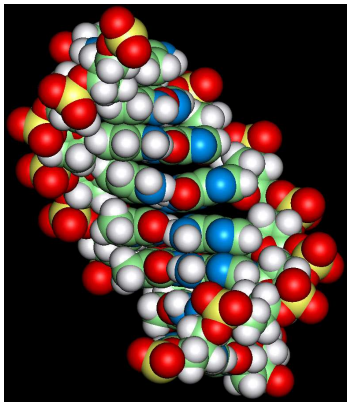

Fig. 3. Molecular-dynamics snapshot of typical DNA conformation for a short 10 bp molecule in solution at room temperature. Reproduced from Ref. [21].

2.2.1. Discrete-segment model of a semiflexible polymer

We can make a simple one-dimensional lattice model of this gradual randomization. If we describe our DNA with a series of tangent vectors $\hat{\mathbf{t}}_j$ that indicate the orientation of the *center axis* of the molecule, then the bending energy associated with two adjacent tangents is $E/(k_BT) = -a\hat{\mathbf{t}}_j \cdot \hat{\mathbf{t}}_{j+1}$.

The dimensionless constant a describes the molecule's bending rigidity: $a >> 1$ means very rigid (adjacent tangent vectors point in nearly the same direction); $a < 1$ means very fbppy. We'll talk more about a below, but just to set your thinking in the correct direction, for the DNA double helix, if we consider adjacent base pairs to be described by successive tangents, the value of a to use is about 150.

Problem: Estimate the typical angle between two adjacent tangent vectors excited thermally in the limit $a >> 1$; your result should be of the form
$\left\langle (\hat{\mathbf{t}}_j - \hat{\mathbf{t}}_{j+1})^2 \right\rangle \propto a^p$
where p is a power. Hint:
$\frac{1}{2}(\hat{\mathbf{t}}_j - \hat{\mathbf{t}}_{j+1})^2 = 1 - \hat{\mathbf{t}}_j \cdot \hat{\mathbf{t}}_{j+1}$ What is the typical single-base bending angle (in degrees) if we take $a = 150$?

We write the (unnormalized) probability distribution of a given conformation of an N-tangent-vector-long chunk of molecule using the Boltzmann distribution:

$$P(\hat{\mathbf{t}}_1, \cdots, \hat{\mathbf{t}}_N) = \Pi_{j=0}^{N-1} e^{a\hat{\mathbf{t}}_j \cdot \hat{\mathbf{t}}_{j+1}} \tag{2.1}$$

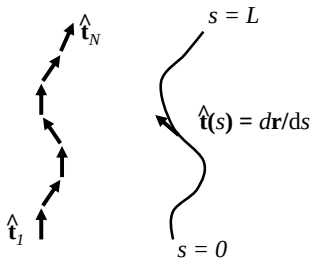

Fig. 4. Discrete-tangent and continuous-tangent models for DNA bending (see text).

Now we compute the thermal correlation of the ends of this segment of polymer:

$$\left\langle \hat{\mathbf{t}}_N \cdot \hat{\mathbf{t}}_1 \right\rangle = \frac{\int d^2t_1 \cdots d^2t_N \hat{\mathbf{t}}_1 \cdot \hat{\mathbf{t}}_N P(\hat{\mathbf{t}}_1, \cdots, \hat{\mathbf{t}}_N)}{\int d^2t_1 \cdots d^2t_N P(\hat{\mathbf{t}}_1, \cdots, \hat{\mathbf{t}}_N)} \tag{2.2}$$

This calculation is not too hard to do if you remember the handy formula (recall decomposition of plane waves into spherical waves):

$$e^{a\hat{\mathbf{t}}\cdot\hat{\mathbf{t}}'} = \sum_{l=0}^{\infty} 4\pi i^l j_l(ia) \sum_{m=-l}^{l} Y_{lm}(\hat{\mathbf{t}}) Y_{lm}^*(\hat{\mathbf{t}}') \tag{2.3}$$

and also if you remember to write the dot product $\hat{\mathbf{t}}_1\hat{\mathbf{t}}_N$ as an $l = 1$ spherical harmonic. (life will be easiest if you recognize that you can choose $\hat{\mathbf{t}}_N$ to be along the $\hat{\mathbf{z}}$ axis).

You will find that the orthogonality of the spherical harmonics results in a 'collapse' of the many sums over l's and m's into one sum. In the numerator only the $l = 1$ term (from the dot product) survives; in the denominator only the $l = 0$ term contributes. The result is:

$$\left\langle \hat{\mathbf{t}}_1 \cdot \hat{\mathbf{t}}_N \right\rangle = \left(\frac{i j_1(ia)}{j_0(ia)} \right)^N = e^{N \ln[\coth(a) - 1/a]} \tag{2.4}$$

The function $\coth(a) - 1/a$ is less than 1 for positive a. Therefore the correlation of direction falls off simply exponentially with contour distance N along our polymer. Small local fluctuations of bending of adjacent tangents build up to big bends over the 'correlation length' of $-\ln[\coth(a) - 1/a]$ tangents.

Problem: For the $a >> 1$ limit, how many segments long is the tangent-vector correlation length?

Problem: Explain the relation between the discrete-tangent model discussed above and the one-dimensional Heisenberg (continuous-spin) model of classical statistical mechanics. Suppose a magnetic field is added: what would that correspond to in the polymer interpretation?

2.2.2. Bending elasticity and the persistence length

We can connect this discrete model to the continuous model for bending of a thin rod, from the theory of elasticity. We note that the bending energy of two adjacent tangents was, in k_BT units, $-a\hat{\mathbf{t}}_j \cdot b\hat{f}t_{j+1}$, which up to a constant equals $\frac{a}{2}(\hat{\mathbf{t}} - \hat{\mathbf{t}}')^2$.

The bending of a thin rod in the theory of elasticity can be described in terms of tangent vectors telling the rod direction, distributed continuously along the rod contour. Using contour length $\hat{\mathbf{t}}(s)$ length, a bent rod has energy which is locally proportional to the square of its bending curvature $d\hat{\mathbf{t}}/ds$.

$$E = \frac{B}{2}\int_0^L ds \left(\frac{d\hat{\mathbf{t}}}{ds}\right)^2 \tag{2.5}$$

where B is the rod bending modulus. For a rod of circular cross section of diameter d made of an isotropic elastic material, $B = \frac{\pi}{64}Yd^4$ where Y is the Young modulus.

Problem: By considering a simple circular arc, find the contour length along a thin rod for which a one-radian bend has energy cost k_BT.

Problem: Pretend that dsDNA is made of a plastic material of Young modulus 3×10^8 Pa. Predict the bending constant B.

The connection between our discrete and the continuous models of bending can now be written, if we introduce the length b of the segments in our discrete model:

$$\begin{aligned} E &= -k_BTa\sum_{j=1}^{N}\hat{\mathbf{t}}_j \cdot \hat{\mathbf{t}}_{j+1} = \frac{k_BTab}{2}\sum_{j=1}^{N} b\left(\frac{\hat{\mathbf{t}}_j - \hat{\mathbf{t}}_{j+1}}{b}\right)^2 \\ &\to \frac{B}{2}\int_0^L ds\left(\frac{d\hat{\mathbf{t}}}{ds}\right)^2 \end{aligned} \tag{2.6}$$

where the final term represents the limit where we make b small, taking the continuum limit to turn the finite difference into a derivative, and the sum into an integral.

The bending elastic constants a and B are connected by the relation $k_BTab = B$; the rod length corresponds to the number of tangents through $Nb = L$. So,

for a rod with bending modulus B, if we wish to use a discrete tangent vector model with segment length b, we need to choose $a = B/(k_B T b)$. This makes sense: as b is chosen shorter, we need to make the local stiffness a larger.

If we now go back to the correlation function (2.4), we can write it in the continuum limit where a becomes large, replacing $\ln[\coth a - 1/a] \to -1/a$ and obtaining

$$\left\langle \hat{\mathbf{t}}(s)\hat{\mathbf{t}}(s')\right\rangle = e^{-k_B T|s-s'|/B} = e^{-|s-s'|/A} \tag{2.7}$$

The final term introduces the continuum version of the correlation length 5A of (2.4) $A = B/(k_B T)$, called the *persistence length*. For the double helix, a variety of experiments show that $A = 50$ nm (150 bp) in physiological aqeuous solution (roughly, water containing between 0.01 and 1 M univalent salt, and with pH between 7 and 8, at temperature between 15 and 30 C).

Problem: Starting with the persistence length $A = 50$ nm, estimate the bending modulus B, and the 'effective Young modulus' Y of the DNA double helix.

A point about a and B is that really both represent effective elastic constants, and the bending energies being discussed are really free energies (as in the theory of elasticity, we consider deformations at fixed temperature). The 'real', microscopic internal energy has to do with the thermal energy and chemical binding energies of the atoms, but as in many other areas of condensed matter physics we'll choose to ignore atomic details and use coarse-grained models for DNA and DNA-protein interactions. This is not to say that atomic detail is not important: a great deal of insight into the double helix and DNA-protein interactions can be obtained through the complementary approach of numerical simulation of all the atoms involved (see Fig. 3), although this is very challenging.

2.2.3. End-to-end distance

The tangent vector $\hat{\mathbf{t}}(s)$ can be used to compute the distance between two points on our polymer, using the relation $\mathbf{r}(L) - \mathbf{r}(0) = \int_0^L ds\hat{\mathbf{t}}(s)$. This relation can be used to compute the mean-square distance between contour points a distance L apart:

$$\left\langle |\mathbf{r}(L) - \mathbf{r}(0)|^2 \right\rangle = 2AL + 2A^2\left(e^{-L/A} - 1\right) \tag{2.8}$$

In the limit where we look at points closer together than a persistence length, $L/A << 1$, we have a mean-square distance $= L^2 + \mathcal{O}(L/A)$; in this limit, the polymer doesn't bend very much, so its average end-to-end distance is just L.

In the opposite limit of a polymer many persistence lengths long, $L/A >> 1$, we have a mean-square-distance of $2AL$, just the size expected for a random-

walk of $L/(2A)$ steps each of length $2A$. We sometimes talk about the *statistical segment length* or *Kuhn segment length* in polymer physics: for the semiflexible polymer this segment length is $2A$. For the double helix, $2A$ is about 100 nm or 300 bp [2].

2.2.4. DNA loop bending energies

We'll hear in Section 5 about proteins which stabilize formation of DNA loops. Often, looping of DNA occurs so that sequences roughly 10 to 1000 bp away from the start of a gene can regulate (repress or enhance) that gene's transcription. Formation of such a loop requires DNA bending, and now we can estimate the bending free energy associated with this.

Suppose we form a loop of length L. The simplest model is a circle of circumference L, with radius $L/(2\pi)$ and bending curvature $2\pi/L$, and bending energy

$$\frac{E_{\rm circle}}{k_BT} = \frac{A}{2}L\left(\frac{2\pi}{L}\right)^2 = 2\pi^2\frac{A}{L} \tag{2.9}$$

For $L = 300$ bp and $A = 150bp$, this is a big energy - close to 10 k_BT. A 100 bp circle would have a bending free energy 9 times larger than this!

You might be interested in the *lowest* energy necessary to bring two points a contour length L along a rod together. The optimal shape of the rod is of course not circular, but is instead tear-drop-shaped. The exact energy can be computed in terms of ellipic functions to be

$$\frac{E_{\rm teardrop}}{k_BT} = 14.055\frac{A}{L} \tag{2.10}$$

about 71% of the energy of the circle. In either teardrop or circle case, the energy of making a loop diverges as $1/L$ for small L.

Problem: Carry out an approximate calculation of the tear-drop shape and energy, by using a circular arc combined with two straight segments. Use energy minimization (with fixed total length) to find the angle at the base of the tear-drop (you should only have one parameter to minimize over) and the tear-drop configuration energy.

Problem: In a protein-DNA structure called the nucleosome, 146 bp of DNA make 1.75 helical turns with helix radius of 5 nm, and helical pitch (spacing of turns along the helix axis) of 3 nm. Using the simple models of this section, estimate the bending free energy of the DNA in k_BT.

2.2.5. Site-juxtaposition probabilities

These bending energies are not by themselves enough to accurately predict the probability that a DNA segment of length L forms a loop; we must also sum over bending fluctuations, thermally excited changes in shape. For the simple bending model described above, sophisticated calculations have been done for the probability of forming a loop.

Calculations of Stockmayer, Shimada and Yamakawa tell us the probability density for finding the two ends of a DNA brought smoothly together (with the same orientation):

$$J_{\text{circle}} = \frac{\pi^2}{(2A)^3}\left(\frac{2A}{L}\right)^6 e^{-E_{\text{circle}}/k_BT + 0.257L/A} \tag{2.11}$$

If we relax the condition that the ends come together smoothly, the same authors find

$$J_{\text{teardrop}} = \frac{28.01}{(2A)^3}\left(\frac{2A}{L}\right)^5 e^{-E_{\text{teardrop}}/k_BT + 0.246L/A} \tag{2.12}$$

The units of these expressions are density (inverse volume), i.e. concentration of one end, at the position of the other. It is worth noting that 1 Mol/litre is close to nm^{-3}, and that since the double helix can bend only over roughly 100 nm, the natural scale for J is very roughly $J \approx (100)^{-3}\ \text{nm}^3 = 10^{-6}\ \text{nm}^{-3} \approx 10^{-6}$ M.

The empirical results provide an accurate interpolation between the two limits where bending energy ($L/A < 1$), and entropy ($L/A > 1$) dominate, including the experimentally- and numerically-established result that the *peak* probability of juxtaposition occurs for molecules about $L = 170$ nm (500 bp) long [2].

For $L >> A$ we reach the long-distance limit, where we may estimate the probability of finding the two ends of a long DNA close together, using the average end-to-end distance (2.8), which is $R \approx \sqrt{2AL}$ for long L. For long L, the two ends are somewhere in a volume of about $(2AL)^3/2$. Therefore, the probability of finding the two ends together for $L/A >> 1$ decays as $J \approx 1/(AL)^{3/2}$.

This formula does not account for self-avoidance, but because the double helix has a segment length $2A$ so much longer than its diameter (only 3 nm even when electrostatic repulsion is taken into account) that self-avoidance effects can be neglected for molecules as large as 10^4 bp in length.

2.2.6. Permanent sequence-driven bends

We've focused on thermally excited bends, using a model which has as its 'ground state' a perfectly straight conformation. You need to keep in mind that the average shape of any DNA molecule depends on its sequence: different sequences

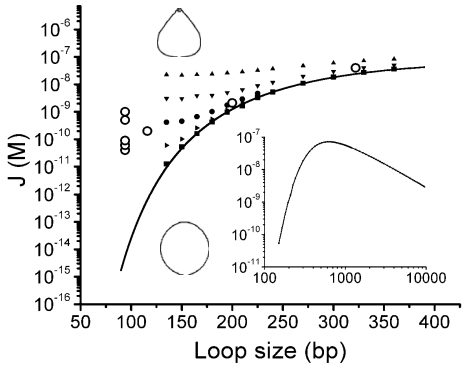

Fig. 5. Juxtaposition probability ($J_{\rm circle}$) including experimental data of Cloutier and Widom [4]. Inset shows $J_{\rm circle}$ for large distances, showing peak near 500 bp and $L^{-3/2}$ decay. Main figure focuses on energetically-dominated small-L behavior, showing strong suppression of probability. Experimental data (circles) show anomalously large probability of juxtaposition for 94 bp. The other symbols correspond to a theory of DNA site juxtaposition including the effect of kinks in the DNA: we'll hear more about this in Sec. 3.1.3. Units of $J_{\rm circle}$ in the figure are Mol/litre.

have slightly different average distortions. A remarkable discovery is that it is possible, by 'phasing' sequences that generate kinks, one can obtain DNAs with strong permanent bends along them. Some of these strong permanent bends are implicated in biological processes, for example facilitation of the binding of proteins that bend or wrap DNA.

2.3. Stretching out the double helix

One type of single-molecule experiment which has become widely studied is the stretching of DNAs using precisely calibrated forces. Early experiments showed that the double helix displayed polymer stretching elasticity of exactly what was expected from the semiflexible polymer model introduced above. This has been important to the design of many experiments focusing on the effects of proteins or other molecules binding to, or moving along DNA. This subsection reviews the basic polymer stretching elasticity of a long ($L >> A$) double helix DNA.

A force f applied to a single DNA molecule of length L appears in the Boltzmann factor coupled to the end-to-end vector along the force direction (which we

take to be z). Our Hamiltonian becomes:

$$E = \frac{k_B T A}{2} \int_0^L ds \left(\frac{d\hat{\mathbf{t}}}{ds} \right)^2 - f\hat{\mathbf{z}} \cdot [\mathbf{r}(L) - \mathbf{r}(0)] \tag{2.13}$$

We can turn the end-to-end vector into an integral over $\hat{t}$ as before, giving

$$-\beta E = \int_0^L ds \left[\frac{A}{2} \left(\frac{d\hat{\mathbf{t}}}{ds} \right)^2 - \beta f \hat{\mathbf{z}} \cdot \hat{\mathbf{t}} \right] \tag{2.14}$$

There is a single parameter $\beta A f$ which controls this Hamiltonian (write the Hamiltonian with contour length in units of A). We have two regimes to worry about: forces below, and above the characteristic force $k_B T/A$.

For the double helix, $A = 50$ nm, so $k_B T/A$ = 0.02 $k_B T$/nm = 0.08 pN. This is a low force due to the long persistence length of the double helix.

Ideally, we want to calculate the partition function

$$Z(\beta A f) = \int \mathcal{D}\hat{\mathbf{t}} e^{-\beta E} \tag{2.15}$$

and then calculate the end-to-end extension, using

$$\langle \hat{\mathbf{z}} \cdot [\mathbf{r}(L) - \mathbf{r}(0)] \rangle = \frac{\partial \ln Z}{\partial \beta f} \tag{2.16}$$

This can be done in general numerically, but we can find the low- and high-force limits analytically.

2.3.1. Small forces ($< k_B T/A = 0.08$ pN)

For small forces, we can calculate the end-to-end extension using linear response, since we know the zero-force fluctuation of the mean-square end-to-end distance: recall that this was $2AL$. This counted three components; by symmetry we have

$$\left\langle \left(\hat{\mathbf{z}} \cdot [\mathbf{r}(L) - \mathbf{r}(0)] \right)^2 \right\rangle = \frac{2AL}{3} \tag{2.17}$$

The linear force constant will be $k_B T$ divided by this fluctuation, giving a small-extension force law:

$$f = \frac{3 k_B T}{2AL} z + \cdots \tag{2.18}$$

where we use the shorthand $z = \langle \hat{\mathbf{z}} \cdot [\mathbf{r}(L) - \mathbf{r}(0)] \rangle$ to indicate the average end-to-end extension in the force direction.

This is just the usual ideal (Gaussian) low-extension force law familar from polymer physics. The spring constant of the polymer is inversely proportional to the persistence length, and to the total chain length.

2.3.2. Larger forces ($> k_BT/A = 0.08$ pN)

The linear force law shows that our ideal DNA will start to stretch out when forces of $\approx k_BT/A$ are applied to it. We can also calculate the very nonlinear elasticity associated with the nearly fully stretched polymer, using an expansion in $1/\sqrt{f}$.

Suppose that the polymer is quite stretched out, so that $\hat{\mathbf{t}}(s) = \hat{\mathbf{z}}t_{\|} + \mathbf{u}$, where $\mathbf{u}$ is in the xy plane, and has magnitude $<< 1$. Since $\hat{\mathbf{t}}^2 = 1$, $t_{\|} = \sqrt{1 - |\mathbf{u}|^2} = 1 - \frac{1}{2}|\mathbf{u}|^2 + \cdots$ Plugging this into the Hamiltonian (2.14) and expanding to leading order in $|\mathbf{u}|^2$ gives:

$$-\beta E = -\beta f L + \frac{1}{2}\int_0^L ds \left[A\left(\frac{d\mathbf{u}}{ds}\right)^2 + \beta f |\mathbf{u}|^2 \right] \tag{2.19}$$

In this limit, the fluctuations can be seen to slightly reduce the length, generating the final energy cost term.

Introducing Fourier modes $\mathbf{u}_q = \int_0^L ds e^{iqs}\mathbf{u}(s)$ diagonalizes the Hamiltonian:

$$\frac{E}{k_BT} = -\beta f L + \frac{1}{2L}\sum_q (Aq^2 + \beta f)|\mathbf{u}_q|^2 \tag{2.20}$$

where $q = \pm 2\pi n/L$ for $n = 0, \pm 1, \pm 2, \cdots$. The fluctuation amplitude of each mode is therefore

$$\left\langle |\mathbf{u}_q|^2 \right\rangle = \frac{2L}{Aq^2 + \beta f} \tag{2.21}$$

where the leading 2 comes from the two components (x and y) of $\mathbf{u}$. Now we can compute the real-space amplitude:

$$\left\langle |\mathbf{u}(s)|^2 \right\rangle = 2\int_{-\infty}^{\infty} \frac{dq}{2\pi}\frac{1}{(Aq^2 + \beta f)} = \frac{1}{\sqrt{\beta A f}} \tag{2.22}$$

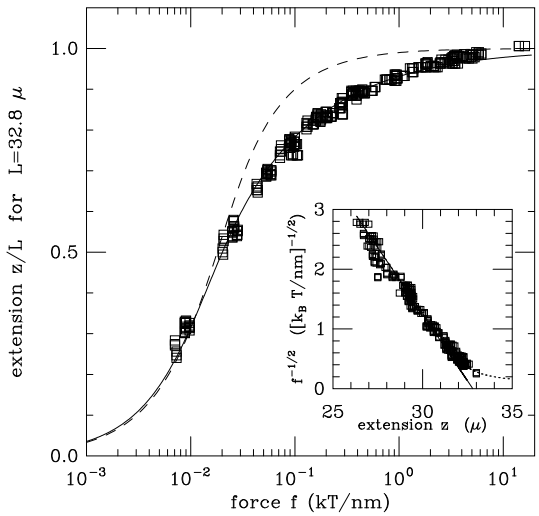

Fig. 6. Experimental data and models for stretching of the double helix. Main figure shows experimental data of Smith et al (1992) and a fit to the semiflexible-polymer model, for a persistence length $A = 53$ nm. The units of force are k_BT/nm; recall $1k_BT/\text{nm} = 4.1$ pN, for $T = 300$ K. Inset shows a plot of extension versus inverse square root of force, showing the linear relation between these two quantities.

and finally the extension in the force direction

$$z = L\left\langle t_{\|}\right\rangle = L\left(1 - \frac{1}{2}\left\langle |\mathbf{u}|^2\right\rangle + \cdots\right) = L\left(1 - \frac{1}{\sqrt{4\beta A f}} + \cdots\right) \qquad (2.23)$$

The semiflexible polymer shows a distinct $1/\sqrt{f}$ behavior as it is stretched out. Also note that the form of the Hamiltonian in wavenumbers indicates that there is a force-dependent correlation length for the bending fluctuations, given by $\xi = \sqrt{A/(\beta f)}$. Experiments on double helix DNAs show this relation.

The asymptotic linear relation between z and $1/\sqrt{f}$ is quite useful. It turns out this holds well theoretically for the exact solution of the semiflexible polymer model under tension, for $z/L > 0.5$. If you have experimental data for stretching a semiflexible polymer, you can plot z versus $1/\sqrt{f}$ and fit a line to the $z/L > 0.5$, the z-intercept of the linear fit estimates the molecular length L, and the $1/\sqrt{f}$ intercept gives an estimate of $\sqrt{4\beta A}$, i.e. a measurement of persistence length. The agreement between different kinds of single-DNA experiments gives quantitatively strong evidence that for long molecules, most of the elastic response comes from thermal bending fluctuations.

2.3.3. Free energy of the semiflexible polymer

It is useful to compute the free energy difference between unstretched and stretched polymer from the extension in the force direction, by integrating (2.16):

$$\ln Z(f) = \beta \int_0^f df' z(f') + \ln Z(0) \tag{2.24}$$

We'll drop the constant $\ln Z(0)$, which amounts to taking the relaxed random coil as a 'reference state' with free energy defined to be zero. For the semiflexible polymer the free energy takes the form

$$\ln Z(f) = \frac{L}{A}\gamma(\beta A f) \tag{2.25}$$

which is the exact scaling form of the partition function for the semiflexible polymer in the limit $L/A >> 1$.

The dimensionless function $\gamma(x)$ can be computed precisely numerically for the semiflexible polymer (or for variations of it), but for us it will be sufficient to consider the limits:

$$\gamma(x) = \begin{cases} x - \sqrt{x} + \cdots & x > 1 \\ 3x^2/4 + \cdots & x < 1 \end{cases} \tag{2.26}$$

The free energy we are computing here, the log of the partition function at fixed force, can be converted to the work done extending the polymer to a given extension, $W(z)$, by the Legendre transformation $W(z) = -k_B T ln Z + fz$.

2.3.4. Really large forces (> 10 pN)

For forces in the range 10 to 40 pN, the double helix starts to stretch elastically. This stretching can be described by adding an term to the result above:

$$\frac{z}{L} = 1 - \frac{1}{\sqrt{4\beta A f}} + \frac{f}{f_0} \tag{2.27}$$

The constant f_0 has dimensions of a force, and represents the stretching elastic constant of the double helix. In terms of the Young modulus, this elastic constant for a circular rod is $f_0 = (\pi/4)d^2 Y$. Experimental data indicate $f_0 \approx 1000$ pN.

Finally, at about 60 to 65 pN, depending a bit on salt concentration, there is an abrupt transition to a new double helix state about 1.7 times longer than B-form. This is sometimes called the S-form of DNA; opinion remains divided on whether this form is base-paired or not. Fig. 7 shows some experimental data for the high-force response of dsDNA (squares and diamonds) from two groups.

Note the elongation of the double helix above the fully extended double helix value of 0.34 nm/bp, and the sharp 'overstretching' transition force 'plateau' near 63 pN.

Problem: Above we saw that $B = (\pi/64)Yd^4$ where d is the diameter of an elastic rod. Compare the values of Y inferred from B and from f_0. Are they consistent?

Problem: Consider longitudinal stretching fluctuations of adjacent base pairs. Compute the energy of a fluctuation of amplitude (length) δ: what is the root-mean-square value of the single-base-pair longitudinal fluctuation $\sqrt{\langle \delta^2 \rangle}$?

Problem: Under some conditions, a *single strand* of DNA will behave like a flexible polymer of persistence length $A_{\rm ss} \approx 1$ nm. Find the characteristic force at which you might expect a single-stranded DNA to become 50% extended.

Problem: Consider the Hamiltonian (2.14) generalized so that it contains a *vector* force $\mathbf{f}$ coupled via dot product to end-to-end extension. Show that $\partial_{\beta f_i} \ln Z = \langle x_i(L) - \mathbf{x}_i(0) \rangle$ and that $\partial^2_{\beta f_i} \ln Z = \left\langle [x_i(L) - x_i(0)]^2 \right\rangle$ where the indices i label the three spatial coordinates. Now verify the following formula relating the end-to-end vector fluctuations transverse to the force direction to the average extension:

$$\frac{k_B T \langle [z(L) - z(0)] \rangle}{\langle [x(L) - x(0)]^2 + [y(L) - y(0)]^2 \rangle} = f \tag{2.28}$$

where the force is assumed to be applied in the z direction.

This exact, nonperturbative relation is often used in magnetic tweezer experiments to calibrate forces applied to single DNA molecules, by measuring all the quantities on the left hand side. Note that this result does not depend on the details of the polymer part of the Hamiltonian - even if it contains long-ranged interactions - as long as it is invariant under space rotation.

Problem: For the semiflexible polymer, consider the approximate force-extension relation $\beta A f = z/L + 1/[4(1 - z/L)^2] - 1/4$. Show that this function reproduces the high- and low-force limiting behaviors derived above (it is not a terribly accurate representation for the exact behavior of $f(z)$). Compute the free energy $W(z)$ using this relation. Hint: partial integration applied to (2.16).

Problem: Consider the 'freely jointed chain' obtained by setting $a = 0$ in the segment model. Calculate the extension, and free energy ($\ln Z$) as a function of force. Also calculate the transverse mean-squared fluctuations as a function of force.

3. Strand Separation

In the previous section we downplayed a biologically and biophysically important feature of the double helix, namely that it consists of two covalently bonded *single-stranded* DNAs (ssDNAs) which are relatively weakly stuck to one another. This makes it possible for the two strands of a double helix to be separated from one another, as occurs in vivo during DNA replication, and transiently, during DNA transcription (RNA polymerase reads one strand of DNA), and DNA repair.

Conversion of dsDNA to ssDNA can be accomplished in a few ways:
Elevated temperature: The double helix is stable in 'physiological' buffer (pH near 7, univalent salt in the 10 mM to 1 M range) for temperatures below about 50 C. Over the range 50 to 80 C, the double helix 'melts', with AT-rich sequences falling apart at the low end of this range, and highly GC-rich sequences holding together until the high end of this temperature range.
Denaturing solution conditions: Too little salt (< 10 mM NaCl), which increases electrostatic repulsion of the negatively charged strands, or pH out of the range 7 to 9, destabilizes the double helix, lowering its melting temperature.
Sufficient 'unzipping force' applied to the two strands: If you pull the two strands apart, they will separate at forces in the 10 to 20 pN range, with force variations reflecting the sequence composition. A process similar to this idealized 'unzipping' process is carried out in the cell during DNA replication: specialized motor enzymes called *DNA helicases* track along the double helix, pushing the two strands apart.

Below we will mainly discuss the last of these three modes of strand separation, unzipping by force.

3.1. Free-energy models of strand separation

In the simplest picture of DNA melting, we ignore base sequence entirely, and consider simply the average free energy difference g per base pair between isolated, relaxed ssDNAs and dsDNA, at room temperature and in physiological solution conditions. Then, for an N-base-pair-long molecule, the free energy difference between ssDNAs and dsDNAs would be just $G_{\rm ssDNAs} - G_{\rm dsDNA} = Ng$. For random DNA sequences, this $g \approx 2.5 k_B T$; its positive value reflects the fact that the double helix is more stable than isolated single strands: very roughly, the probability of observing melted single strands is $e^{-\beta N g}$.

Thermal melting can be most simply thought about by considering the temperature dependence of the base-pairing free energy, breaking it into 'enthalpy' h and 'entropy' s per base pair, i.e. $g = h - sT$. At the melting temperature

$T_m = h/s$, the free energy of isolated ssDNAs is equal to free energy of double helix, making these two states equally probable.

3.1.1. Sequence-dependent models

A number of groups are working on accurate algorithms to predict the melting temperatures of dsDNAs as a function of sequence. One of these classes of models assign a contribution to base-pairing free energy for each *pair* of bases, the idea being that stacking interactions of adjacent base pairs play an important role in determining the stability of the double helix. The raw data behind such models are melting temperature data for a set of different sequence, usually short (10 to 20 bp) dsDNAs.

Table 3.1.1 lists a set of free energies due to SantaLucia [3] for the ten different oriented pairs of bases that occur along a DNA strand (note that all remaining pairs of bases can be obtained from considering the complementary sequence on the adjacent strand, i.e. the contribution of 5'-GA is the same as that of 5'-TC found in the table). The free energy of strand separation for a long $N-$bp molecule is obtained by adding the $N-1$ adjacent-base contributions together. In addition, there are contributions for the ends which we won't discuss - all though they are signifi cant.

The key point of Table 3.1.1 is that AT-rich sequences are lower in strand-separation free energy (the values for AT, AA and TA are all less than $1.7k_BT$), while GC-rich sequences are higher (GG, GC and CG are $3k_BT$ or more). Models of this type are not infallible (in reality, double-helix structure and energy depends on longer than nearest-neighbor sequence correlations) but they do give some idea of sequence dependence of base-pairing free energy.

The data of Table 3.1.1 are for the physiological ionic strength of $150mM$

Base i and $i+1$ (5'→3')	Free energy g_i (150 mM NaCl, pH 7.5, 25 C)
AA	1.68
AT	1.42
AG	2.19
AC	2.42
TA	0.97
TG	2.42
TC	2.12
GG	3.00
GC	3.75
CG	3.68

Table 1

Base-pairing-stacking free energies of Santa Lucia [3]. Free energies are in k_BT units, and are for 25 C, 150 mM NaCl, pH 7.5. For other salt concentrations the values must be corrected (see text).

NaCl; lower ionic strengths reduce the base-pairing free energy. An ionic-strength correction for the base-pairing free energy has been given by Ref. [3]): $\Delta g_i = 0.2\ln(M/0.150)$ where M is the molarity of NaCl.

Problem: For the simple model where the strand separation free energy per base is a constant $g = h - sT$, calculate the probability of finding separated single strands as a function of temperature (Hint: two-state system). How does the *width* of the melting transition as a function of temperature scale with N?

Problem: Calculate the free energy differences between separated ssDNAs and double helicies, for the following sequences: 5'-AATTAATTAATT, 5'-GCGCGCGGCCGG, 5'-AGCTCCAAGGCT. You may want to consult reference [3] to include the end effects.

Problem: In Table 3.1.1 you can see that AT-rich sequences have roughly $2k_BT$ less free energy holding them together than do GC-rich sequences. For a random N-base sequence, there will therefore be a mean free energy of strand separation, and fluctuations of that free energy. Calculate the mean free energy per base pair, and estimate the fluctuations.

3.1.2. Free energy of internal 'bubbles'

The above discussion suggests that thermal melting might be described by a 1-dimensional Ising model with sequence-dependent interactions, i.e. with some quenched 'randomness'. However, this would ignore an important physical effect that acts to suppress opening of bubbles in the interior of a long double helix. This effect is the *entropic cost* of forcing an internal 'bubble' to close. This cost is not included in the strand separation free energy models described above which are fit to data obtained from melting of short double helicies.

This loop free energy is easy to roughly understand - we have already discussed it above indirectly in our discussion of juxtaposition of DNA sequences. We mentioned that the long-molecule limit for DNA juxtaposition probability should be $J \approx N^{-3/2}$ simply from considering the fact that the two molecule ends should be found in a volume of radius $R \approx N^{1/2}$. If we think about this probability in terms of a free energy cost of constraining the ends to be near one another, we obtain the loop free energy cost

$$\Delta G_{\rm loop} = \frac{3}{2} k_B T \ln N \tag{3.1}$$

Since ssDNA has a persistence length of roughly one base (0.7 nm), the N relevant here is simply the number of bases in the loop. For an internal ssDNA bubble formed by opening N base pairs, we should use $2N$ as the loop length.

This additional free energy discourages opening of internal bubbles, eliminating the use of the simple Ising model with short-ranged interactions to describe

DNA melting. In fact, the logarithmic interaction of (3.1) is sufficiently long-ranged to kill the usual argument against a phase transition in a 1d system (if we create bubbles on any sequence scale N, the $3/2 \ln N$ loop entropy cost per bubble exceeds its $\ln N$ translational entropy). A real phase transition occurs in the 'pure' DNA melting model including the logarithmic loop effect; however, variations in local melting temperatures due to sequence variations along long real DNAs wash out a sharp phase transition.

We can estimate the total free energy cost of an N-base-pair internal bubble, adding the base-pairing/stacking free energy to the loop free energy cost:

$$\sum_{i=1}^{N} g_i + \frac{3}{2} k_B T \ln(2N) \tag{3.2}$$

The sequence-dependent term ranges from about $N k_B T$ to $4N k_B T$, making the price of a large, 10 bp bubble from roughly 20 to 45 $k_B T$: i.e. very rare except for the most AT-rich sequences. Larger bubbles are even more costly, making them exceedingly rare excitations.

3.1.3. Small internal bubbles may facilitate sharp bending

Small internal bubbles are not impossibly costly excitations: a 3 bp bubble costs 8 to 15 $k_B T$. Short AT-rich 3 bp sequences are by this reckoning, open roughly 0.1% of the time, and can be expected every few hundred base pairs (e.g. the particularly weak sequence TATA appears once every 256 base pairs in random-sequence DNA).

These small, thermally excited bubbles suggest an explanation for the recent results of Cloutier and Widom [4] (see Fig. 5) showing that the cyclization (circularization) probabilities of dsDNAs less than 300 bp long are far larger than we would expect from the simple elastic bending model 2.5. The experimental data indicate that tight bends of the double helix can occur via an alternative, lower-energy mechanism. One possibility is that via separation of a few base pairs, a 'flexible joint' might appear that could reduce the bending energy of formation of a loop. Although the free energy cost of generating a few-base-pair 'joint' is, according to the above roughly $10 k_B T$, this becomes similar to the bending free energy saved by concentrating much of the bending into a localized 'kink'.

Problem: Consider Fig. 5, which shows experimental data indicating that circular closure of 94 bp DNAs occurs with probability far above the expected value J_{circle}. Suppose that for some free energy ϵ we can *kink* the DNA so that it can still close smoothly, but now with the tear-drop shape which minimizes the bending energy. Estimate what ϵ should be to explain the 94 bp data (Hint: use

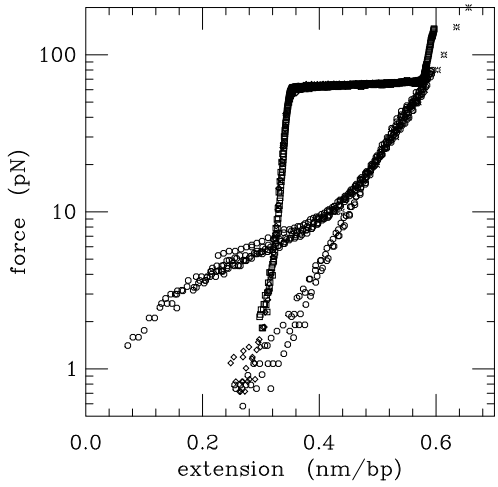

Fig. 7. Force versus extension of double helix and ssDNA. Squares show experimental dsDNA data of Léger et al [7, 9] for 500 mM NaCl buffer, diamonds show experimental dsDNA data of Smith et al [11] for 1 M NaCl buffer. Data for physiological salinity (150 mM NaCl) are similar, but have a plateau shifted a few pN below the 500 mM result, see Refs. [9, 12]. Circles show experimental data of Bustamante et al [10] for ssDNA; stars show high-force ssDNA data of Rief et al [6]. The left, lower-extension curve is for 150 mM NaCl, while the right, higher-extension curve is for 2.5 mM NaCl. The two ssDNA datasets converge at high force, to the behavior $x \approx \ln f$.

the Yamakawa-Shimada loop formation probabilities; don't forget that the kink can appear at any base pair position along the molecule).

Fig. 5 shows how the juxtaposition probability is affected by the inclusion of flexible joints with energy cost 9, 10, 11 and 12 k_BT, via a detailed calculation [13]. The experimental data are described well by joints which cost 11 k_BT, close to the value expected for localized strand separation of a few base pairs.

3.2. Stretching single-stranded nucleic acids

Single-stranded DNA has also been studied in single-molecule stretching experiments, and shows remarkably different polymer elasticity from double-stranded DNA, following from the factors:
ssDNA has twice the contour length per base of the double helix since the helical backbones of the double helix contain about 0.7 nm per base, about half of the

double helix contour length of 0.34 nm per base pair,
ssDNA has a persistence length of roughly a nanometer since the stiffness of the double helix is generated by the base pairing and stacking; once isolated, the ssDNA backbone is very flexible,
ssDNA can stick to itself by base-pairing and stacking interactions between bases along the same molecule.
These features are illustrated in Fig. 7 which plots experimental data for double helix and ssDNA side by side. The double helix, with a persistence length of 50 nm, is extended to its full contour length of about 0.34 nm/bp by forces of a few pN, and then shows a stiff force response, and finally the ≈ 60 pN force plateau. By comparison, ssDNA (open circles) only gradually stretches out, showing no stiff response near 0.34 nm/bp, and no force plateau. The force required to half-extend ssDNA is more than 3 pN; this reflects its short persistence length ≈ 1 nm (recall that the force needed to stretch out a polymer is roughly k_BT/A).

Fig. 7 also shows the strong dependence of ssDNA on salt concentration (open circles, left and right branches). At 150 mM NaCl ('physiological' salt, left set of data), ssDNA sticks to itself at low extensions, leading to an ≈ 1 pN force threshold to start opening the molecule. At low salt concentration (10 mM NaCl, right set of data) electrostatic self-repulsion eliminates this sticking effect, and the force threshold for initial extension.

For low salt concentration, the extension is well described by a logarithmic dependence on force, $\ln f/f_0$. This behavior can be understood in terms of a scale-dependent persistence length resulting from electrostatic effects. At low forces, electrostatic self-repulsion effectively stiffens the polymer, helping to stretch it out; at higher forces, this effect is less pronounced (the monomers are farther away from one another) and the chain becomes harder to stretch. This effect is much more pronounced for ssDNA than for dsDNA since the backbone persistence length ≈ 1 nm is comparable to, or even less than, the screening length for electrostatic interactions (recall the Debye screeing length is $\lambda_D = 0.3\text{ nm}/\sqrt{M}$ for NaCl at M Mol/litre).

Problem: Force-extension data of Fig. 7 at low ionic strength are described by $x(f) \approx x_0 \ln(f/f_0)$ where x_0 and f_0 are constants. Use the high-force limit calculation of the force-extension response for the semiflexible polymer 2.22, and the scale-dependent persistence length

$$A(q) = \begin{cases} D/q & q < q_0 \\ 0 & q > q_0 \end{cases} \tag{3.3}$$

to obtain a similar force-extension response. A more realistic model of scale-dependence of persistence length, based on Coulomb self-interactions, gives rise to similar behavior; see Refs. [14, 15].

3.3. Unzipping the double helix

We now have all the pieces to analyze unzipping of the double helix by a force which pulls the two strands apart (Fig. 3.3). We will compare the free energy of two paired bases, g, to the free energy at constant force for two unpaired and extended bases. The free energy per base is given via 2.24 as

$$\gamma(f) = \int_0^f df' x(f') \tag{3.4}$$

where $x(f')$ is the length per base of the ssDNA data of Fig. 7. The function $\gamma(f)$ increases with f. The threshold for unzipping occurs when this two times this free energy – for the two bases – equals the base-pairing energy g:

$$2\gamma = g \tag{3.5}$$

Treating the ssDNA as a harmonic 'spring' we can write $\gamma(f) \approx (\ell f)^2/(2k_BT)$ where $\ell \approx 0.4$ nm (this roughly matches the integral of the 150 mM force curve of Fig. 7 for forces below 20 pN), gives an unzipping force:

$$f = \frac{\sqrt{k_BTg}}{\ell} \tag{3.6}$$

Plugging in g from 1 to 4 k_BT, we see that the unzipping force varies from 10 to 20 pN, depending on sequence. Experiments of Bockelmann and Heslot have observed this range of forces in double-helix unzipping experiments [16].

Problem: For the harmonic model of ssDNA extensibility, calculate the force-extension relation. Compare the results for forces between 1 and 20 pN with the 150 mM NaCl ssDNA data in Fig. 7.

Problem: We can alternately describe unzipping using *extension* as a control parameter. Suppose one has a partially unzipped dsDNA, where n base pairs have been separated. The free energy is made up of elastic stretching energy, and base pairing energy:

$$F = \frac{k_BT(2x)^2}{2(2n)b^2} + ng \tag{3.7}$$

Note that opening n base pairs results in a $2n$-base-long ssDNA (see Fig. 3.3). Note also that $n \geq 0$. Find the equilibrium number of base pairs unzipped, as a function of extension. For a partially unzipped molecule, also calculate the *fluctuation* in the number of bases that are unzipped. What are the corresponding *extension* fluctuations?

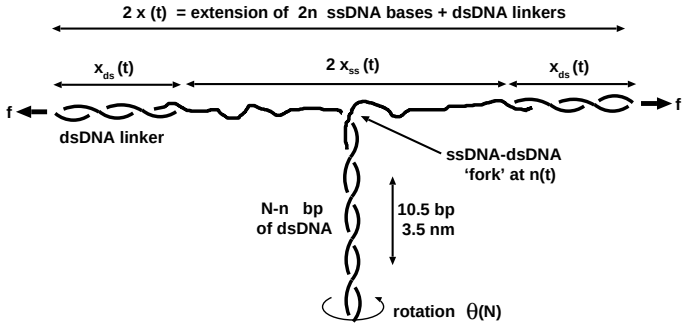

Fig. 8. Unzipping of DNA by force. Note that a torque can be applied to the end of the dsDNA region, coupled to the rotational angle θ.

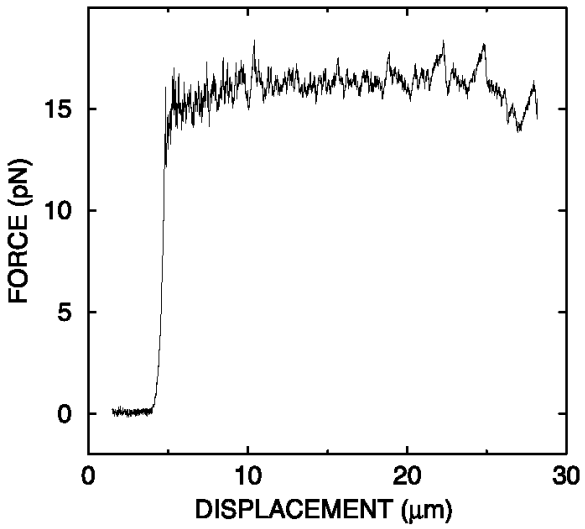

Fig. 9. Experimental data of Bockelmann et al [16] for unzipping of DNA at 0.02 μm/sec. Sequence-dependent variations in force occur, around an average force of about 15 pN

3.3.1. Effect of torque on dsDNA end

As unzipping proceeds, the dsDNA region must rotate to allow the two ssDNAs to be pulled out. If a torque is applied at the end of the dsDNA region, it can affect the unzipping force. This rotation is $\theta_0 = 2\pi/10.5 = 0.60$ radians per base pair unzipped. Adding the work $\tau\theta_0$ that must be done against the torque for each unzipped base pair, the equation for unzipping becomes

$$2\gamma = g - \tau\theta_0 \tag{3.8}$$

For the sign convention of Fig. 3.3, right-handed torque reduces the stability of the double helix, while left-handed torque acts to stabilize it. Using our harmonic approximation, we can obtain a torque-dependent unzipping force:

$$f = \frac{\sqrt{k_B T(g - \tau\theta_0)}}{b} \tag{3.9}$$

As torque becomes more positive, the unzipping force threshold is decreased. When the torque becomes positive enough to unwind the DNA on its own, the unzipping force threshold becomes zero: this point is given by $\tau = g/\theta_0$, which ranges from $1.6k_BT$ for weakly bound (AT-rich) sequences, to $7k_BT$ for the most strongly bound (GC-rich) sequences.

If unzipping is done rapidly, the rotation of the dsDNA will generate a drag torque. In the simplest model for this where the DNA is supposed to spin around its axis, the drag torque is roughly

$$\tau = -4\pi\eta r^2 L_{\rm ds}\frac{d\theta}{dt} \tag{3.10}$$

where $L_{\rm ds}$ is the length of the dsDNA region, $r \approx 1$ nm is the dsDNA cross-section hydrodynamic radius, and viscosity $\eta = 10^{-3}$ Pa·sec for water and most buffers. Effects of the drag associated with dsDNA rotation have been observed in experiments of Bockelmann and Heslot (see Fig. 3.3.1) [17]. P. Nelson has argued that there is an appreciable contribution to the rotational drag by permanent bends along the DNA contour [20]. The shape of the DNA gives rise to an effective increase in its cross-section radius r and thus the rotational drag coeffi cient.

Problem: Estimate the number of base-pairs per second that should be unzipped in order that rotational drag can push the unzipping force up by 5 pN (assume a uniform molecule with $g = 2.5k_BT$).

3.3.2. Fixed extension versus fixed force for unzipping

In single-molecule stretching experiments, like any experiment on a small system, choosing whether force or extension are controlled can be critical to the

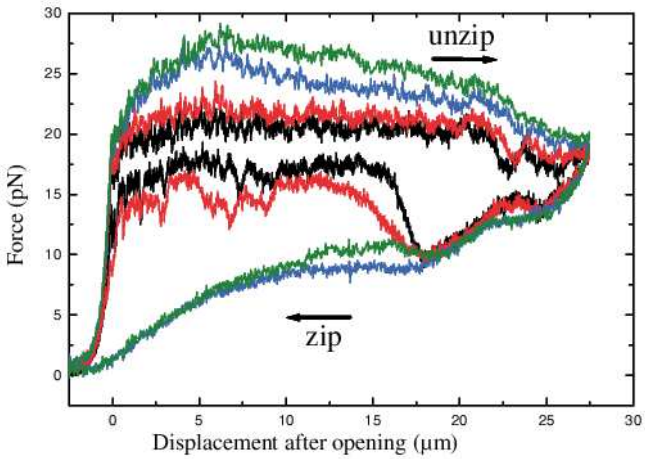

Fig. 10. Experimental results of Ref. [17] showing unzipping force rate-dependence. The two ssDNA ends are forced apart at velocities of 4, 8, 16 and 20 μm/sec. Force versus ssDNA extension (see Fig. 3.3) is plotted. During unzipping, higher velocities generate higher unzipping forces.

results of an experiment. For example, laser tweezers and atomic force microscopes essentially control the position of the end of a molecule; magnetic tweezer setups by contrast provide fixed force. Unzipping of DNA provides a very good example of how these two types of experiments give different kinds of data. Fixed-extension unzipping experiments push the ssDNA-dsDNA 'fork' along, and observe jagged force 'stick-slip' events. Each stick event corresponds to the momentary stalling of the fork at a GC-rich 'barrier': the force then increases to a level where a 'slip', or barrier-crossing event occurs.

Conversely, in a fixed-force experiment, one observes the increase of extension as a function of time. For unzipping, this typically takes the form of a series of extension *plateaus*. These plateaus again correspond to the stalling of the fork at a GC-rich barrier region; however, now the force is constant, and one must wait for a thermal fluctuation for unzipping to proceed. If one is well below the maximum unzipping force for GC-rich sequences (see 3.6), the barriers can be immense: even a fraction of a k_BT per base pair required to cross a long, slightly GC-rich region can give rise to an immense barrier. This effect has been theoretically emphasized by D. Lubensky and D. Nelson [18] and the constant-force extension plateaus have been observed in experiments by the group of Prentiss et al [19].

4. DNA Topology

The topological properties of DNA molecules are important biologically. The linking number of the two strands in the double helix is particularly important to DNA structure in bacterial cells, and controls 'supercoiling', or wrapping of the double helix around itself. The entanglement of the double helix with itself (knotting), and with other molecules (braiding) is also important since DNA molecules (chromosomes) must be separated from one another during cell division.

4.1. DNA supercoiling

The phenomenon of supercoiling is familiar from dealing with twisted strings or wires: twist strain in a string can be relaxed by allowing the string to wrap around itself. For DNA molecules, description of this behavior requires one more ingredient, thermal fluctuation of the molecule conformation.

The physical feature of the double helix that gives rise to supercoiling is the wrapping of the two strands around one another. Neglecting bending for the moment, the relaxed double helix has one link between strands for each 10.5 bp along the molecule. This 'relaxed linking number' can be expressed as $\mathrm{Lk}_0 = N/10.5$ bp for an N-bp double helix. The relaxed helix repeat of 10.5 bp can be expressed as a length $h = 3.6$ nm, allowing us to also write $\mathrm{Lk}_0 = L/h$.

4.1.1. Twist rigidity of the double helix

Still avoiding bending, if we twist the double helix so that one end is rotated by an angle Θ relative to the other, the number of links between the strands will be changed by an amount $\Theta/(2\pi)$. In this case where there is no bending, the change in linking number of the double helix, $\Delta\mathrm{Lk}$, equals the change in twist, $\Delta\mathrm{Tw}$.

It costs some energy for this twist distortion: a simple harmonic model is

$$\frac{E}{k_B T} = \frac{C}{2L}\Theta^2 = \frac{2\pi^2 C}{L}(\Delta\mathrm{Tw})^2 \tag{4.1}$$

This 'twist' energy is controlled by an elastic constant C with dimensions of length. This *twist persistence length* is about 100 nm for double helix DNA based on recent single-molecule experiments [26]; note that this is appreciably larger than the estimate of $\approx$ 75 nm that is the result of a number of solution-phase experiments. We'll see a possible explanation for this disagreement later when we discuss twist rigidity of DNA.

Problem: Consider the harmonic twist energy. Calculate the thermal expectation value of Θ^2: your result will depend on the molecule length L. Why is C called the twist persistence length?

Problem: Assuming the double helix to be composed of a uniform isotropic elastic medium, use A and C to determine the two Lamé coeffi cients, and equivalently the Young modulus and the Poisson ratio (you will want to consult to Landau and Lifshitz' *Theory of Elasticity* [25] unless you are really an expert in elastic theory; also recall that we have already fi gured out the Young modulus from both the bending persistence length *and*, independently, from the stretching force constant).

Problem: What torque is necessary to twist a DNA of length L by angle Θ? For left-handed twisting, for what angle Θ will the twisting build up enough torque to start unwinding AT-rich sequences (see Sec. 3.3.1)?

4.1.2. Writhing of the double helix

When we allow bending of the double helix to occur, the linking number is no longer equal to the twisting number. However, as long as the bending radius is large compared to the radius of the double helix, there is a simple relation between twisting and bending contributions to the total linking number of the double helix:

$$\Delta\mathrm{Lk} = \Delta\mathrm{Tw} + \mathrm{Wr} \tag{4.2}$$

The quantity Wr, or 'writhe', is dependent only on the bending of the double helix backbone. Very roughly, Wr measures the signed number of crossings of the molecule axis over itself, when the molecule shape is projected onto a plane.

Formally, linking number of the two strands can only be defi ned if the double helix is *circular*, i.e. if both of the strands are closed circles. I will be slightly loose with this, and sometimes talk about linking number of *open* molecules. If you want to make linking number of a linear molecule precise, you can just imagine extending the strand ends straight off to infi nity, and closing them there. This will not lead to large corrections in the situations we will be interested in.

We consider the situation where as bending occurs, the linking number remains fi xed. This is most relevant to *circular* double helix molecules (with no breaks or 'nicks' in their backbones), the linking numbers of which are constant. Circular DNAs are found in bacteria: both the large 4.5 Mb chromosome and small plasmids (typically 2 to 15 kb in circumference) are normally found in closed circular form.

A second situation where $\Delta\mathrm{Lk}$ can be considered constant is when one is holding onto the two ends of a DNA molecule, and forcing them to be parallel

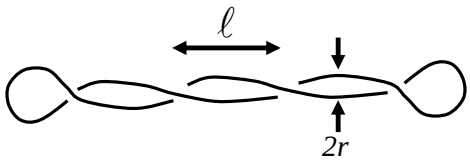

Fig. 11. Plectonemic supercoiled form of circular DNA, showing length between crossings ℓ, and cross-sectional radius r. Note that the only appreciable DNA bending occurs at the ends.

and unable to rotate. This case can be studied experimentally in single-DNA micromanipulation experiments, most notably in elegant magnetic tweezer experiments [22].

By rearranging 4.2 to $\Delta\mathrm{Tw} = \Delta\mathrm{Lk} - \mathrm{Wr}$ we can see the mechanism for buckling of a twisted wire: twisting without bending will change $\Delta\mathrm{Tw}$ away from zero, costing twist energy. However, now if the wire is allowed to buckle so that it wraps around itself, the Wr from the wrapping can cancel the $\Delta\mathrm{Lk}$, and reduce the twisting energy. By braiding the molecule with itself, the bending energy can be small as well. This self-wrapping of DNA is called *plectonemic supercoiling*.

For the plectonemic structure shown above, the magnitude of the writhe is and equal to the number of crossings: $|\mathrm{Wr}| \approx L/(2\ell)$. The sign of the writhe for the right-handed coiling shown in Fig. 4.1.2 is negative; for a left-handed plectonemic supercoil, the writhe would be positive. For achiral conformations, $\mathrm{Wr} = 0$.

4.1.3. Simple model of plectonemic supercoiling

We can write down a simple model for the free energy of the plectoneme:

$$\frac{F}{k_B T} = \frac{2\pi^2 C}{L}\left(\Delta Lk \pm \frac{L}{2\ell}\right)^2 + \frac{AL}{2}\left(\frac{r}{\ell^2}\right)^2 + \frac{L}{(Ar^2)^{1/3}} \tag{4.3}$$

The first term is just 4.1 with 4.2 rearranged and plugged in, using the plectonemic writhe $\mathrm{Wr} = \mp L/2\ell$; the top sign is for a right-handed plectnome, the bottom is for left-handed. The second term is the bending energy 2.5, using r/ℓ^2 as the curvature.

The third term arises from confinement of the DNA inside the 'tube' of the supercoil, of radius r. We can think about this in terms of a correlation length

λ for thermally excited bending fluctuations: the smaller this wavelength, the smaller the transverse fluctuations. For bending with transverse displacement r over wavelength λ, the curvature is r/λ^2; the energy of this bend is k_BTAr^2/λ^3. Using the equipartition theorem this energy will be k_BT, giving us the relation $\lambda = A^{1/3}r^{2/3}$. Finally, the confinement free energy density will be k_BT/λ, giving the third term of 4.3.

The free energy model 4.3 needs to be minimized to determine the equilibrium values of r and ℓ. First, we can determine r:

$$r \approx \frac{\ell^{3/2}}{A^{1/2}} \tag{4.4}$$

Then we can plug this result in to 4.3; simplifying some numerical factors we have

$$\frac{F}{k_BTL} = 2\pi^2 C \left(\frac{|\Delta \mathrm{Lk}|}{L} - \frac{1}{\ell}\right)^2 + \frac{1}{\ell} \tag{4.5}$$

The sign has been chosen so that the writhe has the same sign as $\Delta\mathrm{Lk}$, which always reduces the free energy. Minimizing this with respect to $1/\ell$ gives the result:

$$\frac{1}{\ell} = \frac{|\Delta \mathrm{Lk}|}{L} - \frac{1}{4\pi^2 C} \tag{4.6}$$

There is no solution for positive ℓ when linking number is too small: when $|\Delta\mathrm{Lk}| < L/(4\pi^2 C)$, the confinement free energy is too expensive, so the DNA does not supercoil. Then, as $|\Delta\mathrm{Lk}|$ is increased beyond this limit, $1/\ell$ becomes gradually smaller and the supercoil tightens up. This threshold indicates that until the added linking number exceeds one per twist persistence length, the DNA molecule will not supercoil.

Linking number is often expressed intensively using $\sigma \equiv |\Delta\mathrm{Lk}|/\mathrm{Lk}_0$ which just normalizes the change in linking to the relaxed linking number. In a more careful calculation where numerical factors and geometrical details are accounted for carefully, the threshold for supercoiling is at $\sigma \approx h/(2\pi C)$; plugging in $h = 3.6$ nm and $C = 100$ nm gives a threshold σ of roughly 0.01. Another feature of the more complete theory is that the transition is 'first-order': the minimizing ℓ jumps from $\ell = \infty$ to a finite value. Electron microscopy experiments [23] indicate that plectonemic supercoiling requires about this level of σ (see Fig. 4.1.3); calculations of structural parameters of plectonemes also are in accord with the results of EM studies. In eubacteria such as E. coli, the chromosome and small circular 'plasmid' DNAs have nonzero $\Delta\mathrm{Lk}$, with a $\sigma \approx -0.05$. This

undertwisting is thought to play a role in gene regulation, since AT-rich promoter regions will be encouraged to open by the torsional stress associated with this amount of unlinking.

An important feature of plectonemically supercoiled DNA is its *branched* structure. Branch points can be thought of as defects in the plectonemic supercoil structure: like the ends, there is some energy cost associated with them. However, there is an entropy gain $\approx k_B \ln L/A$ of having a branch point, since it can be placed anywhere in the molecule. Balance of branch point energy and entropy determines the observed density of a Y-shaped branch point for every 2 kb along a supercoil with $\sigma = -0.05$. Branching is also very important to the internal 'sliding' of DNA sequence around in the interior of a plectonemically supercoiled DNA, is important to some enzymes which bind to two sequences simultaneously, often across a plectnomemic superhelix [24].

Problem: For the plectonemic supercoil model discussed above, find the dependence of $\Delta\mathrm{Tw}/\Delta\mathrm{Lk}$ on $\Delta\mathrm{Lk}$ and σ (note that thermal fluctuations cause incomplete writhe-compensation of the twist energy cost of linking number).

Problem: Find the dependence of r on $\Delta\mathrm{Lk}$ for the model of plectonemic supercoiling discussed above. At what value of σ does r reach $2nm$, roughly the point at which the double helix will run into itself? Qualitatively, what will start to happen to the ratio $\Delta\mathrm{Tw}/\Delta\mathrm{Lk}$ for linking numbers significantly beyond this point?

Problem: Calculate the *torque* in a DNA double helix of length L, as a function of $\Delta\mathrm{Lk}$, for the plectnonemic supercoil described above. For $\sigma < 0$, at what value of σ does unwinding of AT-rich sequences in the double helix start to occur?

4.2. Twisted DNA under tension

It is possible to carry out single-DNA experiments as a function of force, and linking number [22]. The description of this situation along the framework of Sec. 2.3 is quite straightforward. At a fixed force, changing σ away from zero compacts a DNA molecule. If enough torsional stress is placed on a DNA molecule under tension, buckling will occur and plectonemic supercoils will appear along its length, leading to strong reduction in molecule extension. [14,28], as has been observed experimentally [22]. However, if σ is not so large that plectonemic coils appear, a milder compaction occurs which can be treated using a small-fluctuation approach as discussed by Moroz and Nelson [26], and by Bouchiat and Mezard [27]. This region of relatively mild compaction by writhing is a good regime in which to measure the double helix torsional modulus.

To carry out the treatment of this mild compaction analytically, we need the

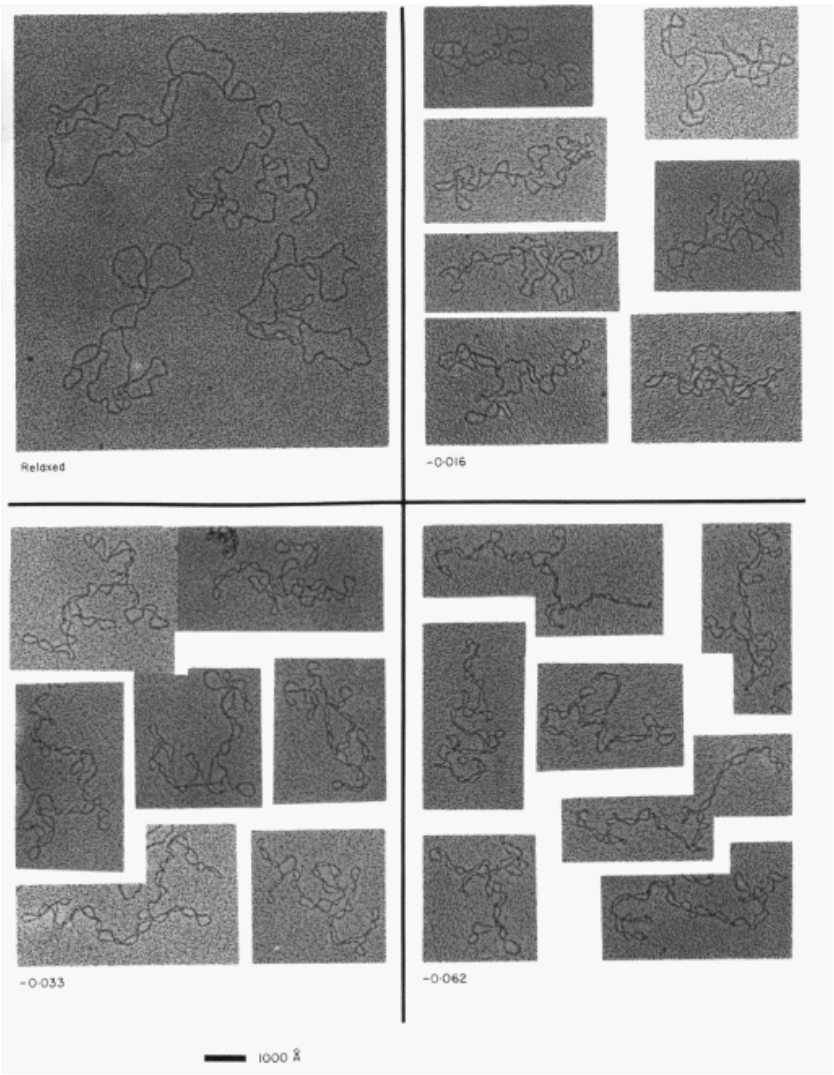

Fig. 12. Electron micrographs of supercoiled DNA at a few different σ values. Scale bar is 100 nm (300 bp); molecules are all 7 kb (2300 nm) in length. Reproduced from Ref. [23].

writhe of a nearly straight DNA, in terms of tangent vector fluctuations. As long as the tangent vector stays in the hemisphere around $\hat{\mathbf{z}}$, we have:

$$\mathrm{Wr} = \int_0^L \frac{ds}{2\pi} \frac{\hat{\mathbf{z}} \cdot \hat{\mathbf{t}} \times \partial_s \hat{\mathbf{t}}}{1 + \hat{\mathbf{z}} \cdot \hat{\mathbf{t}}} \tag{4.7}$$

Now we can write the Hamiltonian for a DNA subjected to tension, plus held at fixed linking number, by just adding the twist energy 4.1 to the stretching Hamiltonian 2.14. The White formula 4.2 allows us to express the twist in terms of linking number and writhe:

$$\begin{aligned} \frac{E}{k_B T} &= \int_0^L ds \left[\frac{A}{2} \left(\frac{d\hat{\mathbf{t}}}{ds} \right)^2 - \frac{f}{k_B T} \hat{\mathbf{z}} \cdot \mathbf{t} \right] \\ &\quad + \frac{2\pi^2 C}{L} \left(\Delta\mathrm{Lk} - \int_0^L \frac{ds}{2\pi} \frac{\hat{\mathbf{z}} \cdot \hat{\mathbf{t}} \times \partial_s \hat{\mathbf{t}}}{1 + \hat{\mathbf{z}} \cdot \hat{\mathbf{t}}} \right)^2 \end{aligned} \tag{4.8}$$

We'll do a harmonic calculation, expanding 4.8 to quadratic order in $\mathbf{u}$, the transverse (xy) components of the tangent vector:

$$\begin{aligned} \frac{E}{k_B T} &= \frac{2\pi^2 C}{L} (\Delta\mathrm{Lk})^2 - \frac{Lf}{k_B T} \\ &+ \int_0^L ds \left[\frac{A}{2} \left(\frac{d\mathbf{u}}{ds} \right)^2 + \frac{f}{2k_B T} \mathbf{u}^2 - \frac{2\pi C \Delta}{h} \sigma \hat{\mathbf{z}} \cdot \mathbf{u} \times \partial_s \mathbf{u} \right] + \mathcal{O}(\mathbf{u}^\ni) \end{aligned} \tag{4.9}$$

Here the $\Delta\mathrm{Lk}$ in the cross term of the twist energy has been converted to the intensive linking number density σ. For $\sigma = 0$, we return to the high-extension limit of the stretched semiflexible polymer; for nonzero σ, the cross-product term will generate chiral fluctuations.

The fluctuation ($\mathbf{u}$-dependent) part of the quadratic Hamiltonian 4.9 can be rewritten in terms of Fourier components of $\mathbf{u}$:

$$\int \frac{dq}{2\pi} \left[\frac{A}{2} \left(q^2 + \frac{f}{k_B T A} \right) |\mathbf{u}_q|^2 + \frac{2\pi C \sigma}{h} i q \hat{\mathbf{z}} \cdot \mathbf{u}_q^* \times \mathbf{u}_q \right] \tag{4.10}$$

Problem: Show that the partition function $Z(f, \sigma)$ for the quadratic-u fluctuations, including the non-fluctuation contributions, has the form, in an expansion in inverse powers of force:

$$\frac{\ln Z}{L} = \frac{f}{k_B T} - \frac{2\pi^2 C \sigma^2}{h^2} - \left(\frac{f}{k_B T A} \right)^{1/2} + \left(\frac{k_B T}{A^3 f} \right)^{1/2} \left(\frac{2\pi C \sigma}{h} \right)^2 + \cdots \tag{4.11}$$

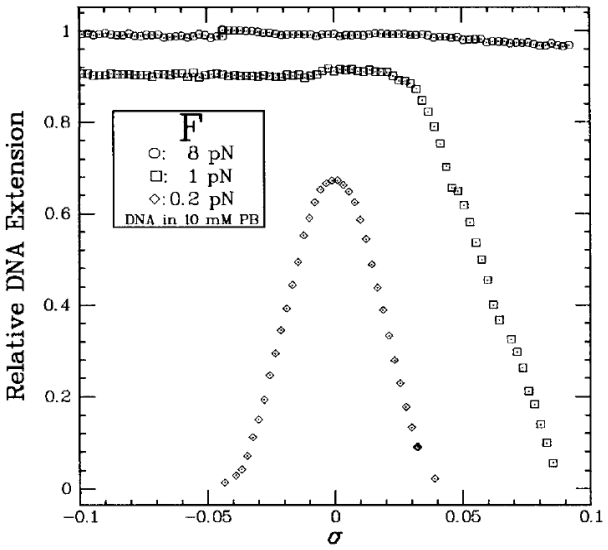

Fig. 13. Extension of DNA as a function of linking number σ, for a few fixed forces. Reproduced from Ref. [22].

You will need to find the normal modes of the fluctuations in order to compute the partition function.

The extension as a fraction of the total molecular length follows via 2.16, as:

$$\left\langle \hat{\mathbf{z}} \cdot \hat{\mathbf{t}} \right\rangle = 1 - \left(\frac{k_B T}{4Af} \right)^{1/2} - \frac{1}{2} \left(\frac{k_B T}{4Af} \right)^{3/2} \left(\frac{2\pi C \sigma}{h} \right)^2 + \cdots \tag{4.12}$$

For this model, twisting the DNA ($\sigma \neq 0$) leads to a reduction in extension. This can be seen in the data of Fig. 4.2; however, note that the quadratic twist-dependence occurs only quite near to $\sigma = 0$, and is clearest in the 0.2 pN data of the figure.

The free energy 4.11 can be used to find the relation between the torque applied at the end of the chain, and the linking number. Since linking number is controlled by rotating the end of the chain, the torque applied at the end of the

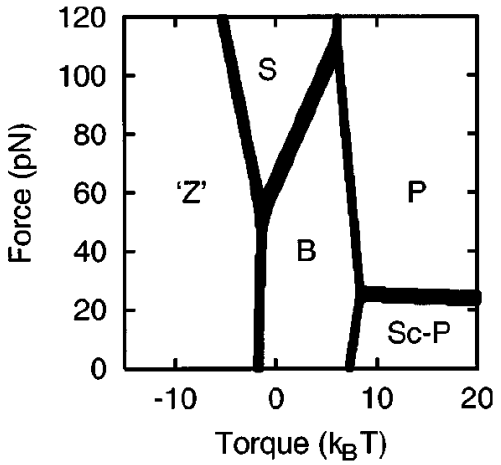

Fig. 14. 'Phase diagram' of double helix as function of external torque and force. Reproduced from Ref. [29].

chain is obtained from the linking number derivative of 4.11:

$$\frac{\tau}{k_B T} = -\frac{1}{2\pi}\frac{d}{d\Delta \mathrm{Lk}} \ln Z = \left[1 - \left(\frac{k_B T}{A^3 f}\right)^{1/2}\right] \frac{2\pi C\sigma}{h} \tag{4.13}$$

Moroz and Nelson have emphasized this result: the effective twist modulus of the double helix goes down as tension in the chain goes down. This effect occurs because for lower forces, more writhing can occur, allowing the twist energy and therefore the torque to be reduced.

4.3. High forces and torques cause structural changes in the double helix

The previous calculations have assumed that the forces and torques were not able to cause structural phase transitions in the double helix. We've already seen that at zero force a left-handed torque of about $-2k_BT$ is sufficient to start unwinding AT-rich regions of the double helix. Experiments show that there may be as many as five different structural states of the double helix which can be accessed by twisting and pulling on DNA [29], and those experiments allow one to predict a 'force-torque phase diagram' (Fig. 4.3). For forces in the < 50 pN range the double helix is stable roughly over the torque range $-2k_BT$ to $+7k_BT$ (the B-DNA region of Fig. [29]).

4.4. DNA knotting

Above we've discussed the effect of supercoiling, which is controlled by the 'internal' linking number of the two ssDNAs inside the double helix. A separate and important property of the double helix is the 'external' entanglement state of the double helix backbone, the more usual case of topology discussed in usual polymer physics. A single circular DNA can carry a knot along its length; alternately two or more circular DNAs can be linked together.

When an initially linear DNA is closed into circular form, there is some possibility that a knot is generated. You might be wondering why all the molecules of Fig. 4.1.3 are all *unknotted*. There are two reasons for this, the first one biological and the second one biophysical.

4.4.1. Cells contain active machinery for removal of knots and other entanglements of DNA

Cells contain enzyme machines called *topoisomerases* which catalyze changes in DNA topology. For example, entanglements (including knots) can be removed or added by *type-II topoisomerases* which are able to cut the double helix, pass another double helix segment through the resulting gap, and then seal the gap up. Type-II topoisomerases require ATP.

However, although existence of type-II topoisomerases tells us that it is *possible* for entanglements to be removed, we still are left wondering how they 'know' how to remove rather than add entanglements. Astonishingly, it has been experimentally demonstrated that type-II topoisomerases by themselves have the capacity to recognize and remove knots and other entanglements along circular DNA molecules [30].

4.4.2. Knotting a molecule is surprisingly unlikely

Let's suppose we take an ensemble of linear DNA molecules of length L at low enough concentration that they do not interact with one another. Now, let's add a small quantity of an enzyme which catalyzes closure of the molecules into circles, and the reverse process (this is possible). Then we can ask what the probability $P_{\mathrm{unknot}}(L)$ is that the molecule is unknotted, as a function of L.

We can argue that $P_{\mathrm{unknot}} \approx \exp -L/(N_0 A)$, for some constant N_0. For small L, there will be a large free energy cost of closing a molecule into a circle making $P_{\mathrm{knot}} \to 1$. However, for larger molecules, the probability of an unknotted configuration should go down. The exponential decay reflects the fact that over some length (N_0 persistence lengths) the probability of having no knot drops to $1/e$: applying this probability to each L_0 along a DNA of length L gives us $P_{\mathrm{unknot}}(L) \approx (1/e)^{L/(N_0 A)}$. This super-rough argument can be made

mathematically rigorous [31].

What is the number N_0? It turns out that even for an 'ideal' polymer which has no self-avoidance interactions, $N_0 \approx 600$. What this means is that to have a 50% probability to find even one knot along a dsDNA, it has to be $600 \times 300 = 180,000$ bp long! (note that the long persistence length of DNA is a big help in making this number so impressive). Even more incredibly, for a self-avoiding polymer, $N_0 \approx 10^6$! This remarkable fact is theoretically understood only on the basis of numerical simulations: see Ref. [32].

Experiments on circular DNAs are in good quantitative agreement with statistical mechanical results for the semiflexible polymer model including DNA self-avoidance interactions. For example, it is found that the probability of finding a knot generated by thermal fluctuations for a 10 kb dsDNA is only about 0.05. [33, 34]. As mentioned above, topoisomerases are by themselves (using ATP) able to push this probability down, by a factor of between 10 and 100 [30].

4.4.3. Condensation-resolution mechanism for disentangling long molecules

Although topoisomerases seem to be able to help get rid of entanglements, there must be other mechanisms acting in the cell to completely eliminate them. Here I'll mention a very simple model that might give you an idea of how this machinery works.

Suppose we have a long dsDNA of length L, in the presence of some proteins which act to fold DNA up along its length. We imagine that these proteins cannot 'cross-link' DNA segments, but that they can only compact the molecule along its length. As these proteins bind, we imagine that they modify the total contour length to be $L' < L$, and the effective persistence length to be $A' > A$.

If these proteins bind slowly in the presence of type-II topoisomerases so that the knotting topology can reach close to equilibrium, then the unknotting probability will have the form:

$$P_{\rm unknot} = \exp\left(-\frac{L'/A'}{N_0}\right) \tag{4.14}$$

As you can see, gradually compacting (decreasing L') while stiffening (increasing A') DNA can drive knotting out of it; unknotting ('entanglement resolution') will occur on progressively larger length scales as this condensation process proceeds.

Problem: Consider a condensation process which gradually condenses DNA a DNA of length L by folding it along its length, to make a progressively thicker fiber, of length L' and cross-section radius r'. If volume is conserved during condensation, and if the effective Young modulus of the fiber is a constant, find

the unknot probability as a function of L'/L.

This simple model gives some idea of how proteins which structure DNA can play a role in controlling its entanglement at short enough scales where entanglements can be thought of using statistical mechanics. However, at the large scale of a whole chromosome cross-links do occur: they are necessary to fit the chromosome into the cell! It is also possible to envision a process where chromosome condensation by cross-linking also can drive out entanglements as long as the cross-links are transient [35].

5. DNA-Protein Interactions

So far we mostly talked about DNA by itself in buffer, focusing on the double helix's physical properties. In the cell, DNA is covered with proteins. You can think of DNA as a long string of thickness 2 nm, and the proteins as little particles of diameters of 1 to 10 nm plastered all along the string's length. The DNA plus all the proteins bound to it make up the biologically active chromosome.

Some proteins which bind to DNA are primarily *architectural*, folding and wrapping DNA so as to package it inside the cell. Other proteins have primarily *genetic* functions, interacting with particular sequences. Of course, these two functions can be mixed: as examples proteins which tightly fold up DNA will likely repress gene expression; the expression of genes likely cannot occur without changes in DNA folding architecture.

Proteins which interact with DNA tend to be sorted into two groups:
Sequence-nonspecific proteins that stick anywhere along the double helix;
Sequence-specific proteins that bind to particular sequences very strongly, and to other sequences only relatively weakly.

Most proteins involved in chromosome architecture have a mainly nonspecific interaction with DNA. Such interactions are often electrostatic in character, and can be disrupted with high salt concentrations. Examples are the histone proteins of the nucleosome, and nonspecific DNA-bending proteins such as HU from E. coli or HMG proteins from eukaryote cells. Note that nonspecifically-interacting proteins usually bind better to some sequences than others, but not a whole lot better. Under physiological conditions, nonspecifically interacting proteins usually bind to DNA once their concentration is in the range 10 to 1000 nM.

Sequence-specific interactions occur via chemical interactions which depend on the structure of the bases. These are not usually electrostatic in character (most of DNA's charge is on the phosphates, which are common to all the bases) although most sequence-specific proteins also have a nonspecific and electrostatic

interaction with DNA. Examples of sequence-specific interactions include transcription factors and restriction enzymes. Sequence-specific proteins can often bind their targets at concentrations well below 1 nM. This high level of affinity is necessary: the concentration of transcription factors in E. coli can be as little as one per cubic micron, which in molar units is $(1/6 \times 10^{23}\ \mathrm{Mol})/(10^{-15}\ \mathrm{litre}) = 1.6$ nM. In the human cell with a nucleus of volume $\approx 10^3\ \mu\mathrm{m}^3$, affinities in the picomolar range are needed to bind sequence-specific proteins to their targets with reasonably high probability.

5.1. How do sequence-specific DNA-binding proteins find their targets?

We might ask the question of how long it takes for a protein to find a single specific target in a large DNA molecule. In fact, there is a long history of experiments studying this in the test tube: the model system for this has been the E. coli protein lac repressor. Let's follow one protein of diameter d as it moves by diffusion to a target of size a; we'll suppose that the targets are present in solution at concentration c.

5.1.1. Three-dimensional diffusion to the target

In the absence of any nonspecific interaction effects, the protein will diffuse through space until it hits the target. To analyze how long it takes for the protein to find one target, let's divide space up into boxes of volume $V = 1/c$ each containing one target. We'll further divide each box up into 'voxels' of volume a^3; one voxel is the target. Since the protein moves by diffusion, its trajectory will be a random walk in the box. The number of steps at scale a that must be taken to move across the box (of edge $V^{1/3}$) is $V^{2/3}/a^2$; the probability of finding the target before the protein leaves the box is therefore $a/V^{1/3}$ (this result is often called 'diffusion to capture'). The time that this search occurs in is just the diffusion time $V^{2/3}/D$, where $D = k_B T/(3\pi\eta d)$ is the protein diffusion constant.

Once the protein leaves one box of volume V, the same search starts over in an adjacent box: this will occur $V^{1/3}/a$ times before a target is found. So, the total time required for our protein to find a target by simple three-dimensional diffusion is

$$\tau_{3\mathrm{d}} = \frac{V^{1/3}}{a} \times \frac{V^{2/3}}{D} = \frac{V}{Da} \tag{5.1}$$

Note that this formula could apply in solution where there is one target per solution volume V, or to targeting in a cell compartment of volume V. In the latter case, the same volume is searched over and over until the target is found.

If we convert V to concentration c, our result is $1/\tau_{3\mathrm{d}} = 4\pi Dac$. The factor of 4π comes from a more detailed calculation of diffusion to capture, originally due to Smolochowski [36]. This rate is proportional to concentration; biochemists usually describe the rate of this type of bimolecular reaction by normalizing the actual rate by concentration, leaving the *association rate*:

$$k_{\mathrm{a}} = 4\pi Da = \frac{4\pi k_B T}{3\pi\eta}\frac{a}{d} \tag{5.2}$$

Since the target size is less than the protein size, we find that the maximum association rate is the prefactor $4\pi k_B T/(3\pi\eta) \approx 4 \times 10^{-18}$ m^3/s $\approx 10^8$ M^{-1}s^{-1}. This rate is often referred to as the 'diffusion-limited reaction rate'.

5.1.2. Nonspecific interactions can accelerate targeting

Experiments in the 1970s showed that lac repressor binds its target at closer to $k_a \approx 10^{10}$ M^{-1}s^{-1} which was initially quite a puzzle. However, Berg, Winter and von Hippel proposed and experimentally supported a solution to this paradox [37]. They realized that lac repressor also had a *nonspecific* interaction with DNA, and that this nonspecific interaction could make the target effectively larger than the protein!

The picture is that when lac repressor first hits DNA, the nonspecific interaction allows it to 'slide' randomly back and forth along the DNA, exploring a region of length ℓ_{sl} before it dissociated. Now, the target size is increased to be ℓ_{sl}, so the time we will have to spend doing three-dimensional diffusion will be reduced to $1/(4\pi D_{\mathrm{sl}}\ell_{\mathrm{sl}}c)$, where D_{sl} is the diffusion constant for the 'sliding' motion.

However, it is not yet clear if this will really accelerate k_a, since each sliding event will eat up a time of roughly $\ell_{\mathrm{sl}}^2/D_{\mathrm{sl}}$, and to be sure to find the target, L/ℓ_{sl} sliding events have to occur, requiring a total one-dimensional diffusion time of $L\ell_{\mathrm{sl}}/D_{\mathrm{sl}}$.

So, the total time required to find the target by this 'facilitated diffusion' process is

$$\tau_{\mathrm{fac}} = \frac{1}{4\pi D\ell_{\mathrm{sl}}c} + \frac{L\ell_{\mathrm{sl}}}{D_{\mathrm{sl}}} \tag{5.3}$$

If we write the ratio of this and the three-dimensional result 5.2, we obtain

$$\frac{k_{\mathrm{a,fac}}}{k_{\mathrm{a,3d}}} = \frac{\tau_{3\mathrm{d}}}{\tau_{\mathrm{fac}}} = \frac{\ell_{\mathrm{sl}}/a}{1 + 4\pi\frac{D}{D_1}\ell_{\mathrm{sl}}^2 Lc} \tag{5.4}$$

As long as $\ell_{\rm sl}$ is not too long, the nonspecific interaction does accelerate the reaction rate. This basic model and its experimental study are discussed in detail by Berg, Winter and von Hippel [37].

An interesting feature of 5.3 is that for fixed total DNA length L and target concentration $c = 1/V$ there is an *optimal* $\ell_{\rm sl}$:

$$\ell_{\rm sl}^* \approx \sqrt{V/L} \tag{5.5}$$

where we have dropped the 4π and the ratio of diffusion constants. For the *E. coli* cell, $V \approx 10^9$ nm^3 and $L \approx 10^6$ nm, indicating $\ell_{\rm sl}^* = 30$ nm = 100 bp. This is the sliding length inferred for lac repressor from biochemical experiments on facilitated diffusion [37]. This suggests that $\ell_{\rm sl}$ for lac repressor is optimized to facilitate its targeting *in vivo* [38].

5.2. Single-molecule study of DNA-binding proteins

A number of groups are working at present on experiments looking at protein-DNA interactions using single-DNA micromanipulation. The basic idea is to study proteins which change DNA mechanical properties, for example, by putting bends or loops into the double helix. Then, the binding of the proteins can be monitored via the force-extension response of the molecule. To give an idea of what can be obtained from this kind of study, I'll consider a few simple models.

5.2.1. DNA-looping protein: equilibrium 'length-loss' model

Consider a protein which binds to a double helix under tension f, resulting in a reduction in contour length available for extension by amount ℓ. If we suppose that the binding energy of the protein in k_BT units is ϵ and its bulk concentration is c, we can write down a simple model for the free energy of the protein-DNA complex as a functiona of the occupation of the protein n ($n = 1$ for protein bound, $n = 0$ for protein free in solution)

$$\ln Z_n = \frac{(L - n\ell)}{A}\gamma(\beta A f) + n(\epsilon + \ln vc) \tag{5.6}$$

The first term is the DNA stretching free energy: when the protein binds, the contour length of extensible DNA is reduced by ℓ. The last two terms give the free energy for a bound protein including the entropy cost of its removal from free solution. A factor of protein volume v is included to make the inside of the log dimensionless.

By just summing over the states of n, we can immediately find the probability for the protein to be bound:

$$P_{\rm on} = \left(\frac{K_d}{c} \exp \left[\frac{\ell}{A} \gamma(\beta A f) \right] \right)^{-1} \approx \left(1 + e^{\beta \ell f - \ln(c/K_d) + \cdots} \right)^{-1} \tag{5.7}$$

where we have defined $K_d = e^{-\epsilon}/v$, the 'dissociation constant' or concentration at which the protein is half-bound for zero force; the final term gives the large-force limit (recall $\gamma(x) = x + \sqrt{x} + \cdots$). If the complex is bound at low force ($c > K_d$), it will stay bound until a force threshold is reached. Roughly, this characteristic force is $f^* \approx \ln c/K_d/\ell$; beyond this force, the DNA-protein complex will open up. In an experiment where extension versus force can be used to monitor the stability of a protein-DNA complex, one can therefore make a measurement of the zero-force K_d – if one can observe equilibrium (on-off fluctuations) at the force-induced dissociation point [39]. Experiments of this general type have recently been done for the gal repressor protein, which is able to trap a DNA loop [40].

5.2.2. Loop formation kinetics

The previous section supposes that one can observe on-off fluctuations in the presence of force. However, it is possible, with very strong protein-DNA complex, that spontaneous dissociation will be essentially unobservable. In this case, one will observe the on-kinetics only. In the case of a small DNA loop, the complex formation rate will involve a barrier made of two components: the free energy cost of pulling in a length ℓ of DNA as discussed above, plus the free energy cost of making the DNA loop, as discussed in Sec. 2.2.5. Putting these together gives a loop formation rate:

$$k_{\rm loop} \approx k_0 \exp\left[-\beta \ell f + \ln J(\ell) \right] \tag{5.8}$$

where J is the juxtaposition 'J-factor' relevant to the reaction (see 2.11, 2.12). The exponential dependence on force will effectively shut off the reaction above $f^* \approx -k_B T \ln J/\ell$; below this force threshold the loop should form irreversibly. If a series of loop binding sites are available on a long DNA, the rate (5.8) will be proportional to the velocity at which the DNA end retracts due to loop formation [41].

PREVIOUS SECTION NEEDS MORE COMPLETE REFERENCES

5.2.3. DNA-bending proteins

FINISH THIS SECTION

Acknowledgements

These lectures include results of research done jointly with Eric Siggia, Didier Chatenay, Jean-Francois Légèr, Abhijit Sarkar, Simona Cocco, Sumithra Sankararaman, Dunja Skoko, Michael Poirier, Steven Halford, Monte Pettitt, and Michael Feig, whose many insights I gratefully acknowledge. I am also grateful for advice and help, including experimental data, from David Bensimon, Vincent Croquette, Terence Strick, Jean-Francois Allemande, Carlos Bustamante and Nick Cozzarelli. I thank NATO and the CNRS for making it possible for me to visit Les Houches to present these lectures. This work was supported in part by the US National Science Foundation, through Grants DMR-0203963 and MCB-0240998.

References

[1] M.D. Wang, M.J. Schnitzer, H. Yin, R. Landick, J. Gelles, S.M. Block, *Science* **282** 902-907 (1998).

[2] P.J. Hagerman Ann. Rev. Biophys. Biophys. Chem. **17** (1988) 265.

[3] J. Santalucia Jr., Proc. Natl. Acad. Sci. USA **95** (1998) 1460.

[4] T.E. Cloutier and J. Widom Mol. Cell **14** (2004) 355.

[5] A.M. Lane and D. Robson, Phys. Rev. **151** (1966) 774, and references therein.

[6] M. Rief, H. Clausen-Schaumann, H.E. Gaub, *Nat. Struct. Biol.* **6**, 346 (1999).

[7] J.-F. Leger, G. Romano, A. Sarkar, J. Robert, L. Bourdieu, D. Chatenay, J.F. Marko, *Phys. Rev. Lett.* 83, 1066 (1999).

[8] H. Clausen-Schaumann, M. Rief, C. Tolksdorf, H.E. Gaub, *Biophys. J.* **78**, 1997 (2001).

[9] J.-F. Leger, Ph.D. Thesis, l'Université Louis Pasteur Strasbourg I, Strasbourg France (2000).

[10] C. Bustamante, D. Smith, S. Smith, *Curr. Opin. Struct. Biol.* **10**, 279 (2000).

[11] S.B. Smith, Y.J. Cui, C. Bustmante, *Science* **271**, 795 (1996).

[12] J.R. Wenner, M.C. Williams, I. Rouzina, V.A. Bloomfield, *Biophys. J.* **82**, 3160 (2002).

[13] J. Yan, J.F. Marko, *Phys. Rev. Lett.* in press (2004).

[14] J.F. Marko, E.D. Siggia, *Macromolecules* **28** 8759 (1995).

[15] J. Yan, A. Sarkar, S. Cocco, R. Monasson, J.F. Marko, *Eur. Phys. J. E* **10**, 249-263 (2003).

[16] U. Bockelmann, B. Essevaz-Roulet, F. Heslot, *Biophys. J.* **82**, 1537 (2002).

[17] P. Thomen, U. Bockelmann, F. Heslot, *Phys. Rev. Lett.* **88**, 248102 (2002).

[18] D. Lubensky, D.R. Nelson, *Phys. Rev. Lett.* **85**, 1572 (2000); D. Lubensky, D.R. Nelson, *Phys. Rev. E* **65**, 031917 (2002).

[19] M. Prentiss, D. Lubensky, D.R. Nelson, *Proc. Natl. Acad. Sci. USA* (2003).

[20] P. Nelson, *Phys. Rev. Lett.* **80**, 5810 (1998).

[21] M. Feig, J.F. Marko, M. Pettitt, Microscopic DNA fluctuations are in accord with macroscopic DNA stretching elasticity without strong dependence on force field choice, *NATO ASI Series: Metal Ligand Interactions*, ed. N. Russo (Kluwer Academic Press), 193-204 (2003).

[22] T.R. Strick, J.-F. Allemand, D. Bensimon, V. Croquette, *Biophys. J.* **74** 2016 (1998).

[23] T.C. Boles, J.H. White, N.R. Cozzarelli *J Mol Biol.* **213**(1990).

[24] J.F. Marko, *Physica A* **296**, 289 (2001); J.F. Marko, *Physica A* **244**, 263 (1997).

[25] L.D. Landau, E.M. Lifshitz, *Theory of Elasticity* (Pergamon, NY, 1986) Ch. II.

[26] J.D. Moroz, P. Nelson *Proc. Natl. Acad. Sci. USA* **94**, 14418 (1997); P. Nelson, *Biophys. J* **74**, 2501 (1998); J.D. Moroz, P. Nelson, *Macromolecules* **31** 6333 (1998).

[27] C. Bouchiat, M. Mezard, *Phys. Rev. Lett.* **80** 1556 (1998); C. Bouchiat, M. Mezard, Eur. Phys. J. **E2**, 377 (2000).

[28] J.F. Marko, E.D. Siggia, *Science* **265**, 506 (1994).

[29] A. Sarkar, J.-F. Leger, D. Chatenay, J.F. Marko, *Phys. Rev. E* **63**, 051903 (2001).

[30] V.V. Rybenkov, C. Ullsperger, A.V. Vologodskii, N.R. Cozzarelli, *Science* **277**, 648 (1997).

[31] D.W. Sumners, S.G. Whittington, *J. Phys. A Math. Gen.* **21** 1689 (1988).

[32] K. Koniaris and M. Muthukumar *Phys. Rev. Lett.* **66** 2211 (1991).

[33] S.Y. Shaw, J.C. Wang *Science* **260**, 533 (1993).

[34] V.V. Rybenkov, N.R. Cozzarelli, A.V. Vologodskii, *Proc. Natl. Acad. Sci. USA* **11** 5307 (1993).

[35] Marko JF, Siggia, *Mol. Biol. Cell* **8** 2217 (1997).

[36] H.C. Berg, *Random Walk in Biology* (Princeton, Princeton, 1993)

[37] O.G. Berg, R.B. Winter, P.H von Hippel, *Biochemistry* **20** 6929 (1981); R.B. Winter, P.H. von Hippel, (1981) *Biochemistry* **20** 6948 (1981); R.B. Winter, O.G. Berg, P.H. von Hippel, *Biochemistry* **20** 6961 (1981).

[38] S.E. Halford, J.F. Marko *Nucl. Acids Res.* **32**, 3040 (2004).

[39] L. Finzi, G. Gelles *Science* **267** 378 (1995).

[40] G. Lia, D. Bensimon, V. Croquette, J.-F. Allemand, D. Dunlap, D.E. Lewis, S. Adhya, L. Finzi *Proc. Natl. Acad. Sci. USA* **20** 11373 (2003).

[41] S. Sankararaman, J.F. Marko, preprint (2004).