

Chapter 2

Sequence dependent elasticity of DNA : Implications for DNA-protein binding

2.1 Introduction

In this chapter, we study the interplay between the chemical kinetics of protein binding and the mechanics of the underlying filament. We focus our attention on binding of proteins to the DNA filament. In the process of binding, the protein distorts the DNA locally [1].

Both in prokaryotes and eukaryotes, initiation of genetic processes such as transcription, replication and site-specific recombination requires the successful binding of a variety of DNA-binding proteins [2]. In most cases, molecular binding to specific or non-specific regions of the DNA is accompanied by a local distortion of the DNA substrate. The binding modes of such molecules are usually investigated through ensemble *in-vitro* measurements; however in order to understand how complexation affects both the structural and mechanical properties of a single DNA, one needs to use a variety of single molecule (force spectroscopy) probes such as AFM [3].

Since DNA is a heterogeneous filament, one would like to understand whether the local DNA sequence affects the binding kinetics of the protein. We start with enumerating various examples where we believe the local DNA sequence significantly affects the binding of the proteins.

(a) Repressors regulate the expression of genes by decreasing the rate of transcription. In the experiments done with 434 repressor [1], it was shown that repressor binding affinity was reduced by at least 50-fold in response to the sequence changes of the underlying DNA substrate. Subsequent analysis of these experiments done in Ref. [4] showed that the variation of the local bending rigidity on the DNA sequence, can quantitatively explain the measured binding rate constants for the repressor.

(b) *Integration host factor* (IHF), a major component of the bacterial nucleoid, is a DNA-bending protein and functions as an architectural factor in prokaryotes [2]. IHF is a small heterodimeric protein that specifically binds to DNA through the sequence-dependent structure and distortability of the DNA rather than via direct side chain - base hydrogen bonds [5]. Based on crystal structure data, it is believed that the DNA is wrapped around the protein and bent by an angle in excess of 160° , thus reversing the direction of the helix axis within a very short distance [5].

(c) The eukaryotic architectural factor, the histone octamer [2], binds to DNA via a combination of (nonspecific) electrostatic and (specific) hydrophobic and direct side chain - base hydrogen bonds, inducing the DNA to wrap around it on a scale of 150 *bp*. The binding to the DNA involves specific local distortions of the DNA chain. Recent time resolved anisotropy

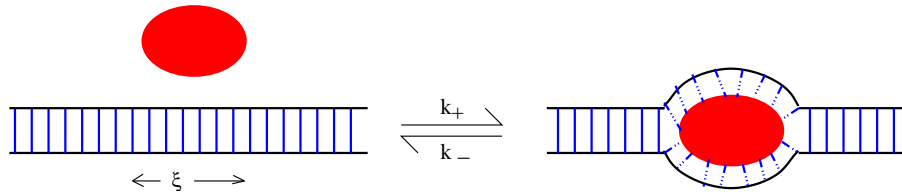


Figure 2.1: Sketch of a DNA-binding protein such as RNA-polymerase binding and distorting the DNA substrate over a scale ξ . The binding/unbinding rates k_+ , k_- depend on the distortion energy over this scale and hence on the sequence.

measurements indicate that the binding may be sequence dependent [6]; indeed short chain intrinsic bendability is highly sequence dependent [7].

(d) Proteins such as rec-A, transcription factors, gene regulatory proteins and RNA-polymerase [2] involved in replication and transcription, are widely *believed* to have two protein-DNA binding modes : a weak, *non-specific* binding to non-cognate DNA, usually *assumed* to be independent of the DNA sequence that the protein is bound to, and a strong *specific* binding at operator sites. For instance, RNA-polymerase, a 20 kD (kilo Dalton) protein, binds to DNA over a scale of 10 bp, the sequence independence of its non-specific binding is based primarily on low resolution experiments describing the ease with which these proteins bind to apparently random sites of the DNA. However even this non-specific binding of RNA-polymerase needs to locally open out and untwist the DNA over a scale of 10 bp — successful binding needs to overcome an energy barrier corresponding to the stiffness of the DNA over this scale.

(e) DNA-binding drugs such as ethidium bromide intercalate into DNA and alter the base-stacking interaction. Successful binding involves local prestretching and unwinding of the ds-DNA. The effects of intercalation has been observed in force-extension measurements [8], particularly in the overstretching plateau regime.

We believe these examples reveal a general principle regarding DNA-protein binding (Fig. 2.1):

1. The DNA substrate is heterogeneous.
2. On binding, the DNA gets distorted over a scale ξ .
3. The binding of the protein is sensitive to the DNA sequence over this scale.

The scale dependent stiffness of the DNA is a consequence of the relative distortions of neighbouring stacking plates and bend, twist and stretch of the sugar-phosphate backbone.

Thus DNA-binding proteins may be viewed as molecular probes which “measure” the local deformability or stiffness of the ds-DNA over the scale ξ . An immediate manifestation of the sequence dependent stiffness of the DNA over the scale ξ , will be in the measurable rates of binding/unbinding of DNA-binding proteins. Thus the binding/unbinding reaction rates of DNA-binding proteins will depend on the local conformational stiffness of the DNA and on the applied tension [4].

We therefore investigate the scale and sequence dependent mechanical properties of the DNA heteropolymer and study how sequence dependent elastic properties of a single filament affects the chemical kinetics of DNA-protein binding.

We study the ‘sequence’-distribution of thermally averaged global and local elastic properties of a DNA random heteropolymer of a fixed length N , within a simple elastic, worm-like chain (WLC) model. In section 2.3, we use a mapping to the disordered Heisenberg chain, to arrive at a number of qualitative results on the form of the distribution function of the thermally averaged end-to-end distance $\langle R^2 \rangle$, and its moments. We find that for long, $N \rightarrow \infty$, chains, this distribution is a gaussian; for shorter chains, there is a crossover to an exponential distribution, with the most probable end-to-end distance deviating significantly from the mean. Further, we find that the distribution of local quantities related to the thermally averaged tangent-tangent correlator are typically broad, even in the thermodynamic limit, *i.e.*, *they do not self average*. In section 2.4, we argue that this scale dependent ‘sequence’ sensitivity should have important biological implications, specifically for the binding of proteins to DNA — we present a simple model calculation of the binding/unbinding kinetics of DNA-binding proteins, with numerical estimates for the human DNA-repair enzyme HOGG1.

2.2 Elastic description of DNA

At long scales, the elastic response of a double stranded DNA heteropolymer may be modeled as a semiflexible polymer with bend and twist elastic coefficients [9] which *vary with position along the polymer*. Recent progress in micromanipulation techniques and confocal fluorescence microscopy on single biomolecules in vitro, allows us to measure the full distribution of molecular (elastic) properties without needing to average over a macroscopic sample [3]; one may thus obtain sample-to-sample variations of the elastic response of DNA hetero-chains. Yet force-extension and torsion experiments on single DNA heteropolymers are typically analysed assuming an *average homogeneous* semiflexible polymer, with one (few) fit parameters such as L/L_p , the ratio of the contour length to the bend persistence

length as mentioned earlier in Chapter 1 (see Fig. 1.2) [10]. This invites the question : are the force-extension measurements sensitive to the heteropolymer nature of DNA, or are the effects of sequence randomness simply averaged out for long chains. This was first addressed theoretically in the context of worm-like chain models (or its equivalent) by [11, 12], followed by [13, 14]. The analysis, done for large enough chains, suggested that randomness simply self-averages. However it may be argued that the effects of quenched randomness were not taken *exactly*; as is known from exact studies of quenched random spin systems, randomness can play a dramatic role, especially in one dimension [15]. We therefore revisit the problem of sequence dependence in the elastic description of the DNA.

2.3 Sequence - distribution in a worm-like chain (WLC) model

The DNA filament is represented by an inextensible space curve $\mathbf{r}(s)$ of total length L , parametrised in terms of the arc length s . We shall discuss the consequence of sequence distribution at the largest scales, when the effective description of elastic deformations of a DNA heteropolymer, free to swivel at one end, is given by a worm-like chain (pure bend) with arc length dependent moduli $\kappa(s)$. The Hamiltonian may be written as,

$$\mathcal{H} = \frac{1}{2} \int_0^L ds \left[\kappa(s) \left(\frac{\partial \hat{\mathbf{t}}}{\partial s} \right)^2 \right], \quad (2.1)$$

where, $\hat{\mathbf{t}} = \frac{\partial \mathbf{r}}{\partial s}$ is the tangent vector with the constraint that $|\hat{\mathbf{t}}(s)| = 1$. The bending moduli $\kappa(s) \geq 0$ are taken to be uncorrelated (quenched) random numbers derived from a bimodal or a gaussian distribution.

Earlier work on the random version of the WLC model [11, 12, 13, 14] computed physical quantities such as the end-to-end distance averaged over both thermal and random distributions. These calculations, restricted to the lowest moments, incorporated the effects of randomness only approximately. In this section we compute the entire *distribution* of global and local (thermally averaged) elastic quantities as a function of the realisation of randomness; our treatment of quenched randomness in most cases is exact. Specifically in the context of single molecule experiments, we would like to know if the average value of physical quantities computed in these earlier studies represent typical measured values.

The bending energy term involves the energy cost associated with the relative orientation of neighbouring tangent vectors, thus the discrete version of the WLC model is identical to

the Heisenberg spin chain,

$$H_{DNA} = - \sum_{i=1}^N j_i \mathbf{t}_i \cdot \mathbf{t}_{i+1} \quad (2.2)$$

with a local ferromagnetic $j_i \equiv \frac{\kappa(s)}{\Delta} > 0$, where Δ is the lattice discretization. A constant force F applied to one end of the heteropolymer appears as a uniform “magnetic field” $f = F\Delta$ in this description,

$$H_{DNA} = - \sum_{i=1}^N j_i \mathbf{t}_i \cdot \mathbf{t}_{i+1} - f \sum_{i=1}^N t_i^z \quad (2.3)$$

Note that in this representation, the tangent vectors are associated with ‘links’ and the energy penalty j_i with the ‘hinge’ between links i and $i + 1$ (Fig. 2.2). The bending moduli $\{j_i\} \geq 0$ are independent random numbers which may be taken from either a bimodal distribution,

$$P(j_i) = p \delta(j_i - J_1) + (1 - p) \delta(j_i - J_2) \quad (2.4)$$

or a gaussian distribution,

$$P(j_i) = (2\pi\sigma)^{-1/2} \exp\left[-(j_i - J)^2/2\sigma\right]. \quad (2.5)$$

Most of our quoted results are for the bimodal distribution.

We first study the sequence dependence of global elastic properties such as the distribution of the thermally averaged end-to-end distance $\langle R^2 \rangle$. Next we study the sequence dependence of local thermally averaged quantities such as the local persistence length defined via tangent-tangent correlators and the local extensional stiffness.

2.3.1 Distribution of global elastic variables

Light scattering in dilute suspensions or measurement of bead fluctuations in single-molecule experiments, provide information on the statistics of the end-to-end vector $\mathbf{R} = \sum_{i=1}^N \mathbf{t}_i$, in particular the thermally averaged end-to-end distance $\langle R^2 \rangle$ of the polymer

$$\langle R^2 \rangle = \sum_{i=1}^N \sum_{j=1}^N \langle \mathbf{t}_i \cdot \mathbf{t}_j \rangle. \quad (2.6)$$

The tangent-tangent correlation function in our Heisenberg-chain representation may be evaluated exactly using transfer matrix methods [17],

$$\langle \mathbf{t}_i \cdot \mathbf{t}_{i+1} \rangle \equiv x_i = \coth(\beta j_i) - \frac{1}{\beta j_i} \quad (2.7)$$

$$\langle \mathbf{t}_i \cdot \mathbf{t}_{i+R} \rangle \equiv C_{iR} = \prod_{m=i}^{i+R} x_m \quad (2.8)$$

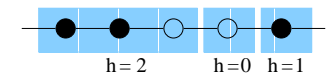
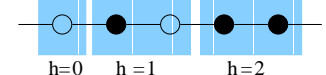
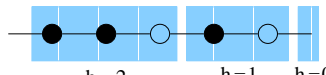
CONFIGURATION	WEIGHT	$\langle R^2 \rangle$
 $h=2$ $h=0$ $h=1$	$p^2 p^0 p^1 (1-p)^2$	$3^2 + 1^2 + 2^2$
 $h=0$ $h=1$ $h=2$	$p^0 p^1 p^2 (1-p)^2$	$1^2 + 3^2 + 2^2$
 $h=2$ $h=1$ $h=0$	$p^2 p^1 p^0 (1-p)^2$	$3^2 + 2^2 + 1^2$

Figure 2.2: Typical configuration of a heteropolymer with a bimodal distribution of ‘soft’ ($J_1 = 0$) and ‘hard’ ($J_2 = \infty$) hinges. Inset shows the two-string H1 representing a ‘markov unit’ (see text).

from which, together with (2.6), we obtain $\langle R^2 \rangle$.

Thus far we have expressed the thermal averaged end-to-end distance as a function of a given sequence of $\{j_i\}$. To determine the probability distribution of this quantity as a function of sequence realisation, we need to *prepare a natural a priori ensemble*. For this we consider a box containing a large number $M \rightarrow \infty$ of hinges of two strengths J_1 and J_2 , with fraction p and $q = 1 - p$, respectively. Initiate a polymerisation reaction to generate all possible heteropolymers of length N ; this set has heteropolymers with different N_1 of J_1 hinges. The probability of encountering a value j_i at the i^{th} hinge is independent of other hinges and is given by the bimodal distribution (2.4).

We first consider a special limit of the model, $J_1 = 0$ (soft hinge) and $J_2 = \infty$ (hard hinge), for which we can obtain several exact statements regarding the properties of the distribution function of global elastic variables.

Let the total number of soft (hard) hinges in the heteropolymer be N_s (N_h), so that $N_s + N_h = N$. Consider a typical configuration of the heteropolymer labelled by the following string $H_1 L_1 H_2 L_2 H_2 \dots L_n H_n$, which represents alternating stretches of $L_\alpha \geq 1$ soft and $H_\alpha \geq 1$ hard hinges (except for H_1 which can be ≥ 0). Clearly $\sum_{\alpha=1}^n L_\alpha = N_s$ and $\sum_{\alpha=1}^n H_\alpha = N_h$.

To compute the thermally averaged $\langle R^2 \rangle$ for the given sequence, we use the following strategy. We may represent each L_α string as $1H_1^\alpha 1H_2^\alpha \dots 1$ where $H_1^\alpha = H_2^\alpha \dots = 0$ and 1 appears L_α number of times. Since $J_1 = 0$, it is clear that the two-string H1 (containing H consecutive hard hinges followed by 1 soft hinge) forms a ‘Markov unit’ (Fig. 2.2), with a

thermally averaged end-to-end distance given by $(H + 1)^2$; the $\langle R^2 \rangle$ for the entire string is then simply the sum of the $\langle R^2 \rangle$ of the individual two-strings.

We can now calculate the probability distribution of $\langle R^2 \rangle$ and its moments for arbitrary N . As before, the probability of encountering a two-string $H1$ is given by $p^H q$. The joint probability distribution for strings of length N to have the value $\langle R^2 \rangle = Y$ and $N_h = K$ hard hinges is given by

$$P(Y, K) \equiv \left\langle \left\langle \delta(Y - \langle R^2 \rangle) \delta(K - N_h) \right\rangle \right\rangle \quad (2.9)$$

where the double angular brackets explicitly read as,

$$P(Y, K) = \frac{\sum_{N_h=0}^N C_{N_h}^N p^{N_h} (1-p)^{N-N_h} \delta(Y - \langle R^2 \rangle) \delta(K - N_h)}{\sum_{N_h=0}^N C_{N_h}^N p^{N_h} (1-p)^{N-N_h}} \quad (2.10)$$

The denominator is easily seen to be unity. The numerator may be rewritten in terms of our ‘Markov unit’ as,

$$P(Y, K) = \sum_{N_h=0}^N \sum_{h_1=0}^{N_h} \dots \sum_{h_{N'_s}=0}^{N_h} \prod_{i=1}^{N'_s} p^{h_i} (1-p) \delta_Y \delta(K - N_h) \delta_{N_h} \quad (2.11)$$

where for convenience we have written $N'_s = N_s + 1 \equiv N - N_h + 1$ and δ_Y and δ_{N_h} are a short-hand notation for the delta functions,

$$\delta_Y = \delta\left(Y - \sum_{j=1}^{N'_s} (h_j + 1)^2\right) \text{ and } \delta_{N_h} = \delta\left(N_h - \sum_{j=1}^{N'_s} h_j\right) \quad (2.12)$$

impose the constraint. Note that (2.11) is automatically normalised. In Fourier representation,

$$P(Y, K) = \int \int \frac{dk_1 dk_2}{(2\pi)^2} e^{ik_1 Y} e^{ik_2 K} [F(k_1, k_2)]^{N'_s} \quad (2.13)$$

where,

$$F(k_1, k_2) = \sum_{h=0}^{N_h} p^h (1-p) e^{-ik_1 (h+1)^2} e^{-ik_2 h} \quad (2.14)$$

Expanding the exponentials in (2.14) and separating the terms upto quadratic order from the rest, $F(k_1, k_2)$ can be rewritten as

$$\left[1 - (ik_1 \bar{u} + ik_2 \bar{w}) - \left(\frac{k_1^2}{2} \bar{u}^2 + \frac{k_2^2}{2} \bar{w}^2 + k_1 k_2 \bar{u} \bar{w} \right) \right] + \left[\left(\frac{k_1^3}{6} \bar{u}^3 + \frac{k_2^3}{6} \bar{w}^3 + \frac{k_2 k_1^2}{2} \bar{w} \bar{u}^2 + \frac{k_1 k_2^2}{2} \bar{u} \bar{w}^2 \right) + \dots \right],$$

where $u = (h + 1)^2$, $w = h$ and the averages are defined as

$$\begin{aligned}\bar{u} &= \sum_{h=0}^{N_h} p^h (1-p)(h+1)^2 \\ \bar{w} &= \sum_{h=0}^{N_h} p^h (1-p)h \\ \overline{uw} &= \sum_{h=0}^{N_h} p^h (1-p)h(h+1)^2,\end{aligned}\quad (2.15)$$

and so on. The expression for $\ln [F(k_1 k_2)]^{N'_s}$ may now be read out easily,

$$\begin{aligned}N'_s \ln \left[1 - (ik_1 \bar{u} + ik_2 \bar{w}) - \left(\frac{k_1^2}{2} \bar{u}^2 + \frac{k_2^2}{2} \bar{w}^2 + k_1 k_2 \overline{uw} \right) \right. \\ \left. + i \left(\frac{k_1^3}{6} \bar{u}^3 + \frac{k_2^3}{6} \bar{w}^3 + \frac{k_2 k_1^2}{2} \overline{uw}^2 + \frac{k_1 k_2^2}{2} \overline{uw}^2 \right) + \dots \right].\end{aligned}$$

We now expand the \ln in a power series and re-arrange the resulting terms in cumulants of powers of u and w ,

$$\begin{aligned}\ln [F(k_1 k_2)]^{N'_s} &= iN'_s (ik_1 \bar{u} + k_2 \bar{w}) - N'_s \left(\frac{k_1^2}{2} \overline{u^2} + \frac{k_2^2}{2} \overline{w^2} + k_1 k_2 \overline{uw} \right) \\ &+ N'_s \left[i \left(\frac{k_1^3}{6} \overline{u^3} + \frac{k_2^3}{6} \overline{w^3} + \frac{k_2 k_1^2}{2} \overline{u^2 w} + \frac{k_1 k_2^2}{2} \overline{u w^2} \right) + \dots \right],\end{aligned}\quad (2.16)$$

where the cumulants are defined in the usual way [18], e.g.,

$$\begin{aligned}\overline{u^2} &= \bar{u}^2 - \bar{u}^2 \\ \overline{w^2} &= \bar{w}^2 - \bar{w}^2 \\ \overline{uw} &= \overline{uw} - \bar{u} \bar{w} \\ \overline{u^3} &= \bar{u}^3 - 3 \bar{u}^2 \bar{u} + 2 \bar{u}^3 \\ \overline{w^3} &= \bar{w}^3 - 3 \bar{w}^2 \bar{w} + 2 \bar{w}^3 \\ \overline{u^2 w} &= \overline{u^2 w} - 2 \overline{uw} \bar{u} + \bar{u}^2 \bar{w} - \bar{u}^2 \bar{w} \\ \overline{u w^2} &= \overline{u w^2} - 2 \overline{uw} \bar{w} + \bar{u} \bar{w}^2 - \bar{u} \bar{w}^2.\end{aligned}\quad (2.17)$$

Re-exponentiating and after a little bit of algebra, we get,

$$\begin{aligned}[F(k_1 k_2)]^{N'_s} &= \exp \left[iN'_s (ik_1 \bar{u} + k_2 \bar{w}) - N'_s \left(\frac{k_1^2}{2} \overline{u^2} + \frac{k_2^2}{2} \overline{w^2} + k_1 k_2 \overline{uw} \right) \right] \times \\ &\left[1 + iN'_s \left(\frac{k_1^3}{6} \overline{u^3} + \frac{k_2^3}{6} \overline{w^3} + \frac{k_2 k_1^2}{2} \overline{u^2 w} + \frac{k_1 k_2^2}{2} \overline{u w^2} \right) + \dots \right].\end{aligned}\quad (2.18)$$

Rescaling by $q_1 = \sqrt{N'_s} k_1$, $q_2 = \sqrt{N'_s} k_2$ we may now express the joint probability distribution function $P(Y, K)$ (2.13) as a summation in powers of $1/\sqrt{N'_s}$,

$$\begin{aligned}
 P(Y, K) = N_0 & \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dq_1 dq_2 \exp\left(\frac{i(q_1 Y + q_2 K)}{\sqrt{N'_s}}\right) \right. \\
 & \exp\left(-i\sqrt{N'_s}(q_1 \bar{u} + q_2 \bar{w}) - \frac{1}{2}(\bar{u}^2 q_1^2 + 2q_1 q_2 \bar{u} \bar{w} + q_2^2 \bar{w}^2)\right) \\
 & \left. \left[1 + \frac{i}{\sqrt{N'_s}} \left(\frac{k_1^3}{6} \bar{u}^3 + \frac{k_2^3}{6} \bar{w}^3 + \frac{k_2 k_1^2}{2} \bar{u}^2 \bar{w} + \frac{k_1 k_2^2}{2} \bar{u} \bar{w}^2 \right) + \dots \right] \right], \quad (2.19)
 \end{aligned}$$

where,

$$N_0 = \frac{1}{4\pi^2 N'_s} \quad (2.20)$$

So far the expression for the joint probability distribution is valid for arbitrary N . We now study the asymptotic, $N_s \rightarrow \infty$, form of (2.19), which gets contributions only from the first term,

$$\begin{aligned}
 P(Y, K) = N_0 & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dq_1 dq_2 \exp\left(\frac{i(q_1 Y + q_2 K)}{\sqrt{N'_s}}\right) \times \\
 & \exp\left[-i\sqrt{N'_s}(q_1 \bar{u} + q_2 \bar{w}) - \frac{1}{2}(\bar{u}^2 q_1^2 + \bar{w}^2 q_2^2 + 2\bar{u} \bar{w} q_1 q_2)\right]
 \end{aligned}$$

The integration is simply carried out after converting the integrand to diagonal form by an appropriate similarity transformation; the final expression for the asymptotic form of $P(Y, K)$ is,

$$P(Y, K) = \frac{e^{-(K - \bar{w} N_s)^2 / 2N_s \bar{w}^2}}{2\pi^2 N_s \sqrt{\mathcal{D}}} \exp\left[-\frac{\bar{w}^2 (Y - (\bar{u} N_s - \frac{\bar{u} \bar{w}}{\bar{w}^2} (N_h - \bar{w} N_s))^2)}{2N_s \mathcal{D}}\right] \quad (2.21)$$

where, $\mathcal{D} = \bar{u}^2 \bar{w}^2 - \bar{u} \bar{w}$. Given the joint probability distribution, one may use the formula,

$$P(Y|K) = \frac{P(Y, K)}{P(K)}, \quad (2.22)$$

to obtain the distribution function of Y for a given K , where $P(K)$ is obtained by integrating $P(Y, K)$ over all possible values of Y —

$$P(Y|K) = \sqrt{\frac{(1-p)^3}{8\pi p^2 N}} \exp\left[-\frac{(1-p)^3}{8p^2 N} \left(Y - \frac{1+p}{1-p} N\right)^2\right] \quad (2.23)$$

In the above expression, we have substituted the calculated values of the cumulants in the asymptotic limit,

$$\bar{u} = \frac{1+p}{1-p^2}$$

$$\begin{aligned}
\overline{w} &= \frac{p}{1-p} \\
\overline{u^2} &= \frac{p(9+10p+p^2)}{(1-p)^4} \\
\overline{w^2} &= \frac{p}{(1-p)^2} \\
\overline{uw} &= \frac{p(3+p)}{(1-p)^2}.
\end{aligned} \tag{2.24}$$

Equation (2.23) is our final result; in the limit $N \rightarrow \infty$, the probability distribution function for the thermally averaged end-to-end distance for a fixed $p = N_h/N$ is a gaussian with mean value and variance given by

$$\begin{aligned}
\overline{\langle R^2 \rangle} &= N \frac{1+p}{1-p} \\
\overline{\langle R^2 \rangle^2} - \overline{\langle R^2 \rangle}^2 &= N \frac{4p^2}{(1-p)^3}
\end{aligned} \tag{2.25}$$

where the above overbars represent an average over the sequence distribution.

In principle, we can systematically calculate the subleading corrections in powers of $1/\sqrt{N}$ for a fixed value of p . We find it more instructive, however, to compute the probability distribution numerically. We numerically prepare our ensemble with fixed $\{N_h, N\}$ as discussed before; for each ‘sequence’ in the ensemble we compute the thermally averaged $\langle R^2 \rangle$ using (2.6). It is then a simple exercise to compute the probability distribution $P(Y|N_h)$. Fig. 2.3 shows the normalised distribution in scaled variables $z = Y - \overline{Y}/\sqrt{N}$, for different values of N and p . It is clear that the numerically computed distribution for increasing N (keeping p fixed between $0 < p < 1$) converges to the analytically computed gaussian (2.23). However there are significant deviations from the gaussian for smaller N ; numerics indicates that the distribution crosses over from roughly an exponential at smaller N (≈ 100) to the gaussian (2.23) at large N , with a crossover that depends on p . Fitting the tail of the distribution $P(z|N_h)$ for arbitrary N to a form e^{-z^α} , we find that α is bounded between 1 (exponential) and 2 (gaussian) (Fig. 2.4). The exponential tail suggests that the distribution is broad, indicating that the most probable value is very different from the mean. This is exemplified in Fig. 2.5, where Δ , the percentage deviation of the most probable from the mean, plotted against N , shows a gradual drop from a value of $O(1)$ to zero for large N (as expected from the asymptotic gaussian). This indicates that *single-molecule measurements of elastic quantities such as $\langle R^2 \rangle$ (or force-extension characteristics) done on short DNA strands with a given sequence realisation are not representative*. Moreover, the skewness, represented by the third cumulant C_3 , shows a non-monotonic behaviour as a function of N ,

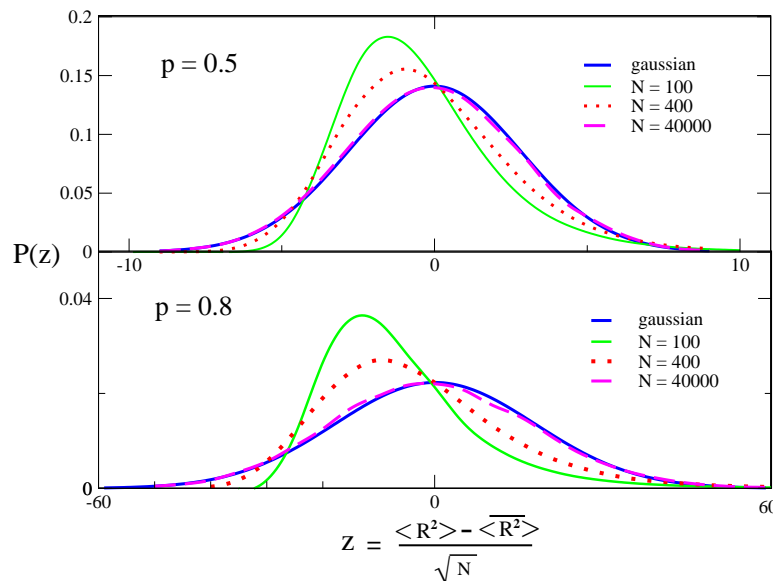


Figure 2.3: Normalised distribution of the scaled variable z , for various chain lengths in the ‘soft-hard’ model. The numerically obtained distribution z agrees with (2.23) for large N . Fraction of hard hinges, $p = 0.5$ (a) and $p = 0.8$ (b).

i.e., at very small N , the skewness shows a maximum before it gradually goes to zero for large N (Fig. 2.6).

This qualitative behaviour of the distribution in this soft-hard hinge model is replicated in the more general case when J_1 and J_2 are arbitrary positive numbers. Much of our results for the distribution are numerical, made more accurate by the use of Kesten variables [19]; nevertheless we do calculate the asymptotic form of the lower moments exactly. Since the results are qualitatively similar to the special case, we only present a few graphs. Our analysis for the general case follows :

To determine the probability distribution of the thermally averaged end-to-end distance for arbitrary J_1 and J_2 , we prepare the ensemble of heteropolymers of length N as before with a fixed number $N_1 = pN$ of J_1 hinges. The probability of encountering a value J at the i^{th} hinge is given by the bimodal distribution, $P(J) = p \delta(J - J_1) + (1 - p) \delta(J - J_2)$, where $q = 1 - p$. Having prepared our ‘natural’ ensemble, we first compute the thermally averaged $\langle R^2 \rangle$ for a typical member of the ensemble and then as before estimate its distribution over realisations of sequence randomness.

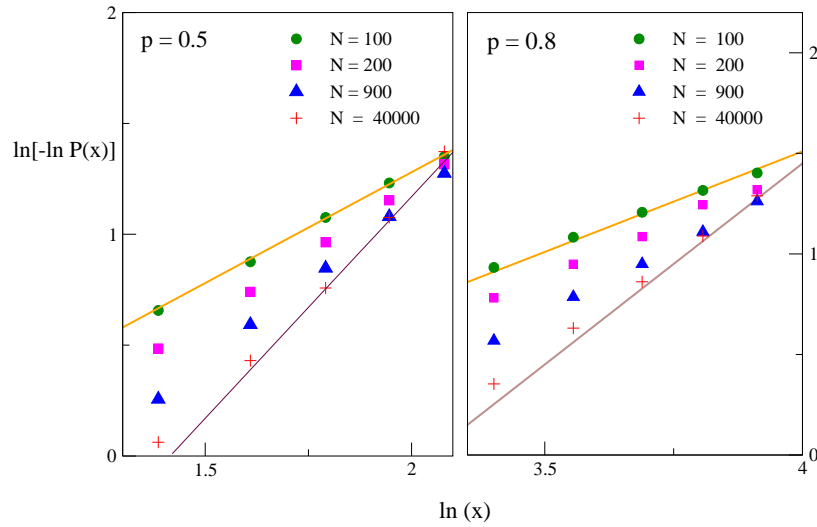


Figure 2.4: The tail of the normalised distribution of the scaled variable z , showing the crossover from an exponential (slope = 1) for short chains to a Gaussian (slope = 2) for longer chains. Note the crossover depends on the value of p .

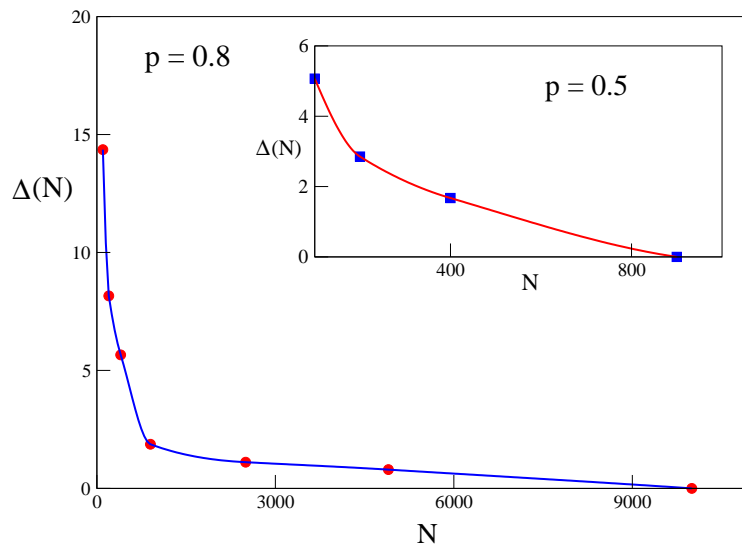


Figure 2.5: Lack of self-averaging for short chains, as demonstrated by the percentage deviation of the most probable from the mean Δ . Self-averaging is restored for larger chains, $\Delta = 0$, in a p dependent manner.

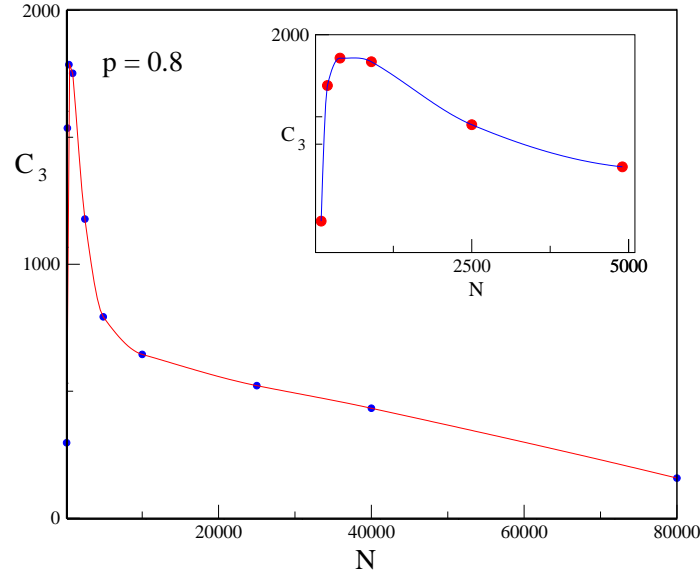


Figure 2.6: The skewness of the distribution, parametrised by its third cumulant C_3 , shows a crossover from an exponential to a gaussian ($C_3 = 0$) at large N .

Note that the expression (2.6) for $\langle R^2 \rangle$ for a given sequence realisation can be recast as

$$\langle R^2 \rangle = N + 2 \sum_{n=1}^{N-1} f_n \quad (2.26)$$

where $f_1 = x_1$ and sequence f_n are the so-called *Kesten variables* [19] which satisfy the recursion relation

$$f_n = x_n(1 + f_{n-1}), \quad (2.27)$$

an observation that greatly simplifies the numerical computation of the desired probability distribution.

Written explicitly for a given realisation,

$$\langle R^2 \rangle \equiv Y = N + 2 \left[\sum_{i=1}^{N-1} x_i + \sum_{i=1}^{N-2} x_i x_{i+1} + \dots + x_1 \dots x_{N-1} \right], \quad (2.28)$$

while the mean over sequence randomness is given by

$$\overline{\langle R^2 \rangle} = N + 2 \left[\sum_{i=1}^{N-1} \bar{x}_i + \sum_{i=1}^{N-2} \overline{x_i x_{i+1}} + \dots + \overline{x_1 \dots x_{N-1}} \right]. \quad (2.29)$$

In the asymptotic $N \rightarrow \infty$ limit, x_i are independent random variables and so simplifies to,

$$\overline{\langle R^2 \rangle} = N \left(\frac{1 + \bar{x}}{1 - \bar{x}} \right) \quad (2.30)$$

where,

$$\bar{x} = pX_1 + (1 - p)X_2 \quad (2.31)$$

with $X_1 = \coth[\beta J_1 - \frac{1}{\beta J_1}]$ and $X_2 = \coth[\beta J_2 - \frac{1}{\beta J_2}]$. This expression for $\overline{\langle R^2 \rangle}$ reduces to (2.25) when $J_1 = 0$, $J_2 = \infty$. To obtain the second moment, note that

$$Y^2 = N^2 + 4NS_0 + 4S_0^2, \quad (2.32)$$

where

$$S_0 = \left[\sum_{i=1}^{N-1} x_i + \sum_{i=1}^{N-2} x_i x_{i+1} + \dots + x_1 \dots x_{N-1} \right]. \quad (2.33)$$

The second moment of the probability distribution of $\langle R^2 \rangle$ is easily worked out by computing $\overline{S_0}$ and $\overline{S_0^2}$; as before, in the asymptotic $N \rightarrow \infty$ limit, x_i are independent random variables, so

$$\begin{aligned} \overline{S_0} &= \sum_{i=1}^N (N - i) \bar{x}^i \\ \overline{S_0^2} &= \sum_{r=1}^{N-1} \sum_{s=1}^{N-1} \epsilon_{rs}. \end{aligned}$$

The summand ϵ_{rs} is given by the expression,

$$\begin{aligned} \epsilon_{rs} &= f^r \bar{x}^{s-r} \left[(s - r + 1)(N - s) + 2 \sum_{i=1}^{r-1} (N - s - i) \bar{x}^{2i} f^{-i} \right] \\ &+ \bar{x}^{r+s} \left[(N - s)(N - s - 1) - \sum_{i=1}^{r-1} (N - s - i) \right] \end{aligned}$$

when $s > r$. In the above formula, $f \equiv \bar{x}^2 = pX_1^2 + (1 - p)X_2^2$. It is easy to perform the sums in the large N limit, we find for the relative fluctuations of $\overline{\langle R^2 \rangle}$,

$$\begin{aligned} \frac{\overline{Y^2} - \bar{Y}^2}{N} &= 4 \left[\frac{2\bar{x}^2}{(1 - \bar{x})^3} + \frac{-2\bar{x}^2 f + 4\bar{x}f}{(1 - f)(1 - \bar{x})^2} + \frac{3\bar{x}f - 2\bar{x}^2 - \bar{x}^4}{(1 - f)(1 - \bar{x}^2)^2} \right. \\ &\left. - \frac{10\bar{x}^3 + 4\bar{x}^4 + 2\bar{x}^5}{(1 - \bar{x})^3(1 + \bar{x})^2} + \frac{4f\bar{x}^3}{(1 - f)(1 + \bar{x})(1 - \bar{x})^2} \right] + O(1/N) \quad (2.34) \end{aligned}$$

In this way one may obtain all the disorder moments of $\langle R^2 \rangle$. To obtain the full probability distribution in the limit $N \rightarrow \infty$, we note that the integral representation of the distribution of $\langle R^2 \rangle$ can be written recursively using the recursion relations (2.27) for the Kesten variables.

Thus,

$$P(Y, N + 1) = \prod_{i=1}^N \sum_{x_i \in X_1, X_2} \int df_i \delta(f_i - x_i(1 + f_{i-1})) p(x_i) \delta\left(Y - N - 2 \sum_{j=1}^{N-1} f_j - 1\right) \quad (2.35)$$

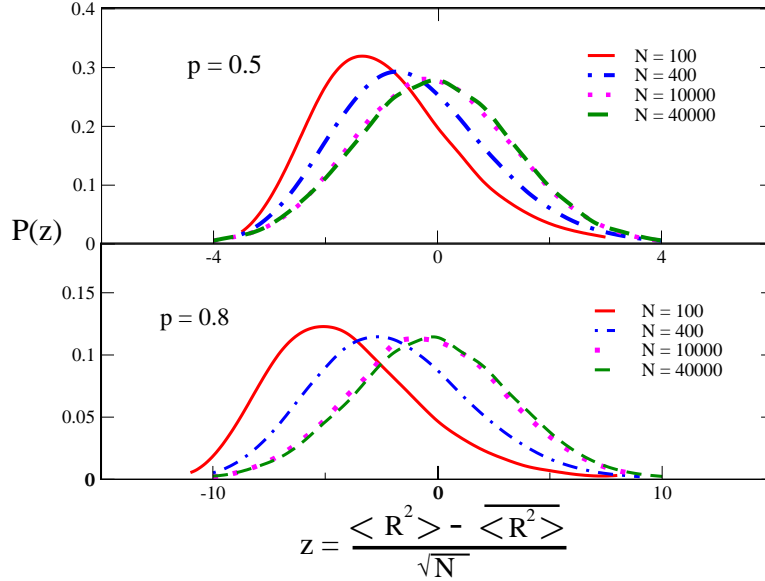


Figure 2.7: Normalised distribution of the scaled variable z , for various chain lengths in the $J_1 - J_2$ model. The numerically obtained distribution reduces to a gaussian for large N . Fraction of J_2 (hard) hinges, $p = 0.5$ (a) and $p = 0.8$ (b).

where we may define $f_0 = 0$. While we have not been able to use this formula to show that the asymptotic distribution is a gaussian, we have compelling numerical evidence that this is so. Moreover the moments obtained from (2.28) in the limit $N \rightarrow \infty$ reduces to the appropriate form (2.23) in the soft-hard limit, when $J_1 = 0, J_2 = \infty$.

As before, to obtain the leading corrections to this asymptotic result, we compute the probability distribution numerically. Fig. 2.7 shows the normalised distribution in scaled variables $z = Y - \bar{Y} / \sqrt{N}$, for different values of N and p . The qualitative trends are exactly as reported earlier, the numerically computed distribution crosses over from roughly an exponential at smaller N (≈ 100) to the gaussian at large N , with a crossover that depends on p . We again display the three measures describing the nature of the distribution and the crossover : tail of the distribution (Fig. 2.8), deviation from self-averaging (Fig. 2.9) and the skewness of the distribution (Fig. 2.10). To give numerical estimates of this crossover, we consider a ds-DNA which has *quenched* random stretches of single-stranded bubbles. Taking $\kappa_1/k_B T = 2, \kappa_2/k_B T = 150$ (in units of bp) corresponding to stretches of ss-DNA and ds-DNA respectively, Fig. 2.11 shows the crossover for different p and N — the strong deviation from a gaussian distribution should be comfortably observable.

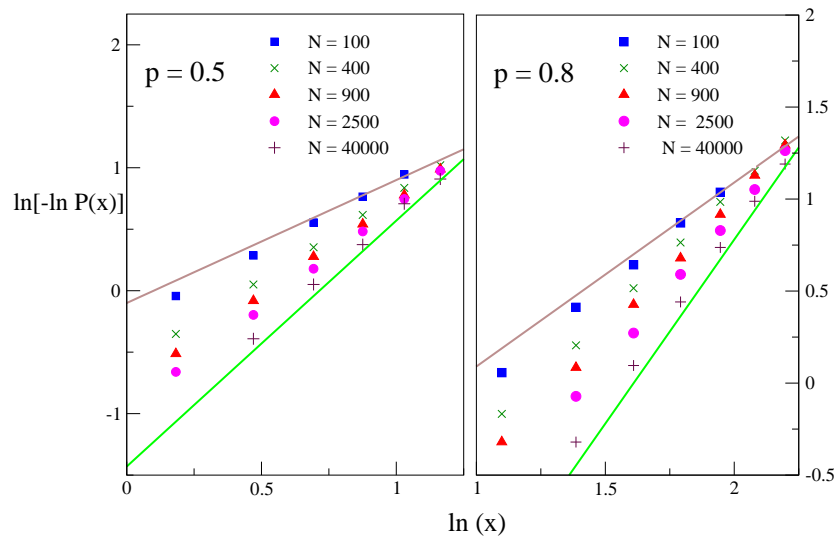


Figure 2.8: The tail of the normalised distribution of the scaled variable z , showing the crossover from an exponential (slope = 1) for short chains to a gaussian (slope = 2) for longer chains. Note the crossover depends on the value of p .

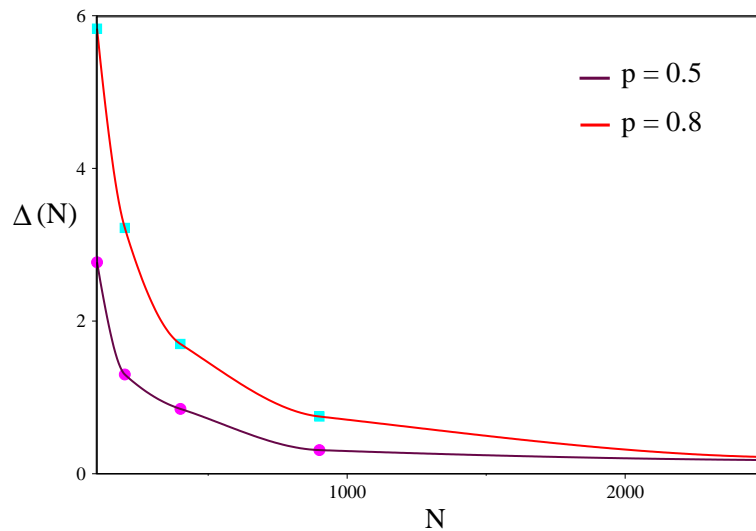


Figure 2.9: Lack of self-averaging for short chains, as demonstrated by the percentage deviation of the most probable from the mean Δ . Self-averaging is restored for larger chains, $\Delta = 0$, in a p dependent manner.

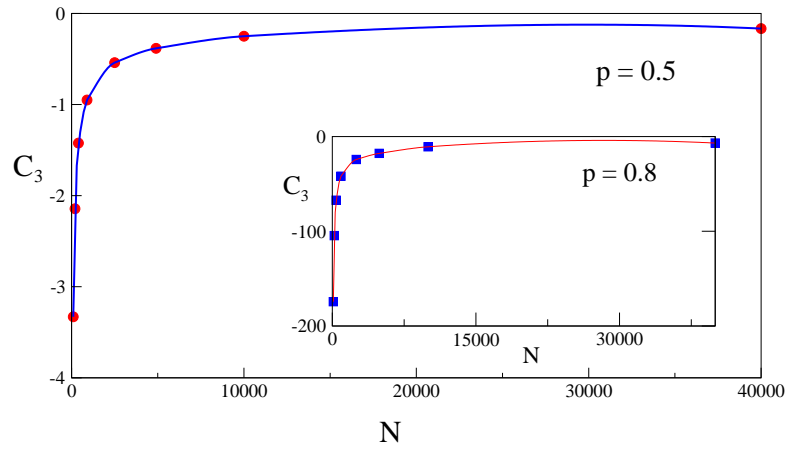


Figure 2.10: The skewness of the distribution, parametrised by its third cumulant C_3 , also shows a crossover from a nongaussian to a gaussian ($C_3 = 0$) at large N .

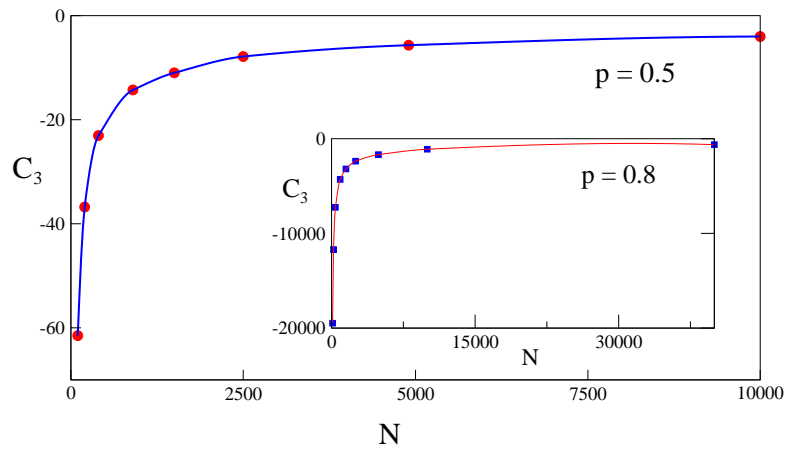


Figure 2.11: The crossover of the third cumulant C_3 when $\kappa_1/k_B T = 2$ and $\kappa_2/k_B T = 150$ as a function of N , for two different values of p , indicating that the deviation from gaussian should be comfortably observable.

The results obtained in this section are new and exact. They differ from the conclusions of earlier authors [12, 13, 20] qualitatively. These authors study slight variants of our 1-dimensional lattice model (2.2) and compute the lowest moments of the distribution of the thermally averaged end-to-end distance within a mean-field approximation; they conclude that the effects of randomness on the measured $\langle R^2 \rangle$ are weak. From our exact analysis, we see that while their conclusions are valid asymptotically, there are significant and observable differences at smaller values of N .

2.3.2 Distribution of local elastic variables

Force microscopy, such as AFM, probes the local stiffness of the heteropolymer, and so would be sensitive to sequence heterogeneity. In this section, we calculate the distribution of local quantities evaluated primarily from the tangent-tangent correlation function.

Recall that the persistence length L_p for a homopolymer is defined as the distance over which the local tangents are decorrelated, i.e., $\langle \mathbf{t}_i \cdot \mathbf{t}_{i+R} \rangle \sim \exp(-R/L_p)$, when $R \gg L_p$. The tangent-tangent correlator is however difficult to measure directly; an operational measure of the persistence length is obtained via a fit of the measured force-extension ($F - x$) curve of a DNA tethered at one end and pulled by means of an optical tweezer at the other, to the expression derived from the worm-like chain model using a single fit parameter L/L_p , where L is the contour length of the DNA [10],

$$F = \frac{k_B T}{L_p} \left[\frac{1}{4} (1 - x/L)^{-2} - \frac{1}{4} + \frac{x}{L} \right] \quad (2.36)$$

For a homopolymer this operational definition coincides with the expression derived from the tangent correlator. We will see that such a connection is less obvious for a random heteropolymer.

We define a sequence dependent persistence length for a random heteropolymer via the mapping onto the random Heisenberg spin-chain; the ‘local’ persistence length over a scale R ($R \gg \xi$),

$$\xi(i, i + R) = \frac{-R}{\ln \left(\prod_{l=i}^{i+R} x_l \right)} \quad (2.37)$$

with the J_l taken from the distribution (2.4). Note that $\xi_N \equiv \xi(i, i+N)$, the ‘global’ persistence length has a unique value (independent of the realisation) as long as N_1 and N_2 are fixed. However the force-extension curves, a graph of F versus $\sqrt{\langle R^2 \rangle}$, is, as we saw in the previous section, very sensitive to the sequence realisation (especially for smaller values of N). Thus a persistence length, extracted by fitting the force-extension curves to a theoretical formula,

would have broad distribution. The two ‘definitions’ in the case of heteropolymers are not compatible.

We next study the probability distribution of three thermally averaged quantities derived from the tangent-tangent correlator; these are the (i) distribution of the space averaged correlation function for a given R , (ii) distribution of the correlation function for specific sites i and j site separated by distance R , and (iii) limiting distribution for $0 \ll R \ll N$. We first note that it is easy to obtain the distribution of

$$Y \equiv \ln C_i(R) = \sum_{l=i}^{i+R} \ln x_l \quad (2.38)$$

in the limit $0 \ll R \ll N$; since $C_i(R)$ is the product and therefore Y , the sum of R independent random variables, the result follows simply from the central limit theorem. The probability distribution of $C_i(R)$ is thus a log-normal distribution, with the property that its most probable value is equal to the *mean of the distribution of Y* , a quantity which is easily calculable. We may thus calculate the deviation of the mean of the distribution of $C_i(R)$ from its most probable value, a measure of its lack of self-averaging.

The mean value of Y is easily seen to be,

$$\begin{aligned} \overline{\ln C_i(R)} &= R \overline{\ln \left(\coth(\beta J) - \frac{1}{\beta J} \right)} \\ &= R (p \ln X_1 + (1 - p) \ln X_2) \end{aligned} \quad (2.39)$$

for J taken from the bimodal distribution (2.4). As just mentioned, this is identical to the most probable value of $C_i(R)$. The mean value of $C_i(R)$ is given by

$$\begin{aligned} \overline{C_i(R)} &= \overline{\coth(\beta J) - \frac{1}{\beta J}} \\ &= pX_1 + (1 - p)X_2. \end{aligned} \quad (2.40)$$

The difference between the mean value and the most probable values of the correlation function $C_i(R)$ can be a large factor, indicating that its distribution is broad, Fig. 2.12, *even in the asymptotic limit*. Explicit simulation for finite chains reinforces this feature — Fig. 2.12 shows the distribution of Y for $N = 3000$ and $N = 20,000$ with $p = 0.8$. The mean value of $C_i(R)$ indicated by the arrow is the same for the different values of N , and is very different from its most probable value given by the mean of $\ln C_i(R)$ (peak of the graph), showing a strong violation of self-averaging *even in the asymptotic limit*.

Inset shows the distribution of the space averaged, $(N - R + 1)^{-1} \sum_i \ln C_i(R) \equiv \ln C(R)$, for $N = 20,000$ and $R = 2000$, which is a gaussian when $0 \ll R \ll N$. However even

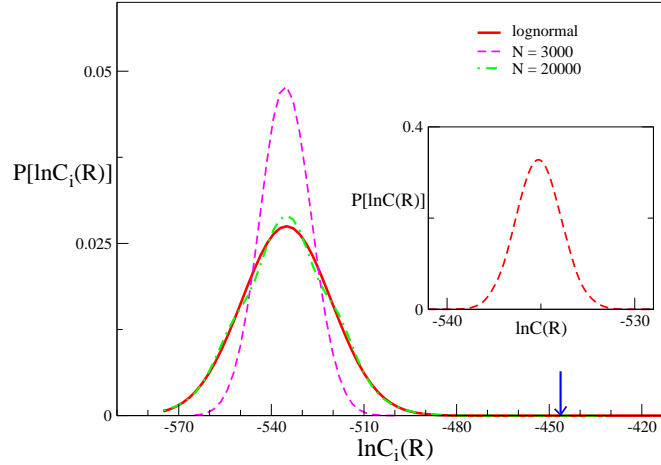


Figure 2.12: Distribution of $\ln C_i(R)$ ($i = 500, R = 2000$), for $x_1 = 0.9$ and $x_2 = 0.4$ chosen from a bimodal distribution with $p = 0.8$. Arrow indicates value of the mean correlator $C_i(R)$, distinct from its most probable value given by the mean of $\ln C_i(R)$ (peak of graph), showing a violation of self-averaging *even for the largest chains*. Inset : distribution of space averaged $\ln C(R)$ (see text).

after space averaging, the distribution of the correlator $C_i(R)$ remains broad in the asymptotic limit. As discussed in [15], the space averaged correlation function remains broad as long as the inequality

$$R \leq \frac{\ln N}{\ln b^2/a^2} \quad (2.41)$$

is satisfied, where $a^2 = \left(\overline{\coth \beta J - \frac{1}{\beta J}}\right)^2$ and $b^2 = \overline{\left(\coth \beta J - \frac{1}{\beta J}\right)^2}$.

2.4 Implications for DNA-protein binding

We have seen that both global and local elastic quantities are strongly sequence dependent for short chains; it is thus reasonable to expect that the binding/unbinding characteristics of DNA-proteins is sensitive to sequence. For as we discussed in the Introduction, DNA-binding proteins may be viewed as molecular probes which “measure” the local deformability or stiffness of the double-stranded DNA. As an example, RNA-polymerase, a 20 kD protein, locally distorts the DNA substrate in order to bind over a scale of 12 bp. An immediate consequence of the sequence dependent stiffness of the DNA over the scale ξ , would be seen in the rates of binding/unbinding (measured by the reaction constants k_d) of DNA-binding proteins.

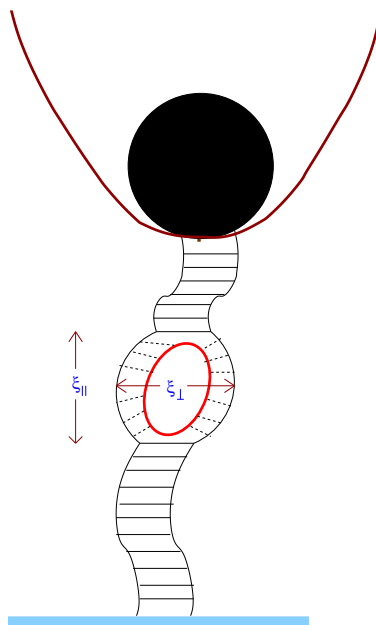
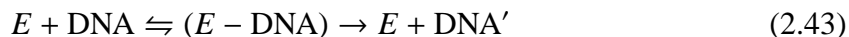


Figure 2.13: Sketch of a DNA-binding protein, binding and distorting the DNA substrate over scales $\xi_{||}$ and ξ_{\perp} . One end of the DNA is grafted to a substrate and the other end held in an optical trap. The binding/unbinding rates k_+, k_- depend on the distortion energy over this scale and hence on the sequence.

Consider a protein P which binds to a stretch of length ξ of the DNA heteropolymer of a given sequence with rate k^+ and unbinds with rate k^- (Fig. 2.1). The two states of the protein are denoted as bound (b) or unbound (u); on complexing with the DNA, the protein distorts the DNA over a scale ξ (denoted by a prime),



Alternatively consider an enzyme E which binds to DNA, forms an intermediate complex ($E - \text{DNA}$), resulting in a product DNA' via a Michaelis-Menten reaction scheme [21],



To address the issues raised in the beginning of this section, we envisage a single-molecule chemical kinetics experiment such as in Ref. [22]. For our purposes, we may think of a ds-DNA with one end attached to a substrate while the other held in a steep optical trap (Fig. 2.13), immersed in a buffer containing titrated amounts of the fluorescently labeled protein P . In this setup the end-to-end distance R of the heteropolymer is held fixed. Measurement of the time series of fluorescence anisotropy within a confocal volume, will provide information about the statistics of bound and unbound concentrations of the protein.

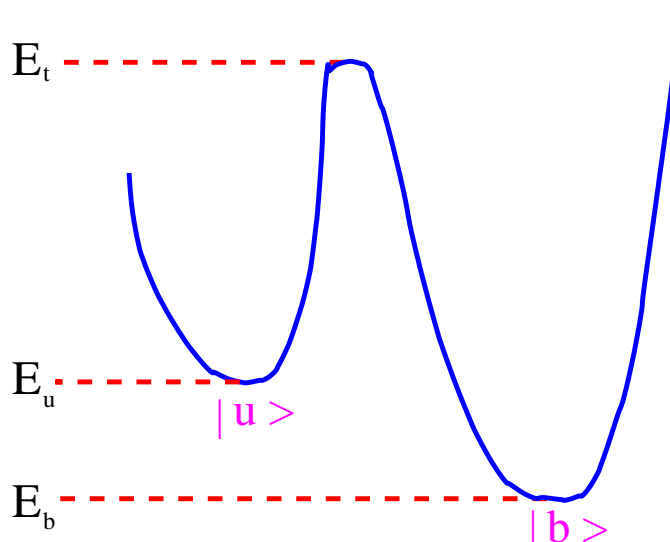


Figure 2.14: Sketch of the free energy landscape of the DNA , with the two minima corresponding to the unbound and the bound state of the DNA

Focussing on (2.42), we construct a free-energy profile (Fig. 2.14) with two minima denoting the bound ($|P_b, \text{DNA}' \rangle \equiv |b \rangle$) and unbound ($|P_u, \text{DNA} \rangle \equiv |u \rangle$) states with energies E_b and E_u respectively, and a maxima E_t corresponding to the intermediate (transition) state ($|t \rangle$). The transition rate for the forward reaction, k^+ is given by $A(\omega_u) \exp -(E_t - E_u)/k_B T$, where the prefactor A depends on the curvature (or oscillator frequency, ω_u) at $|u \rangle$. This is akin to a Fermi golden rule, where the prefactor A is a measure of the density of states, reflecting the number of low energy configurations sampled at $|u \rangle$. This molecular flexibility has been shown to be a key factor in determining binding probabilities in ligand-receptor chemical reactions [23, 24, 25]. In Ref. [26], it has been shown that the breathing modes of the ds-DNA, related to the flexibility of the separation between the complementary strands, is sequence dependent.

Here we will only consider the dominant exponential factor — the task is therefore to compute the energy barrier for distorting the DNA and protein to effect successful binding. We shall make the further simplification of ignoring the distortion of the protein — indeed the example of the DNA-repair protein that we consider is one where the protein distortion upon binding is insignificant.

To describe local deformations of the DNA over the scale $\xi (\simeq 10bp)$, one needs to go beyond the conventional WLC model, equation (2.1). A mesoscale description of the double stranded DNA [27], includes the energies due to base pair interactions and stacking interactions. Since we are interested in those stacking deformations that are strongly sequence

dependent, we will restrict ourselves to roll, tilt and twist [27].

In a further simplification, we will model the above mesoscale distortions by an extension of the WLC model; which in principle includes bend, twist, stretch and base-pair distortions. The energy barrier to be overcome in a typical DNA-protein binding is proportional to the local elastic stiffness, $\Delta E \simeq K\xi$, of the DNA heteropolymer. The binding rate $k^+ \propto \exp(\Delta E/k_B T)$, while the ratio $\frac{P_u}{P_b} = \frac{k^+}{k^-} \simeq \exp(E_b - E_u)/k_B T$.

We consider a simple modification of the ‘railway-track’ model [28, 29] for a *double-stranded* DNA heteropolymer. The two strands of the DNA, represented by $\mathbf{R}_1(s)$ and $\mathbf{R}_2(s)$ with a common arc-length coordinate which runs from 0 to L , can be used to define two local tangent vectors, $\mathbf{t}_\alpha \equiv \partial_s \mathbf{R}_\alpha$, where $\alpha = 1, 2$. The physical length of the two strands of the ds-DNA is given by $L_\alpha = \int_0^L ds(1 + \epsilon_\alpha(s))$ with $\epsilon_\alpha(s)$ being the extensional strain in the two strands, defined by $\epsilon_\alpha(s) \equiv \partial_s u_\alpha(s)$, where $u_\alpha(s) \equiv \mathbf{R}_\alpha(s) - \mathbf{R}_0(s)$ (and $\mathbf{R}_0(s)$ the undistorted position at monomer index s). The hamiltonian for the chains has a bend (E_{bend}) and a stretch (E_{str}) energy; we ignore for simplicity the twist and a symmetry allowed bend-stretch coupling [30],

$$\mathcal{F} = \frac{1}{2} \int ds \left[\sum_{\alpha=1}^2 \left(\kappa(s) \left(\frac{\partial \mathbf{t}_\alpha}{\partial s} \right)^2 + \lambda(s) \epsilon_\alpha^2(s) \right) + V(\mathbf{R}_1(s) - \mathbf{R}_2(s)) \right] \quad (2.44)$$

The force constants $\kappa(s), \lambda(s)$ correspond to the local bend and stretch moduli respectively, which includes the effects of steric interaction between neighbouring base-pairs. In addition we include a base-pair energy (E_{bp}) via a short-range potential V between the two strands, having a parabolic-well form,

$$V(R^-(s)) = \begin{cases} \infty & \text{if } R^-(s) < a \\ -V_0(s) \left[1 - \left(\frac{R^-(s) - R_0}{a - R_0} \right)^2 \right] & \text{if } a < R^-(s) < d \\ 0 & \text{if } R^-(s) > d \end{cases} \quad (2.45)$$

where V is a function of the relative separation $|\mathbf{R}^-(s)| \equiv |\mathbf{R}_1(s) - \mathbf{R}_2(s)| \equiv R^-(s)$ alone and $d = 2R_0 + a$. This potential mimics the short-range hydrogen bonding between complementary base pairs. The well depths $V_0(s)$ are random, reflecting the random occurrence of $A = T$ and $G \equiv C$ pairs. Writing $\mathbf{R}^-(s) = 2R\mathbf{b}(s)$, we note that in the usual formulation of the elasticity of the B-form of DNA, where the distance between the strands is held fixed, one imposes the constraint that $\mathbf{b} \cdot \mathbf{t}_\alpha = 0$; here we allow the sugar-phosphate backbone to be flexible and so do not impose this constraint.

Consider a protein which locally binds to the ds-DNA by distorting it over a scale ξ_\perp normal to and ξ_\parallel along the DNA axis. This distortion has contributions from the bend, stretch

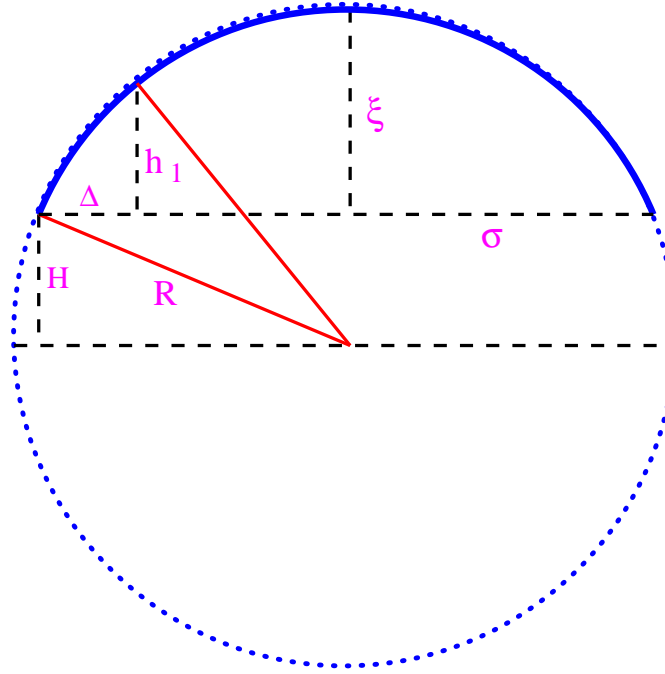


Figure 2.15: Geometric construction, illustrating the relation (2.48)

and V terms; the binding rate depends on the free-energy barrier between the unbound and bound states. A simple way to calculate this barrier is to determine the free-energy for a fixed $\xi_{\parallel}, \xi_{\perp}, R$ from (2.44). For a homopolymer, this has a simple scaling behaviour,

$$\Delta E(\xi_{\parallel}, \xi_{\perp}; R) \sim \kappa \left(\frac{\xi_{\perp}}{\xi_{\parallel}^2} \right)^2 \xi_{\parallel} + V(\xi_{\perp}) \xi_{\parallel} + k \xi_{\perp}^2 \xi_{\parallel}. \quad (2.46)$$

For the heteropolymeric ds-DNA, it is more convenient to work with the discrete spin representation of (2.44), namely,

$$H = \sum_{i=1}^N \left[V(\mathbf{R}_i^-) + \sum_{\alpha=1}^2 \left(\kappa_i \mathbf{t}_{\alpha i} \cdot \mathbf{t}_{\alpha i+1} + \lambda_i \epsilon_{\alpha i}^2 \right) \right] \quad (2.47)$$

The number of segments that are deformed over the scale ξ_{\parallel} is $M = \xi_{\parallel}/d$, where d is the undistorted distance between neighbouring base pairs. Each of the segment has suffered a length extension of $\ell_i = R_0/\cos\theta_i \approx R_0(1 + \theta_i^2/2)$, where θ_i is the slope of the distorted segment i (going from $1 \dots M$) with respect to the undistorted axis. The local slope (θ_i) and transverse separation between the two strands can be easily computed from the relation, (Fig. 2.15)

$$(\sigma - i\Delta)^2 + (H + h_i)^2 = R^2 \quad (2.48)$$

where $R = (\sigma^2 + \xi^2)/2\xi$ and $H = (\sigma^2 - \xi^2)/2\xi$ are defined in terms of $\xi \equiv \xi_{\perp} - R_0$ and $\sigma \equiv \xi_{\parallel}/2$.

We explicitly compute this energy scale, for the DNA binding protein Human OGG1 (HOGG1) enzyme, involved in DNA-damage repair [31]. The presence of reactive oxygen species in the cell can lead to DNA-damage via the formation of an oxidised guanine (8-oxoG); the resulting structural distortion of the DNA is found to be negligible. However this oxidised purine is highly mutagenic, since it mispairs with adenine during replication. Cells have evolved repair enzymes such as HOGG1, which recognise the damaged site and bind to it. The subsequent repair of the damaged nucleotide is a subject of major study.

The binding of HOGG1 to the damaged site is accompanied by a major distortion of the DNA. Since the crystal structure of the DNA-HOGG1 complex is unavailable, we do not know whether the enzyme undergoes a concomitant distortion; however crystal structure studies on a related enzyme FOGG reveal that the structure of this enzyme is not appreciably distorted on DNA binding. We will therefore assume that the energy scale for distortion only comes from the distortion of the DNA over the scale of HOGG1-DNA binding. From the Protein Data Bank, we find that the the DNA distorts by opening a bubble of size, $\xi_{\parallel} \approx 8\text{bp} = 2.72\text{ nm}$ and $\xi_{\perp} = 1.78\text{ nm}$. Taking the undistorted DNA structural parameters to be $R_0 = 1.08\text{ nm}$, $\Delta = 0.34\text{ nm}$ and $a = 0.79\text{ nm}$, it is easy to read out the rest of the parameters from (2.48), thus computing the local slope changes and longitudinal and transverse extensions. We can now readily compute the energies for these distortions. Thus the bend energy is simply

$$\frac{\delta E_{bend}}{k_B T} = \sum_{i=1}^{\xi_{\parallel}} \frac{\kappa_i}{2\Delta} (t_{i+1} - t_i)^2 = 1.296 \frac{\bar{\kappa}}{\Delta} \quad (2.49)$$

where $\bar{\kappa}$ is the average sequence dependent bending stiffness over the scale $\xi = 8$. Thus, if the base sequence over the scale ξ_{\parallel} , is chosen from a uncorrelated random sequence, then $\bar{\kappa} = 175 \Delta$ [4] so that the bend energy in units of $k_B T$ is ≈ 228.5 . On the other hand, if the sequence over which HOGG1 binds is GC rich, then , $\bar{\kappa} = 350 \Delta$, so that, the bend energy in units of $k_B T$ is ≈ 457 .

The stretch deformation energy can be calculated from,

$$\frac{\delta E_{str}}{k_B T} = \sum_{i=1}^{\xi_{\parallel}} \frac{\omega_{0i}^2 \lambda \Delta}{2} \epsilon_i^2 \quad (2.50)$$

where, $\epsilon_i = 1 - \frac{1}{\cos \theta_i}$. The undeformed stretch moduli for the homopolymer are given by the following parameters, $\omega_o = 1.85\text{nm}^{-1}$, $\lambda = 78\text{nm}$. We have found that there is not much sequence dependent variation in λ .

Next we consider the energy cost δV , the change in the base-pair energy E_{bp} , associated

with breaking the hydrogen bonds. For distortions greater than $d = 2R_0 - a$, the energy cost due to breakage of the hydrogen bonds is V_0 . For the parameter values appropriate to HOGG1, all the bonds suffer the breakage. Therefore,

$$\frac{\delta V}{k_B T} = 8V_0 \approx 24 \quad (2.51)$$

Although in principle, the base-pair energy is sequence dependent, its contribution relative to bend distortion, δE_{bend} is negligible in the case of HOGG1 binding.

The chemical reactivities, k^\pm , are related to the exponential of the thermally averaged energy of distortion of a ds-DNA over a scale ξ_{\parallel} , ξ_{\perp} , valid when the chemical reaction rates are slower than the thermal relaxation rates of ds-DNA over the scale ξ_{\parallel} . Note that we have not included an excess polymeric entropy which may be associated with the opening of a bubble of size ξ_{\perp} . Our analysis shows that the binding kinetics for HOGG1 depends sensitively on sequence primarily through the bend distortion. Thus for a given DNA heteropolymer chain of length $L > \xi_{\perp}$, the binding kinetics will depend on the position along the polymer.

2.5 Conclusions

In this chapter, we have attempted to study simple physical situations wherein sequence randomness along a DNA heteropolymer plays an important role. We have defined sequence randomness in very broad terms; for instance randomness could arise as a result of the differing chemistry (both the nature of monomers and chemical bonds) along the polymer or as a result of random denaturation of a DNA heteropolymer in a thermal bath. In either case, we model the random heteropolymer at large scales by an elastic (worm-like) chain model with position dependent bending modulus κ ; the randomness in κ is quenched.

We explicitly compute the *distribution* of thermally averaged elastic quantities such as the end-to-end distance, as a function of the realisation of randomness and show that while this distribution is indeed a *gaussian* in the asymptotic ($L \rightarrow \infty$) limit, it crosses over to roughly an *exponential* for shorter chains. Thus while the end-to-end distance *self averages* for longer chains, the effect of quenched sequence randomness is significant for shorter chains. In addition, we find that the distribution of physical quantities related to the tangent-tangent correlator are very broad *even in the thermodynamic limit*.

We have argued that this broad distribution manifests in a sequence sensitivity of the cyclisation time and probability of loop formation in single/double -stranded DNA measured via fluorescence quenching [16] and cyclisation assays using biochemical ligation [33]

respectively. While we have not calculated it explicitly, our analysis may be modified to compute the mean cyclisation time or the propensity for looping. The former is a first-passage-time calculation, related to the exponential of the energy barrier to be traversed from an open to a closed configuration. One would expect therefore that the distribution of first passage times would be very broad [34]. The latter is related to the question : what is the probability for the end-to-end distance $R = 0$ for a given realisation of randomness, and how sensitive is this to the distribution of randomness [35]. However our main message, with clear biological implications, is in the kinetics of binding of proteins onto substrates such as ds-DNA. Since molecular binding to specific or non-specific regions of the DNA is accompanied by a local distortion of the DNA substrate, an immediate manifestation of the sequence dependent stiffness of the DNA over the scale ξ , is in the measurable rates of binding/unbinding of DNA-binding proteins. We have argued that the binding/unbinding reaction rates of DNA-binding proteins will depend on the local conformational stiffness of the DNA.

Bibliography

- [1] G.B. Koudelka, S.C. Harrison and M. Ptashne, *Nature* **326**, 886 (1987).
- [2] J.D. Watson et al., in *Molecular Biology of the Gene* (Benjamin/Cummings Pub., California, 2002, 5th ed).
- [3] R. H. Austin et al., *Phys. Today* **50**, 32 (1997); H. Clausen-Schaumann et al., *Curr. Opin. Chem. Biol.* **4**, 524 (2000).
- [4] M. E. Hogan and R. H. Austin, *Nature* **329**, 263 (1987).
- [5] T.W. Lynch et al., *J. Mol. Biol.* **330**, 493 (2003); K.K. Swinger, K.M. Lemberg, Y. Zhang and P.A. Rice, *EMBO J.* **22** (2003).
- [6] D. Bhattacharya, A. Mazumder, S. A. Miriam and G.V. Shivashankar, *Biophys. J.* **91**, 2326 (2006) .
- [7] T. E. Cloutier and J. Widom, *Mol. Cell* **14**, 355 (2004).
- [8] H.E. Gaub lab website, <http://www.biophysik.physik.uni-muenchen.de/>.
- [9] J. Marko and E. Siggia, *Macromolecules* **28**, 8759 (1995).
- [10] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, *Science* **265**, 1599 (1994); T. Strick et al., *Science* **271**, 1835 (1996).
- [11] P. Nelson, *Phys. Rev. Lett.* **80**, 5810 (1998).
- [12] D. Bensimon, D. Dohmi and M. Mezard, *Europhys. Lett.* **42**, 97 (1998).
- [13] D. Garrivier and B. Fourcade, *Europhys. Lett.* **49**, 390 (2000).
- [14] P. Debnath and B. J. Cherayil, *J. Chem. Phys.* **116**, 4330 (2002); **118**, 1970 (2003).
- [15] B.Derrida and H. Hilhorstz, *J. Phys. C* **14**, L539 (1981).

- [16] N.L. Goddard, G. Bonnet, O. Krichevsky and A. Libchaber, Phys. Rev. Lett. **85**, 2400 (2000).
- [17] M. Fisher, Am. J. Phys. **32**, 343 (1964); M.F. Thorpe, Phys. Rev. B **13**, 2186 (1976); G.A. Baker, G.S. Rushbrooke and P.W. Wood, in *Phase Transitions and Critical Phenomena*, eds. C. Domb and M. S. Green, Vol. 3 (Academic Press, NY, 1984).
- [18] S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).
- [19] H. Kesten, M. Koslov, and F. Spitzer, Compos. Math. **30**, 145 (1975).
- [20] Y. Kafri, D.K. Lubensky and D.R. Nelson, Phys. Rev. E **71**, 041906 (2005).
- [21] B. Choi et al., Phys. Rev. Lett. **95**, 078102 (2005).
- [22] B. P. English et al., Nat. Chem. Biol. **2**, 87 (2006).
- [23] S. Qi et al., Proc. Nat. Ac. Sc. **103**, 4416 (2006).
- [24] J. SantaLucia, Jr., Proc. Natl. Acad. Sci. U.S.A. **95**, 1460 (1998).
- [25] J. F. Leger et al., Proc. Natl. Acad. Sci. U.S.A. **95**, 12295 (1998).
- [26] G.A. Bonnet, A. Libchaber and O. Krichevsky, Phys. Rev. Lett. **90**, 138101 (2003)
- [27] M.G. Munteanu et al., TIBS **23**, 341 (1998).
- [28] R. Everaers, R. Bundschuh and K. Kremer, Europhys. Lett. **29**, 263 (1995); T.B. Liverpool, R. Golestanian and K. Kremer, Phys. Rev. Lett. **80**, 405 (1998).
- [29] H. Zhou, Y. Zhang, and Z.-c. Ou-Yang, Phys. Rev. Lett. **82**, 4560 (1999); Phys. Rev. E **62**, 1045 (2000).
- [30] This would have allowed the helical axis to move away from a straight line, see C.S. O'Hern, R.D. Kamien, T.C. Lubensky and P. Nelson, MRS Proceedings Vol. 463 (MRS, Pittsburgh, 1997).
- [31] S. Boiteux and J. Radicella, Biochimie, **81**, 59 (1999).
- [32] G.V. Shivashankar, M. Feingold, O. Krichevsky and A. Libchaber, Proc. Natl. Acad. Sci. USA **96**, 7912 (1999).

- [33] J. Widom, *Annu. Rev. Biophys. Biomol. Struct.* **27**, 285 (1998); *Quart. Rev. Biophys.* **34**, 269 (2001).
- [34] This is a hard calculation even for a flexible chain, see I. M. Sokolov, *Phys. Rev. Lett.* **90**, 080601 (2003).
- [35] P. A. Wiggins, R. Phillips and P. C. Nelson, *Phys. Rev. E* **71**, 021909 (2005); J. Yan and J. F. Marko, *Phys. Rev. Lett.* **93**, 108108 (2004); P. Ranjith, P.B. Sunil Kumar and G.I. Menon, *Phys. Rev. Lett.* **94**, 138102 (2005).