Statistical Properties of Metastable Intermediates in DNA Unzipping

J. M. Huguet,[1] N. Forns,[1,2] and F. Ritort[1,2,*]

[1]*Departament de Física Fonamental, Facultat de Física, Universitat de Barcelona, Diagonal 647, 08028 Barcelona, Spain*
[2]*CIBER-BBN, Networking Centre on Bioengineering, Biomaterials and Nanomedicine, ISCIII, Madrid, Spain*

We unzip DNA molecules using optical tweezers and determine the sizes of the cooperatively unzipping and zipping regions separating consecutive metastable intermediates along the unzipping pathway. Sizes are found to be distributed following a power law, ranging from one base pair up to more than a hundred base pairs. We find that a large fraction of unzipping regions smaller than 10 bp are seldom detected because of the high compliance of the released single stranded DNA. We show how the compliance of a single nucleotide sets a limit value around 0.1 N/m for the stiffness of any local force probe aiming to discriminate one base pair at a time in DNA unzipping experiments.

The mechanical response of biomolecules to externally applied forces allows us to investigate molecular free energy landscapes with unprecedented accuracy. Single molecule experiments with optical tweezers, atomic force microscope (AFM), and magnetic tweezers are capable of measuring forces in the pN range and energies as small as tenths of kcal/mol. An experiment that nicely illustrates the potential applications of single molecule manipulation is molecular unzipping [1–5]. By applying mechanical force to the ends of biopolymers such as DNA, RNA, and proteins, it is possible to break the bonds that hold the native structure and measure free energies and kinetic rates. In unzipping experiments, a DNA double helix is split into two single strands by pulling them apart and the force vs distance curve (FDC) measured. A typical FDC shows a force plateau around 15 pN with a characteristic sawtooth pattern corresponding to the progressive separation of the two strands. Mechanical unzipping is also a process mimicked by motor proteins that unwind the double helix. In fact, anticorrelations between unzipping forces and unwinding rates have been found in DNA helicases suggesting that such enzymes unzip DNA by exerting local stress [6]. DNA unzipping experiments have several applications such as identifying specific locations at which proteins and enzymes bind to the DNA [5]. Moreover, the strong dependence of the shape of the sawtooth pattern with the sequence might be used for DNA sequencing [7], i.e., a way to infer the DNA sequence from the unzipping data. A limitation factor in these applications is the accuracy at which base pair (bp) locations along the DNA can be resolved. This is mainly determined by the combined stiffness of the force probe and the large compliance of the released single stranded DNA (ssDNA) [1,8]. The unzipping process, even if carried out reversibly (i.e., infinitely slowly), shows a progression of cooperative unzipping-zipping transitions that involve groups of bps of different sizes. These cooperatively unzipping-zipping transitions regions (CUR) of bps breath in an all-or-none fashion hindering details about the individual bps participating in such transitions. Unzipping experiments pose challenging questions to the experimentalist and the theorist. What is the typical size of these CUR? What is the smallest size of the CUR that can be detected with single molecule techniques? Under what experimental conditions might be possible to resolve large CUR into individual bps? There have been several DNA unzipping studies with controlled force using magnetic tweezers. Because at constant force the unzipping transition is abrupt, this setup is not suitable to answer such questions [3,9].

We carried out DNA unzipping experiments with optical tweezers [10,11] and determined the distribution of CUR sizes in DNA fragments a few kbp long. For the experiments, a 2.2 kbp DNA molecular construct was synthesized [11]. In a typical unzipping experiment one bead is held fixed at the tip of a micropipette and the other bead is optically trapped and the force exerted on the molecule measured. By moving the center of the optical trap at a very low speed (10 nm/s) double stranded DNA (dsDNA) is progressively and quasireversibly converted into ssDNA through a succession of intermediate states corresponding to the successive opening of CUR [Fig. 1(a)]. The experimentally measured FDC shows a sawtoothlike pattern [Fig. 1(b)] that alternates force rips and gentle slopes. Slopes correspond to the elastic response of the molecule while the force rips correspond to the release of CUR. The slope is due to the combined elastic response of the optical trap and the released ssDNA. The size of a CUR can be inferred from the difference of slopes that precede and follow a given force rip. However, the identification of the CUR sizes is not straightforward as often the slopes cannot be isolated because the experimental FDC exhibits noise. Here we extract the different sizes of the CUR that separate contiguous intermediate states along the unzipping pathway. For that we adopt a Bayesian approach where for each experimental data point (distance, force) we determine the most probable intermediate state to which the data point belongs.
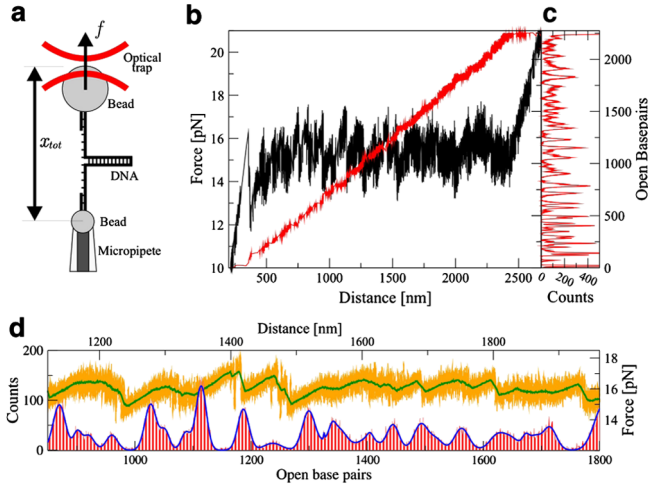
FIG. 1 (color online). Identifying CUR in DNA unzipping experiments. (a) Experimental setup. A 2.2 kbp sequence of DNA is unzipped using an optical trap and a micropipette. (b) Black curve shows the raw data of a typical FDC in an unzipping experiment. Red or gray curve shows the number of open bps n^* corresponding to each experimental data point (y axis of this curve is shown in panel c). (c) Histogram of the values for n^* shown in panel (b). (d) The lower curve shows a detailed view of the histogram overlapped with a fit to a sum of Gaussians. The upper curve shows the FDC (raw data and 1 Hz low pass filtered data) corresponding to that region of the histogram.

To this end we consider the molecular system as composed of different elements: the optical trap, the dsDNA handles, the released ssDNA, and the hairpin at the intermediate state I_n where n bases are open. We express the total distance between trap and pipette x_{tot} at a given force f as the sum of the extensions of each element at that force:

$$x_{\text{tot}}(f, n) = x_b(f) + x_h(f) + x_s(f, n) + \frac{\phi_b}{2} \qquad (1)$$

where x_b is the position of the bead with respect to the center of the optical trap; x_h is the extension of the flanking dsDNA handles; x_s is the extension of the released ssDNA, and ϕ_b is the diameter of the bead. The extension of the ssDNA depends on the number of open bases at the intermediate state I_n. The different contributions to Eq. (1) are calculated by using well-known elastic models for biopolymers [11]. For each experimental data point of the FDC (x, f), the intermediate state I_{n^*} that passes closest to that point for a fixed force f is determined by

$$|x - x_{\text{tot}}(n^*, f)| = \min_n(|x - x_{\text{tot}}(n, f)|). \qquad (2)$$

In this way each experimental data point (x, f) is associated to a value of n^* [red or gray curve in Fig. 1(b)]. The histogram built from all values n^* results in a series of sharp peaks that can be identified with the many intermediate states I_n [Fig. 1(c)]. The histogram contains information about the stability of the intermediate states: the higher the peak, the higher the stability of that state and the larger the GC content of that part of the sequence (data not

shown). The histogram can be fit to a sum of Gaussians each one characterized by its mean, variance, and statistical weight [Figs. 1(d) and 2(a)]. Finally, the size of the CUR is obtained by calculating the difference of the means (in bps) between consecutive Gaussians. The experimental distribution of CUR sizes is shown in Fig. 2(b). Sizes range from a few bps up to 90 bp with a maximum number of detected CUR sizes between 20 and 50 bp.

To better understand the distribution of CUR sizes we have computed the sequence dependent free energy profile using a mesoscopic model for DNA based on nearest neighbor bp interactions that includes the different elements of the experimental setup [12,13]. The model is defined by the total free energy of the system, $G(x_{\text{tot}})$, which gets contributions from the partial free energies $G(x_{\text{tot}}, n)$ of the many intermediates I_n: $G(x_{\text{tot}}) = -k_B T \log\{\sum_n \exp[-G(x_{\text{tot}}, n)/k_B T]\}$. From $G(x_{\text{tot}})$ we can determine the theoretical FDC by using the relation, $f(x_{\text{tot}}) = \frac{\partial G(x_{\text{tot}})}{\partial x_{\text{tot}}}$. We have used this model to compute the partial free energies $G(x_{\text{tot}}, n)$ of all intermediates I_n. For a given value of x_{tot} we identify the most stable intermediate I_{n^*} corresponding to the value of n^* for which $G(x_{\text{tot}}, n)$ is
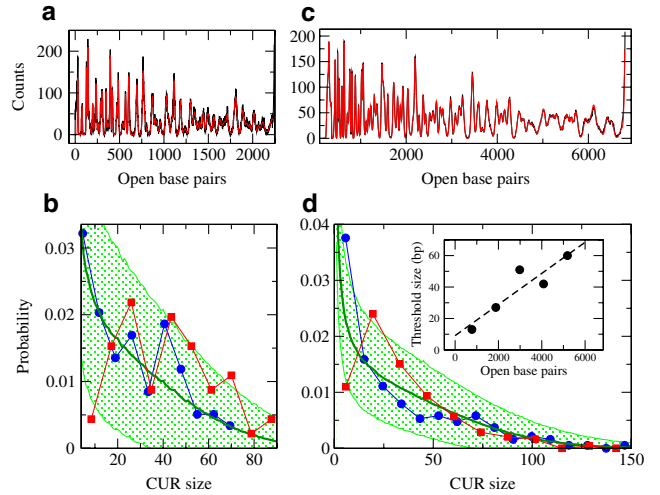

FIG. 2 (color online). CUR size distributions. (a) Histogram of n^* values corresponding to the different intermediate states for the 2.2 kbp DNA sequence. Red or gray curve shows the experimentally measured histogram. Black curve shows the fit to a sum of Gaussians. (b) Distribution of CUR sizes for the 2.2 kbp sequence. Red or gray curve shows the experimentally measured distribution. Blue or dark gray curve shows the distribution predicted by the mesoscopic model for DNA. Green or smooth and thick curve shows the distribution predicted by the toy model (see text) and the shaded area shows the standard deviation from different sequence realizations of the same length. (c) Histogram of intermediate states for the 6.8 kbp sequence. Same color code as in panel (a). (d) Distribution of CUR sizes for the 6.8 kbp sequence. Same color code as in panel (b). [Inset of (d)] Threshold size n^{thr} as a function of the number of open bps n. The dashed line is a linear fit, $n^{\text{thr}} = 9.1 + 0.01n$. For both molecular constructs, 6 molecules have been measured and analyzed.

the absolute minimum [i.e., $G(x_{tot}, n^*) \leq G(x_{tot}, n), \ \forall \, n$].
Integer values of n^* change in a stepwise manner as x_{tot} is
continuously varied according to the following scheme:

$$\ldots \Leftrightarrow I_{n_a^*} \Leftrightarrow I_{n_b^*} \Leftrightarrow I_{n_c^*} \Leftrightarrow \ldots \qquad (3)$$

where n_a^*, n_b^*, n_c^* indicate the number of open bps corre-
sponding to consecutive intermediates. Differences be-
tween consecutive values of n^* provide the sizes of the
CUR. The resulting size distributions are shown in
Fig. 2(b). The good agreement between the experimental
and the theoretical size distributions shows that our method
of analysis is capable of discriminating the metastable
intermediates during unzipping. There are two remarkable
facts in Fig. 2(b). First, the mesoscopic model predicts a
large fraction of CUR of size smaller than 10 bp that are
not experimentally observed. Second, size distributions are
not smooth but have a rough shape in agreement with the
prediction by the mesoscopic model. In order to check the
generality of these results we have repeated the same
analysis by unzipping a different and longer molecular
construct of 6.8 kbp [Figs. 2(c) and 2(d)]. The agreement
between experiments and theory remains good. Again a
large fraction of predicted CUR sizes smaller than 10 bp
are not detected [11]. However, the CUR size distributions
are now smoother suggesting that a monotonically decreas-
ing continuous distribution could describe the distribution
of CUR in the thermodynamic (infinite DNA length) limit.
The fact that CUR sizes show a long tailed distribution
indicates that large sizes occur with finite probability.
However, large-sized CUR hinder their internal DNA se-
quence limiting the possibility of sequencing DNA by
mechanical unzipping. Under what experimental condi-
tions is it possible to break up large-sized CUR into indi-
vidual bps?

   In order to answer this question we have developed a toy
model useful to elucidate the mathematical form of the
CUR size distribution. Similar distributions have been
investigated in the context of DNA thermal denaturation
[14,15] and DNA unzipping experiments in the constant
force ensemble [9]. Our model contains only two elements:
the bead in the optical trap and the DNA construct to be
unzipped. The latter is composed of the DNA duplex and
the released ssDNA [Fig. 3(a)]. The optical trap is modeled
by a harmonic spring with energy, $E_b(x_b) = \frac{1}{2}kx_b^2$. The
DNA duplex is modeled as a one-dimensional random
model with bp free energies ϵ_i along the sequence [16].
The free energy of a given intermediate I_n is given by
$G_{DNA}(n) = -\sum_{i=1}^{n} \epsilon_i$. The ϵ_i are distributed according to
a normal distribution $\mathcal{N}(\mu, \sigma)$, where $\mu(<0)$ and σ are
the mean and the standard deviation of the energies, re-
spectively (other more realistic energy distributions give
similar results). The released ssDNA is taken as inexten-
sible: its extension (x_m) is given by $x_m = 2dn$, where d is
the interphosphate distance, n is the number of open bps,
and the factor 2 stands for the two strands of ssDNA. By
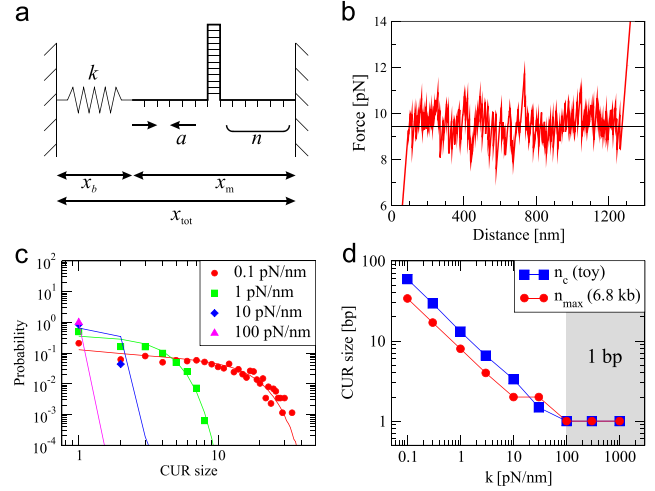using the relation $x_b = x_{tot} - 2dn$ [Fig. 3(a)], the total

FIG. 3 (color online).   Toy model. (a) The unzipping experi-
ment is modeled with the minimum elements of the experimental
setup. (b) Red or gray curve shows the FDC for one random
realization. The horizontal black line shows the analytical ap-
proximation when disorder is neglected. (c) CUR size distribu-
tions in log-log scale for some values of k using the toy model.
Data plotted with points show the CUR size distribution for the
6.8 kbp sequence. Data plotted with lines, show the average CUR
size distribution over 10^4 realizations ($k = 60$ pN/μm, $d =
0.59$ nm, $\mu = -1.6$ kcal/mol, $\sigma = 0.5$ kcal/mol). (d) The fit
of the average CUR size distributions in panel (c) to Eq. (6) give
the cutoff size n_c. It decreases like $n_c \simeq k^{-2/3}$. Blue or dark gray
curve shows n_c vs k. Red or gray curve shows the maximum
CUR size (n_{max}) predicted by the toy model for the 6.8 kbp
sequence. For $k > 100$ pN/nm, both curves level off to CUR
sizes of 1 bp.

energy of the system can be written as

$$E(x_{tot}, n) = \frac{1}{2}k(x_{tot} - 2dn)^2 - \sum_{i}^{n} \epsilon_i. \qquad (4)$$

At fixed x_{tot}, the system will occupy the state (n^*) that
minimizes the total energy of the system, i.e., $E(x_{tot}, n^*) \leq
E(x_{tot}, n), \ \forall \, n$. The function $n^*(x_{tot})$ gives the thermody-
namic energy function at the minimum, $E_m(x_{tot})$, and the
FDC, $f(x_{tot}) = \frac{\partial E_m(x_{tot})}{\partial x_{tot}}$. The FDC obtained from this model
reproduces the sawtooth pattern that is experimentally
observed [Fig. 3(b)].

   Equation (4) can be approximated by neglecting the
disorder and taking $\epsilon_i = \mu, \ \forall \, i$. This gives,

$$E(x_{tot}, n) \simeq \frac{1}{2}k(x_{tot} - 2dn)^2 - \mu n. \qquad (5)$$

From this approximation we immediately get the following
results: $n^*(x_{tot}) = \frac{1}{2d}(x_{tot} + \frac{\mu}{2dk})$, $E_m(x_{tot}) = -\frac{\mu}{2d}(x_{tot} + \frac{\mu}{4kd})$,
and $f_m = -\frac{\mu}{2d}$. These expressions capture the dependence
of the averaged number of open bps, energy, and force on
the external parameters μ, σ [11].

   Finally, we have numerically computed the CUR size
distribution. We find that this mostly depends on σ and k.
For several combinations of σ and k we simulated 10^4

realizations (i.e., sequences) of 10^4 bp sequences, while d and μ were kept constant. The size distributions are excellently fit by a power law with a superexponential cutoff [11]:

$$P(n) = An^{-B}\exp[-(n/n_c)^C], \qquad (6)$$

where $P(n)$ is the probability of observing a CUR of size n; A, B, C, and n_c (cutoff size) are positive fitting parameters. How much can the toy model predict the experimental results? For the 2.2 kbp sequence the parameters that best fit the experimental histograms are $\mu = -2.80$ kcal/mol, $\sigma = 2.2$ kcal/mol, and $k = 60$ pN/μm (equal to the stiffness of the trap that we can measure independently). This gives $A = 0.058$, $B = 0.42$, $C = 2.95$, and $n_c = 69$ [fit shown in Fig. 2(b)]. For the 6.8 kbp sequence we find $\sigma = 3.3$ kcal/mol while k and μ have the same value. This gives $A = 0.050$, $B = 0.43$, $C = 3.0$, and $n_c = 91$ [fit shown in Fig. 2(d)]. The values of μ and σ are not far from the actual mean and standard deviation of the energies of the nearest neighbor model for DNA, $\mu = -1.6$ kcal/mol, $\sigma = 0.44$ kcal/mol [16]. Having not included the elastic effects of the ssDNA in the toy model we should not expect a good match between the fitting and the experimental values.

What is the limiting factor in detecting small-sized CUR? A look at Figs. 1(c), 2(a), and 2(c), and Figs. S14 and S15 in [11] shows that histograms become smoother as the molecule is progressively unzipped. The increased compliance of the molecular setup as ssDNA is released markedly decreases the resolution in discriminating intermediates. In fact, for the 6.8 kbp construct we found that along the first 1500 bp of the hairpin only 30% of the total number CUR smaller than 10 bp are detected whereas beyond that limit no CUR smaller than that size is discriminated. If we define the threshold size n^{thr} as the size of the CUR above which 50% of the predicted CUR are experimentally detected we find that n^{thr} increases linearly with the number of open bps putting a limit around 10 bp for the smallest CUR size that we can detect [Fig. 2(d), inset]. What is the limiting factor in resolving large-sized CUR into single bps? Only by applying local force on the opening fork (thereby avoiding the large compliance of the molecular setup) and by increasing the stiffness of the probe might be possible to shrink CUR size distributions down to a single bp [8]. Figures 3(c) and 3(d) show how the CUR size distributions shrink and the largest CUR size decreases as the stiffness increases. Its value should be around 50–100 pN/nm for all CUR sizes to collapse into a single bp. Remarkably enough this number is close to the stiffness value expected for an individual DNA nucleotide stretched at the unzipping force [11]. Any probe more rigid than that will not do better. Similarly to the problem of atomic friction between AFM tips and surfaces we can define a parameter η (defined as the ratio between the rigidities of substrate and cantilever) that controls the transition from stick slip to continuous motion [17]. For DNA unzipping we have $\eta = \frac{|\mu|}{kd^2}$ where μ is the average free energy of formation of a single bp, k is the probe stiffness, and d is the interphosphate distance. The value $\eta = 1$ determines the boundary where all CUR are of size equal to one bp ($\eta < 1$). In our experiments we have $\eta \simeq 500$ and to reach the boundary limit $\eta = 1$ we should have $k \sim 100$ pN/nm consistently with what is shown in Figs. 3(c) and 3(d). It is remarkable that the elastic properties of ssDNA lie just at the boundary to allow for one bp discrimination. This suggests that molecular motors that mechanically unwind DNA can locally access the genetic information one bp at a time [11].

In summary, we have measured the distribution of sizes of unzipping regions of DNA. A toy model reproduces the experimental results and can be used to infer the experimental conditions under which the unzipping is done one bp at a time. This is achieved when the stiffness of the probe is higher than 100 pN/nm, which coincides with the stiffness of one base of ssDNA at the unzipping force.

*To whom correspondence should be addressed.
fritort@gmail.com
http://www.ffn.ub.es/ritort/

[1] U. Bockelmann *et al.*, Biophys. J. **82**, 1537 (2002).
[2] M. T. Woodside *et al.*, Science **314**, 1001 (2006).
[3] C. Danilowicz *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 1694 (2003).
[4] M. Rief, H. Clausen-Schaumann, and H. E. Gaub, Nat. Struct. Biol. **6**, 346 (1999).
[5] S. J. Koch *et al.*, Biophys. J. **83**, 1098 (2002).
[6] D. S. Johnson *et al.*, Cell **129**, 1299 (2007).
[7] V. Baldazzi *et al.*, Phys. Rev. Lett. **96**, 128102 (2006).
[8] N. K. Voulgarakis *et al.*, Nano Lett. **6**, 1483 (2006).
[9] D. K. Lubensky and D. R. Nelson, Phys. Rev. E **65**, 031917 (2002).
[10] Experiments were done in a dual-beam miniaturized optical tweezers with fiber-coupled diode lasers (845 nm wavelength) that produce a piezocontrolled movable optical trap and measure force using conservation of light momentum. The experimental setup is based on C. Bustamante and S. B. Smith, U.S. Patent No. 7 133 132 B2 (2006).
[11] See EPAPS Document No. E-PRLTAO-103-149952 for the appendixes. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[12] M. Manosas and F. Ritort, Biophys. J. **88**, 3224 (2005).
[13] M. Manosas *et al.*, Biophys. J. **92**, 3010 (2007).
[14] S. Ares and G. Kalosakas, Nano Lett. **7**, 307 (2007).
[15] T. Ambjörnsson *et al.*, Phys. Rev. Lett. **97**, 128105 (2006); Biophys. J. **92**, 2674 (2007).
[16] J. J. SantaLucia, Proc. Natl. Acad. Sci. U.S.A. **95**, 1460 (1998).
[17] A. Socoliuc *et al.*, Phys. Rev. Lett. **92**, 134301 (2004).