Supplementary information of the paper

Statistical properties of metastable intermediates in DNA unzipping

J. M. Huguet

Departament de Física Fonamental, Facultat de Física,
Universitat de Barcelona, Diagonal 647, E-08028, Barcelona

N. Forns and F. Ritort*

Departament de Física Fonamental,
Facultat de Física, Universitat de Barcelona,
Diagonal 647, E-08028, Barcelona and
CIBER de Bioingeniería, Biomateriales y Nanomedicina,
Instituto de Salud Carlos III, Madrid

(Dated: November 16, 2009)

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

1. **Experimental details** 3
   1.1. DNA sequences 3
   1.2. Measurements calibration and data acquisition 3
   1.3. Statistics and reproducibility of measurements 4

2. **Experimental errors** 4
   2.1. Parameters used in the model 4
   2.2. About using force-distance curves (FDCs) instead of force-extension curves (FECs) 6
   2.3. Base pair determination 7
   2.4. Reproducibility of intermediate histograms 8
   2.5. Error in CUR size distributions 8
   2.6. Discrepancies between experimental and theoretical CUR size distributions 9

3. **The toy model** 10
   3.1. Approximated solution to the toy model 10
   3.2. Size distribution of CUR 11
      3.2.1. Dependence on $\sigma$ 12
      3.2.2. Dependence on $k$ 12

4. **Stiffness of one nucleotide** 14

5. **Protein-DNA interaction** 15

6. **CUR and genes** 16

References 17
1. EXPERIMENTAL DETAILS

1.1. DNA sequences

The two DNAs (2.2 and 6.8 kb) unzipped in the experiments are obtained by gel extraction from λ-DNA. They are synthesized following similar procedures: 1) A restriction enzyme is used to cleave the λ-DNA molecule. The SphI restriction enzyme (BamHI) is used for the 2.2 kb (6.8 kb) sequence. 2) A fragment of 2216 bp (6770 bp) is isolated from the resulting digestion by gel extraction. 3) Two handles of 29 bp are added by hybridization and annealing of short complementary oligonucleotides to one end of the 2.2 kb (6.8 kb) fragment. Another oligonucleotid that forms a tetraloop (5’-acta-3’) is also annealed at the other end of the 2.2 kb (6.8 kb) fragment. The handles were labeled with biotin and digoxigenin that specifically attach to coated polystyrene beads. Figure S1 shows the resulting sequences. Experiments were done in aqueous buffer containing 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 500 mM NaCl and 0.01% Sodium Azide.

1.2. Measurements calibration and data acquisition

The instrument has a force resolution below 1 pN, which represents about 6% uncertainty at the mean unzipping force (16.5 pN). The force is inferred by measuring the deflection of the scattered light by the bead. The offset of the deflected light is measured using a Position Sensitive Detector (PSD), which is converted into force by multiplying it with a calibration factor. The force is calibrated using three different methods and all agree within the aforementioned uncertainty: power spectrum measurements, the Stokes law and the equipartition theorem. The distance measurements have a resolution of 1 nm which represents about 3%. The distance is measured with a light-lever. A small amount of the laser beam is split before it enters the focusing objective and forms the optical trap. The light is redirected to a Position Sensitive Detector (PSD) that measures the position of the center of the optical trap. The PSD is calibrated using a motorized stage with known pitch distance.

The analog signals from the PSDs (position and force) are filtered using an analog low pass filter of bandwidth 1 kHz. The resulting signal is sampled at 4 kHz producing the raw data that we obtain in the unzipping experiment.
1.3. Statistics and reproducibility of measurements

Six different molecules were analyzed for the 2.2 kb and the 6.8 kb DNA sequences.

In fig S2 we show force-distance curves measured for 3 different molecules corresponding to the 2.2 kb and 6.8 kb sequences. As can be seen our measurements are reproducible. Slight differences between different traces are due to the variability of the molecular setup and instrumental drift effects.

2. EXPERIMENTAL ERRORS

2.1. Parameters used in the model

Here we give the numerical values of the parameters used in the model (Eq. (1) main text) to extract the histogram of intermediate states. The bead in the optical trap is modeled by
a Hookean spring, $f = k x_b$, where $k$ is the measured stiffness of the optical trap and $x_b$ is the position of the bead with respect to the center of the optical trap. The measured stiffness of the optical trap is $k = 60 \pm 5$ pN/nm (error is due to bead size heterogeneity). The elastic response of the handles is described by the worm-like chain (WLC) model and parameters are obtained from the literature [1]. The elastic response of the released ssDNA is described using a freely-jointed chain (FJC) model [2]. We use $d = 0.59$ nm for the ssDNA because this value fits well the elastic response of the ssDNA in our data. This value is similar to the one found by Dessinges et al. [3] ($d=0.57$ nm at 1 and 10 mM phosphate buffer) and to Johnson et al. [4] ($d=0.537$ nm at 50 mM NaCl, 20 mM Tris-HCl, pH 7.5). Moreover we consider that interphosphate distance is an effective parameter of an elastic model which does not need to be equal to the parameter measured by crystallography. In order to determine the Kuhn length of the ssDNA, we have proceeded as follows (see fig S3): We fix the interphosphate distance at $d=0.59$ nm and then we determine the Kuhn length of the ssDNA by fitting the last part of the FDC (where the dsDNA duplex is fully unzipped and the elastic response of the ssDNA can be measured) to a Freely Jointed Chain. The best value among the 12 molecules (6 molecules of each) for the Kuhn length is $b = 1.2 \pm 0.3$ nm. We assume an inextensible Freely Jointed Chain model because the effect of a stretch modulus on the elastic
response of the ssDNA is barely negligible below 20 pN, where the unzipping is observed.

![Graph showing force vs. distance for ssDNA unzipping](image)

**FIG. S3**: Fit of the Kuhn length of the ssDNA using the last part of the FDC to one 2.2 kb molecule.

### 2.2. About using force-distance curves (FDCs) instead of force-extension curves (FECs)

We define the distance $x_{\text{tot}}$ as the length between the bead of the micropipette and the center of the optical trap (see Fig. 1a for an illustration of how $x_{\text{tot}}$ is defined and Eq. (1) for a mathematical expression). This magnitude is a measurement that we collect directly from the instrument as the optical trap is moved up and down along the fluidics chamber. This is the control parameter in the experiment, i.e. the variable that does not fluctuate and the parameter that determines the statistical ensemble (what we call mixed ensemble). Since we know the trap stiffness ($k$) and we measure the total distance ($x_{\text{tot}}$) and the force ($f$), it is straightforward to convert the force-distance curve into a force-extension curve using the following relation: $x_m = x_{\text{tot}} - f/x_b$, where $x_m$ is the molecular extension. As we show in the figure S4 we do not appreciate any significant difference when computing the histogram using a FEC or a FDC.
FIG. S4: Histograms of intermediate states calculated using the FDC and the FEC. Note that, apart from some peaks in the beginning, there are no significant differences in the position of the peaks. Eventually, the distribution of CUR sizes will hardly be affected by the choice of FDC or FEC.

2.3. Base pair determination

Here we will discuss what level of uncertainty is introduced in calculating the number of base pairs unzipped through the use of model parameters. The contribution of the handles can be neglected as they behave almost like rigid rods (their contour length is much shorter than the persistence length). We have calculated the same histograms varying the Kuhn length of the ssDNA. When the Kuhn length is modified, the peaks of the histogram are located in a different position. The error introduced might be as large as $\sim 60$ bp when the Kuhn length is varied from 1.2 nm to 1.5 nm. However, the difference between the position of two correlative peaks is weakly affected by the Kuhn length ($\sim 3$ bp). Fig. S5 illustrates these results. Finally, we do not expect important differences in the determination of the CUR sizes by using a slightly different value of the interphosphate distance $d$ because a correction in that value is somehow equivalent to a change in the value of the Kuhn length.
FIG. S5: Dependence of the histogram on the model parameters. Here we show a region of two histograms of intermediate states that have been calculated using different values of the Kuhn length ($b$) for the ssDNA (see Eq. (1) in the main text). We highlight two peaks (green and red arrows) of each histogram that correspond to two consecutive intermediate states. The error in the CUR size due to the Kuhn length is less than 4% (about 3 bps error in a 80 bp-sized CUR).

### 2.4. Reproducibility of intermediate histograms

Histograms in fig S6 show the probability of intermediate states for the 6 different molecules of 6.8 kb (each molecule corresponds to one color). Despite of the fact they look very similar, there are some differences at the beginning (between 0 and 650 bp), mainly due to two reasons: 1) Molecular frying. Some molecules are not capable of completely refold into the DNA duplex and sometimes the first 50-100 bases of the stem remain open. 2) Adhesion between the two beads. The two beads must be very close each other when the DNA is fully zipped because the handles of the molecular construct are very short. Sometimes the beads get stuck and the firsts rips of the unzipping curve cannot be detected.

### 2.5. Error in CUR size distributions

The experimental error of CUR size distributions is negligible as intermediate histograms are fully reproducible among different molecules (see fig. S6). However we can estimate
FIG. S6: Histograms of intermediate states for six different molecules of 6.8 kb. They are depicted in red, green, blue, magenta, orange and dark green. Although the height of each peak is different for the six histograms, the position of the peak is almost the same (±10 bp). The histograms for the 2.2 kb molecules have similar reproducibility.

the error committed in measuring the CUR size distribution among different sequence realizations of the same DNA length. Although that error should vanish for infinitely long sequences (CUR size distributions are self-averaging in the thermodynamic limit) there are large fluctuations for finite length molecules. An experimental measurement of unzipping curves for many DNA sequences is beyond our capabilities. However to estimate that error we can use the toy model to determine the expected standard deviation of the CUR size distributions (see Fig. S7).

2.6. Discrepancies between experimental and theoretical CUR size distributions

Discrepancies between the experimental results and the mesoscopic model are attributed to two factors: 1) Small CUR are missed due to limited instrumental resolution as described in the paper; 2) Medium and large CUR sizes are prone to large error because less than 10 bp CUR are seldom detected. Indeed, the power law describing the CUR size distributions indicates that the majority of CUR is small sized. However if one small sized CUR is missed then medium or large sized CUR will be overcounted as they should split into smaller pieces
FIG. S7: CUR size distributions calculated with the toy model. Upper (lower) panel shows the results for a 2252 (6838 kb) sequence. The black curve shows CUR size distribution averaged over $10^4$ realizations. The green region represents the upper and lower limits of the error bars that correspond to the standard deviation of those realizations. Red, blue and orange curves show 3 different realizations. Note the large deviations from the average histogram due to the finite length of the sequences.

whenever they contain a small CUR. It is a difficult math problem to evaluate the final effect of all missed CUR in the resulting CUR size histogram. These two effects concur to modify the shape of the power law for the 6.8kb molecule in Fig. 2d. Yet it is remarkable that the general trend of the experimental data shown in Figs. 2c and 2d follows reasonably well the predictions of the mesoscopic model. This is specially true for the 2.2 kb data shown in Fig. 2b where the non-monotonic oscillations observed in the experimental distribution are captured by the mesoscopic model.

3. THE TOY MODEL

3.1. Approximated solution to the toy model

The toy model (Eq. (4) in the main text) gathers all the relevant features of a DNA unzipping experiment. The most interesting of them are the force rips in the FDC and
FIG. S8: Average behavior and one realization of the toy model. Black curves in all the panels show the approximated solution (see main text for the expressions) and colored curves show the solution for one disorder realization. The following parameters have been used: $k=60$ pN/µm, $d=0.59$ nm, $\mu=-1.6$ kcal/mol, $\sigma=3.20$ kcal/mol. (a) Energy minimum $E_m(x_{tot})$. (b) Number of open base pairs vs. total distance. Inset shows a detailed view of the stair-shaped character of the curve. (c) FDC.

3.2. Size distribution of CUR

This section shows the results of the simulation of the toy model introduced in the main text. The distribution of sizes of the CUR depends on the parameters of the model. The aim of this section is to characterize such dependency. Starting from the energy contribution of the model (see Eq. 4 in main text for further details), we generate random realizations and obtain the CUR size distribution for each realization of the disorder. After collecting all the simulated data we obtain the averaged size distribution of the CUR for the chosen
parameters. By varying the parameters of the model along a wide range we observe how the shape of the CUR size distribution changes.

In all our simulations we took $d = 0.59$ nm and $\mu = -1.6$ kcal/mol constant, since the distribution of CUR sizes weakly depends on them. Therefore we only changed $\sigma$ (the standard deviation of the random distribution of energies) and $k$ (the stiffness of the optical trap). We simulated sequences of $10^4$ base pairs and we made $10^4$ realizations for each value of $\sigma$ and $k$.

3.2.1. Dependence on $\sigma$

We fixed the trap stiffness at $k=60$ pN/μm. The distribution of CUR obtained for each value of $\sigma$ is shown in fig S9. The data was fit to Eq. (6) in main text, where a set of 4 parameters ($A, B, C, n_c$) was obtained for each value of the parameter $\sigma$. Fig. S10 shows the dependence of these parameters with $\sigma$.

3.2.2. Dependence on $k$

We fixed the amount of disorder at $\sigma=3.20$ kcal/mol. Figure S11 shows the distribution of CUR for some values of $k$ and their fit to Eq. (6) in main text. Note that in the low $k$
FIG. S10: Fit parameters plotted versus $\sigma$.

FIG. S11: CUR size distributions for different values of $k$. The black curves show the fit to Eq. (6) in main text. The tails of the CUR distributions for trap stiffnesses higher than 3 pN/nm need many realizations to be accurately inferred. The distributions are too narrow and the fit of Eq. (3) does not converge easily. Note that at $k = 100$ pN/nm all CUR are one base pair sized.

range the CUR size distributions are wide and have good statistics to extract the values for $A, B, C, n_c$. However, for $k > 5$ he CUR size distributions are too narrow to be reliably fit to Eq. (6). Figure S12 shows the dependence of the four parameters ($A, B, C, n_c$) on the trap stiffness.
FIG. S12: Fit parameters plotted versus $k$. The cutoff CUR size ($n_c$) vs. $k$ follows a power law behavior (see Fig. 3d in main text for a log-log plot).

4. STIFFNESS OF ONE NUCLEOTIDE

Here we calculate the expected stiffness of one nucleotide of ssDNA. The numerical value has been calculated from the elastic response of Freely Jointed Chain (FJC) model for semiflexible polymers, which is given by the following Extension vs. Force curve,

$$x_s(f) = L_0 \left( \coth \left( \frac{bf}{k_B T} \right) - \frac{k_B T}{bf} \right)$$  \hspace{1cm} (1)

where $x_s$ is the extension, $f$ is the force applied at the ends of the polymer, $L_0$ is the contour length, $b$ is the Kuhn length, $k_B$ is the Boltzmann constant and $T$ is the temperature. In the case of a polymer, the contour length ($L_0$) can be written in terms of the number of monomers ($n$) times the length of one monomer ($d$) according to

$$L_0 = n \cdot d$$  \hspace{1cm} (2)

In the case of a ssDNA molecule, $n$ is the number of bases and $d$ is the interphosphate distance of one nucleotide. The FJC model assumes that the elastic response of the polymer scales with the number of bases. Therefore, the resulting Extension vs. Force expression is a homogeneous function with respect to the number of bases. The stiffness of the polymer at each stretching force is the derivative of the force with respect to the extension $k_s(f) =$
\[
\frac{df}{dx_s} = (dx_s/df)^{-1}. \text{ For the FJC model, the stiffness is given by the following expression}
\]

\[
k_s(f) = \left[ n \cdot d \left(-\frac{b}{k_BT} \text{cosech}^2 \left( \frac{bf}{k_BT} \right) + \frac{k_BT}{bf^2} \right) \right]^{-1}
\]  

(3)

Using the parameters from section 2.2.1 \((b = 1.2 \text{ nm}, d = 0.59 \text{ nm})\) for one nucleotide \((n = 1)\) we get a stiffness of \(k_s = 113 \text{ pN/nm} \) at \(f = 15 \text{ pN}\) and \(k_s = 127 \text{ pN/nm} \) at \(f = 16 \text{ pN}\) (see Fig. S13).

![Graph](image)

FIG. S13: Stiffness of one nucleotide vs. force. Inset shows a zoomed section of the curve around the unzipping force.

5. PROTEIN-DNA INTERACTION

In the cell, the function of helicases is to unzip DNA during the replication process. Although their mechano-chemistry is not clear [5] we interpret that helicases pull directly on the ssDNA. In this simplified view of the process, we visualize the helicase as a clamp that slides along one strand of the DNA and applies local force at the unzipping fork. In a more general scheme, the helicase applies force on the DNA by means of an effective stiffness \(k^{-1}_{\text{eff}} = k^{-1}_{h} + k^{-1}_{s}\), where \(k_h\) is the stiffness of the helicase and \(k_s\) is the stiffness of one base of ssDNA. From the conclusions of our work, we know that \(k_s\) is high enough to locally unzip DNA one bp at a time. Therefore, the unzipping process will be one bp at a time as long as \(k_h\) is higher than \(k_s\). Indeed, when the helicase pulls directly on DNA the stiffness of the helicase can be assumed to be very large (proteins are indeed very rigid objects) compared
to the stiffness of a single base pair \((k_h^{-1} \ll k_s^{-1})\) and the effective stiffness between the helicase and the DNA is approximately equal to the stiffness of ssDNA \((k_{\text{eff}} \sim k_s)\).

The previous explanation can be extended to proteins that interact with DNA. If a protein increases the stiffness of one bp of ssDNA, the local unzipping still could be done one bp at a time. On the other hand, if a protein decreases the ssDNA stiffness below the boundary of \(k_s \sim 100 \text{ pN/nm}\) the local unzipping would show CUR of sizes larger than one bp. As far as we know, there is no protein with high compliance bound to the ssDNA between the helicase and the unzipping fork when the replication complex (helicase, polymerase, etc.) is set. However the full scenario of what might happen for different biological models under varied conditions remains to be seen.

6. CUR AND GENES

As an extra information to the reader, here we show the position of the genes that are localized in the two fragments of \(\lambda\)-DNA used in this work. The 2.2 kb fragment contains partially one gene and one complete gene. Upper panel in Fig. S14 shows the localization of the genes along the Force Distance Curve. Lower panel in Fig. S14 shows the position of the genes superimposed on the histogram of number of unzipped base pairs. On the other hand, the 6.8 kb fragment contains partially one gene and 15 complete genes. Figure S15 shows the location of these genes superimposed on the histogram of number of unzipped base pairs. It can be clearly observed that the lengths of most of these genes span over several rips.

We should not expect correlations between the CUR and the genes because the CUR depend on the trap stiffness used in the experimental setup. In other words, a different trap stiffness produces a different distribution of CUR on the same DNA molecule.
FIG. S14: Localization of genes for the 2.2 kb sequence. (a) Start and end points of each gene. The arrow shows the transcription direction. Each gene is depicted in a different color. (b) Position of the genes along the histogram of number of unzipped base pairs.

FIG. S15: Localization of genes for the 6.8 kb sequence. Same color code as in Fig. S14.