

# Inferring DNA sequences from mechanical unzipping data: the large-bandwidth case.

V. Baldazzi<sup>1,2,3</sup>, S. Bradde<sup>2,4</sup>, S. Cocco<sup>2</sup>, E. Marinari<sup>4</sup>, R. Monasson<sup>3</sup>

<sup>1</sup> *Dipartimento di Fisica, Università di Roma Tor Vergata, Roma, Italy*

<sup>2</sup> *CNRS-Laboratoire de Physique Statistique de l'ENS, 24 rue Lhomond, 75005 Paris, France*

<sup>3</sup> *CNRS-Laboratoire de Physique Théorique de l'ENS, 24 rue Lhomond, 75005 Paris, France*

<sup>4</sup> *Dipartimento di Fisica and INFN, Università di Roma La Sapienza, P.le Aldo Moro 2, 00185 Roma, Italy*

The complementary strands of DNA molecules can be separated when stretched apart by a force; the unzipping signal is correlated to the base content of the sequence but is affected by thermal and instrumental noise. We consider here the ideal case where opening events are known to a very good time resolution (very large bandwidth), and study how the sequence can be reconstructed from the unzipping data. Our approach relies on the use of statistical Bayesian inference and of Viterbi decoding algorithm. Performances are studied numerically on Monte Carlo generated data, and analytically. We show how multiple unzippings of the same molecule may be exploited to improve the quality of the prediction, and calculate analytically the number of required unzippings as a function of the bandwidth, the sequence content, the elasticity parameters of the unzipped strands.

## I. INTRODUCTION

As DNA molecules are the support for the genetic information, the knowledge of their sequence content is very important both from the biological and medical points of view. Over the last decade the sequencing of various genomes, in particular the human one, was done at the price of intense efforts. A traditional strategy for reading a DNA molecule is based on the so-called Sanger method [1, 2]. The DNA molecule is divided into fragments (with  $N \sim 100 - 1000$  base pairs); each fragment is amplified through PCR. The copies of each fragment are denaturated, and double-stranded DNA subfragments are synthesized under the action of DNA polymerases. The key point is that each of the four nucleotides  $A, T, C, G$  is present in solution under its normal form at high concentration and under a modified form, tagged with a base-specific fluorescent label and inadequate for further polymerization, at low concentration. At the end of the polymerization step many copies of each fragment are obtained. The copies of a fragment have a common extremity and have various lengths  $L$ , with a base-specific fluorescent base  $B$  at the end. The entire population of copies is sorted by length using gel electrophoresis and the sequence of the fragment is reconstructed from the list of terminal bases  $B(L)$ ,  $1 \leq L \leq N$ . The method correctly predicts 99.9% of the bases of a fragment, but additional errors may arise during the reconstruction of the whole sequence from its fragments.

Despite the success of conventional sequencing the quest for alternative (faster or cheaper) methods is an active field of research. Recently various single molecule experiments were carried out, allowing a direct investigation of DNA mechanics and protein-DNA interaction [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. These experiments provide dynamical information usually hidden in large scale bulk experiments, such as intermediate metastable states or fluctuations at the scale of the individual molecule. Remarkably, these dynamical effects are largely sequence-dependent in various experimental situations e.g. the opening of the double helix under a mechanical stress [8, 9, 10, 11, 12, 13, 14, 15], the digestion of a DNA molecule by an exonuclease [16, 17], DNA polymerization [18, 19, 20], translocation through nanopores [22, 23]. Understanding how much information about the sequence is contained in the measured signals is important.

Hereafter, we focus on mechanical unzipping experiments, first introduced by Bockelmann and Heslot in 1997 [8]. The complementary strands are pulled apart at a constant velocity while the force necessary to the opening is measured. The average opening force for the  $\lambda$  phage is of about 15 pN, with fluctuations around this value that depend on the particular sequence content. In a more recent experiment, Bockelmann, Heslot and collaborators have shown that the force signal is correlated to the average sequence on the scale of ten base pairs but could be affected by the mutation of one base pair adequately located along the sequence [10].

Liphart et al. [13] and Danilowicz et al. [14] have performed an analogous experiment, using a constant force setup, on a short RNA and a long DNA respectively. As sketched in Fig 1, the distance between the two strands extremities is measured as a function of the time while the molecule is submitted to a constant force. The dynamics is characterized by rapid zipping or unzipping jumps followed by long pauses where the unzipped length remains constant. Several repetitions have shown that positions and duration of these plateaus are largely reproducible, thus providing a 'fingerprint' of the sequence. The theoretical description of the DNA mechanical unzipping, at constant velocity and constant force, has been extensively developed [9, 12, 24, 25, 26, 27, 28, 29, 30, 31, 32]. Models have been able to reproduce the force (for constant velocity experiments) or position (for constant force experiments) signals given the DNA sequence. It is a natural question to ask whether one could, inversely, get information about the sequence from experimental data [33].

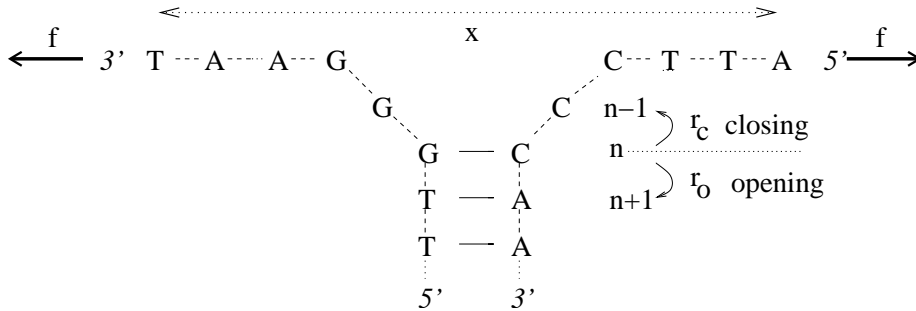


FIG. 1: Sketch of a fixed-force unzipping experiment: the adjacent 5' and 3' extremities of a DNA molecule are submitted to a constant force  $f$ . The distance between the extremities,  $x$ , is measured as a function of time.  $x$  is proportional to the number  $n$  of open base pairs (bp) up to some fluctuations due to the floppiness of the unzipped strands. The number  $n$  of open bp increases or decreases by one with rates  $r_o$  and  $r_c$  respectively, see dynamical model in Section II A.

This question was addressed by us in a recent letter [34]. It was found that the error in the prediction e.g. the probability that a base is erroneously predicted decreases exponentially with the amount of available data. The decay rate was shown to depend on the sequence content, the applied force, the time and space resolution, ... The goal of the present paper is to provide a complete presentation of the numerical and analytical work supporting the results of [34] in the idealized case of perfect time and space resolutions. Though this case is not realistic from an experimental point of view, it can be studied in great detail. We show that the most important result, the exponential decay of the probability of misprediction with the amount of collected data, holds in more realistic situation where the bandwidth and the fluctuations in the extension of the DNA strands are taken into account. Our analysis focuses on the fixed force device data which is somewhat simpler from a theoretical point of view.

In Section II we first introduce the dynamical model that, given a sequence, determines the unzipping signal. The inverse problem is then introduced and treated within the Bayesian inference framework. Section III reports the numerical results for the quality of prediction from numerical data obtained from the Monte Carlo simulation of the unzipping of a  $\lambda$ -phage DNA. The analytical study of inference performances is presented in section IV. While the above study assumed the existence of infinite temporal and spatial resolution over the fork location the effects of realistic limitations are studied in Section V. A summary and discussion of the results is presented in Section VI.

## II. BAYESIAN INFERENCE FRAMEWORK

The direct problem of fixed-force DNA unzipping is to determine, given the sequence of the molecule, the distribution of the stochastic measured signal, that is, the extension between the two strands extremities as a function of time. The direct problem is considered in Section II A, and results are used in Section II B to address the inverse problem, that is, the prediction of the sequence given a measured extension signal.

Throughout this section we consider that the experimental signal gives access to the number of open bases itself rather than the distance between the extremities of the unzipped strands. This is merely an approximation since, due to the fluctuations in the extension of strands, the number of open bases is not in one-to-one correspondence with the distance between the strands. Corrections to this simplifying assumption will be discussed in Section V B.

### A. From sequence to signal: the direct problem

In a previous work we have developed a theoretical description of the dynamics of DNA and simple RNA molecules under a constant unzipping force [28]. Despite its simplicity this model is capable of reproducing the unzipping data for a given sequence [13, 14] and the rezipping dynamics of a partially unzipped DNA [11].

Let  $b_i = A, T, C$ , or  $G$  denote the  $i^{\text{th}}$  base along the  $5' \rightarrow 3'$  strand (the other strand is complementary), and  $B = \{b_1, b_2, \dots, b_N\}$ . The free energy excess when the first  $n$  bp of the molecule are open with respect to the closed configuration ( $n = 0$ ) is

$$G(n, f; B) = \sum_{i=1}^n g_0(b_i, b_{i+1}) - n g_s(f). \quad (1)$$

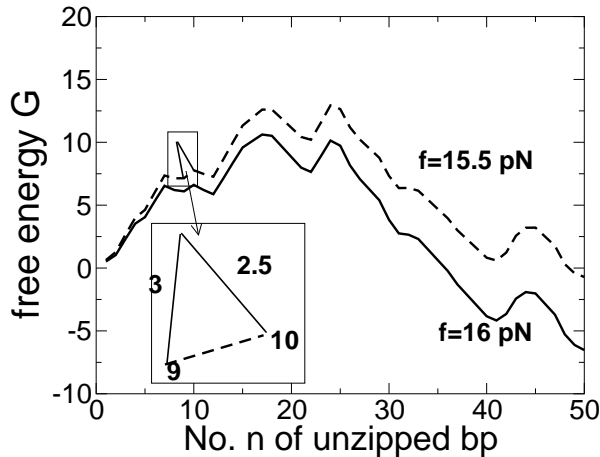


FIG. 2: Free energy  $G$  (units of  $k_B T$ ) to open the first  $n$  base pairs, for the first 50 bases of the DNA  $\lambda$ -phage at forces 15.9 (dashed curve) and 16.4 pN (full curve). For  $f = 15.9$  pN the two minima at bp 1 and bp 50 are separated by a barrier of 12  $k_B T$ . Inset: additional barrier representing the dynamical rates (3) to go from base 10 to 9 (barrier equal to  $g_s = 2.5$   $k_B T$ ), and from base 9 to 10 (barrier equal to  $g_0(b_9, b_{10}) = 3$   $k_B T$ ), see text.

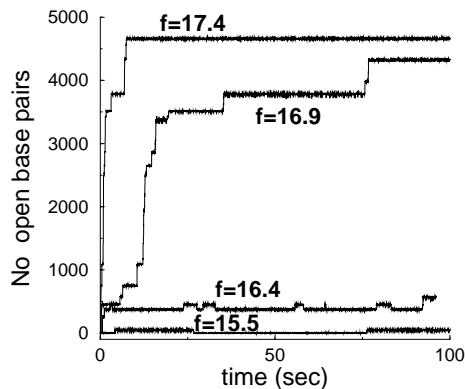


FIG. 3: Number of open base pairs as a function of the time for various forces (shown on Figure). Data show one numerical unzipping (for each force) obtained from a Monte Carlo simulation of the random walk motion of the fork with rates (3).

and involves two contributions. The first free energy, called  $g_0(b_i, b_{i+1})$  is the binding energy of base pair (bp) number  $i$ ; it depends on  $b_i$  (pairing interactions) and on the neighboring bp  $b_{i+1}$  due to stacking interactions.  $g_0$  is obtained from the MFOLD server [35, 36], and listed in Table I. The second contribution, called  $g_s(f)$  is the work to stretch the two opened single strands when one more bp is opened. The elasticity of DNA strands is described by a modified freely jointed chain with a Kuhn length  $\ell_0 = 15 \text{ \AA}$  and an effective nucleotide length  $\ell = 5.6 \text{ \AA}$  [7]. The corresponding

$g_0$	A	T	C	G
A	1.78	1.55	2.52	2.22
T	1.06	1.78	2.28	2.54
C	2.54	2.22	3.14	3.85
G	2.28	2.52	3.90	3.14

TABLE I: Binding free energies  $g_0(b_i, b_{i+1})$  (units of  $k_B T$ ) obtained from the MFOLD server [35, 36] for DNA at room temperature, pH=7.5, and ionic concentration of 0.15 M. The base values  $b_i, b_{i+1}$  are given by the line and column respectively.

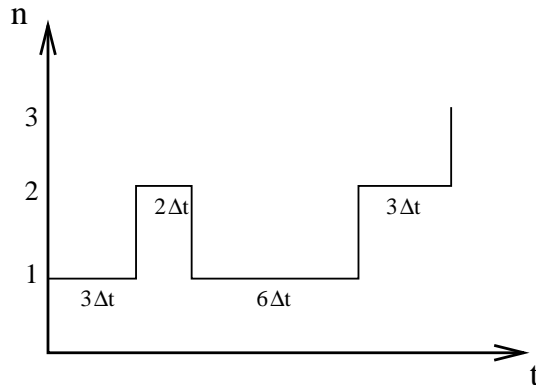


FIG. 4: Fork position  $n$  as a function of time  $t = i \times \Delta t$  with  $i$  integer-valued; the sojourn times on each base are given. We call  $t_i$  the total time spent on base  $i$ , and  $u_i, d_i$  the numbers of  $i \rightarrow i + 1, i \rightarrow i - 1$  transitions respectively. Assuming the fork does not come back to  $n = 1$  or  $2$  at later times, we have:  $t_1/\Delta t = 9$ ,  $u_1 = 2$ ,  $d_1 = 0$ , and  $t_2/\Delta t = 5$ ,  $u_2 = 1$ ,  $d_2 = 1$ .

free energy for forces up to 20 pN is

$$g_s(f) = 2 f \ell \ln [\sinh(z)/z] / z \quad \text{with} \quad z \equiv f \ell_0 / (k_B T). \quad (2)$$

As an illustration the free energy  $G(n, f; \Lambda)$  of the first 50 bases of the  $\lambda$  phage sequence,  $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_N)$ , is plotted in Fig 2 for forces  $f = 15.9$  and  $16.4$  pN. At these forces the two global minima are located in  $n = 1$  (closed state) and  $n = 50$  (partially open state). Experiments on a small RNA molecule, called P5ab, [13] have shown that, at the critical force  $f_c$  such that the closed state has the same free energy than the open one:  $G(0, f_c; B) = G(N, f_c; B)$ , the barrier between these two minima is not too high, the molecule then switches between these two states. For long molecule e.g.  $\lambda$ -DNA the barrier between the closed and open states may become very large e.g.  $\sim 3000 k_B T$  for the  $\lambda$ -DNA at the critical force  $f_c = 15.5$  pN [28]. The time it takes to cross this barrier is huge and full opening of the molecule never happens during experiments (unless the force is chosen to be much larger than its critical, infinite time value). The experimental opening signal is characterized by pauses at local minima of the free energy  $G(n, f; \Lambda)$  and rapid jumps between them [14]. This dynamical behavior is reproduced (Fig 3) when one considers that the fork separating the closed from the open regions along the molecule undergoes a random walk motion in the free energy landscape  $G(n, f; \Lambda)$  [28]. The fork, located at position  $n$ , can move forward ( $n \rightarrow n + 1$ ) or backward ( $n \rightarrow n - 1$ ) with rates (probability per unit of time) equal to, respectively,

$$r_o(b_n, b_{n+1}) = r \exp [g_0(b_n, b_{n+1})], \quad r_c = r \exp [g_s(f)] \quad (3)$$

see Fig 1. The value of the attempt frequency  $r$  is of the order of  $10^6$  Hz [11, 28, 30]. Notice that the free-energies are measured in units of  $k_B T$ .

The expression (3) for the rates is derived from the following assumptions. First the rates should satisfy detailed balance. Secondly we impose that the opening rate  $r_o$  depends on the binding free energy, and not on the force, and vice-versa for the closing rate  $r_c$ . This choice is motivated by the fact that the range for base pairs interaction is very small: the hydrogen and stacking bonds are broken when the bases are kept apart at a fraction of  $\text{\AA}$ , while the force work is appreciable on the distance of the opened bases ( $\approx 1$  nm). On the contrary, to close the base pairs, one has first to work against the applied force, therefore the closing rate  $r_c$  depends on the force but not on the sequence. This physical origin of the rates is reported in the inset of Fig 2. Notice that, as room temperature is much smaller than the thermal denaturation temperature, we safely discard the existence of denatured bubble in the zipped DNA portion.

## B. From signal to sequence: the inverse problem.

We consider here the ideal case where the experimental setup is not affected by any instrumental noise: data are acquired with a infinite temporal resolution, and, in addition, the unzipped strands do not fluctuate in length. The latter assumption will be lifted in Section V B, while the case of a large but not infinite bandwidth will be studied in Section V A.

In the absence of DNA strands fluctuations the distance between extremities is exactly proportional to the number  $n$  of unzipped bases. The measured signal is thus the time trace  $T = (i_0, i_1, i_2, \dots, i_M)$  where  $i_m$  is the position of

the fork at time  $m \times \Delta t$ , and  $t_{exp} = M \Delta t$  is the duration of the experiment. The infinite bandwidth assumption amounts to postulate that the delay  $\Delta t$  between two measures is smaller than the sojourn time on a base. Therefore successive positions  $i_m, i_{m+1}$  differ by  $\pm 1$  at most. A typical result of this idealized experimental situation is sketched in Fig. 4. The signal is stochastic due to the thermal motion of the fork in the landscape of Fig 2: two repetitions of the experiment do not yield the same time-traces. The probability of a time-trace  $T$ , given the sequence  $B$ , reads

$$\mathcal{P}(T|B) = \prod_{m=1}^{M-1} \begin{cases} \Delta t r_o(b_{i_m}, b_{i_{m+1}}) & \text{if } i_{m+1} = i_m + 1 \\ \Delta t r_c & \text{if } i_{m+1} = i_m - 1 \\ 1 - \Delta t (r_o(b_{i_m}, b_{i_{m+1}}) + r_c) & \text{if } i_{m+1} = i_m \end{cases} . \quad (4)$$

This probability can be conveniently rewritten through the introduction of the numbers  $u_i$  and  $d_i$  of, respectively, up ( $i_m = i \rightarrow i_{m+1} = i + 1$ ) and down ( $i_m = i \rightarrow i_{m+1} = i - 1$ ) transitions from base  $i$ , as well as the total time  $t_i$  spent on base  $i$  (number of sojourn events  $i_m = i \rightarrow i_{m+1} = i$ , multiplied by  $\Delta t$ ) in the time-trace  $T$ ,

$$\mathcal{P}(T|B) = \prod_i [\Delta t r_o(b_i, b_{i+1})]^{u_i} [\Delta t r_c]^{d_i} [1 - \Delta t (r_o(b_i, b_{i+1}) + r_c)]^{t_i/\Delta t} = C(T) \times \prod_i M(b_i, b_{i+1}; u_i, t_i) \quad (5)$$

where

$$M(b_i, b_{i+1}; t_i, u_i) = \exp [g_0(b_i, b_{i+1}) u_i - r e^{g_0(b_i, b_{i+1})} t_i] \quad (6)$$

and  $C(T) = \Delta t^{u+d} r_c^d \exp(-r_c t_{exp})$ ,  $u = \sum_i u_i$ ,  $d = \sum_i d_i$ , and we have used the fact that  $\Delta t$  is small with respect to the average sojourn time on a base,  $(r_o + r_c)^{-1}$ . Up to the multiplicative factor  $C(T)$  (which does not depend on the sequence  $B$ ), the probability  $\mathcal{P}(T|B)$  is equal to the product of terms  $M$  expressing the interactions between adjacent bases (6).

The probability that the DNA sequence is  $B$  given the observed time-trace  $T$  is, in the Bayesian inference framework [37],

$$\mathcal{P}(B|T) = \frac{\mathcal{P}(T|B) \mathcal{P}_0(B)}{\mathcal{P}(T)} \quad (7)$$

The value  $B^*(T)$  of the sequence maximizing this probability, for a given time-trace  $T$ , is our prediction for the sequence. In the absence of any knowledge over the sequence  $B$  the *a priori* distribution over the sequences,  $\mathcal{P}_0$ , is uniform and equal to  $4^{-N}$ . A straightforward albeit important consequence of (7) is that  $B^*(T)$  can be found from the maximization of  $\mathcal{P}(T|B)$  (5). We will briefly see in Section III B an alternative way of predicting sequences from the probability (7).

In practice  $B^*(T)$  can be exactly found in a time growing linearly with  $N$  only with the Viterbi algorithm [37, 38]. The principle of the algorithm is equivalent to a zero temperature transfer matrix technique. We start from the first base and choose the optimal value of this base for each possible value of the second one; in this way we assign a probability  $P_2$  to each value  $b_2$  of the second base through  $P_2(b_2) = \max_{b_1} M(b_1, b_2; t_1, u_1)$ . Then we optimize on the second base, and obtain  $P_3(b_3) = \max_{b_2} M(b_2, b_3; t_2, u_2) P_2(b_2)$ , and so on,

$$P_{i+1}(b_{i+1}) = \max_{b_i} M(b_i, b_{i+1}; t_i, u_i) P_i(b_i) \quad (8)$$

until we reach the last base  $N$  of the sequence. At each step, the maximum of (8) is reached for some base  $b_i^{max}(b_{i+1})$  that depends on the choice of the next base  $b_{i+1}$ . Once the value  $b_N^*$  that optimize  $P_N(b_N)$  has been calculated, one obtains the whole optimal sequence using the recursive relation  $b_{i-1} = b_{i-1}^{max}(b_i^*)$  until the first base of the chain.

A direct application of the procedure may produce substantial numerical errors due to the product of a large number of terms. It turns out convenient to introduce the logarithms of the probabilities,  $\pi_i(b_i) = -\ln P_i(b_i)$ , and solve the recurrence relation

$$\pi_{i+1}(b_{i+1}) = \min_{b_i} [\pi_i(b_i) - g_0(b_i, b_{i+1}) u_i + r e^{g_0(b_i, b_{i+1})} t_i] , \quad (9)$$

obtained from (8).

If more than one unzippings are performed on the same molecule, several time-traces  $T_1, T_2, \dots, T_R$  are available. As all unzippings are independent of each other we have

$$\mathcal{P}(T_1, T_2, \dots, T_R|B) = \prod_{\rho=1}^R \mathcal{P}(T_\rho|B) \quad (10)$$

where the distribution of a single time-trace is given by (5). It is immediate to check that equations (8) and (9) are still valid provided  $u_i$  and  $t_i$  are, respectively, the total number of transitions  $i \rightarrow i + 1$  and the total time spent on base  $i$ . Total means that these numbers have to be computed from the all  $R$  time-traces taken together.

### C. Estimators of performances

As in the previous Section, we consider a time-trace  $T$ , and call  $B^*(T)$  the sequence with maximal probability given those data. The true sequence is denoted by  $B^L$ ; in most applications  $B^L = \Lambda$ , the phage sequence but we will consider other e.g. repeated sequences. We focus on the indicators

$$v_i(T) = \begin{cases} 1 & \text{if base } i \text{ is correctly predicted } i.e. b_i^*(T) = b_i^L \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

As the time-trace  $T$  is stochastic, so are the  $v_i(T)$ s. Our numerical and theoretical analysis aim at calculating some statistical properties of these indicators. For instance the probability that base  $i$  is not correctly predicted is given by

$$\epsilon_i = 1 - \langle v_i(T) \rangle, \quad (12)$$

where the average value  $\langle . \rangle$  is taken over the probability  $\mathcal{P}(T|B^L)$  of time-traces given the true sequence  $B^L$ . The two-points connected correlation function,

$$\chi_{i,j} = \langle v_i(T) v_j(T) \rangle - \langle v_i(T) \rangle \langle v_j(T) \rangle, \quad (13)$$

tells us how much a correct prediction on base  $i$  influences the quality of prediction on base  $j$ . From this local quantities we define the global error and correlation functions through, respectively,

$$\epsilon = \frac{1}{N} \sum_{i=1}^N \epsilon_i, \quad \chi_d = \frac{1}{N-d} \sum_{i=1}^{N-d} \chi_{i,i+d}. \quad (14)$$

Note that the zero-distance correlation function is simply  $\chi_0 = \epsilon(1 - \epsilon)$  in the limit of large sequences.

## III. NUMERICAL ANALYSIS

### A. Maximum probability prediction

To test this inference method we have generated ideal opening data from the sequence  $\Lambda$  of the  $\lambda$ -phage with a Monte Carlo procedure. Once a time-trace  $T$  has been produced a second program ignoring the phage sequence and based on the Viterbi algorithm allows us to make a prediction on the sequence,  $B^*(T)$ .

#### 1. Generation of numerical time-traces

The unzipping signal  $T$  is obtained through a Monte Carlo (MC) simulation with opening and closing rate defined by the model (3). To save time, at each MC step, the fork moves by one base pair, either forward or backwards, without remaining on the same base. Prior to the move the sojourn time  $t$  on the base where the fork is, say,  $i$ , is randomly chosen according to an exponential distribution with characteristic time  $\tau = 1/(r_o(i) + r_c)$ . Then, the fork moves backward ( $i \rightarrow i - 1$ ) with probability  $q = r_c \tau$ , and forward ( $i \rightarrow i + 1$ ) with probability  $1 - q$ .

The total number of open base pairs increases with the duration of the opening experiments *i.e.* with the number of MC steps as shown in Fig 5. The higher the force the more tilted the free energy landscape, and the larger is the number of open bases. With  $10^7$  MC steps we typically open 290 bp at 15.9 pN, 450 bp at 16.4 pN, and 4700 bp at 17.4pN; each numerical unzipping lasts for  $\sim 15$  sec.

The temporal resolution is introduced by filtering the output dynamics with a time step  $\Delta t$ . Fork positions  $n_i$  are registered at times  $t_i = i \times \Delta t$ . Each time-trace is then preprocessed to obtain the numbers  $u_i$  of  $i \rightarrow i + 1$  transitions and the set of times  $t_i$  spent on each base  $i$ . The set of data  $\{u_i, t_i\}$  is then passed to the Viterbi procedure.

#### 2. Results for global estimators

We show in Fig 6 the average fraction of mispredicted bases,  $\epsilon$  (14), as a function of the force. For each time-trace we calculate the fraction of the opened bases that were incorrectly predicted, and then average over MC time-traces (samples).  $\epsilon$  increases with the force because the number of predicted (open) base pairs (Fig 5) increases, and the

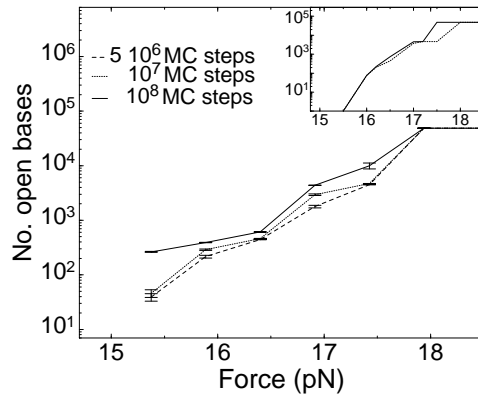


FIG. 5: Number of open bases as a function of applied force, and for  $5 \times 10^6$ ,  $10^7$ ,  $10^8$  Monte Carlo steps. Data are averaged over 100 samples. The durations of the unzippings are, respectively, of 7, 15, and 140 seconds. The DNA  $\lambda$ -phage includes 48,502 bp. In inset we report the theoretical estimate of the number of open base pairs, for  $10^7$  and  $10^8$  MC steps, of Section IV D 1.

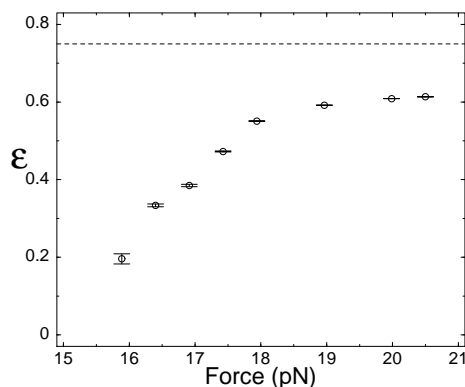


FIG. 6: Fraction  $\epsilon$  (14) of mispredicted bases as a function of the force for the  $\lambda$ -phage sequence. Data are averaged over 100 samples and shown with standard deviations. The dotted line  $\epsilon = 0.75$  shows the failure rate for a random choice of one base among the four base values.

time the opening fork spends on each base decreases. At a force of 16 pN 80% of the predicted bases are correct. As the force increases  $\epsilon$  approaches 0.75, which corresponds to a random guess among four possible bases.

The quality of prediction is, not surprisingly, greatly improved by the repetition of the numerical unzipping on the same molecule. Let  $R$  denote the number of time-traces (of the same duration) available. We show in Fig. 7 how the error  $\epsilon$  decreases with  $R$ . Notice that the error is calculated over the bp that have been opened at least once in all  $R$  unzippings. When opening and closing several times the molecule, the opening fork makes multiple passages through the same portion of the sequence; in this way more information on the waiting and transition times on each base are collected, and processed altogether by the Viterbi algorithm. Figure 7 indicates that the error decreases exponentially with  $R$ , an observation that will find theoretical support in Section IV.

### 3. Results for local estimators

Figure 8A (dashed curve) show the errors  $\epsilon_i$  for the first 450 bases of the  $\lambda$ -phage at  $f = 16.4$  pN. Comparison with the free energy landscape  $G(n, f; \Lambda)$  (1) at the same force shows that the best predicted bases correspond to valleys (Fig 9 top), in which the fork spends a lot of time, while prediction for bp located on the top of barriers are much poorer. In addition Fig. 8A shows that the errors  $\epsilon_i$  sharply decrease when the prediction is made from  $R = 40$  unzippings.

We have investigated in detail the decay of the error  $\epsilon_i$  with  $R$  for two arbitrarily selected bases  $i = 6$  and  $i = 27$ . Figure 9(top) shows that bp 6 is located in a valley of the free energy landscape at force  $f = 16.4$  pN while base pair 27 is located on a barrier at the same force. Figure 10 shows that the error decays exponentially

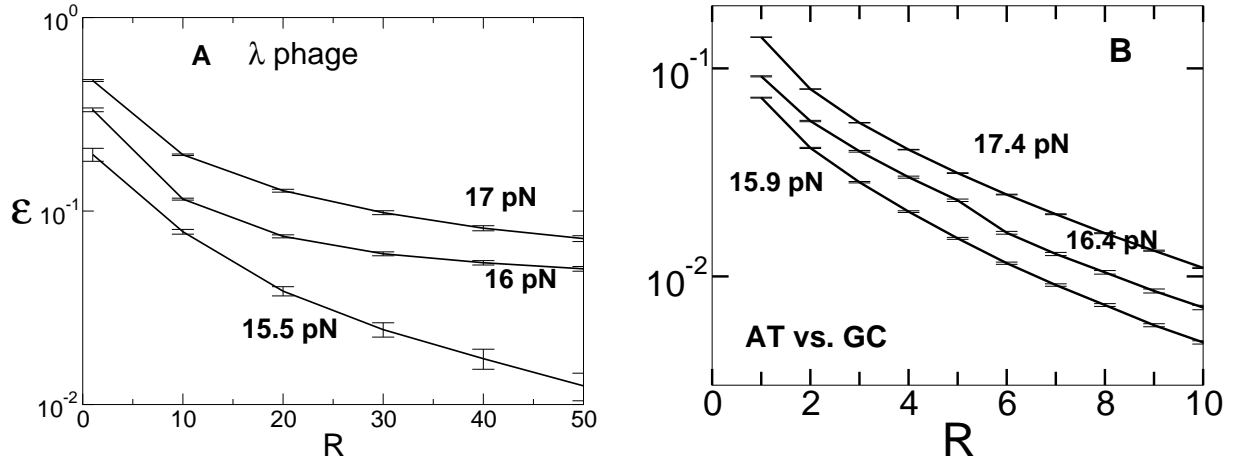


FIG. 7: **A.** Error  $\epsilon$  as a function of the number of unzippings for the phage. **B.** Same as A but without distinguishing A from T and G from C, see text.

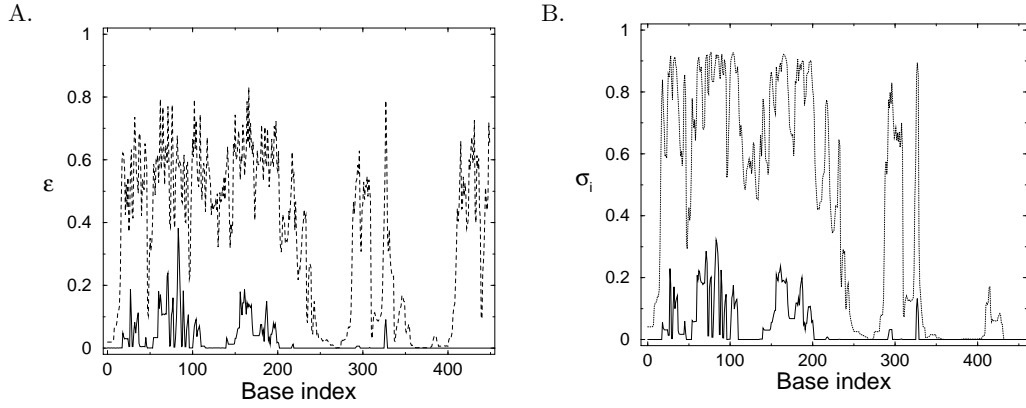


FIG. 8: Probability  $\epsilon_i$  (**A**) that bp  $i$  is not correctly predicted and Shannon entropy  $\sigma_i$  (**B**) for the first 450 bp of the DNA  $\lambda$ -phage. Inference is made from  $R = 1$  unzipping (dashed line) and  $R = 40$  unzippings (full line). The force is  $f = 16.4$  pN, and data are averaged over 1000 MC samples.

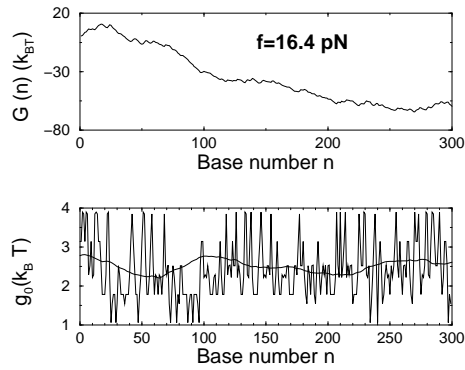


FIG. 9: Top: free energy landscape for unzipping at force  $f = 16.4$  pN. Local minima correspond to the portion of the sequence that are best predicted. Bottom: pairing free energy as a function of the base pair index, without and with window-average (Gaussian weight over 20 base pairs).

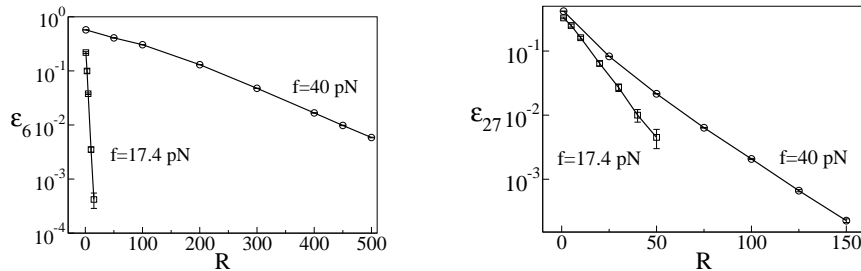


FIG. 10: Error rate  $\epsilon_i$  (semilog scale) as a function of the number of repeated unzippings for base pairs  $i = 6$  (left) and  $i = 27$  (right) arbitrarily selected, for forces  $f = 17.4$  and  $40$  pN. Numerical data are averaged over 25000 to  $10^7$  samples, see error bars.

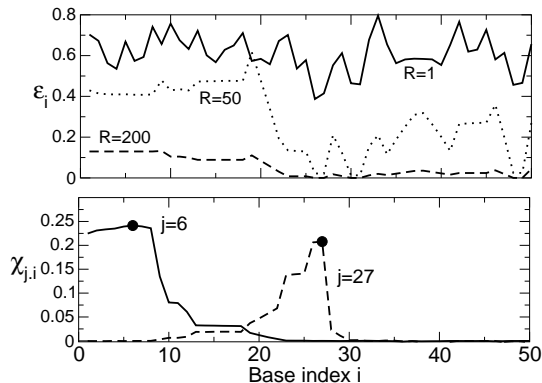


FIG. 11: Top: error  $\epsilon_i$  for the first 50 bases of the  $\lambda$ -DNA for  $R = 1, 50, 200$  unzippings. Bottom: connected correlation  $\chi_{j,i}$  for bases  $j = 6$  and  $j = 27$  (black dots) for  $R = 50$  unzippings.  $\chi_{27,i}$  is multiplied by 10 to be more visible; data correspond to  $f = 40$  pN (large force).

with  $R$ ,  $\epsilon_i \simeq \exp(-R/R_c(f, i))$ . The value of the decay constant  $R_c(f, i)$  strongly depends on the force and the bp index. At large force,  $f = 40$  pN, bp 27 is more easily predicted than bp 6. Fitting of the numerical data yields  $R_c(f = 40, i = 6) = 113 \pm 2$  and  $R_c(f = 40, i = 27) = 25 \pm 1$ . Correspondingly about 400 and 75 unzippings, collected and analyzed together, are needed to make the error smaller than 1%. At moderate force,  $f = 17.4$  pN, predictions for bp 6 require less unzippings than for bp 27. We obtain  $R_c(f = 17.4, i = 6) = 2.2 \pm 0.1$ , meaning that about 6 unzippings are sufficient to reduce the failure rate to 1%, while  $R_c(f = 17.4, i = 27) = 13 \pm 1$  and about 40 unzippings are needed to reduce the error to the same amount.

The quality of predictions exhibit strong correlations from base to base. We show in Fig 11(top) the error  $\epsilon_i$  for the first 50 bases of  $\lambda$ -DNA at high force  $f \geq 40$  pN. We observe that groups of neighboring bases are locked-in in that their errors decay at the same rate when increasing the number  $R$  of unzippings. See for instance in Fig 11 the blocks containing base 6, extending from bases 1 to 9, and base 27, including bases 26 and 27 only. All the bases  $i$  in a block have the same decay constant  $R_c(f, i)$ . The lock-in phenomenon is visible from the connected correlation function  $\chi_{j,i}$  (13), shown for bases  $j = 6$  and  $j = 27$  in Fig 11(bottom).  $\chi_{i,j}$  is essentially a step-wise function, with highest value for the bases  $i$  in the same block as  $j$ , and smaller values for neighboring blocks. The values of the decay constants at finite force as well as the blocks of locked-in bases will be found back analytically by the theory.

#### 4. Entropy of predictions on a base

The error  $\epsilon$  is defined from the exact knowledge of the true sequence. In practice one would like to be able to assess the quality of prediction  $b_i^*$  over base  $i$  without referring to the unknown true sequence. To do so we calculate the four optimal sequences for each of the four possible choices of  $b_i = A, T, G, C$  using the above Viterbi algorithm, starting from base  $i$  and going backward until the first base  $b_1$  is reached and optimized over; we call  $P_1(b_1^*|b_i)$  the probability (8) corresponding to this left part of the sequence. Then we repeat the process starting from base  $i$

and going forward until the last base of the molecule is reached and optimized over, and we obtain the probability  $P_N(b_N^*|b_i)$  corresponding to the right part of the sequence. Hence we obtain the most likely sequence constrained to have base  $i$  equal to  $b_i$ , together with its weight  $W(b_i) = P(b_0^*|b_i) \times P(b_N^*|b_i)$ . After a proper normalization we define the probability

$$\mu(b_i) = \frac{W(b_i)}{W(A) + W(C) + W(T) + W(G)} \quad (15)$$

for each of the four base values at location  $i$ . The base with the highest value of  $\mu$  is the one predicted by the usual Viterbi procedure. The Shannon entropy, once averaged over MC data,

$$\sigma_i = -\left\langle \sum_{b_i} \mu(b_i) \log_4 \mu(b_i) \right\rangle \quad (16)$$

is small when one of the four possible bases has much higher probability than the other ones, and high (close to 1) when bases are equiprobable. Figure 8B shows that the behavior of  $\sigma_i$  follows the one of  $\epsilon_i$  along the sequence (fig 8A). In other words, if a base has a much higher probability  $\mu$  than the other three bases it is very likely to be the correct one. The Shannon entropy is a good estimator of the quality of the prediction.

### B. Average Bayesian prediction

Instead of the maximum likelihood probability  $\mu(b_i)$  we can compute the probability  $\mu_i^A(b)$  that base  $i$  is of type  $b = A, T, C, G$  through the expression (7),

$$\mu_i^A(b) = \sum_{B'|b'_i=b} \mathcal{P}(B'|T) \quad (17)$$

where we have summed over all sequences constrained to have the value  $b$  for base  $i$ . This corresponds to an average Bayesian prediction in contrast with the maximum probability prescription of Section III A. We construct our predicted sequence  $B^A$ , assigning to each base  $i$  the argument  $b$  which maximizes probability  $\mu_i^A$ .

As in Section III A we have studied the quality of the prediction for different values of the applied force and of the number of unzippings. The fraction of mispredicted bases in  $B^A$  as a function of the force and of the number of unzippings shows a similar behaviour (not shown) to its maximum probability case counterpart (Fig. 6 and 7); a theoretical discussion of this equivalence in the case of homogeneous sequences will be given in Section IV A 2. In order to better understand this similarity for the  $\lambda$ -phage we have considered three ten bp long portions of its sequence,  $B_i^{(10)} = (b_i, b_{i+1}, b_{i+2}, b_{i+3}, b_{i+4}, b_{i+5}, b_{i+6}, b_{i+7}, b_{i+8}, b_{i+9})$ , located at  $i = 200$ ,  $i = 140$ , and  $i = 90$ . The choice of the locations corresponds to low ( $\sigma \simeq 0$ ), medium ( $\sigma \simeq 0.5$ ) and high ( $\sigma \simeq 1$ ) entropy regions (Fig 8B). We obtain complete sequences of length  $N$  by setting the bases outside the 10 bp window to the values they have in  $B^*$ . For each of the three locations we have calculated the probability (7) of the  $4^{10} \simeq 10^6$  sequences  $B$  with the recursive formula (8), divided by the largest probability *i.e.* the one of the sequence  $B^*$ . These ratios  $r(B) \leq 1$  are called relative probabilities. Even in a high entropy region most of the sequences have a very small relative probability  $r(B) \ll 1$ , meaning that the average sequence  $B^A$  is actually very close to the most likely one,  $B^*$ . It is interesting to notice that smaller and smaller relative probabilities  $r$  do not necessarily correspond to higher and higher ‘mutations’ from  $B^*$ . The average Hamming distance (number of bases  $b_i$  not equal to their values  $b_i^*$  in  $B^*$ ) of sequences with relative probabilities in  $[r; r + dr]$  is not a monotonic function of  $r$ . Less and less likely sequences are not obtained from the ground sequence through the mutation of a larger and larger number of bases. Due to stacking interactions, in fact, bases are not independent and it can be energetically favorable to flip a group of bases instead of a single one.

## IV. ANALYTICAL STUDY OF INFERENCE PERFORMANCES

In this section, we present the theoretical studies carried out to better understand how the quality of the prediction depends on parameters e.g. force, sequence content, number of repetitions of the unzipping on the same molecule, ... We start with the high force case where closing basically never occurs. The analytical study of this situation is performed first in the absence of stacking interactions between bases, then in the presence of stacking interactions. We show that the overall quality of the prediction crucially depends on the number of repetitions of the unzipping. Later on we turn to the case of finite force where closing and opening both take place, and show how the finite force study can be exactly reduced to the high force one with a stochastic number of unzippings whose distribution is calculated.

Throughout Section IV A and Section IV B 1 only two types of bases, called weak ( $W$ ) and strong ( $S$ ) have been considered instead of the four types  $A, T, G, C$ . The real case of four type of bases is taken back into account from Section IV C. Considering two instead of four base types allows us to make calculation shorter; we however stress that there is, in principle, no obstacle to the extension of our calculation to the four bases case. It is also justified *a posteriori* by our finding. The error in predicting the true value of a base  $b$ , say,  $b = A$ , is the sum of the probabilities of predicting the other three bases, here  $b = G$ ,  $b = T$ , and  $b = C$ . We show that, when a large amount of data is collected, one of these three probabilities, say,  $b = G$ , is much larger than the other two probabilities, turning the four base type problem into an effective two base types problem.

### A. High force theory: no stacking interactions

A quick calculation shows that, for forces equal or larger than 40 pN, the fork separating open and closed regions never goes backward in the course of unzipping. Indeed,  $g_s(f = 40 \text{ pN}) \simeq -8.6$ , and thus even for strong bases with pairing free energy  $g_0 \simeq -3.6$ , the ratio of closing over opening rates equals  $\exp(g_s(f) - g_0) \simeq e^{-5}$ , and is less than one percent. Bases essentially never close, and the matrix  $M(b_i, b_{i+1}; u_i, t_i, d_i)$  (6) simplifies since  $d_i = 0$ , and  $u_i = 1$  for all open base pairs. We hereafter calculate the quality of prediction in this case.

Let us simplify further the problem and assume that base pair interactions are essentially due to the presence of hydrogen bonds, and not to stacking effects. In other words, we replace  $g_0(b_i, b_{i+1})$  with  $g_0(b_i)$  where  $b_i$  can take two values:  $W$  (weak) or  $S$  (strong). The free energies are  $g_0(S) < g_0(W) < 0$ , and  $\Delta = g_0(W) - g_0(S) > 0$  denotes their difference.

Consider an unzipping experiment (one run of our Montecarlo program) which opens  $N$  base pairs:  $d_i = 0$  for all  $i$ ,  $u_i = 1$  for  $i < N$  and  $u_i = 0$  for  $i \geq N$ . The times  $t_i$  spent on the bases  $i = 1, \dots, N$  are uncorrelated and exponentially distributed:

$$P(t_i | b_i^L) = r e^{g_0(b_i^L)} \exp(-r e^{g_0(b_i^L)} t_i) \quad (18)$$

The distributions corresponding to  $W$  and  $S$  bases are plotted in Fig. 12. We define the mean sojourn time on base  $i$ ,

$$\langle t_i \rangle = \frac{1}{r} \exp(-g_0(b_i^L)) . \quad (19)$$

and the normalized time

$$\tau_i = \frac{t_i}{\langle t_i \rangle} . \quad (20)$$

Obviously neither  $\langle t_i \rangle$  nor  $\tau_i$  are accessible from the measure which gives access to  $t_i$  only. From (18), the distribution of the normalized time is exponential with average value unity,

$$P_1(\tau_i) = \exp(-\tau_i) . \quad (21)$$

#### 1. Maximum a posteriori prediction

Given a random value for  $\tau_i$  drawn from distribution (21), the most likely value for the base,  $b_i^*$ , is obtained from Bayes formula (7) by maximizing

$$P(b_i | \tau_i) \propto r e^{g_0(b_i)} \exp(-r e^{g_0(b_i)} \langle t_i \rangle \tau_i) \propto \exp(g_0(b_i) - e^{g_0(b_i) - g_0(b_i^L)} \tau_i) \quad (22)$$

An immediate calculation leads to the conclusion that a weak base (respectively a strong base) will be correctly predicted if  $\tau_i < \tau^W$  (resp.  $\tau_i > \tau^S$ ) where

$$\tau^W = \frac{\Delta}{1 - e^{-\Delta}} \quad \text{and} \quad \tau^S = \frac{\Delta}{e^{\Delta} - 1} \quad (23)$$

Therefore, the probability that a base is wrongly predicted depends on whether the base is weak or strong, and reads

$$\begin{aligned} \epsilon_1^W &= \int_{\tau^W}^{\infty} d\tau P_1(\tau) = \exp\left(-\frac{\Delta}{1 - e^{-\Delta}}\right) \\ \epsilon_1^S &= \int_0^{\tau^S} d\tau P_1(\tau) = 1 - \exp\left(-\frac{\Delta}{e^{\Delta} - 1}\right) . \end{aligned} \quad (24)$$

Plots of  $\epsilon_1^W$  and  $\epsilon_1^S$  as functions of the free energy difference  $\Delta$  shows that the latter probability is smaller than the former. At high force, maximum likelihood prediction works better on weak bases than on strong bases. The two limiting cases are:

- $\Delta \rightarrow 0$ : we find  $\epsilon_1^W = \frac{1}{e} = 0.368$ , while  $\epsilon_1^S = 1 - \frac{1}{e} = 0.632$ . This result is, at first sight, surprising since both bases should become equivalent when the free energy difference tends to zero. It is a consequence of the maximal likelihood principle: the reduced time  $\tau$  has a higher probability to be smaller than its average value ( $\tau^W = \tau^S = 1$  when  $\Delta \rightarrow 0$ ), and therefore weak bases are predicted with higher probabilities than strong bases independently of the true base  $b_i^L$ . We shall see in Section IV A 2 that this artifact disappears when prediction are carried out from the average Bayesian framework of Section III B.
- $\Delta \rightarrow \infty$ : when the difference in free energies between both bases gets very large, both are asymptotically perfectly predicted. The convergence to 100% correct prediction is faster for weak than for strong bases:  $\epsilon_1^W \simeq e^{-\Delta}$ ,  $\epsilon_1^S \simeq \Delta e^{-\Delta}$ .

The above analysis can straightforwardly be extended to the case of predictions made from repeated experiments. Let us call  $R$  the number of unzippings, and  $\tau_i^{(1)}, \tau_i^{(2)}, \dots, \tau_i^{(R)}$  the (normalized) times spent on base  $i$ . Using formula (10), we have to maximize

$$\begin{aligned} P_R(b_i | \{\tau_i^{(1)}, \tau_i^{(2)}, \dots, \tau_i^{(R)}\}) &\propto \left[ r e^{g_0(b_i)} \right]^R \exp \left[ -r e^{g_0(b_i)} \langle t_i \rangle (\tau_i^{(1)} + \tau_i^{(2)} + \dots + \tau_i^{(R)}) \right] \\ &\propto \exp \left[ R g_0(b_i) - r e^{g_0(b_i) - g_0(b_i^L)} \tau_i \right] \end{aligned} \quad (25)$$

where

$$\tau_i = \tau_i^{(1)} + \tau_i^{(2)} + \dots + \tau_i^{(R)} \quad (26)$$

is the total time spent on base  $i$ . The maximization over  $b_i$  is very similar to the one carried out from eqn (22). We find that formula (24) for the probabilities of correct prediction holds for  $R$  unzippings provided the single time distribution  $P_1$  is replaced with the distribution  $P_R$  of the total time  $\tau_i$  (see Appendix B 1),

$$P_R(\tau_i) = \frac{\tau_i^{R-1}}{(R-1)!} \exp(-\tau_i), \quad (27)$$

and the times  $\tau^W, \tau^S$  (23) are multiplied by  $R$ . The distribution of (not normalized) sojourn times after  $R$  unzippings are shown in Fig. 12 for  $W$  and  $S$  sequences. An important remark is that the distributions become more and more concentrated as  $R$  grows; in other words the times become less and less stochastic and are faithful signatures of the thermodynamic nature of the attached base. The probabilities that weak and strong bases are not correctly predicted after  $R$  unzippings are given by

$$\begin{aligned} \epsilon_R^W &= \int_{R\tau^W}^{\infty} d\tau P_R(\tau) = \gamma \left( R, \frac{R\Delta}{1 - e^{-\Delta}} \right) \\ \epsilon_R^S &= \int_0^{R\tau^S} d\tau P_R(\tau) = 1 - \gamma \left( R, \frac{R\Delta}{e^{\Delta} - 1} \right). \end{aligned} \quad (28)$$

where

$$\gamma(a, x) = \int_x^{\infty} dt \frac{t^{a-1} e^{-t}}{(a-1)!} \quad (29)$$

is the normalized incomplete Gamma function.

To better understand how the quality of predictions improves with the number of unzippings, we have analytically calculated the asymptotic expansion of  $\epsilon$  in Appendix E. From expression (28), we have when  $R \gg 1$ ,

$$\epsilon_R \simeq \frac{e^{-R(\tau-1-\ln \tau)}}{\sqrt{2\pi R} (\tau-1)} \quad (30)$$

with  $\tau = \tau^W$  or  $\tau^S$  (23) depending on the type of base. As a consequence, achieving good recognition requires a number of unzippings (much) larger than

$$R_c = \frac{1}{\tau - 1 - \ln \tau}. \quad (31)$$

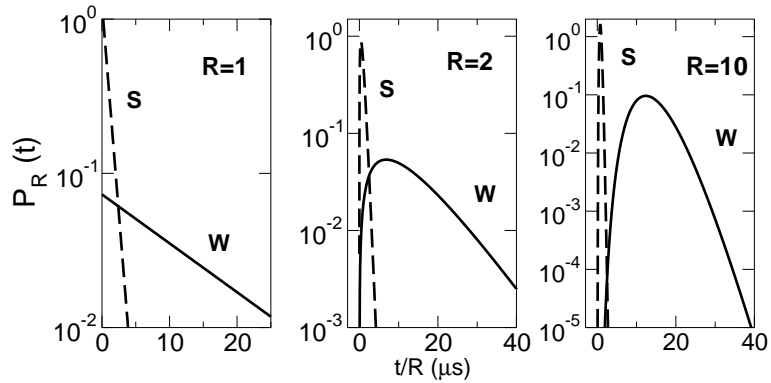


FIG. 12: Probability distribution  $P_R$  of the sojourn time  $t$  spent on a weak ( $g_0(W) = -1.06$ ,  $\langle t \rangle_W = 0.8\mu\text{s}$ , dashed line) and strong ( $g_0(S) = -3.9$ ,  $\langle t \rangle_S = 13.7\mu\text{s}$ , full line) bases. Time is rescaled by  $1/R$  (see horizontal axis). The number of unzippings is  $R = 1$  (left),  $R = 2$  (middle), and  $R = 10$  (right). The probability  $\epsilon$  (12) that a  $W$  (resp.  $S$ ) base is not correctly predicted is the area under the dashed (resp. full) curve right (resp. left) to the crossing point. As  $R$  increases time distributions are more and more concentrated, and the error gets smaller and smaller.

This crossover number depends on the free energy difference  $\Delta$ , but not on the type of base:  $R_c(\tau^W) = R_c(\tau^S)$ . Fig 13 shows that  $R_c$  is all the more large than  $\Delta$  is small. Definitions (31) for  $R_c$  and (23) for  $\tau^W, \tau^S$  yield

$$R_c \simeq \frac{8}{\Delta^2} \quad , \quad \Delta \rightarrow 0 \quad . \quad (32)$$

This expression is a good quantitative approximation for  $R_c$  up to  $\Delta \simeq 3$ . We have checked the validity of these theoretical results through numerical experiments using the Viterbi procedure of Section IV B, where the free energy matrix  $g_0$  was modified to avoid stacking interaction. Figure 13 shows the perfect agreement between numerical and theoretical results.

That the effort (number of unzippings) necessary to ensure an excellent prediction essentially depends on the difference of pairing free energies between the two types of bases one wishes to distinguish justifies *a posteriori* the simplification of taking into account only two types of bases. The cases of interest are:

- Weak bases represent  $A$  or  $T$ , and strong bases  $G$  or  $C$ : the free energy difference is estimated to be  $\Delta \simeq 2.8$  (obtained from  $g_0(T, A) = -1.06, g_0(G, T) = -3.9$ ). The probability of wrong prediction for strong bases,  $\epsilon_R^S$ , is plotted in Fig 13, as a function of the number  $R$  of unzippings.  $R = 5$  unzippings are enough to achieve excellent base recognition.
- Weak bases are  $A$ , strong bases are  $T$ : the free energy difference is  $\Delta \simeq 0.5$  (obtained from  $g_0(T, A) = -1.06, g_0(A, T) = -1.55$ ). Figure 13 shows it takes about 100 unzippings to reach 99% confidence in the prediction. Thus, the number of unzippings considerably increases if we want to precisely resolve all base pairs.

Sequence prediction can be then done in a hierarchical manner. A small number of unzippings  $R \simeq 5$  is sufficient to distinguish between  $A, T$  and  $G, C$  bases, in agreement with numerical simulation data shown in Fig 7A&B, while more unzippings  $R \simeq 100$  are necessary to clearly separate  $A$  from  $T$ , and  $G$  from  $C$  bases. In this regard, our prediction procedure always amounts to distinguish between two types of bases.

## 2. Average Bayesian prediction

Average Bayesian prediction consists in estimating the the probability of the correct base  $P(b_i^L|t_i)$  (thermal average) and averaging over  $t_i$  (quenched average) rather than looking for the most likely base  $b_i$  given the time  $t_i$  spent on base  $i$  (III B). This procedure gives, in the general case of  $R$  unzippings,

$$\epsilon_R^A = \int_0^\infty d\tau \frac{P_R(\tau)}{1 + \exp(-R\Delta + \tau(e^\Delta - 1))} \quad . \quad (33)$$

We stress that the above expression gives the value of  $\epsilon_R^A$  for both  $W$  and  $S$  bases. The quality of prediction does not depend on base  $b_i^L$ , in contradistinction with the maximal likelihood case, see eqn (28). This independence is a direct

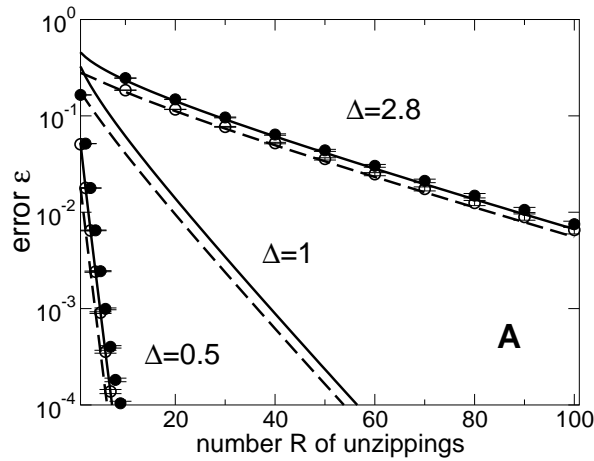


FIG. 13: Errors on sequences of, respectively, strong (full line) and weak (dashed line) bases as a function of the number  $R$  of unzippings in the infinite force limit and without stacking interaction. The difference of pairing free-energies  $\Delta$  is, from bottom to top, 0.5, 1, and 2.8. We show the results of numerical simulations for  $\epsilon_R^W, \epsilon_R^S$  with the error bars for  $\Delta = 0.5, 2.8$  (full dots:  $S$  sequence, empty dots:  $W$  sequence).

consequence of Bayes inference formula. By definition indeed,

$$\epsilon^{W,A} = \int_0^\infty d\tau P(\tau|W) P(S|\tau) = \int_0^\infty d\tau P(\tau|W) \frac{P(\tau|S)}{P(\tau|W) + P(\tau|S)}. \quad (34)$$

This expression is left unchanged when we exchange  $S$  and  $W$ . Therefore

$$\epsilon^{S,A} = \epsilon^{W,A} \quad (35)$$

Notice that this proof is quite general: it not only holds for any number  $R$  of unzippings, but also for any microscopic model yielding an explicit expression for  $P(\tau|b^L)$ . In particular, it remains true at finite force. As the number  $R$  of unzippings increases, the prediction approaches perfection, see Appendix E,

$$\epsilon_R^A \simeq \frac{\pi\sigma}{\sin(\pi\sigma)} \frac{e^{-R(\tau-1-\ln\tau)}}{\sqrt{2\pi R}(1-\tau)} \quad (36)$$

with

$$\tau = \frac{\Delta}{e^\Delta - 1} \quad \text{and} \quad \sigma = \frac{1}{\Delta} - \frac{1}{e^\Delta - 1}. \quad (37)$$

This asymptotic scaling is, to the exponential order, identical to the one obtained in the maximum likelihood case (30). Therefore average and maximum likelihood predictions are asymptotically equivalent.

### 3. Relationship with Shannon entropy

The above findings explains the similarity between the error (12) and the Shannon entropy (16) observed in Fig. 8A&B. Let us call  $\epsilon$  and  $1 - \epsilon$  the probabilities that the prediction on a base is correct and erroneous respectively. The Shannon entropy reads

$$\sigma = -\epsilon \ln \epsilon - (1 - \epsilon) \ln(1 - \epsilon) \simeq -\epsilon \ln \epsilon \simeq \text{cst} \times \sqrt{R} e^{-R/R_c} \quad (38)$$

when the number of unzippings is large with respect to  $R_c$ . This explains why the error and the Shannon entropy on a base roughly behave in the same way, and essentially vanish when the number of unzippings is far above its critical value  $R_c$ . This result is left unchanged in the case of four, and not two base types.

## B. High force theory: stacking interactions

Let us now study how the presence of stacking interactions modify the above findings. With two kinds of bases, the pairing free energy matrix is a  $2 \times 2$  matrix  $g_0(b, b')$ . Strong bases ( $S$ ) are chosen to be 'average' bases from a repeated GCGGC... sequence while weak bases ( $W$ ) represent a repeated ATATAT... sequence. The values of the interactions are the average values of the pairing free energy in each of the four quadrants of the original  $4 \times 4$  matrix:  $g_0(W, W) = -1.42$ ,  $g_0(S, W) = g_0(W, S) = -2.39$ , and  $g_0(S, S) = -3.50$ . We define the free energy differences

$$\Delta^W = |g_0(W, W) - g_0(W, S)|, \quad \Delta^S = |g_0(W, S) - g_0(S, S)|. \quad (39)$$

whose values are  $\Delta^W = 0.98$ ,  $\Delta^S = 1.11$ . The calculation of the probability of correct base prediction is more difficult than in the absence of stacking but can be carried out using techniques issued from the statistical mechanics of one dimensional disordered systems [39, 40].

We start from the recursive eqn (8) for the probability  $P_i(b_i)$  that the  $i^{\text{th}}$  base of the sequence is equal to  $b_i$ . As in the no-stacking case, we introduce the normalized time  $\tau_i$  through eqn (20) where the average sojourn time on base  $i$  now reads

$$\langle t_i \rangle = \frac{1}{r} \exp(-g_0(b_i^L, b_{i+1}^L)) \quad (40)$$

Defining  $\pi_i(b_i) = -[\ln P_i(b_i)]/R$  and introducing the local fields,

$$h_i = \pi_i(S) - \pi_i(W) \quad (41)$$

we rewrite eqns (8,9) under the form

$$h_{i+1} = F_i(h_i, \tau_i) \quad (42)$$

where function  $F_i$  depends on base  $b_i^L$  through the average sojourn time (40),

$$\begin{aligned} F_i(h, \tau) = & \max \left[ h + g_0(W, W) - g_0(S, W) - r \frac{\langle t_i \rangle}{R} (e^{g_0(W, W)} - e^{g_0(S, W)}) \tau, 0 \right] \\ & + \min \left[ -h, g_0(W, S) - g_0(S, S) - r \frac{\langle t_i \rangle}{R} (e^{g_0(W, S)} - e^{g_0(S, S)}) \tau \right] \end{aligned} \quad (43)$$

As  $\tau_i$  is a stochastic variable with distribution  $P_R$  (27) (for  $R$  repetitions of the experiment),  $h_i$  is itself a stochastic variable. Its probability distribution,  $Q_i$ , obeys the recursion

$$Q_{i+1}(h_{i+1}) = \int_0^\infty d\tau_i P_R(\tau_i) \int_{-\infty}^\infty dh_i Q_i(h_i) \delta(h_{i+1} - F_i(h_i, \tau_i)). \quad (44)$$

### 1. Repeated sequences

The stationary solution  $Q = Q_i$  of eqn (44) is calculated in Appendix C for the three repeated sequences  $B^L = WWWW\dots$ ,  $SSSS\dots$ , and  $SWSW\dots$  referred to as  $WW$ ,  $SS$ , and  $SW$  sequences respectively. These sequences differ from each other through their sojourn times  $\langle t \rangle$  (40). When the condition  $\Delta^W \leq \Delta^S$  is fulfilled as is the case for the example considered above, the stationary field distribution is better written in terms of its cumulative function

$$\hat{Q}(h) \equiv \int_h^\infty dh' Q(h'), \quad (45)$$

with the result

$$\hat{Q}(h) = \begin{cases} A(h) & \text{if } h < -\Delta^S \\ \frac{A(h) - A(-h)B(h)}{1 - B(-h)B(h)} & \text{if } -\Delta^S < h < \Delta^S \\ 0 & \text{if } h > \Delta^S \end{cases} \quad (46)$$

where

$$A(h) = 1 - \gamma \left( R, \frac{R(\Delta^S - h)}{x(1 - e^{-\Delta^S})} \right), \quad B(h) = \gamma \left( R, \max \left( \frac{R(\Delta^W - h)}{x(e^{\Delta^W} - 1)}, 0 \right) \right) - \gamma \left( R, \frac{R(\Delta^S - h)}{x(1 - e^{-\Delta^S})} \right), \quad (47)$$

and  $\gamma$  is the incomplete Gamma function (29). The parameter  $x$  is defined as the ratio of the average sojourn time  $\langle t \rangle$  over its value for the  $SW$  sequence,

$$x = \frac{\langle t \rangle}{\langle t \rangle^{SW}}. \quad (48)$$

Knowledge of the field distribution allows us to calculate the average fraction  $\epsilon$  of mispredicted bases (14) and the nearest-neighbor ( $d = 1$ ) disconnected correlation function

$$\chi_1^{dis} = \chi_1 + (1 - \epsilon)^2 \quad (49)$$

where the connected correlation function is defined in eqn (13). The calculations are reported in Appendix D. Results are

- *WW sequence*: we have  $x = e^{-\Delta^W}$ , and

$$\epsilon_R^{WW} = 1 - \int_{-\Delta^S}^{\Delta^S} dh \hat{Q}(-h) Q(h), \quad (\chi_1^{dis})_R^{WW} = \int_0^\infty d\tau P_R(\tau) \hat{Q} \left( -\Delta^W + \tau R(1 - e^{-\Delta^W}) \right)^2. \quad (50)$$

- *SS sequence*: we have  $x = e^{\Delta^S}$ , and

$$\epsilon_R^{SS} = \int_{-\Delta^S}^{\Delta^S} dh \hat{Q}(-h) Q(h), \quad (\chi_1^{dis})_R^{SS} = \int_0^\infty d\tau P_R(\tau) \left[ 1 - \hat{Q} \left( -\Delta^S + \frac{\tau}{R}(e^{\Delta^S} - 1) \right) \right]^2. \quad (51)$$

- *SW sequence*: we have  $x = 1$ ; the probabilities that bases  $S$  and  $W$  are not correctly predicted are, respectively,

$$\epsilon_R^{SW,S} = \int_{-\Delta^S}^{\Delta^S} dh \hat{Q}(-h) Q(h), \quad \epsilon_R^{SW,W} = 1 - \epsilon_R^{SW,W}, \quad (52)$$

while the correlation function reads

$$(\chi_1^{dis})_R^{SW} = \int_0^\infty d\tau P_R(\tau) \left[ \hat{Q} \left( -\Delta^S + \frac{\tau}{R}(1 - e^{-\Delta^S}) \right) - \frac{1}{2} \hat{Q} \left( -\Delta^S + \frac{\tau}{R}(1 - e^{-\Delta^S}) \right)^2 - \frac{1}{2} \hat{Q} \left( -\Delta^W + \frac{\tau}{R}(e^{\Delta^W} - 1) \right)^2 \right] \quad (53)$$

The subscript ‘R’ reminds us that the above expressions hold for data collected from  $R$  unzippings of the experiment. Let us stress that the field distributions  $Q$  (and their cumulative functions  $\hat{Q}$ ) appearing in the expressions of  $\epsilon$  and  $\chi_1^{dis}$  above depend on the sequence through the ratio  $x$ , see eqns (46,47,48).

The above theoretical predictions are shown in Fig. 14 and Fig. 15 for the three sequences, and perfectly agree with numerical experiments. For  $SS$  and  $WW$  sequences, we find that the quality of predictions tends to 100% accuracy as the number  $R$  of unzippings increases. It is shown in Appendix E that the asymptotic scaling of  $\epsilon_R$  is given by

$$\epsilon_R \simeq \frac{\tau^2 e^{-2R(\tau-1-\ln\tau)}}{\sqrt{4\pi R}(\tau-1)} \quad (54)$$

where  $\tau$  equals

$$\tau^{WW} = \frac{\Delta^W}{1 - e^{-\Delta^W}} \quad \text{and} \quad \tau^{SS} = \frac{\Delta^S}{e^{\Delta^S} - 1} \quad (55)$$

for  $WW$  and  $SS$  sequences respectively. The above formula shows that the number of unzippings must exceed

$$R_c = \frac{1}{2(\tau - 1 - \ln \tau)} \quad (56)$$

in order to achieve good recognition; we find  $R_c \simeq 4.3$  and  $R_c \simeq 3.3$  for  $WW$  and  $SS$  sequences respectively. The nearest-neighbor correlation function  $\chi_1$  in Fig. 16 is very small, even for  $R = 1$  unzipping. The quasi-independence of predictions can be understood from the analytical calculation of Appendix D, and is essentially due to the fact that

	A	T	C	G
A	18	75	72	51
T	8	14	14	13
C	13	51	50	39
G	14	72	69	50

$b = A$

	A	T	C	G
A	51	44	12	13
T	59	51	13	14
C	14	13	11	8
G	12	12	7	7

$b = C$

TABLE II: Single base mutation decay constant  $R_c^{sm}(xby)$ , that is, value of the number of unzippings necessary for a good prediction at high force of a base  $b$  as a function of the contiguous bases  $x$  (row) and  $y$  (column). See equation (60) for a precise definition. Left: the central base is  $b = A$ ; the most dangerous mutation is  $b = A \rightarrow b' = T$  for all contiguous bases, except for  $xy = AA$  where  $b' = G$ . Right: the central base is  $b = C$ ; the most dangerous mutation is  $b = C \rightarrow b' = G$  for all contiguous bases, except for  $xy = CC$  where  $b' = A$ .

	A	T	C	G
A	151	151	89	89
T	15	32	118	118
C	78	78	22	16
G	139	139	14	21

TABLE III: Decay constant  $R_c(xb \rightarrow xb')$ , that is, number of unzippings necessary for a good prediction, at high force, of a bond between base  $x$  (fixed as in the sequence, value indicated in the leftmost column) and base  $b$  (value reported in the top line), potentially predicted to be of  $b'$  type. The most dangerous (requiring the largest number of unzippings) mutation  $b \rightarrow b'$  are given by  $b'$  equal to the complementary base of  $b$ , except for the cases  $TT \rightarrow TC$ ,  $CC \rightarrow CA$ ,  $GG \rightarrow GT$ .

the sums of the diagonal and off-diagonal elements of the  $g_0$  matrix are equal. We have numerically checked that the correlation function is very small at all distances  $d$ , not only at high forces, but for all forces above criticality.

The above findings can be easily understood from the findings of Section IV A 1. Consider for instance the  $WW$  sequence. When  $R$  gets very large, very few bases  $S$  are (wrongly) predicted to be in the sequence. Call  $\epsilon$  the probability that a single base  $S$  is predicted. The predicted event  $WSW$  violates two stacking interactions (bonds) with respect to the correct event  $WWW$ . Let us make the simplifying hypothesis that these two violations are independent:  $\epsilon = \mu^2$ , where the probability  $\mu$  of one bond violation depends on the free energy excess  $\Delta^W$  (39) of the erroneous bond  $WS$  (or  $SW$ ) with respect to the true bond  $WW$ . We estimate the value of  $\mu$  from the theory of Section IV A 1:  $\mu = \epsilon_R$  (30) with  $\tau = \tau^{WW}$ , see (23,55). This simple argument explains why the quality of predictions is much closer to 100% success in presence than in absence of stacking (for the same number of unzippings). In particular, the cross-over number of unzippings  $R_c$  required to achieve good recognition is twice smaller in the former case (56) than in the latter case (32).

The behavior of the error  $\epsilon$  for the alternate  $SW$  sequence is slightly more subtle to interpret, see Fig. 15. From expressions (52,53), we find (see Appendix E), in the infinite  $R$  limit,

$$\epsilon_R^{SW,S} \ \& \ \epsilon_R^{SW,W} \ \rightarrow \ \epsilon_\infty^{SW} = \frac{1}{2} \quad \text{and} \quad (\chi_1)_R^{SW} \ \rightarrow \ (\chi_1)_\infty^{SW} = \frac{1}{2}. \quad (57)$$

The limit value of  $\epsilon$  is at, first sight, disappointing. There is 50% probability that a  $S$  or  $W$  is predicted at a given position  $i$  along the sequence, showing that our prediction is not better than a purely random guess! However, the nearest-neighbor correlation function  $\chi$  is much higher than the value  $(1-\epsilon)^2$  it would have if there were no correlation. Indeed, we find that the probability that base  $i+1$  is correctly predicted provided its neighbor at position  $i$  is equals

$$\frac{\langle n_i n_{i+1} \rangle}{\langle n_i \rangle} \rightarrow \frac{\chi_\infty^{SW}}{1 - \epsilon_\infty^{SW}} = 1 \quad (58)$$

as the number of unzippings increases. In other words, only two sequences can be predicted, either the correct one  $SWSWSW\dots$  or its mirror sequence  $WSWWSW\dots$ . Actually, both sequences produce identical unzipping signals since the pairing matrix  $g_0$  is symmetric, which is not the case for the true matrix (Table I).

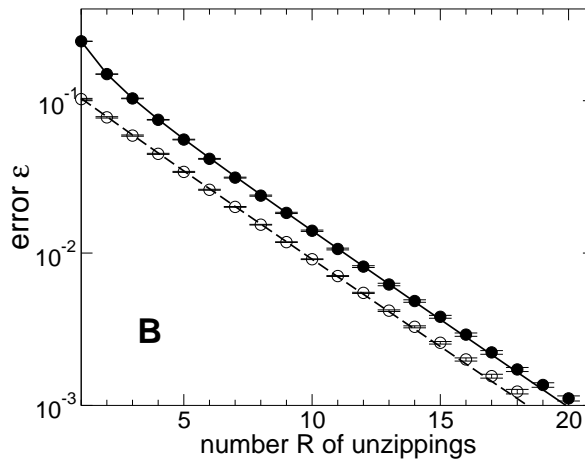


FIG. 14: Probability of misprediction for repeated  $WW$  (full line) and  $SS$  (dashed line) sequences as a function of the number  $R$  of unzippings in the infinite force limit and in presence of stacking interactions. Here,  $g_0(W, W) = -1.5$ ,  $g_0(S, W) = g_0(W, S) = -2.5$ ,  $g_0(S, S) = -3.5$ . The strong and weak sequences are repeated  $SS$  and  $WW$  sequences respectively. Simulation results are shown with the error bars. Remark that the slope of  $\ln \epsilon$  is about twice the one for the non-stacking case with  $\Delta = 1$  (Fig. 13), see eqn (56) and attached discu

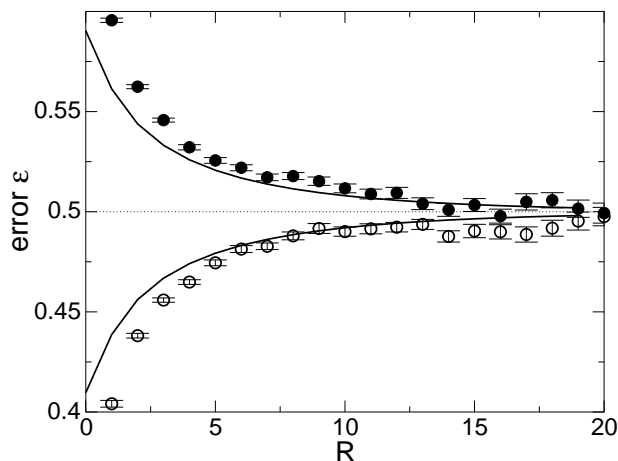


FIG. 15: Probabilities  $\epsilon_R^{SW,S}$  and  $\epsilon_R^{SW,W}$  of mispredicting, respectively, a  $S$  (black dots, full curve) and  $W$  (empty dots, dashed curve) base in a repeated  $SW$  sequence as a function of the number  $R$  of unzippings in the infinite force limit. The stacking interactions are  $g_0(W, W) = -1.5$ ,  $g_0(S, W) = g_0(W, S) = -2.5$ ,  $g_0(S, S) = -3.5$ . Simulation results are shown with the error bars, while continuous curves correspond to the theoretical expression (52). As  $R$  grows the prediction on a single base becomes essentially random ( $\epsilon \rightarrow \frac{1}{2}$ ) since  $SWSW\dots$  and  $WSWS\dots$  sequences cannot be distinguished from one another.

### C. High force theory: decay constants $R_c$ for heterogeneous sequences

Let us turn to the realistic case of a non-repeated sequence with four base types, and stacking interactions between neighbouring bases. From the numerical findings of Section III A and the theoretical analysis of repeated sequences of Section IV B we expect the error on a base to decay exponentially with the number  $R$  of unzippings. In a first step we estimate the decay constant within a single mutation assumption: all bases are assumed to be correctly predicted but the one under study [34]. However this single mutation assumption is not always correct. We will show that the decay of the error in predicting one base is often due to the difficulty in predicting a whole block of co-mutated bases, and give the corresponding expression of the decay constant  $R_c$ .

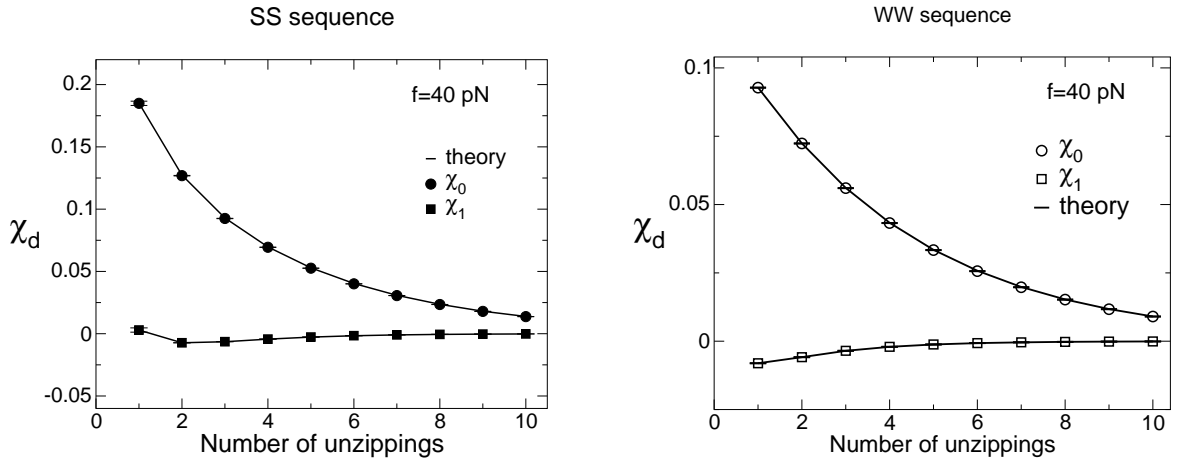


FIG. 16: Connected correlation function  $\chi_1$  at distances  $d = 1$  for, respectively, repeated *SS* (left panel) and *WW* (right panel) sequences as a function of the number  $R$  of unzippings in the infinite force limit ( $f = 40$  pN in simulations). For comparison we show the  $d = 0$  correlation function,  $\chi_0 = \epsilon(1 - \epsilon)$ .

### 1. Decay constant in the single base mutation assumption

Consider a triplet of contiguous bases along the sequence,  $xyy$  and let us start by calculating the error due to a predicted sequence with a single base mutation e.g.  $b \rightarrow b'$  when keeping bases  $x$  and  $y$  to the correct values. In this case the argument following eqn (56) and obtained in the case of repeated sequences is still valid. As a result of stacking interactions the probability  $\epsilon^{b \rightarrow b'}$  of this mistake is the product of the probabilities  $\epsilon^{xb \rightarrow xb'}$  and  $\epsilon^{by \rightarrow b'y}$  of either bond violation. The large  $R$  behavior of the error probability

$$\epsilon_R^b \sim e^{-R/R_c^{sm}(xyy)} \quad (59)$$

on base  $b$  is then obtained by selecting the worst value for the mutation  $b'$ ,

$$\frac{1}{R_c^{sm}(xyy)} = \min_{b'(\neq b)} \left[ \frac{1}{R_c(xb \rightarrow xb')} + \frac{1}{R_c(by \rightarrow b'y)} \right] \quad (60)$$

where  $R_c(xb \rightarrow xb')$  is the decay constant of the error obtained in the no-stacking theory of Section IV A 1 (applied here to a bond and not to a base violation); it is given by formula (31) with  $\Delta = g_0(x, b) - g_0(x, b')$  and  $\tau = \Delta/(e^\Delta - 1)$ . The values of  $R_c$  obtained from formula (60) are given in Table II (after rounding to the closest integer) for base triplets  $xyy$  with central base  $b = A$  and  $b = C$  respectively. The values of  $R_c$  for triplets with central bases  $b = T$  and  $b = G$  can be deduced from the decay constants of the complementary triplets, expressed in reversed order, due to the symmetry of the interaction matrix  $g_0$  of Table I e.g.  $R_c^{sm}(ATT) = R_c^{sm}(AAT)$ . The value  $b'$  of the most difficult base to distinguish from  $b$ , see (60), is  $T$  when the central base is  $A$  and  $G$  when the central base is  $C$ , except in the  $AAA$ ,  $CCC$  cases where  $b' = G$ ,  $b' = A$  respectively.

### 2. Propagation of errors, and blocks of locked-in bases

The above single base mutation offers only a lower bound to the true value of the decay constant  $R_c(i)$  of the error  $\epsilon_i$  in predicting base pair  $i$ . Strictly speaking, to calculate  $R_c(i)$ , one must consider all the  $3 \times 4^{N-1}$  sequences where base  $i$  differ from its value in the true sequence, and find among those sequences the one which requires the largest number of unzippings to be discarded. In other words errors on bp  $i$  may result from the difficulty of correctly predicting a block of more than one bp located around bp  $i$  rather than this bp alone.

We start by defining the decay constant for the large  $R$  behavior of the single bond misprediction probability  $\epsilon^{xy \rightarrow x'y'}$  for two contiguous mutations ( $xy \rightarrow x'y'$ ),

$$\epsilon_R^{xy \rightarrow x'y'} \sim e^{-R/R_c(xy \rightarrow x'y')} \quad (61)$$

where  $R_c(xy \rightarrow x'y')$  is given by eqn (23) with  $\Delta = g_0(x, y) - g_0(x', y')$  and  $\tau = \Delta/(e^\Delta - 1)$  (31). We then define, in the maximum likelihood framework, the probabilities (with respect to the random variables  $t_i$ )  $\mu_i^\rightarrow(b)$  and  $\mu_i^\leftarrow(b)$  of predicting base pair  $i$  to be of  $b$ -type when, respectively, the bases located to the right and the left of  $i$  are ignored. We assume that

$$\mu_i^\rightarrow(b) = e^{-R\pi_i^\rightarrow(b)} \quad \text{and} \quad \mu_i^\leftarrow(b) = e^{-R\pi_i^\leftarrow(b)} \quad (62)$$

for a large number  $R$  of unzippings, with boundary conditions  $\pi_1^\rightarrow(b) = 0$  and  $\pi_N^\leftarrow(b) = 0$  for all  $b$ . These probabilities can be evaluated from the probabilities of the most dangerous subsequence to the left and right of base pair  $i$ , according to the recurrence equations

$$\begin{aligned} \pi_i^\rightarrow(b') &= \min_b \left( \pi_{i-1}^\rightarrow(b) + \frac{1}{R_c(b_{i-1}^L b_i^L \rightarrow b b')} \right) \\ \pi_i^\leftarrow(b') &= \min_b \left( \pi_{i+1}^\leftarrow(b) + \frac{1}{R_c(b_i^L b_{i+1}^L \rightarrow b' b)} \right), \end{aligned} \quad (63)$$

remember  $b_i^L$  denotes the true type of bp  $i$ . These recurrence equations have a simple meaning. The probability that bp  $i$  is of  $b'$  type, when there is no base to the right of  $i$ , is simply given by the sum over  $b$  of the probability that bp  $i-1$  is of  $b$  type times the probability of predicting the bond  $bb'$  instead of  $b_{i-1}^L b_i^L$ . Notice that recurrence eqns (63) are simply the asymptotic counterpart of eqn (44) in the large  $R$  limit (for four and not two base types). They can be obtained from eqn (9) and (D3) by choosing for  $t_i$  the time having equal probabilities with the true bond  $b_{i-1}^L b_i^L$  and the erroneous bond  $bb'$  distributions [42].

The decay constant  $R_c(i)$  of the error on bp  $i$  is obtained by selecting the most dangerous value for the type  $b$ ,

$$\frac{1}{R_c(i)} = \min_{b \neq b_i^L} (\pi_i^\leftarrow(b) + \pi_i^\rightarrow(b)). \quad (64)$$

In general  $R_c(i)$  differs from the single mutation value,  $R_c^{sm}(i)$ . The latter depends only on the base and its two neighbors while the former depends on the whole sequence. Equations (63) and (64) can be interpreted by considering  $\pi_i^\leftarrow(b) + \pi_i^\rightarrow(b)$  as the free energy for the lowest excited state (sequence) with the base  $i$  fixed to a value,  $b$ , distinct from the one,  $b_i^L$ , in the ground state (real sequence). If the base  $i$  has a very large value for  $R_c^{sm}(i)$ , because both the bonds on the right and on the left of the base have a large  $R_c$  (see eqn 60), the most dangerous sequence is exactly this 'single mutation' sequence. In this case the minimum over  $b$  in (63) is exactly obtained for  $b = b_{i-1}^L$  and  $b = b_{i+1}^L$ , and the recursion halts after the nearest neighbors. However, when the bond constant  $R_c^{sm}$  is small, we can expect that it is less costly, in terms of free energy, to propagate the excitation at site  $i$  in a configuration where the base and its neighboring base are both mutated into their complementary values. The decay constant  $R_c$  for such a bond is indeed large because it is difficult to distinguish two bases from the complementary ones (Table III). This 'defect' propagates, in the recurrence eqn (63), until an interface with a large value for  $R_c$  is found. Obviously this propagation mechanism takes place on both sides of bp  $i$ . The most dangerous excitations are thus blocks of complementary bases of the real sequence. The bases in a block have then roughly the same  $R_c$  and are locked-in together (Fig 11).

The high force behaviour of the errors  $\epsilon_i$  (for  $R = 1, 50, 200$ ), obtained by the numerical inference and shown in Fig 11 agree with these theoretical results. The theoretical values for the decay constants  $R_c(f \geq 40pN, i)$  obtained from (63,64) are shown in Fig. 20 (dotted line). By solving eqn (63) we find that bp  $i = 6$  belongs to a block extending from bp 1 to 9. The boundary bp 1 has  $R_c$  on the left equal to  $\infty$  and bp 8 has  $R_c(GA \rightarrow CA) = 139$ . From eqn (63) we obtain  $R_c = 114$  for the whole block 1-9. This value coincide with the decay of the error at large  $R$  found from simulations and shown in Fig 10. We obtain  $R_c^{num} = 113 \pm 2$  from a fit of  $\log \epsilon_i$  vs.  $R$  at  $f = 40$  pN. [43] Base pair 27 belongs to a block on the right spreading over the whole sequence down to base 1, while the block on the left stops on the base itself. The number of unzippings needed for a good prediction of bp 27 is smaller: we obtain from theory  $R_c = 24$ , and from simulation  $R_c^{num} = 25 \pm 1$ . Note that the propagation of the error by blocks of complementary bases in this section go beyond the single mutation approximation reported in [34].

## D. Moderate force theory

### 1. On the number of single-base openings

We now investigate the case of unzipping under a finite force. The opening fork may go backward, closing a previously open base pair, and reach this base pair later. Therefore the number  $u_i$  of opening transitions  $i \rightarrow i+1$ ,  $u_i$ ,

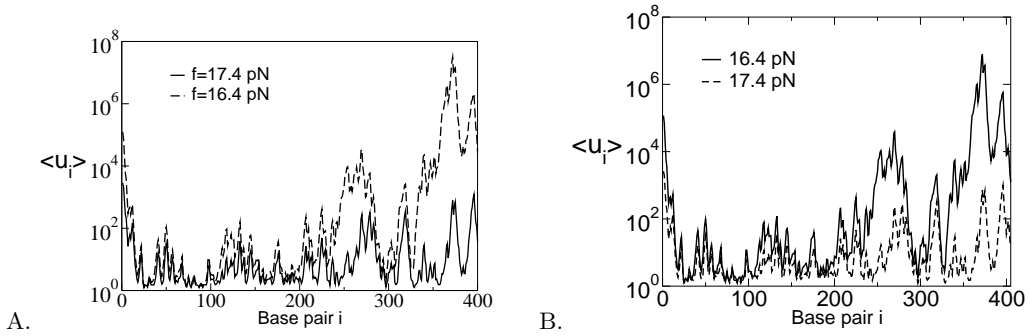


FIG. 17: Average number  $\langle u_i \rangle$  of openings of bp  $i$  for the  $\lambda$ -phage sequence during one unzipping for forces 16.4 and 17.4 pN. **A.** theoretical values in the limit of infinite time. **B.** numerical values from MC simulations with  $M = 10^7$  steps. Note that the infinite time theoretical values coincide with the numerical values up to some base index  $i_{max}$  such that  $\sum_{i < i_{max}} \langle u_i \rangle \ll M$  e.g.  $i_{max} \simeq 200$  for  $f = 16.4$  pN and  $M = 10^8$  steps.

is not always equal to unity but is stochastic and varies from experiment to experiment, and base to base. To calculate the distribution of  $u_i$  it is convenient to think of the opening and closing process as an unidimensional random walk where, at each move, the probability to go backward and forward (closing and opening transitions respectively) are equal to  $q_i$  and  $1 - q_i$  respectively, with

$$q_i = \frac{e^{g_s(f)}}{e^{g_s(f)} + e^{g_0(b_i, b_{i+1})}}. \quad (65)$$

For forces larger than the critical force, we have  $q_i < \frac{1}{2}$ : the random walk is submitted to a forward drift and is transient. We define the probability of escape,  $E_i$ , as the probability of never reaching back position  $i$  starting from position  $i + 1$ . The case of infinite force corresponds to  $E_i = 1$ . For a homogeneous sequence the free energy landscape  $G(n, f)$  in which the random walk takes place is simply a tilted line;  $E = (1 - 2q)/(1 - q)$  depends on the force and on the sequence type. For a heterogeneous sequence the free energy landscape  $G(n, f)$  is more complex (Fig. 9),  $E_i$  depends not only on the force and on the base type  $b_i$  (and on its neighbor  $b_{i+1}$ ) but also on its environment e.g. whether base  $i$  is located in a local minimum or in a local maximum of the free-energy landscape. We show how to calculate  $E_i$  in Appendix F for any given sequence.

The distribution  $\rho_1$  of the number  $u_i$  of opening transitions  $i \rightarrow i + 1$  during a single unzipping is simply obtained from  $E_i$  and reads

$$\rho_1(u_i) = (1 - E_i)^{u_i - 1} E_i \quad (66)$$

From equation (66) we have that the average number of openings of bp  $i$  is

$$\langle u_i \rangle = \frac{1}{E_i}. \quad (67)$$

$\langle u_i \rangle$  is shown in Fig. 17 for forces  $f = 16.4, 17.4$  pN for the first 400 bases of the  $\lambda$  phage DNA sequence. Theoretical values for  $\langle u_i \rangle$  are obtained in the limit of infinite time while MC simulations (or experiments) duration is finite. Call  $t_i^{last}$  the expectation value of the last-passage time of the fork at site  $i$ ;  $t_i^{last}$  is finite since the random walk is transient. Clearly theoretical and MC values for  $\langle u_i \rangle$  will coincide for bases of indices  $i < i_{max}$  where  $t_i^{last}$  is equal to the duration of the simulation. In practice we estimate  $i_{max}$  through the condition  $\sum_{i < i_{max}} \langle u_i \rangle \simeq M$ , where  $M$  is the number of MC moves. The outcome for  $i_{max}$  is plotted in the inset of Fig 5. For instance, as shown in Fig. 17,  $i_{max} \simeq 200$  for  $f = 16.4$  pN and  $M = 10^8$ .  $\langle u_i \rangle$  varies a lot from base to base, and reaches values up to  $10^8$  (for the considered force).

The generalization of the calculation of the distribution  $\rho_R(u_i)$  of the number of openings of base pair  $i$  to the case of  $R$  unzippings is immediate (Appendix B 2). The result is the  $R^{th}$  convolution power of  $\rho_1$ , and reads

$$\rho_R(u_i) = \binom{u_i - 1}{R - 1} (1 - E_i)^{u_i - R} E_i^R. \quad (68)$$

## 2. Error in predicting a base in the absence of stacking

The number of opening transitions of a base at finite force,  $u_i$ , plays the same role as the number  $R$  of repetitions of the unzippings at large force. As the fork visits again and again the same base pair more and more data are collected on the sojourn time  $t_i$  on this base and the prediction error becomes smaller and smaller. However, contrary to  $R$ ,  $u_i$  is a stochastic variable. The error in predicting base pair  $i$  of type  $b_i = W, S$ , in the absence of stacking is then obtained by averaging the error on this bond at large force and after  $u_i$  unzippings,  $\epsilon_{u_i}^{b_i}$  (28), over the distribution  $\rho_R$  (68),

$$\epsilon_{f,R}^{b_i} = \sum_{u_i \geq 1} \rho_R(u_i) \epsilon_{u_i}^{b_i}, \quad (69)$$

where the  $f$  subscript indicate that the above formula holds for a finite force. A detailed derivation of eqn (69) is given in Appendix G 1. In the limit of large force  $E_i \rightarrow 1$  from (65),  $\rho_R(u_i) \rightarrow \delta_{u_i,R}$  from (68), and  $\epsilon_{f,R}^{b_i} \rightarrow \epsilon_R^{b_i}$  as expected.

Error (69) can be easily computed when the error  $\epsilon_{u_i}^{b_i}$  is replaced with asymptotic expression (30). Using the expression for the generating function of the probability  $\rho_R$  with argument  $\exp(-1/R_c)$  given in Appendix B 2 we obtain

$$\epsilon_{f,R}^{b_i} \simeq e^{-R/R_c(f,i)} \quad \text{with} \quad R_c(f,i) = \left[ \ln \left( 1 + \langle u_i \rangle (e^{1/R_c} - 1) \right) \right]^{-1} \quad (70)$$

The above decay constants  $R_c$  can be approximated with the simpler expression

$$R_c(f,i) \simeq \frac{R_c}{\langle u_i \rangle} \quad (71)$$

which are quantitatively accurate unless the number of required unzipping at large force,  $R_c$ , becomes much smaller than  $\langle u_i \rangle$  *i.e.* close to the critical force. This formula simply expresses that the effective number of unzippings to correctly predict base  $i$  at finite force is  $R \times \langle u_i \rangle$  rather than  $R$ . Recall that the value of the decay constant of the error at high force,  $R_c$ , depends only on the free energy difference between  $W$  and  $S$  bases. At finite force this decay constant is roughly divided by  $\langle u_i \rangle$ . The latter depends on the whole free energy landscape around the base. Therefore at finite force, even in the absence of stacking interaction, the error on a base depends on the whole sequence of bases. Moreover bases with a large  $R_c$  that are in a valley of the free energy landscape can be better predicted than bases with a small  $R_c$  located on the top of barriers in the landscape.

Let us apply the above result to the case of a homogeneous sequence, with two base types,  $b = W, S$ . The decay constant  $R_c$  (31) at high force depends only on the free energy difference  $\Delta$  between  $W$  and  $S$  bases. For a homogeneous sequence the average number of openings of each base is simply  $\langle u \rangle = \frac{1-q}{1-2q}$ , where  $q$  is obtained from formula (65) with  $g_0(b_i, b_{i+1}) = g_0(b)$ . In Fig 18 we plot the error for  $W$  bases for  $\Delta = 2.8$  (to distinguish a sequence of bases  $A$  or  $T$  from a sequence of bases  $G$  or  $C$ ) and  $\Delta = 0.5$  (to distinguish a sequence of  $A$  bases from one of  $T$  bases, or a sequence of  $C$  bases from one of  $G$  bases). The plot for a repeated sequence of  $S$  bases is similar. As shown in Fig 18 the error sharply decreases when the force reaches its critical value from above e.g.  $f_c = 9.25$  pN for  $g_0(W) = -1.1$   $k_B T$ . As shown in Fig 18 the decay constant (70)

$$R_c(f) = \left[ \ln \left( \frac{(1-q)e^{1/R_c} - q}{1-2q} \right) \right]^{-1} \quad (72)$$

obtained by approximating  $\epsilon_u^b$  with a pure exponential is in perfect agreement with the numerical calculation of formula  $\epsilon_{f,R}^W$ . The simplified expression (71)

$$R_c(f) = R_c \times \frac{1-2q}{1-q}. \quad (73)$$

is in very good agreement with  $R_c(f)$ , except in the case  $\Delta = 2.8$ ,  $f = f_c + 2$  pN for which the decay constant is very small.

The value of  $R_c(f)$  is plotted as a function of the force in Fig 19 for various sequences, and allows us to draw the phase diagram for the prediction in the force vs. number of unzippings plane. The prediction becomes perfect,  $\epsilon_{f,R}^b \ll 1$ , if the number  $R$  of unzippings is (much) larger than some crossover value  $R_c$  (72). It appears that  $R_c(f)$  is always smaller than its infinite force value  $R_c$ , and vanishes when the force reaches the critical unzipping force from

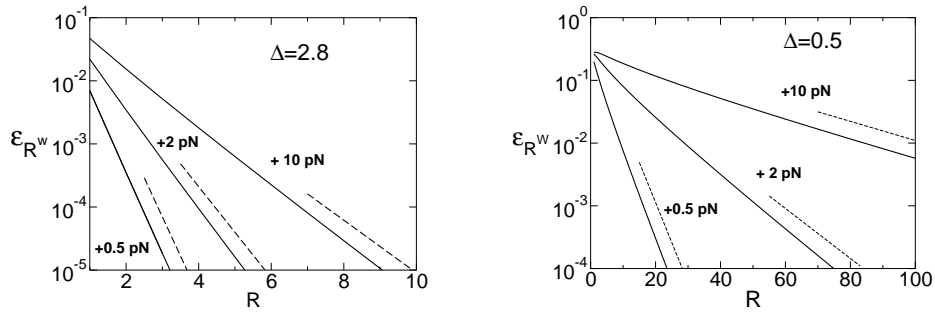


FIG. 18: Probability  $\epsilon$  of misprediction on repeated sequences of  $W$  (empty dots, dashed lines) and  $S$  (black dots, full lines) bases for pairing free-energy differences  $\Delta = 2.8$  (A) and  $\Delta = 0.5$  (B) in the absence of stacking. For each case we show the error as a function of the number  $R$  of unzippings for forces above the critical force by 0.5, 2 and 10 pN. The decay constants have for  $\Delta = 2.8$  the following values:  $R_c(f = \infty) = 32$ ; for  $f = f_c + 10$  pN,  $R_c = 28.5$ ; for  $f = f_c + 2$  pN,  $R_c = 10.9$ ; for  $f = f_c + 0.5$  pN,  $R_c = 3.4$ .

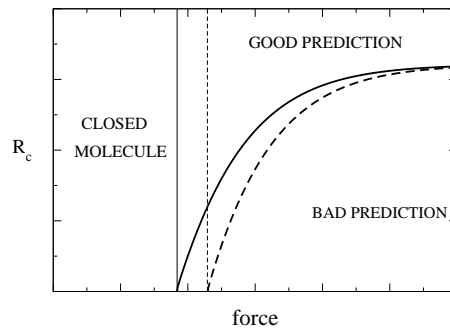


FIG. 19: Phase diagram in the number of unzippings vs. force plane. Efficient prediction is possible above the critical line  $R_c(f)$  (72). Here  $g_0(W) = -1.06$ ,  $g_0(S) = -1.55$ . The full line indicates the repeated  $W$  sequence, the dashed line corresponds to the repeated  $S$  sequence. For forces smaller than the critical value  $f_c \simeq 9$  pN for the  $W$  sequence,  $f_c \simeq 12$  pN for the  $S$  sequence (vertical lines) the molecule remains closed. At large force the number of required unzippings reaches a common value  $R_c \simeq 30$ .

above,  $f \rightarrow f_c^+$ . In this limit,  $q \rightarrow \frac{1}{2}$ : the motion of the opening fork becomes purely diffusive, and each base is visited a very large number of times going to infinity for an infinite duration of the experiment. Predictions made from a single unzipping are reliable provided  $R_c(f) < 1$  *i.e.* the force  $f$  does not exceed by a large amount its critical value  $f_c$ ,

$$f - f_c \leq \frac{\Delta^2}{8d_c} \quad (74)$$

where  $d_c = |dg_s/df(f_c)|$  is twice the extension of a DNA single strand monomer at the critical force, and we have used expression (31) for  $R_c$ . Typically,  $d_c \sim 1$  nm  $\simeq 0.25$   $k_B T$ /pN, leading to  $f - f_c < \frac{1}{2}\Delta^2$  pN with  $\Delta$  expressed in units of  $k_B T$ . Notice that this theoretical result does not consider the actual number of open base pairs, which decreases as the force is lowered to its critical value, but only the quality of their prediction.

### 3. Results for heterogeneous sequence in presence of stacking interactions

The above theory tells us how many unzippings are necessary to recognize a base type from another at moderate force, when the pairing free energies of these two base types differ by  $\Delta$  and when the fork opens the base  $\langle u_i \rangle$  times in each unzipping. It can be applied to the case of bond and not base recognition as we have done at large force in Section IV C 1. The number of unzippings  $R_c(f, i, b_i^L b_{i+1}^L \rightarrow b b')$  necessary to recognize that the bond between base pairs  $i$  and  $i + 1$  is not  $b b'$  is given by expression (70) or (71) with  $R_c$  substituted with  $R_c(b_i^L b_{i+1}^L \rightarrow b b')$ , see Section IV C 1, which depends on the biochemical parameters  $g_0(b_i^L, b_{i+1}^L) - g_0(b, b')$  given in Table II.

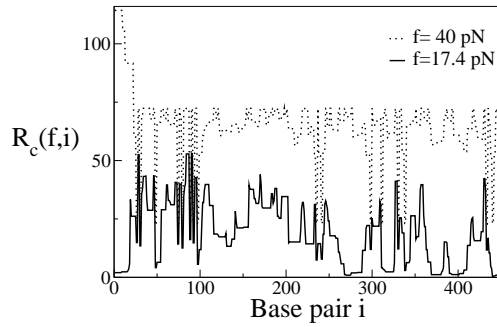


FIG. 20: Theoretical values for the number  $R_c(f, i)$  of unzippings necessary for a good prediction of base  $i$  at force  $f = 17.4$  (full line) and  $f \geq 40$  pN (dashed line) for the first 400 bases of the  $\lambda$  phage sequence obtained from formula (70)

The decay constant of the error on base  $i$  at finite force,  $R_c(f, i)$ , is calculated by applying the recursive formula (63) and the minimization formula (64) after replacing the bond decay constants at infinite force with the ones at finite force,

$$\begin{aligned} \pi_{i,f}^{\rightarrow}(b') &= \min_b \left( \pi_{i-1,f}^{\rightarrow}(b) + \frac{1}{R_c(i, f, b_{i-1}^L b_i^L \rightarrow b b')} \right) \\ \pi_{i,f}^{\leftarrow}(b') &= \min_b \left( \pi_{i+1,f}^{\leftarrow}(b) + \frac{1}{R_c(i, f, b_i^L b_{i+1}^L \rightarrow b' b)} \right), \end{aligned} \quad (75)$$

with boundary condition  $\pi_{1,f}^{\rightarrow}(b) = \pi_{N,f}^{\leftarrow}(b) = 0$ . The minimization condition then reads

$$\frac{1}{R_c(f, i)} = \min_{b \neq b_i^L} \left( \pi_{i,f}^{\leftarrow}(b) + \pi_{i,f}^{\rightarrow}(b) \right). \quad (76)$$

Figure 20 shows the values of  $R_c(f, i)$  at  $f = 17.4$  pN (full line) for the first 400 base pairs of the  $\lambda$ -phage derived from (70).  $R_c(f, i)$  is in very good agreement with the decay constant of the error  $\epsilon_i$  obtained through the numerical inference procedure and shown in Fig 8A. Indeed, roughly, for all bases with  $R_c(f, i) \leq 15$  the numerical inference errors goes to zero with  $R = 40$  unzippings. For a more precise comparison we have focused on two specific bases (Fig 10).

Base pair 6 is located in a valley of the landscape  $G$  at force of 17.4 pN, hence the number of openings of the base,  $\langle u_i \rangle$ , and of its neighbors,  $\langle u_j \rangle$  with  $j$  close to  $i$ , are large *e.g.*  $\langle u_1 \rangle = 28000$ ,  $\langle u_6 \rangle = 60$  as shown in Fig 17. The decay constant of the error quickly decreases with the force from  $R_c(f \geq 40 \text{ pN}, i = 6) = 114$  to  $R_c(f = 17.4 \text{ pN}, i = 6) = 2$ ; these theoretical values are in very good agreement with the numerical findings of Fig 10. Moreover the connected correlation function  $\chi_{i,6}$  at  $f = 17.4$  pN has non-zero value up to the base  $i = 20$ . Solving the recursive eqns (75,76) we found that the decay of the prediction error on  $i = 6$  originates from a 20 defect-sequence where bases 1-20 are locked-in into their complementary values with respect to the true sequence.

Base pair 27 lies, on the contrary, on a barrier of the free energy landscape and the numbers of openings (at a force of 17.4 pN) of this base (and its neighbors) is smaller:  $\langle u_{27} \rangle = 1.5$  as shown in Fig. 17. The decay constant decreases slightly when the force diminishes, from  $R_c(f \geq 40 \text{ pN}, i = 27) = 24$  to  $R_c(f = 17 \text{ pN}, i = 27) = 15$ . These theoretical values agree very well with the fit of the numerical simulations in Fig 10. Moreover the decay of the prediction error on base 27 at  $f = 17.4$  pN came from a two-defect excitation of bases 26-27. Note that numerical results are limited by the finite number of samples from which the error  $\epsilon_i$  is calculated. The number of samples  $M_p$  necessary to estimate accurately the error must be much larger than the inverse of the probability of misprediction. With  $M_p = 2 \cdot 10^4$  (Fig 10) errors smaller than  $\epsilon = 10^{-3}$  cannot be measured. As  $\epsilon$  decreases exponentially with  $R$ ,  $M_p$  must scale as  $\exp(R\mu)$  with  $\mu > R_c$  to reach a good estimate of  $R_c$ . Finite sampling could also lead to statistical bias due to the large deviation fluctuations of  $u_i$ . We show that these effects are negligible in Appendix I.

### E. Inference from two-way unzippings

We hereafter consider that the molecule can be unzipped from both extremities (two-way opening) and want to infer its sequence from the data collected in both directions. This investigation is motivated by the observation that

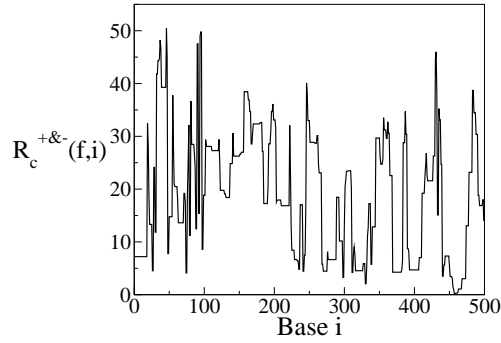


FIG. 21: Decay constant  $R_c^{+&-}(f, i)$  of the prediction error on base  $i$  for the first 450 base pairs of the  $\lambda$  phage DNA, at force  $f = 17.4$  pN from the two-way unzipping numerical unzipping.

the free energy landscape is flipped *i.e.* multiplied by  $-1$  when the molecule is opened from the other extremity. Bases that were located in local maxima in the landscape, hence poorly predicted, become local minima in the new landscape, and are much better predicted.

Let us denote  $+$  the normal direction of unzipping of the molecule: the  $i^{\text{th}}$  base (along the  $5' \rightarrow 3'$  strand of molecule) in this direction is simply  $b_i$ . The free energy to open the first  $n$  bases of the molecule is  $G^+(n, f; B)$ , equal to  $G$  defined in (1). In the reverse direction, denoted by  $-$ , we denote by  $b_i^-$  the  $i^{\text{th}}$  base along the  $5' \rightarrow 3'$  direction:  $b_i^- = \text{compl}(b_{N+1-i})$  where  $\text{compl}(b)$  denotes the complementary base of  $b$ . The free energy to open the first  $n \geq 0$  bases of the molecule in the  $-$  direction is

$$G^-(n, f; B) = \sum_{i=0}^{n-1} g_0(b_i^-, b_{i+1}^-) - n g_s(f) = \sum_{i=N-n+1}^N g_0(b_{N-i}, b_{N-i+1}) - n g_s(f) = -G^+(N-n, f; B) + G(N, f; B) \quad (77)$$

where we have used the symmetry  $g_0(b, b') = g_0(\text{compl}(b'), \text{compl}(b))$  of the  $g_0$  interaction matrix (Table I) [44]. Therefore, up to an irrelevant additive constant, the free energy to open  $n$  bp in the  $-$  direction is simply the opposite of the free energy to open  $N - n$  bp in the  $+$  direction.

If we unzip  $R$  times the molecule in the  $+$  direction the error in predicting base  $i$  will decay exponentially with  $R$  with a decay constant equal to  $R_c^+(f, i)$  given by eqn (73). We may instead open  $R$  times the molecule in the  $-$  direction, and infer the value of base  $i$  (labeled  $N + 1 - i$  in the  $-$  nomenclature). The probability of a mistake is again an exponentially decreasing function of  $R$  with decay constant  $R_c^-(f, i)$  (73), calculated from the number of openings of base  $i$  in the  $-$  direction (Appendix G 2).

Assume now that the unzip  $R/2$  times the molecule in the  $+$  direction and  $R/2$  times in the  $-$  direction. We show in Appendix G 2 that the probability of predicting that the bases attached to the bond  $i, i + 1$  are  $b, b'$  decays exponentially with  $R$  with a decay constant equal to

$$R_c^{+&-}(f, i, b_i^L b_{i+1}^L \rightarrow b b') = \left[ \ln \left( 1 + \langle u_i^+ \rangle (e^{1/2 R_c(b_i^L b_{i+1}^L \rightarrow b b')} - 1) \right) + \ln \left( 1 + \langle u_{i+1}^- \rangle (e^{1/2 R_c(b_i^L b_{i+1}^L \rightarrow b b')} - 1) \right) \right]^{-1} \\ \simeq 2 R_c(b_i^L b_{i+1}^L \rightarrow b b') / (\langle u_i^+ \rangle + \langle u_{i+1}^- \rangle) \quad (78)$$

We have taken into account the effects of stacking interactions between nearest neighbor base pairs as done in Section IV C. The decay constant of the error  $\epsilon_i$  in the two-way unzipping at force  $f$ ,  $R_c^{+&-}(f, i)$ , is obtained using recurrence eqn (76) upon substitution of  $R_c(f, i, b_i^L b_{i+1}^L \rightarrow b b')$  with  $R_c^{+&-}(f, i, b_i^L b_{i+1}^L \rightarrow b b')$ . The results for  $R_c^{+&-}(f, i)$  are shown in Fig 21. A comparison with Fig 20 shows that the number of unzippings necessary for a good prediction greatly decreases with the two-way unzipping procedure with respect to the one-way unzipping (for the same amount of collected data).

## V. TOWARDS MORE REALISTIC DATA MODELING

### A. Finite-bandwidth inference

So far we have assumed that the temporal resolution was infinite. A time-trace contains a perfect information on the opening dynamics *i.e.* on the motion of the fork (set of numbers  $u_i, d_i$ ) and on the sojourn times  $t_i$  for every

base  $i$  of the chain. Real experiments obviously do not have such a perfect sensitivity: actual feedback systems and detectors are limited to delays between measures of about  $\Delta t \sim 0.1 - 1$  ms. This temporal resolution is a major limitation: during the delay  $\Delta t$  the fork can explore up to 100-1000 bases around the starting position, depending on the local structure of the free energy landscape. The true dynamics of the fork is therefore unknown and the prediction algorithm has to consider all the trajectories of the fork (in a  $\sim 100$  bp window). This problem is studied in detail in [42]. Hereafter we limit ourselves to the case of a finite but very large bandwidth where the delay  $\Delta t$  between two measures is of the order of the opening time of a bp (and not much smaller as considered so far).

### 1. Typical jump between two measures

Rates (3) define the non zero (off diagonal) elements of the elementary transitions matrix

$$\hat{H}_{i',i} = r_o(i) \cdot \delta_{i',i+1} + r_c(f) \cdot \delta_{i',i-1} - (r_o(i) + r_c(f)) \cdot \delta_{i',i} \quad (79)$$

The evolution operator after a time  $\Delta t$  is given by the matrix exponential

$$\hat{U} = \exp [\Delta t \hat{H}] \quad (80)$$

The entry  $\hat{U}_{i',i}$  represents the probability of going from base  $i$  to base  $i'$  in the time interval  $\Delta t$ . In principle all transitions are allowed and  $\hat{U}$  is therefore a  $N \times N$  matrix. In practice jumps  $j = i' - i$  are unlikely to exceed (in absolute value) the ratio  $\Delta t/\tau$  where  $\tau$  is the typical time to open a bp. The probability distribution of jumps  $j$ , averaged over the starting base  $i$ , is shown in Fig 22 for  $f = 16.4$  and  $17.4$  pN, and  $\Delta t$  ranging between  $10^{-5}$ s and  $10^{-3}$ s. As the force and the sampling interval increases the distribution gradually spreads over larger jump values, and long tails appear. Nevertheless, long jumps seem to be rare events, restricted to particular regions of the landscape. Most of the information on the opening dynamics can therefore be kept when discarding displacements larger than some threshold  $J$  e.g.  $J = 10$  in Fig 22. To do so, given the starting base  $i$ , we construct a reduced  $(2J+1) \times (2J+1)$  matrix  $\hat{H}^{(J,i)}$  as follows,

$$\hat{H}^{(J,i)} = \begin{pmatrix} -r_o(i-J) - r_c & r_c & 0 & \dots & 0 \\ r_o(i-J) & -r_o(i-J+1) - r_c & r_c & \dots & 0 \\ 0 & r_o(i-J+1) & -r_o(i-J+2) - r_c & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & r_o(i+J-1) & -r_o(i+J) - r_c \end{pmatrix} \quad (81)$$

and the associated evolution operator  $\hat{U}^{(J,i)} = \exp [\Delta t \hat{H}^{(J,i)}]$ , which encodes all the jumps from base  $i$  of amplitude less or equal to  $J$ . There are  $4^{(2J+1)}$  different  $\hat{U}^{(J)}$  matrices, one for each possible choice of the  $2J+1$  bases involved.

### 2. Extended Viterbi algorithm

Given a sequence  $B$  for the molecule the probability of a time-trace  $T$  (where the number of open bp is measured at times multiple of  $\Delta t$ ) is given by a product of  $4^J \times 4^J$  transfer matrices

$$\mathcal{P}^{(J)}(T|B) = \prod_i M^{(J,i)}(b_i, \dots, b_{i+J}) \quad (82)$$

with

$$M^{(J,i)}(b_i, \dots, b_{i+J}) = (\hat{U}_{i,i}^{(J,i)})^{k_i^{(0)}} \prod_{j=1}^J (\hat{U}_{i+j,i}^{(J,i)})^{k_i^{(j)}} (\hat{U}_{i,i+j}^{(J,i+j)})^{k_{i+j}^{(-j)}} \quad (83)$$

and  $k_i^{(j)}$  is the number of transitions  $i \rightarrow i+j$ , with  $j = -J, -J+1, \dots, J-1, J$  in  $T$ . Notice that  $k_i^{(0)}$ ,  $k_i^{(1)}$  and  $k_i^{(-1)}$  coincide with  $t_i/\Delta t$ ,  $u_i$  and  $d_i$  respectively.

An extended Viterbi algorithm allows us to find the most probable sequence. We now have to consider the probability of a sequence of  $J$  contiguous base, starting from  $i$ , and write a recursion equation for this probability,

$$P_{i+1}^{(J)}(b_{i+1}, \dots, b_{i+J-1}) = \max_{b_i} [M^{(J,i)}(b_i, \dots, b_{i+J}) \times P_i^{(J)}(b_i, \dots, b_{i+J-1})], \quad (84)$$

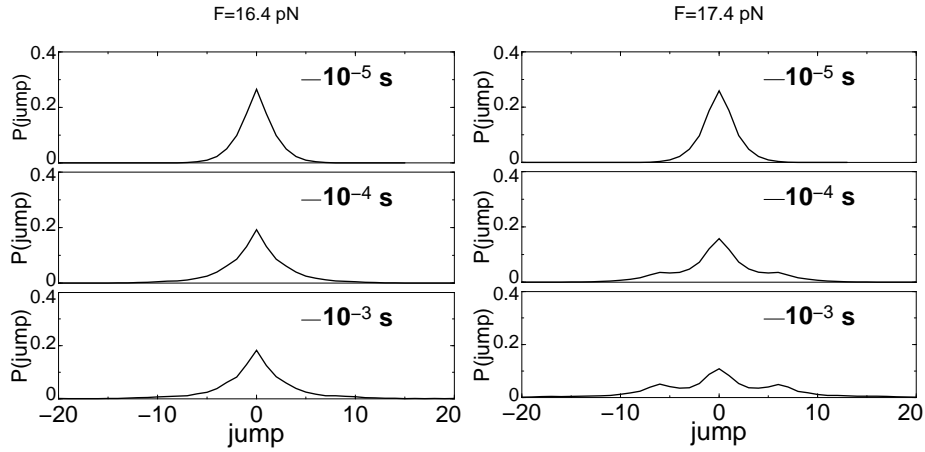


FIG. 22: Probability distribution of a  $j$ -base jump for  $\Delta t$  between  $10^{-5}$  s and  $10^{-3}$  s and for forces  $f = 16.4$  and  $17.4$  pN. Notice that the probability is not necessarily a monotonously decreasing function of  $|j|$ , see extra humps in the right column, due to sequence effects.

which extends eqn (8) to  $J \geq 2$ . For the first base  $i = 1$  the optimization is simply

$$P_2^{(J)}(b_2, \dots, b_{J+1}) = \max_{b_1} M^{(J,2)}(b_1, b_2, \dots, b_{J+1}) \quad (85)$$

The optimal choice for  $b_1$  depends on the  $J$  next base values,  $b_1^* = b_1^{max}(b_2, \dots, b_{J+1})$ . Then we find the next base,  $b_2^*$  as a function of  $b_3, \dots, b_{J+2}$  through (84), and so on, until the last base of the chain is reached. Its most probable value is selected and the whole optimal sequence is recursively reconstructed from the  $b_i^{max}$  functions.

### 3. Numerical study

We first generate a set of numerical data by recording the MC output (fork position) at discrete times multiple of a sampling interval  $\Delta t$ ; intermediate states are simply ignored as the instrument does not have the resolution to appreciate them. Then we preprocess this partial time-trace to obtain the transition number  $k_i^{(j)}$ , and make a prediction for the sequence using the above extended Viterbi algorithm.

Figure 23A shows the quality of prediction as a function of the delay  $\Delta t$  at fixed range  $J = 2, 3, 4, 6$  and for a single unzipping ( $R = 1$ ). Data shows that, for a given range  $J$ , there exists a threshold value for  $\Delta t$  above which the maximum displacement permitted becomes too small to properly describe the unzipping dynamics. The information collected is no longer sufficient for a reliable prediction and the error  $\epsilon$  rapidly increases (see Fig 23A). As expected the threshold  $\Delta t$  increases with the range, meaning that larger ranges are better suited to deal with longer sampling intervals. When  $\Delta t$  is small, comparable with the elementary sojourn time on a base ( $\tau \simeq 1 \mu\text{s}$  for a weak base), the performances are equivalent to the one of the  $J = 1$  case.

The relationship between the range  $J$  and the largest delay  $\Delta t$  it can sustain is better seen on the case of uniform sequences. The characteristic sojourn time on a base,  $\langle t \rangle$  (19), is then uniform throughout the sequence e.g.  $\langle t \rangle \simeq 1 \mu\text{s}$  for a repeated sequence of  $W$  bases. Fig 23B shows that the prediction is perfect up to a temporal resolution  $\Delta t \simeq J \times \langle t \rangle$ , where  $\langle t \rangle$  is the characteristic sojourn time on a base pair, and  $J$  is the range of the algorithm. The existence of a threshold for the delay is clearer at high  $R$  than for  $R = 1$  (Fig 23A) due to the presence of larger fluctuations in the sojourn time in the latter case.

Figure 24 (left) shows that the quality of the prediction betters when the information from several opening experiments is collected. As long as the typical jump associated to a delay  $\Delta t$  is smaller than the range  $J$  (Fig 22) the error  $\epsilon$  can be reduced and values of order  $10^{-2}$  are reached after 50 unzippings for the  $\lambda$ -phage sequence at force  $f = 16.4$  pN. Once the threshold  $\Delta t$  is crossed, however, the loss of information can not be ‘repaired’ and repetitions of the experiment appear to be useless. The fork has moved too far away during the delay  $\Delta t$  and a lot of information falls out the window of size  $J$  our algorithm is based on, an effect which cannot be compensated with multiple experiments. The effect is qualitatively similar for the weak/strong (AT/GC) distinction shown in Figure 24, but is somewhat less dramatic from a quantitative point of view.

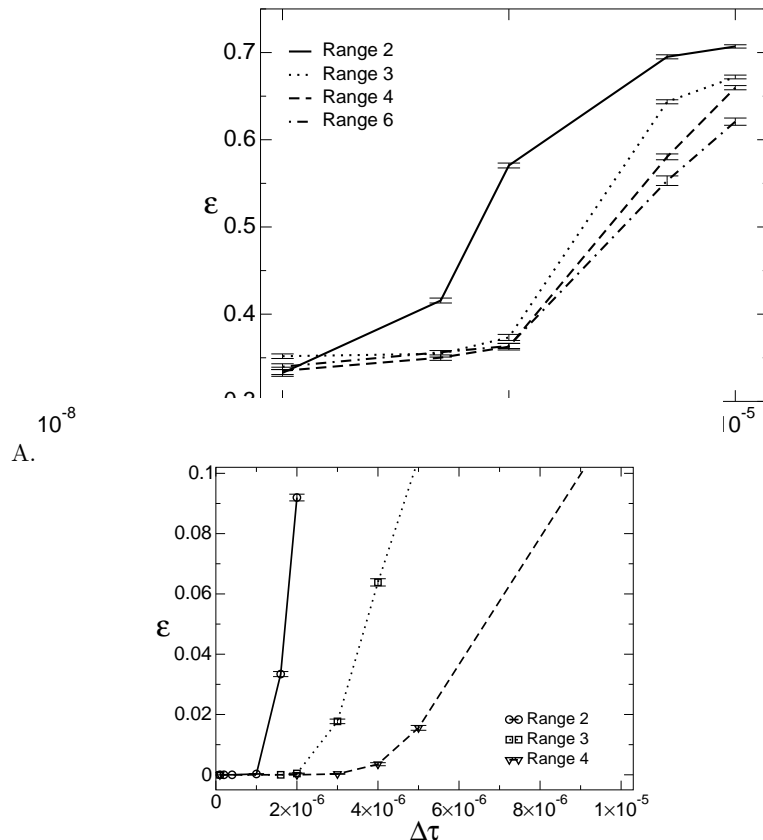


FIG. 23: Error  $\epsilon$  as a function of the delay  $\Delta t$  between measures for various ranges (shown on Figure). **A.** Case of one unzipping ( $R = 1$ ) of a  $\lambda$ -phage DNA molecule at  $f = 16.4$  pN. **B.** Case of  $R = 20$  unzippings of a uniform sequence of weak bases at  $f = 11.8$  pN. Results are averaged over 50 samples in both panels.

## B. Fluctuations of the unzipped DNA strands

Real experiments give access to the extension  $x$  of the open DNA (ssDNA) strands, and not to the number  $i$  of open bp (Fig. 1). Due to the intrinsic elasticity of the strands  $x$  fluctuates even at fixed  $i$ , and these fluctuations grow with  $i$ . Indeed a strand is made of  $i$  monomers, each acting as a spring with stiffness constant  $K \simeq 170$  pN/nm at  $f = 16$  pN and room temperature [24]. The distribution  $A(x|i)$  of the extension  $x$  for a given  $i$  is roughly Gaussian, with mean  $i x_0$  where  $x_0 = dg_{ss}/df \simeq .9$  nm is twice the average extension of a ssDNA monomer, and standard deviation  $\sqrt{i} \delta x$  where  $\delta x = \sqrt{2k_B T/K} \simeq .2$  nm (Fig 25). Distribution  $A$  could be precisely measured through a combination of optical trap and single-molecule fluorescence techniques [21].

### 1. Effect of ssDNA fluctuations on the Bayesian inference

We hereafter study the effects of these fluctuations on the inference problem in the absence of stacking interactions and at high force. We start by making more precise the notion of the time spent on a base:

- *the real time  $t_i^r$* : this is the time really spent by the fork on bp  $i$ , simply denoted by  $t_i$  so far. This number is stochastic since the fork undergoes a random walk motion, with a distribution depending on the nature of base  $i$  (18). The absence of stacking ensures that real times attached to distinct bases are uncorrelated; the probability of the set of real times  $T^r = \{t_i^r\}$  given a sequence  $B$  is, up to a sequence-independent multiplicative factor,

$$\mathcal{P}(T^r|B) \propto \prod_i \exp [g_0(b_i) - r e^{g_0(b_i)} t_i^r] \quad (86)$$

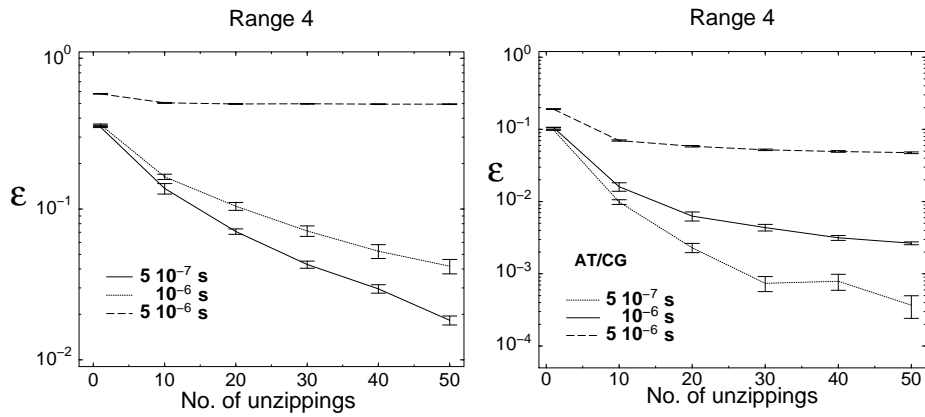


FIG. 24: Left: Fraction of mispredicted bases  $\epsilon$  as a function of the number of unzippings for different temporal resolutions  $\Delta t$ . The value of the range is  $J = 4$ . Right: same as right but we only discriminate among strong and weak bases. Data refer to the opening of a  $\lambda$ -phage sequence at  $f=16.4$  pN and they are averaged over 50 samples.

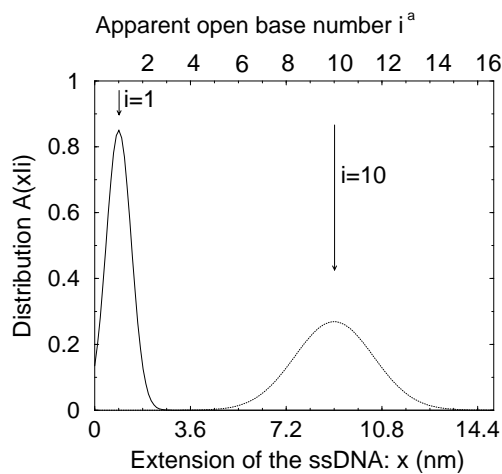


FIG. 25: Distribution  $A(x|i)$  of the extension  $x$  of the open ssDNA at fixed position of the opening fork,  $i = 1$  and  $i = 10$ . The r.m.s. of the distribution (at a force of 16 pN) increases as  $\sqrt{i}$ . The apparent value of the number of opened bases corresponding to a given  $x$ ,  $i^a$  (88), is shown on the top axis.

which corresponds to (5,6) in the limiting case of high force and no stacking. Given a set of real times the best sequence  $B^*(T^r)$  is the one maximizing  $\mathcal{P}(T^r|B)$ . The probability of predicting sequence  $B$  is, given the true sequence  $B^L$ ,

$$\mathcal{Q}^r(B) = \int dT^r \mathcal{P}(T^r|B^L) \prod_{B'(\neq B)} \theta(\mathcal{P}(T^r|B) - \mathcal{P}(T^r|B')) \quad (87)$$

where  $\theta$  is the Heaviside function,  $\theta(x) = 1$  if  $x > 0$ , 0 otherwise. In practice, however, one has no access to the real times.

- *the apparent time  $t_i^a$* : Given a measure for the extension  $x$  of the ssDNA we define the apparent position of the fork through

$$i^a = \text{Closest integer to } \frac{x}{x_0} . \quad (88)$$

The value of  $i^a$  is stochastic, with a probability  $A$  depending on the real position of the fork,  $i^r$ . Considering Rouse dynamics for the monomers [41] the longest relaxation time of a strand is, denoting the viscosity of the solvent by  $\zeta$ ,  $t_r(n) \sim \zeta/(K\pi^2) \times (2n)^2 \sim 100 n^2$  ps. For molecules with  $< 100$  bp ssDNA reaches equilibrium

faster than the fork moves. The probability to observe  $i^a \geq 1$  at some instant thus depends only on the true position  $i^r$  of the fork at the same time, and reads, when  $i^r \geq 1$ ,

$$A_{i^a, i^r} = \int_{i^a - \frac{1}{2}}^{i^a + \frac{1}{2}} \frac{d\nu}{\sqrt{2\pi i^r \sigma^2}} \exp\left[-\frac{(\nu - i^r)^2}{2 i^r \sigma^2}\right] \quad (89)$$

with  $\sigma^2 = \delta x/x_0$ ; the expression for  $i^a = 0$  is obtained from (89) upon replacement of the lower integration limit with  $-\infty$ . When the molecule is entirely closed ( $i^r = 0$ ) all values of  $i^a$  have zero probability except  $i^a = 0$  ( $A(0|0) = 1$ ); this choice amounts to neglect the fluctuations in the extension of the DNA linkers.

We call  $t_i^a$  the time apparently spent by the fork on bp  $i$ , that is, the number of measures in a time-trace in which the fork appears to be at location  $i$  according to (88), divided by the delay  $\Delta t$  between two measures. Matrix  $A$  (89) implicitly define the probability distribution of a set of apparent times  $T^a = \{t_i^a\}$  given a set  $T^r$  of real times, see Appendix H for more details. Multiplicating by (86) and integrating over the real times formally defines the probability  $\mathcal{P}^a(T^a|B)$  of a set  $T^a$  of apparent times given a sequence  $B$ . Given an apparent signal  $T^a$  the best sequence  $B^*(T^a)$  is the one maximizing  $\mathcal{P}^a(T^a|B)$ . The probability of predicting sequence  $B$  is, given the true sequence  $B^L$ ,

$$\mathcal{Q}^a(B) = \int dT^a \mathcal{P}^a(T^a|B^L) \prod_{B'(\neq B)} \theta(\mathcal{P}^a(T^a|B) - \mathcal{P}^a(T^a|B')) . \quad (90)$$

Consider first the ideal case where the delay  $\Delta t$  between successive measures is vanishingly small. In this limit, given the set of real times, the apparent times  $t_i^a$  are not stochastic but simply obtained through the convolution of the  $t_i^r$ 's with matrix  $A$  (89):  $T^a = A \cdot T^r$  in vectorial notation. Starting from the probability (90) of predicting a sequence from the apparent times and performing the change of variable  $T^r = A^{-1} \cdot T^a$  we obtain  $\mathcal{Q}^a(B) = \mathcal{Q}^r(B)$  (87). The probability, within Bayes framework, of predicting the true sequence  $B^L$  is the same as in the absence of fluctuations. In particular the values for  $R_c$  calculated in the previous Section are unaffected by the presence of ssDNA elasticity.

This result does not hold for finite delays  $\Delta t$  where, given a set  $T^r$  of real times, the apparent times  $t_i^a$  are stochastic due to the finite number of samplings during the sojourn time on each base. Let us assume that the delay  $\Delta t$  between successive measures is small with respect to the sojourn time  $\langle t \rangle$  on a base pair but non zero. The Bayesian probability  $\mathcal{Q}^a(B)$  of a sequence now depends on the fluctuation matrix  $A$ . For the sake of simplicity we consider only the case of a large number of unzippings, and a repeated sequence of bases  $S$  with a unique  $W$  base at location  $i$ . Let

$$\rho = \frac{\Delta t}{\langle t \rangle^S} = r e^{g_0(S)} \Delta t \quad (91)$$

denote the ratio of the delay over the average time spent on a  $S$  base; by hypothesis  $\rho \ll 1$ . The probability that the  $W$  base is not correctly predicted reads (Appendix H),

$$\epsilon_{R,i} = \mathcal{Q}^a(B^S) \sim e^{-R/R_c(i)} \quad \text{where} \quad R_c(i) \simeq \frac{8}{\Delta^2 (A^T \beta^{-1} A)_{i,i}}, \quad \beta_{j,k} = (1 - \rho) (A A^T)_{j,k} + \rho I d_{j,k} . \quad (92)$$

and  $A^T$  denotes the transposed matrix of  $A$ . The above formula holds for a small difference  $\Delta$  of free energies between the weak and strong bases, see (32). The outcome for  $R_c(i)$  is shown in Fig 26A for  $\rho = 0.1$  and grows as the square root of  $i$  [34]. More precisely we find  $R_c(i) \propto \sigma \sqrt{i}$  where  $\sigma = \sqrt{\delta x/x_0}$ , and the proportionality factor depends on  $\rho$ . Perfect prediction is still possible, but at the price of a number of unzippings growing with the base index.

## 2. Sequence prediction through deconvolution

The above results do not tell us how to make a prediction for the sequence given an apparent signal  $T^a$ . The expression for  $\mathcal{P}^a$  is highly non local: the probability of the time  $t_i^a$  does not depend on the type  $b_i$  of base at location  $i$  but also on its neighbors. A practical procedure consists in calculating, once the apparent times  $T^a$  are measured, the set of deconvoluted times  $T^d = \{t_i^d\}$  through the formula

$$t_i^d = \sum_j D_{i,j} t_j^a \quad (93)$$

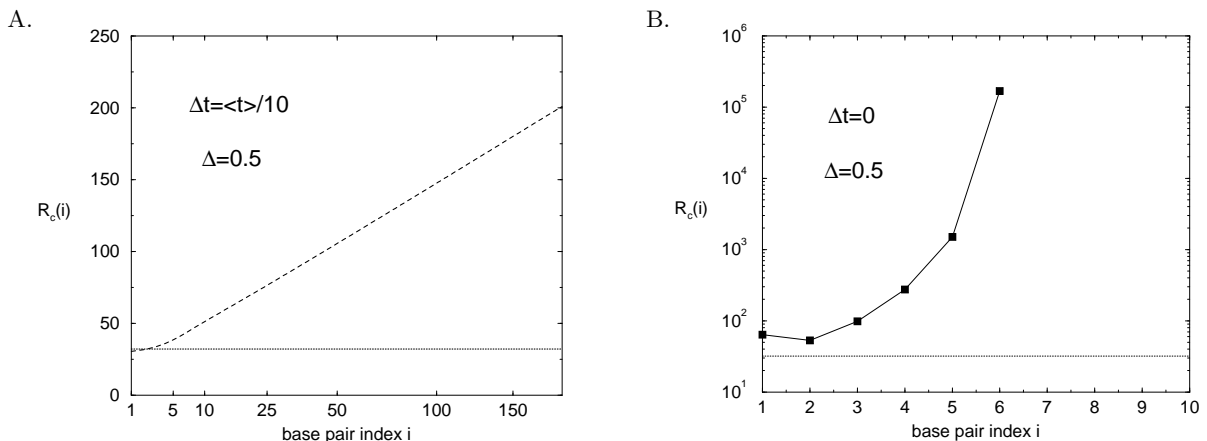


FIG. 26: Value of the number of unzippings controlling the decay of the error in predicting a base,  $R_c(i)$ , as a function of the base index  $i$ . The sequence is made of bases  $S$  with a single  $W$  base at position  $i$ . The dotted line shows the value of  $R_c$  in the absence of ssDNA fluctuation, for a difference of free energy between  $S$  and  $W$  bases equal to  $\Delta = 0.5$ . **A.** Case  $\Delta t = \langle t \rangle / 10$ . The decay constant  $R_c(i)$  for the Bayesian error (92) grows as  $\sqrt{i}$  (dashed line). **B.** Case  $\Delta t \rightarrow 0$ . The full line shows  $R_c(i)$  for the Viterbi procedure without deconvolution; for  $i \geq 7$   $R_c(i)$  is infinite, meaning that the  $W$  base is almost surely predicted to be of  $S$  type. With appropriate deconvolution the dotted line value for  $R_c$  is recovered.

where  $D$  is an appropriate deconvolution kernel to be specified later. Ideally, after deconvolution, the probability of  $T^d$  given the sequence  $B$  should coincide with the local probability (86). The prediction for the sequence is then done through the maximization of  $\mathcal{P}$  (86) over  $B$ , given the set  $T^d$  of deconvoluted times.

We start by showing how the performances of the inference procedure are dramatically worsened by fluctuations if no deconvolution is performed ( $D = Id$ ), and then show how the effects of fluctuations are cured when deconvolution is performed. We focus here on the cases  $R = 1$  and  $R \gg 1$  only, and concentrate on the case  $\Delta t \rightarrow 0$  first. Consider the base at location  $i$ , which we suppose to be, say, of type  $W$ . The error in predicting this base reads, see Appendix H,

$$\epsilon_{1,i}^W = \sum_i \prod_{j(\neq k)} \left(1 - \frac{C_{i,j}}{C_{i,k}}\right)^{-1} e^{-\tau^W/C_{i,k}}, \quad (94)$$

where

$$C_{i,j} = \exp(g_o(b_i) - g_o(b_j)) (DA)_{i,j}, \quad (95)$$

and  $\tau^W, \tau^S$  are defined in (23). The subscript 1 refers to the value  $R = 1$  of the number of unzippings. Figure 27 shows  $1 - \epsilon_{1,i}^W$  as a function of  $\sqrt{i}$  for a repeated sequence  $SSSS\dots$ , and for an alternate sequence  $SWSW\dots$  in the absence of deconvolution ( $D = Id$ ). The error increases from a value for  $i = 1$  essentially equal to its counterpart  $\epsilon_{1,i}^W$  (24) in the absence of strand fluctuation, to reach unity at large  $i$ . This behavior is easily interpreted: in the absence of deconvolution the apparent time  $t_i^a$  (more precisely, the reduced time  $\tau_i^a$  (20)) on base  $i$  is the sum of the real times  $t_j^r$  spent on each base  $j$ , weighted with the probability  $C_{i,j}$  (95). As  $i$  grows more and more bases  $j$  contribute to the sum with smaller and smaller weights, with a number of contributing terms scaling as  $\sqrt{i}$ . The law of large numbers tells us that the distribution of  $\tau_i^a$  is asymptotically concentrated around a single value, equal to  $\tau_\infty^a = e^\Delta$  and to  $\tau_\infty^a = \frac{1}{2}(1 + e^\Delta)$  for the  $SSSWSSS\dots$  (where the unique  $W$  base is located at position  $i$ ) and  $SWSW\dots$  sequences respectively. As these values exceed  $\tau^W$  (23) the base is almost never correctly predicted[45]. The very tiny probability of success is due to the tail of the times below  $\tau^W$ , which decreases exponentially with  $\sqrt{i}$  (Fig 27).

In the limit of a large number  $R$  of unzippings the error decreases as (Appendix H)

$$\epsilon_{R,i} \sim e^{-R/R_c(i)} \quad \text{where} \quad R_c(i) = \begin{cases} \frac{2 \sum_j C_{i,j}^2}{(1 + \frac{\Delta}{2} - \sum_j C_{i,j})^2} & \text{if } \sum_j C_{i,j} < 1 + \frac{\Delta}{2} \\ +\infty & \text{if } \sum_j C_{i,j} \geq 1 + \frac{\Delta}{2} \end{cases}. \quad (96)$$

The above expression was derived when the free energy difference  $\Delta$  between  $W$  and  $S$  bases is small, the hardest case from the inference point of view. In the absence of fluctuation  $A = D = Id$  we find back result (31) as expected. Notice

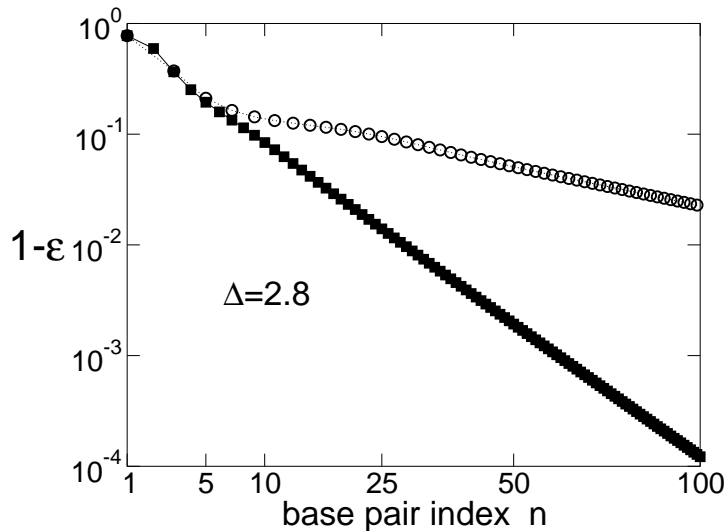


FIG. 27: Probability that a base is correctly predicted,  $1 - \epsilon_i^W$ , as a function of its location  $i$  in the case of: a repeated sequence of  $S$  bases with a single  $W$  base at position  $i$  (black dots), an alternate sequence  $SWSW \dots$  (empty dots). In both cases the rate of success decreases exponentially with the square root of  $i$ . The difference of free energies between  $S$  and  $W$  bases is  $\Delta = 2.8$ .

$R_c = \infty$  simply means that the error does not converge to zero when  $R$  increases. An illustration of this situation is given in Fig 26A. The number  $R_c(i)$  of unzippings necessary to correctly predict a unique  $W$  base located at position  $i$  inside a repeated  $SSSS \dots$  sequence increases with  $i$ , and diverges for  $i \geq 7$  in the absence of deconvolution. The reason for this failure is the same as in the above  $R = 1$  case: the apparent time on base  $i$  is corrupted by too many  $S$  bases and the true nature of the base cannot be recognized.

Fortunately the situation drastically improves when the signal is deconvoluted with the kernel

$$D = A^\dagger \quad (97)$$

equal to the pseudo-inverse of matrix  $A$ . We have not encountered any numerical problem to calculate this pseudo-inverse from the inverse of  $A^T A$  for sequences with a few hundred bases. The matrix  $C$  in (95) then reduces to the identity matrix, and the errors for a single (94) and a large number (96) of unzippings decrease to their respective values in the absence of fluctuations. In particular the number of unzippings necessary to correctly predict a base is simply  $R_c \sim 8/\Delta^2$ , independently of  $i$ . As a conclusion, through an adequate and sequence-independent deconvolution procedure, we have been able to completely remove the effect of ssDNA fluctuations.

In the case of a finite delay  $\Delta t$  we expect that an appropriate deconvolution with the kernel (97) is sufficient to correctly infer the sequence with the extended Viterbi algorithm of Section V A [42].

## VI. SUMMARY AND CONCLUSION

In this paper we have studied the inference of a DNA sequence from Monte-Carlo generated unzipping signals. Inference is made uneasy by the fact that unzipping signals are largely affected by thermal noise, due to the fact that the free energy to open a base pair (the loss in binding free energy plus the work to stretch the unpaired DNA strands) are of the order of  $k_B T$ . The main goal of the present work was precisely to reach a theoretical understanding of how to cope with thermal noise in the inference process.

The present study is in part numerical and in part analytical. From the numerical side we have first generated, from a given sequence, unzipping data by a Monte Carlo algorithm based on a previously introduced dynamical model of the unzipping [28]. We have then implemented algorithms to reconstruct the most probable sequence from the unzipping signal. The prediction error on each base can be simply evaluated through the comparison between the true and the predicted sequences. From a theoretical side we have calculated the error (probability of misprediction) with the aim to understand its dependence on the sequence, the intrinsic parameters *i.e.* the biochemical base pair free energies, and the extrinsic parameters *i.e.* the unzipping force, the number of repetitions of the unzipping, the collection of unzippings from both sides of the molecule, .... Numerical results compare very well with analytical

calculations. Our main analytical finding is that the average prediction error on a base  $i$  decreases exponentially with the number  $R$  of unzippings. The decay constant  $R_c(i)$  gives the number of unzippings required to achieve an excellent prediction of the base. We have analytically calculated the value of  $R_c$  in the following cases: (high force) repeated sequences without (30,31) and with (56) stacking interactions, heterogeneous sequences (64); (moderate force) with (70,71) and without stacking interactions (76), for two-way unzippings (78), and taking into account the fluctuations of the extension of the unzipped strands (92,96).

We have first considered the ideal case in which it is possible to follow directly the dynamics of the opening fork with a perfect temporal resolution; in this limit all base pair opening and closing events are detected. The only source for stochasticity is the thermal motion of the fork. In the absence of stacking interaction the decay constant  $R_c(f, i)$  for the base  $i$  and at a force  $f$  can be obtained, in this case, as the ratio of the decay constant at large force,  $R_c(f = \infty, i)$ , over the average number of openings of the base during a single unzipping,  $\langle u_i \rangle$ . The average number of openings of a base,  $\langle u_i \rangle$ , depends on the free energy landscape of the molecule, determined by the force and the sequence content, and was computed in Appendix F. In the presence of stacking interactions  $R_c(f, i)$  depends on the whole sequence and was calculated through an asymptotic version of the Viterbi algorithm (Section IV D 3). Base pairs exhibit a lock-in phenomenon : there exist blocks of neighbouring bases with the same decay constant  $R_c(f, i)$ , while bases in different blocks have much weaker correlations. We also show that much better predictions on the value of a base can be obtained from the same amount of collected data if the molecule is unzipped from both extremities rather than from one extremity (as done so far).

The assumption of infinite temporal bandwidth and precise knowledge of the fork position dynamics allows us to start from the simplest case for the sequence prediction analysis. The advantage is that Bayesian inference can be done exactly with a fast procedure, the so-called Viterbi algorithm. The most likely sequence, given a measured unzipping signal, is found in a time scaling linearly with the number of the bases. The existence of a fast, exact algorithm allowed us to check analytical results; the latter are indeed always obtained for the optimal sequence, irrespectively of the existence of a practical algorithm capable of finding this sequence.

In the second part of the paper we have made a step forward toward the analysis of real experimental data and have included in the inference analysis two major sources of instrumental limitations: the finite data acquisition bandwidth, and the elastic fluctuations of the unzipped DNA strands.

The finite resolution in time is such that during the time interval between two data acquisitions the opening fork can move by (much) more than one base. The exact Viterbi algorithm has been generalized to the case of a large but finite bandwidth, by considering all the forward and backward transitions of the opening fork which can take place, within a range  $J$ , during the time interval  $\Delta t$  between two measures. This new algorithm is able to reconstruct the sequence when the range  $J$  is of the order of the ratio between  $\Delta t$  and the typical sojourn time  $\langle t \rangle$  on a base pair. Though our extended Viterbi algorithms still runs in a time growing linearly with the number of bases, it is exponential in the range  $J$ , and is limited in practice to  $J \leq 10$ . This algorithm is thus implementable for  $\Delta t \sim 10 \langle t \rangle$ , *i.e.* up to about 10  $\mu$ s. In other word the bandwidth frequency should be larger than 100 KHz, a larger value than the current value for the bandwidth in real experiments of the order of 1-10 KHz. Other algorithms presumably not guaranteed to reach the most likely sequence, but with a running time polynomial in the range  $J$ , should be implemented.

In addition we have considered the effects of the fluctuations in the extension of the DNA strands. Indeed, even if the distance between the extremities of the unzipped strands is typically known within  $< 1$  nm accuracy [3, 10], thermal fluctuations in the strand length (and possibly in the linkers) are responsible for a larger uncertainty over the position of the opening fork. We have, in particular, extended our theoretical formalism to calculate the decay constant of the error with the number of unzippings  $R_c$  at high force, without stacking, in presence of DNA strand fluctuations and with an interval  $\Delta t$  between two measures finite but small with respect to the sojourn time  $\langle t \rangle$ . We have obtained that the decay constant  $R_c$  for the error on base  $i$  is multiplied by  $\sqrt{i}$  with respect to its counterpart in the absence of DNA fluctuations. The further from the beginning of the sequence a base is, the larger is the number of unzipping to reach a good prediction.

The theoretical formalism for  $\Delta t \rightarrow 0$  suggests a way to preprocess the signal by deconvoluting it with the pseudo-inverse of the (sequence-independent) DNA fluctuation matrix (89). This signal can then be processed with the usual Viterbi algorithm, and the quality of the prediction is the same as in the absence of strand fluctuations. A natural question is whether the same deconvolution procedure could be applied to the realistic case of a finite bandwidth or not. We are currently working on this problem, and are developing a formalism for the calculation of  $R_c$  in the presence of DNA strand fluctuations and for experimental value of  $\Delta t \sim 0.1$  ms [42]. The design of efficient inference algorithms in this realistic case is a challenging issue.

An implicit but not well justified assumption we have so far is to have a perfect knowledge of the pairing free energies and dynamics of unzipping *i.e.* of the conditional probability  $P(T|B)$ . In practice, however, modeling cannot be perfect and any functional form for  $P(T|B)$  will be only approximate for a given experimental setup. Numerical investigations show, not surprisingly, that the quality of prediction deteriorate when the rates used by the Viterbi procedure differ too much from their values in the data generating Monte Carlo procedure. A possible way out should

be based on a learning principle: in a first stage unzipping data corresponding to a known sequence ( $\lambda$ -phage) are collected to calibrate the rates, in a second stage predictions are made for new sequences. Last of all we have here considered unzipping at constant force. Investigation of the constant velocity case [10] would be very interesting. Local minima are well predicted and remarkably the force signal may be affected by the substitution of one base pair [10].

Let us finally mention a related albeit more complex problem, the analysis of RNA unzipping data. The non complementarity of single strands in RNA molecules give rise to complex folded secondary structures with multiple helices. Gerland and collaborators have suggested a way to reconstruct RNA secondary structure by combining the recording of the force-extension curve and the passage through a nanopore [29]. The passage through the nanopore would indeed force the helices to open one after the other with a sequence-specific order. In this respect, thanks to the nanopore geometry, the RNA unzipping problem is reduced to a unidimensional problem for which the inference methods presented here could be of interest.

**Acknowledgments.** We thank U. Bockelmann for repeated and useful discussions, and F. Zamponi for a critical reading of the manuscript. We are grateful to H. Isambert for his suggestion of two-way unzipping at the origin of Section IV E. This work has been partially sponsored by the European EVERGROW (IST-001935) and STIPCO (HPRN-CT-2002-00319) programs, and the French ACI-DRAB & PPF Biophysique-ENS actions.

- 
- [1] P.C. Turner, A.G. McLennan, A.D. Bates, M.R.H. White, *Molecular Biology*, Springer-Verlag (2000)
  - [2] V. A. Bloomfield, D. M. Crothers, I. Tinoco J, *Nucleic Acids: Structures, Properties and Functions*, University Science Books, Sausalito, CA (2000)
  - [3] C. Bustamante, Z. Bryant and S. B. Smith *Nature* **421**, 423 (2003).
  - [4] S. Cocco, J.F. Marko, *Physics World* **16**, 37 (2003)
  - [5] S.B. Smith, L. Finzi, C. Bustamante, *Science* **258**, 1122 (1992)
  - [6] P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J.L. Viovy, D. Chatenay, F. Caron, *Science* **271**, 792 (1996)
  - [7] S.B. Smith, Y. Cui, C. Bustamante, *Science* **271**, 795 (1996)
  - [8] B. Essevaz-Roulet, U. Bockelmann, F. Heslot, *Proc. Natl. Acad. Sci. (USA)* **94**, 11935 (1997)
  - [9] U. Bockelmann, B. Essevaz-Roulet, F. Heslot, *Phys. Rev. E* **58**, 2386 (1998)
  - [10] U. Bockelmann, P. Thomen, B. Essevaz-Roulet, V. Viasnoff, F. Heslot, *Biophys. J.* **82**, 1537 (2002)
  - [11] U. Bockelmann, P. Thomen, F. Heslot, *Biophys. J.* **87**, 3388 (2004)
  - [12] M. Manosas, D. Collin, F. Ritort submitted to *Phys. Rev. Lett.* (2006); M. Manosas, F. Ritort in preparation (2006)
  - [13] J. Liphardt, B. Onoa, S.B. Smith, I. Jr. Tinoco, C. Bustamante, *Science* **297**, 733 (2001)
  - [14] C. Danilowicz *et al.*, *Proc. Natl. Acad. Sci. (USA)* **100**, 1694 (2003); *Phys. Rev. Lett.* **93**, 078101 (2004)
  - [15] S. Harlepp *et al.*, *Eur. Phys. J. E* **12**, 605 (2003)
  - [16] A.M. Van Oijen, P.C. Blainey, D.J. Crampton, C.C. Richardson, T. Ellemberg, X. Sunney Xie, *Science* **301**, 123 (2003)
  - [17] T.T. Perkins, R.V. Dalal, P.G. Mitis, S.M. Block *Science* **301**, 1914 (2003).
  - [18] GC Wuite, S.B. Smith, M. Young, D Keller, Bustamante *Nature* **404**, 103 (2000).
  - [19] B. Maier, D. Bensimon, V. Croquette *Proc. Natl. Acad. Sci. (USA)* **97**, 12002 (2000).
  - [20] M.J. Levene, J Korlach J, SW Turner, M Foquet, HG Craighead, WW Webb *et al. Science* **299**, 682 (2003).
  - [21] M.J. Lang, P.M. Fordyce, S.M. Block *J. Biol.* **2**, 6 (2003)
  - [22] A.F. Sauer-Budge, J.A. Nyamwanda, D.K. Lubensky, D. Branton *Phys. Rev. Lett.* **90**, 238101 (2003)
  - [23] J. Mathé, H. Visram, V. Viasnoff, Y Rabin, A. Meller *Biophys. J.* **87**, 3205 (2004).
  - [24] S. Cocco, R. Monasson, J. Marko, *Comptes rendus de l'Académie des Sciences Physiques* **3**, 569 (2002)
  - [25] R. Bundschuh, U. Gerland, *Eur. Phys. J.E* **19**, 319 (2006).
  - [26] D.K. Lubensky, D.R. Nelson. *Phys. Rev. Lett.* **85**, 1572 (2000); *Phys. Rev. E* **65**, 031917 (2002).
  - [27] U. Gerland, R. Bundschuh, T. Hwa. *Biophys. J.* **81**, 1324 (2001).
  - [28] S. Cocco, R. Monasson, J. Marko, *Eur. Phys. J. E* **10**, 153 (2003)
  - [29] U. Gerland, R. Bundschuh, T. Hwa *Phys. Biol* **1**, 19 (2004).
  - [30] M. Manosas, F. Ritort, *Biophys. J.* **88**, 3224 (2004).
  - [31] D. Marenduzzo *et al. Phys. Rev. Lett.* **88**, 028102 (2002).
  - [32] R.E. Thompson, E.D. Siggia, *Europhys. Lett.* **31**, 335 (1995)
  - [33] S.M. Bhattacharjee, D. Marenduzzo *J. Phys. A* **35**, L349 (2002)
  - [34] V. Baldazzi, S. Cocco, E. Marinari, R. Monasson, *Phys. Rev. Lett.* **96**, 128102 (2006)
  - [35] M. Zuker, *Curr. Opin. Struct. Biol* **10**, 303 (2000)
  - [36] Santa Lucia, *Proc. Nat. Aca. Sci. USA* **95**, 1460-1465 (1998)
  - [37] D.J. McKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003)
  - [38] A.J. Viterbi, *IEEE Trans. Inf. Th.* **13**, 260 (1967)
  - [39] J.M. Luck, *Systèmes désordonnés unidimensionnels*, Alea-Saclay, (1992)

- [40] F.J. Dyson, *Phys. Rev.* **92**, 1331 (1953)
- [41] K.L. Sebastian *Phys. Rev. E* **62** 1128 (2000)
- [42] S. Cocco, R. Monasson, in preparation
- [43] A fit of the slope of the curve in figure 10 gives  $R = 100 \pm 1$  while a better fit  $R = 113 \pm 2$  is obtained by taking into account the multiplicative  $1/\sqrt{R}$  term in the error in formula (30).
- [44] To define properly the change in the free energy  $G$  (1) of the molecule when its last base  $i = N$  is opened we have added a fictitious  $i = N + 1$  base; the contribution to the free energy is symbolized by  $\Delta g = g_0(b_N, b_{N+1})$ . In practice  $\Delta g$  is not given by Table I but may have a more complicated origin. For instance the molecule may end with a loop,  $\Delta g$  will then be equal to the gain in entropy when the loop opens.
- [45] The same argument indicate that the probability to mispredict base  $b_i = W$  base among a repeated  $WWWW \dots$  sequence vanishes when  $i$  tends to infinity. The reason is that the apparent time on base  $i$  converges to the average time on the neighbors which are all of the right type  $W$ .

## APPENDIX A: IMPLEMENTATION OF THE EXTENDED VITERBI ALGORITHM

A time trace  $T$  of the unzipping signal, produced by the Monte Carlo procedure, is first encoded in a vector  $\mathcal{K} = \{k_i^{(-J)}, k_i^{(-J+1)} \dots k_i^{(0)}, k_i^{(+1)}, \dots k_i^{(J)}\}$  where  $k_i^{(j)}$  is the number of transitions  $i \rightarrow i + j$ .  $J$  fixes a cutoff on the displacement taken into account: only jumps by  $|j| < J$  bases are considered. The information on the opening dynamics *i.e.* the vector  $\mathcal{K}$ , the applied force  $f$  and the temporal resolution  $\Delta t$  is used to construct the transfer matrix  $M^{(J,i)}$  (83) for the  $i^{th}$  base.

The matrix exponentiation, needed to compute  $\hat{U}$  (80), is carried out by solving the set of  $2J + 1$  coupled differential equations

$$\frac{dy_j}{dt} = \sum_{j'} \hat{H}_{j,j'}^{(J,i)} y_{j'} \quad (\text{A1})$$

where  $j = -J, \dots, J$ , and  $H^{(J,i)}$  is defined in (81). The initial conditions read

$$\begin{aligned} y_i^0 &= 1 \\ y_j^0 &= 0 \quad j \neq i. \end{aligned} \quad (\text{A2})$$

The value of  $y_j$  at time  $\Delta t$  is the matrix element  $\hat{U}_{i+j,i}^{(J,i)}$  of the truncated evolution operator. The operation is repeated for the various values of the starting base index  $i$  to obtain the whole operator. From a numerical point of view we solve (A1) using a classical 4<sup>th</sup> order Runge-Kutta method for integration of ordinary differential equations.

Once the matrix  $\hat{U}^{(J,i)}$  is computed, the transfer matrix  $\hat{M}^{(J,i)}$  can be easily evaluated knowing the unzipping dynamics *i.e.* the vector  $\mathcal{K}$ . The probability of a sequence  $B$  given the unzipping signal  $T$  is then maximized via a transfer-matrix-like algorithm. To avoid errors due to small numbers we apply the recursive procedure (84) to the logarithm of the probability instead of the probability itself. The general optimization step is therefore

$$\ln P_{i+1}^{(J)}(b_{i+1}, b_{i+2}, \dots, b_{i+J-1}) = \max_{b_i} [\ln P_i^{(J)}(b_i, \dots, b_{i+J-1}) + \ln M^{(J,i)}(b_i, \dots, b_{i+J})]$$

At each step, the type of the  $i^{th}$  base that maximizes  $\ln P_i^{(J,i)}$ ,  $b_i^*$ , is stored for each of the  $4^J$  possible choices of following  $J$  bases  $(b_{i+1}, b_{i+2}, \dots, b_{i+J})$ .  $4^J$  possible sequences are thus constructed and kept in memory. When the algorithm reaches the end of the sequence the maximization over the last base type selects the best sequence and all previous bases can be simply reconstructed from the  $b_i^*$ , going backwards from the last base to the first one.

Some problems in memory allocation and state labeling must be faced. The dimension of each vector  $b_i^*(b_{i+1}, b_{i+2}, \dots, b_{i+J})$  grows as  $4^J$  and there are  $N$  (up to 48,502 for a  $\lambda$ -phage DNA) different vectors. When the range  $J$  is large, the memory space required to store this information becomes huge. To circumvent this problem we have reduced the space complexity of the algorithm by increasing its time complexity. To do so we apply the algorithm more times, memorizing, and reconstructing, only a portion of length  $D$  of the sequence during each execution. During the first execution only the last  $D$  bases of the sequence are reconstructed. In the second execution the algorithm stops at base  $N - D$ , where  $N$  is the total number of open base pairs, and another set of  $D$  bases are predicted. This procedure goes on until the first base of the molecule is reached.  $D$  is of course an adjustable parameter and the number of times the algorithm is repeated is chosen consequently.

Our code is written in a range independent way. The user simply sets  $J$  at the beginning of the program, without changing anything else. The  $4^{(2J+1)}$  choices of the variables  $(b_{i-J}, \dots, b_{i+1}, b_{i+2}, \dots, b_{i+J})$  that define a specific reconstruction ‘state’ are represented by a bit string whose length depends on the fixed range  $J$ . The string is assigned

$i + J$	$i + J - 1$		$i - J + 1$	$i - J$	$s$	Sequence
00	00	...	00	00	<b>0</b>	AA ... AA
00	00	...	00	01	<b>1</b>	AA ... AT
00	00	...	00	10	<b>2</b>	AA ... AC
00	00	...	00	11	<b>3</b>	AA ... AG
00	00	...	01	00	<b>4</b>	AA ... TA
00	00	...	01	01	<b>5</b>	AA ... TT
00	00	...	01	10	<b>6</b>	AA ... TC
...	...	...	...	...	⋮	...

TABLE IV: Table of variable labeling for a set of  $(2J + 1)$  bases. Each sequence is identified by a label  $s$  in its binary writing: 2 bits identify the type assigned to each base, the lower bit being corresponding to the base with the lower index along the sequence.

in the following way: 2 bits identify the type selected for a base, the lower bits referring to the base with the lower index along the chain, see Table A. The binary number  $s$  encoding a string of  $2J + 1$  bases is called its label.

The largest range we could test is  $J \sim 10$ . Like the memory cost, the execution time of the program scales linearly with  $N$  but exponentially with the range  $J$ . The time needed to perform a single unzipping (without considering the statistics over samples) increases as  $n_{RK} \times 4^J \times (2J + 1)^3$ , where  $n_{RK}$  is the number of integration steps in the Runge-Kutta subroutine.

## APPENDIX B: CONVOLUTION PRODUCTS FOR $R$ UNZIPPINGS

### 1. Distribution of the sojourn time

The distribution  $P_R$  of the total sojourn time  $\tau$  (26) spent on a base for  $R$  unzippings is defined as

$$P_R(\tau) = \int_0^\infty d\tau^{(1)} P_1(\tau^{(1)}) \int_0^\infty d\tau^{(2)} P_1(\tau^{(2)}) \dots \int_0^\infty d\tau^{(R)} P_1(\tau^{(R)}) \delta\left(\tau - (\tau_i^{(1)} + \tau_i^{(2)} + \dots + \tau_i^{(R)})\right) \quad (\text{B1})$$

where  $P_1$  is defined in eqn (21). Taking the Laplace transform, we obtain

$$\int_0^\infty d\tau P_R(\tau) e^{-s\tau} = \left( \int_0^\infty d\tau P_1(\tau) e^{-s\tau} \right)^R = (1 + s)^{-R} . \quad (\text{B2})$$

It is a simple check that this expression coincides with the Laplace transform of the right hand side of eqn (27), hence proving identity (27) for  $P_R$ .

### 2. Distribution of the number of fictitious unzippings

To calculate the  $R^{\text{th}}$  power (for the convolution product) of  $\rho_1$  (66) we introduce the generating function

$$g(x) = \sum_{u=1}^{\infty} \rho_R(u) x^u = \left( \sum_{u=1}^{\infty} \rho_1(u) x^u \right)^R = \left( \frac{E_i x}{1 - (1 - E_i)x} \right)^R . \quad (\text{B3})$$

Thus  $\rho_R(u)$  is the coefficient of  $x^u$  in the above rightmost expression. It is convenient to define

$$\tilde{g}(x) = \sum_{i=1}^{\infty} \rho_R(u) x^{u-R} = \left( \frac{E_i}{1-x(1-E_i)} \right)^R \quad (\text{B4})$$

We then obtain expression (68) from the identity

$$\rho_R(u) = \frac{1}{(u-R)!} \left. \frac{\partial^{u-R} \tilde{g}}{\partial x^{u-R}} \right|_{x=0}. \quad (\text{B5})$$

### APPENDIX C: STATIONARY DISTRIBUTION OF LOGLIKELIHOOD FIELDS

Assume that the sequence is repeated; hence we can drop the base index  $i$  in the definition of function  $F_i$  (43) and in the distribution  $Q_i$  of the loglikelihood. We rewrite eqn (44) under the form

$$Q(h') = \int_{-\infty}^{\infty} dh T_R(h', h) Q(h), \quad (\text{C1})$$

where the kernel  $T_R$  is defined through

$$T_R(h', h) = \int_0^{\infty} d\tau P_R(\tau) \delta(h' - F(h, \tau)). \quad (\text{C2})$$

In addition we define

$$\tau_1(h) = \frac{h + \Delta^W}{x(e^{\Delta^W} - 1)}, \quad \tau_2(h) = \frac{h + \Delta^S}{x(1 - e^{-\Delta^S})}. \quad (\text{C3})$$

where we have used definition (48) for parameter  $x$ . We now rewrite

$$F(h, \tau) = -h + x(e^{\Delta^W} - 1) \max(\tau_1(h) - \frac{\tau}{R}, 0) + x(1 - e^{-\Delta^S}) \min(\tau_2(h) - \frac{\tau}{R}, 0) \quad (\text{C4})$$

The value of above function of  $\tau$  depends on the relative values of  $\tau_1$  and  $\tau_2$ . Let us make the hypothesis (H1) :  $e^{\Delta^W} + e^{-\Delta^S} > 2$ . Then,  $\tau_1 < \tau_2$  if and only if  $h > h_0$  with

$$h_0 = \frac{\Delta^W (1 - e^{-\Delta^S}) - \Delta^S (e^{\Delta^W} - 1)}{e^{\Delta^W} + e^{-\Delta^S} - 2}. \quad (\text{C5})$$

Assume in addition that (H2) :  $\Delta^W \leq \Delta^S$ . Then

$$h_0 \leq \frac{\Delta^S (1 - e^{-\Delta^S}) - \Delta^S (e^{\Delta^W} - 1)}{e^{\Delta^W} + e^{-\Delta^S} - 2} = -\Delta^S. \quad (\text{C6})$$

We obtain from (C4),

$$F(h, \tau) = \begin{cases} \Delta^W - \frac{\tau}{R} x (e^{\Delta^W} - 1) & \text{if } h > -\Delta^W \text{ and } \tau < \tau_1(h) \\ -h & \text{if } h > -\Delta^S \text{ and } \tau_1(h) < \tau < \tau_2(h) \\ \Delta^S - \frac{\tau}{R} x (1 - e^{-\Delta^S}) & \text{if } \tau > \tau_2(h) \end{cases} \quad (\text{C7})$$

and the following expression for the kernel  $T_R$  (C2),

$$T_R(h', h) = \begin{cases} P_R(-R\tau_1(-h'))/(Rx)/(e^{\Delta^W} - 1) & \text{if } h > -\min(h', \Delta^W) \\ \delta(h' + h) \times [\gamma(R, R \max(0, \tau_1(h))) - \gamma(R, R \tau_2(h))] & \text{if } h > -\Delta^S \\ P_R(-R\tau_2(-h'))/(Rx)/(1 - e^{-\Delta^S}) & \text{if } h' < \min(-h, \Delta^S) \\ 0 & \text{if } h' > \Delta^S \text{ or } \Delta^W \leq -h < h' \leq \Delta^S \end{cases} \quad (\text{C8})$$

where  $\gamma$  is the incomplete Gamma function (29) and distribution  $P_R$  is defined in (27). We then inject expression (C8) for  $T_R$  in the fixed point eqn (C1), and integrate both sides over  $h'$  over the interval  $H \leq h' \leq \infty$ . As a result we obtain the remarkably simple identity

$$\hat{Q}(H) = A(H) - B(H) \hat{Q}(-H) \quad (\text{C9})$$

where the cumulative distribution  $\hat{Q}$  is defined in (45), and functions  $A, B$  in (47).

From (C8) (fourth line)  $Q(h')$  vanishes when  $h' > \Delta^S$ . Hence  $Q(H) = 0$  for  $H > \Delta^S$  (third line of (46)). Choose now  $H < -\Delta^S$ ; then  $\hat{Q}(-H) = 0$  and, from (C9),  $Q(H) = A(H)$  (first line of (46)). Then we iterate (C9) to obtain

$$\hat{Q}(H) = A(H) - B(H) [A(-H) - B(-H) \hat{Q}(H)] , \quad (\text{C10})$$

from which we extract the expression of  $\hat{Q}(H)$  in the range  $-\Delta^S \leq H \leq \Delta^S$  (second line of (46)). It is easy to check that  $\hat{Q}$  is a continuous function of its argument both in  $-\Delta^S$  and  $+\Delta^S$ . Notice that hypothesis (H1,H2) hold for typical values of the binding free-energies.

It is quite remarkable that an exact analytical expression for  $Q(h)$  is available for our model. Indeed the recurrence equation for the field distribution of most disordered one-dimensional cannot be solved in a closed form [39]. Dyson noticed in his original paper [40] that a case of solvable model is obtained when the site disorder (here, the time  $t_i$  spent on each base) is exponentially distributed. The present study generalizes this observation to the case of the convolution of exponentials.

#### APPENDIX D: CALCULATION OF THE ERROR $\epsilon$ AND THE CORRELATION FUNCTION $\chi^{dis}$

Assume the sequence is very long ( $N \gg 1$ ), and consider the base at location  $i$  far away from the extremities ( $1 \ll j \ll N$ ). Base  $i$  can be predicted to be  $b$  ( $= W$  or  $S$ ), with probability

$$P_i^\dagger(b_i) = \exp(-R \pi_i^\dagger(b)) \quad (\text{D1})$$

depending on the stochastic set of times  $\{t_i\}$  spent on the bases. We look for the distribution of the loglikelihoods of base  $i$ ,

$$Q^\dagger(h^\dagger) = \text{Probability}[h^\dagger = \pi_i^\dagger(S) - \pi_i^\dagger(W)] \quad (\text{D2})$$

where the probability is calculated over the sets of sojourn times  $\{t_i\}$ . Notice that we do not expect  $Q^\dagger$  to vary with  $j$  in the bulk of the repeated sequence (see calculation of the correlation function below).

$Q^\dagger$  does not coincide with the distribution  $Q$  of fields used in the iteration equation (44). Indeed the latter merely expresses the dependence of the loglikelihood over base  $i+1$  upon the choice for base  $i$ , independently of what happens at site  $i+2$ . In other words, eqn (9) is a propagation equation for the left-to-right likelihoods  $\pi_i^{\leftarrow}$ ; the  $\rightarrow$  subscript has been omitted so far to lighten notations. The direction of propagation is arbitrary: it corresponds to the choice of running the Viterbi algorithm from the first to the last base, determining the value of this last base, and then deducing the values of all bases from the last one to the first one. Clearly, we could have decided to run the Viterbi procedure in the opposite direction. The recurrence equation for the right-to-left likelihoods  $\pi_i^{\leftarrow}$  is straightforwardly established, and reads

$$\pi_i^{\leftarrow}(b_i) = \min_{b_{i+1}} (\pi_{i+1}^{\leftarrow}(b_{i+1}) - g_0(b_i, b_{i+1}) + r e^{g_0(b_i, b_{i+1})} t_i / R) . \quad (\text{D3})$$

When the binding energy matrix is symmetric, the above recursion can be rewritten as

$$\pi_i^{\leftarrow}(b_{i+1}) = \min_{b_i} (\pi_{i+1}^{\leftarrow}(b_i) - g_0(b_i, b_{i+1}) + r e^{g_0(b_i, b_{i+1})} t_i / R) , \quad (\text{D4})$$

and is identical to the recurrence equation (9) for  $\pi^{\rightarrow}$ . We deduce that the stationary probability distribution of right-to-left fields,  $h_i^{\leftarrow} = \pi_i^{\leftarrow}(S) - \pi_i^{\leftarrow}(W)$ , is equal to the left-to-right field distribution  $Q$ .

Obviously, the actual prediction for base  $i$  is the base  $b_i$  maximizing  $P_i^\dagger$  (D1) and depend on the bases located on both left and right sides, that is, on left-to-right and right-to-left likelihoods,

$$P_i^\dagger(b_i) = P_i^{\rightarrow}(b_i) \times P_i^{\leftarrow}(b_i) \quad i.e. \quad \pi_i^\dagger(b_i) = \pi_i^{\rightarrow}(b_i) + \pi_i^{\leftarrow}(b_i) \quad , \quad (\text{D5})$$

when taking the logarithm. Translating the above equation in terms of fields we obtain

$$h_i^\dagger = h_i^\rightarrow + h_i^\leftarrow \quad . \quad (\text{D6})$$

A symbolic representation of the above equality is proposed in Fig. 28A. The distribution of 'true' likelihoods is thus given by

$$Q^\dagger(h^\dagger) = \int dh^\rightarrow Q(h^\rightarrow) \int dh^\leftarrow Q(h^\leftarrow) \delta(h^\dagger - h^\rightarrow - h^\leftarrow) \quad . \quad (\text{D7})$$

The error in predicting base  $i$  is therefore,

$$\epsilon^W = \int_{-\infty}^0 dh^\dagger Q^\dagger(h^\dagger) = 1 - \int_{-\Delta^S}^{\Delta^S} dh Q(h) \hat{Q}(h) \quad , \quad \epsilon^S = \int_0^\infty dh^\dagger Q^\dagger(h^\dagger) = \int_{-\Delta^S}^{\Delta^S} dh Q(h) \hat{Q}(h) \quad (\text{D8})$$

for repeated sequences of  $W$  or  $S$  bases respectively, see formulae (50,51,52). We have here used definition (45) for the cumulative distribution  $\hat{Q}$  of fields.

A similar approach can be used to calculate the disconnected nearest neighbor correlation function  $\chi^{dis}$  (49). Assume for simplicity that the true sequence is a repeated sequence of  $S$  bases, and consider the two bases at locations  $i$  and  $i+1$ . Call  $h_i^\rightarrow$  and  $h_{i+1}^\leftarrow$  the left-to-right and right-to-left likelihoods incoming onto bases  $i$  and  $i+1$  respectively, see Fig. 28B. Let  $n_i = 1$  if base  $i$  is (correctly) predicted to be  $S$ , 0 if the prediction is  $W$ . We define a similar variable,  $n_{i+1}$ , attached to site  $i+1$ . Finally call  $\tau$  the normalized sojourn time on base  $i$  with distribution (27). Given a pair of incoming likelihoods  $(h_i^\rightarrow, h_{i+1}^\leftarrow)$  and the sojourn time  $\tau$ , the Bayesian prediction for  $(n_i, n_{i+1})$  is

$$(n_i(h_i^\rightarrow, h_{i+1}^\leftarrow, \tau), n_{i+1}(h_i^\rightarrow, h_{i+1}^\leftarrow, \tau)) = \underset{(n, n')}{\operatorname{argmax}} \Xi^{SS}(n, n', h_i^\rightarrow, h_{i+1}^\leftarrow, \tau) \quad (\text{D9})$$

where

$$\Xi^{SS}(n, n', h, h', \tau) = h(1-n) + h'(1-n') + \bar{g}^{SS}(n, n') - \frac{\tau}{R} \exp \bar{g}^{SS}(n, n') \quad (\text{D10})$$

and

$$\begin{aligned} \bar{g}^{SS}(n, n') &= g_0(S, S)(nn' - 1) + g_0(W, W)(1-n)(1-n') + g_0(W, S)[n(1-n') + n'(1-n)] \\ &= nn'(\Delta^W - \Delta^S) + (n+n')\Delta^W + \Delta^W + \Delta^S \quad . \end{aligned} \quad (\text{D11})$$

The correlation function between  $n_i, n_{i+1}$  is

$$\langle n_i n_{i+1} \rangle = \int d\tau P_R(\tau) \int dh_i^\rightarrow Q(h_i^\rightarrow) dh_{i+1}^\leftarrow Q(h_{i+1}^\leftarrow) \delta_{n_i(h_i^\rightarrow, h_{i+1}^\leftarrow, \tau), 1} \delta_{n_{i+1}(h_i^\rightarrow, h_{i+1}^\leftarrow, \tau), 1} \quad (\text{D12})$$

where  $\delta_{a,b} = 1$  if  $a = b$ , 0 otherwise. An inspection of (D10) shows that both bases are correctly predicted to  $S$  when  $h_i^\rightarrow$  and  $h_{i+1}^\leftarrow$  are both smaller than  $-\Delta^S + \frac{\tau}{R}(e^{\Delta^S} - 1)$ . Hence formula (51). Formulae (50) and (53) for repeated WW and SW sequences can be obtained along the same lines through substitution of (D10) and (D11) with, respectively,

$$\begin{aligned} \Xi^{WW}(n, n', h, h', \tau) &= hn + h'n' + \bar{g}^{WW}(n, n') - \frac{\tau}{R} \exp \bar{g}^{WW}(n, n') \\ \bar{g}^{WW}(n, n') &= nn'(\Delta^W - \Delta^S) + (n+n')\Delta^S - \Delta^W - \Delta^S \quad , \end{aligned} \quad (\text{D13})$$

and

$$\begin{aligned} \Xi^{SW}(n, n', h, h', \tau) &= h(1-n) + h'n' + \bar{g}^{SW}(n, n') - \frac{\tau}{R} \exp \bar{g}^{SW}(n, n') \\ \bar{g}^{SW}(n, n') &= nn'(\Delta^S - \Delta^W) - n\Delta^S + n'\Delta^W \quad . \end{aligned} \quad (\text{D14})$$

## APPENDIX E: LARGE $R$ ASYMPTOTIC

A saddle-point calculation of the incomplete Gamma function (29) gives the following large  $R$  asymptotic for  $z \neq 1$ ,

$$\gamma(R, Rz) \simeq \theta(1-z) + \frac{\exp[-R(z-1-\ln z)]}{\sqrt{2\pi R}(z-1)} \quad (\text{E1})$$

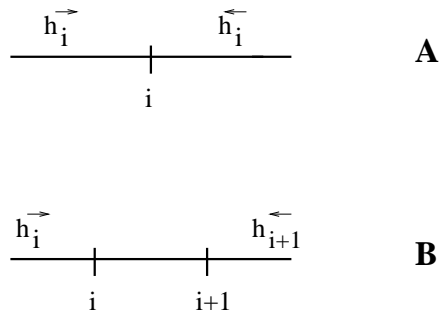


FIG. 28: Symbolic representation of the calculation of the distribution of likelihood at site  $i$  (**A**), and the joint distribution at sites  $i$  and  $i + 1$  (**B**). The left part of the sequence induces a left-to-right likelihood  $h_i^{\rightarrow}$  on base  $i$ , while the right part contribution is  $h_i^{\leftarrow}$  (**A**) or  $h_{i+1}^{\leftarrow}$  (**B**).

where  $\theta$  is the Heaviside function:  $\theta(1 - z) = 1$  if  $z < 1$ , 0 if  $z > 1$ . Application of this formula to the error (28) in the no-stacking case yields the large  $R$  scaling of  $\epsilon$  in (30).

Consider now the case of stacking interactions between neighboring bases. We first calculate the cumulative distribution  $\hat{Q}$  (46) of likelihoods in the  $R \rightarrow \infty$  limit, then derive finite  $R$  corrections. With definitions (47,48) we obtain, in the infinite  $R$  limit,

$$A(h) \rightarrow \theta(h^S - h) , \quad B(h) \rightarrow \theta(h - h^W) - \theta(h - h^S) \quad (\text{E2})$$

where

$$h^W = \Delta^W - x(e^{\Delta^W} - 1) , \quad h^S = \Delta^S - x(1 - e^{-\Delta^S}) . \quad (\text{E3})$$

For repeated sequences of, respectively, bases  $W$  and  $S$ , we have  $x = e^{-\Delta^W}$  and  $x = e^{\Delta^S}$ . It is a simple check that, whatever the value of  $b$ ,  $h^W$  and  $h^S$  have the same sign (positive for the  $WW$  sequence, negative for the  $SS$  sequence). Thus the product  $B(h)B(-h)$  in (46) vanishes. We find that the cumulative distribution  $\hat{Q}(h)$  of fields is a step function. More precisely,

$$Q(h) \rightarrow \delta(h - h_\infty^b) \quad \text{where} \quad h_\infty^W = \Delta^W - 1 + e^{-\Delta^W} , \quad h_\infty^S = -\Delta^S + 1 + e^{-\Delta^S} , \quad (\text{E4})$$

from which we deduce that the error in predicting a base vanishes in the large  $R$  limit. The case of the alternate sequence  $SW$  is more complicated. Setting  $x = 1$  in (E3) we have  $h^W < 0$  and  $h^S > 0$ . Using (E2) and (46) we merely obtain  $\hat{Q}(h) = 1$  for  $h < h^W$ ,  $\hat{Q}(h) = 0$  for  $h > -h^W$  and

$$\hat{Q}(h) = 1 - \hat{Q}(-h) \quad (h^W < h < -h^W) . \quad (\text{E5})$$

Though (E5) is not sufficient to characterize the likelihood distribution it allows us to calculate the error from (52), with the result (57).

Let us now calculate the corrections to the infinite  $R$  limit. The calculation of the error  $\epsilon$  is similar for  $WW$  and  $SS$  sequences, and is thus reproduced below in the  $WW$  case only. Let us introduce

$$\alpha(h) = \frac{\Delta^S - h}{x(1 - e^{-\Delta^S})} , \quad \beta(h) = \max\left(0, \frac{\Delta^W \pm h}{x(e^{\Delta^W} - 1)}\right) \quad (x = e^{-\Delta^W}) . \quad (\text{E6})$$

Using the large  $R$  expansion (E1) for the functions  $A$  and  $B$  in (47) we obtain from (46) the asymptotic expression for the cumulative distribution of loglikelihoods

$$\hat{Q}(h) = 1 - \frac{\exp[-R(\beta(h) - 1 - \ln \beta(h))]}{\sqrt{2\pi R}(\beta(h) - 1)} \quad (\text{E7})$$

and, through differentiation with respect to  $h$ ,

$$Q(h) = \sqrt{\frac{R}{2\pi}} \frac{\beta(h)}{1 - e^{-\Delta^W}} \exp[-R(\beta(h) - 1 - \ln \beta(h))] . \quad (\text{E8})$$

These expressions hold when  $\beta(h) < \alpha(h)$ . This condition happens to be fulfilled for the choice of parameters of Section IV B, and in the vicinity of  $h = 0$ . From (50) we have

$$\begin{aligned} \epsilon^{WW} &= \int_{-\Delta^S}^{\Delta^S} dh Q(h) [1 - \hat{Q}(-h)] \\ &= \frac{\exp[-R(\beta(0) - 1 - \ln \beta(0))]}{2\pi(1 - e^{-\Delta^W})} \int_{-\Delta^S}^{\Delta^S} dh \frac{\beta(h)}{\beta(-h) - 1} \exp \left[ R \ln \left( 1 - \left( \frac{h}{\Delta^W} \right)^2 \right) \right]. \end{aligned} \quad (\text{E9})$$

The dominant contribution to the integral comes from the  $h \simeq 0$  region. Expanding the integrand to the second order in  $h$  and calculating the Gaussian integral we obtain expression (54).

Finally we consider the case of finite temperature prediction of Section (IV A 2) for a base  $b$  ( $S$  or  $W$ ), in the absence of stacking. Let  $\Delta$  be the difference of free-energy between the two base types, and  $\tau$  given by (37). Integrating (33) by part and performing the change of variable  $\tau = R(x + \Delta)/(e^\Delta - 1)$ , we obtain the following expression for the error,

$$\epsilon_R(T = 1) = \int_0^\infty dx \left[ 1 - \gamma \left( R, \frac{R(x + \Delta^S)}{e^{\Delta^S} - 1} \right) \right] \frac{R e^{-Rx}}{(1 + e^{-Rx})^2} \quad (\text{E10})$$

$$= \sqrt{\frac{R}{2\pi}} \int_0^\infty dx \frac{\exp[-RG(x)]}{1 - (\Delta^S + x)/(e^{\Delta^S} - 1)} \quad (\text{E11})$$

where we have made use of (E1) to obtain (E11) from (E10), and have defined

$$G(x) = \frac{\Delta^S + x}{e^{\Delta^S} - 1} - 1 - \ln \left( \frac{\Delta^S + x}{e^{\Delta^S} - 1} \right) + |x|. \quad (\text{E12})$$

The maximal contribution to the integral comes from the  $x = 0$  region, with  $G(0) = \tau - 1 - \ln \tau$ . Defining  $\tilde{x} = Rx$  and expanding  $G$  around  $x = 0$  to the first order, we obtain

$$\epsilon_R(T = 1) = \frac{e^{-R(\tau - 1 - \ln \tau)}}{\sqrt{2\pi R} (1 - \tau)} \int_{-\infty}^\infty d\tilde{x} \frac{e^{-\tilde{x}(1 - \sigma)}}{(1 + e^{-\tilde{x}})^2} = \frac{e^{-R(\tau - 1 - \ln \tau)}}{\sqrt{2\pi R} (1 - \tau)} \frac{\pi \sigma}{\sin(\pi \sigma)} \quad (\text{E13})$$

where  $\sigma = |G'(0)|$  is given by (37).

## APPENDIX F: CALCULATION OF THE ESCAPE PROBABILITY $E_i$

In this appendix we calculate the escape probability  $E_i$  that the fork moves away from base pair  $i$  (never reaches it back) after its first visit. Assume the fork starts its motion from base  $j$ . We define  $p_j^{(i)}$  as the probability that the fork will never reach position  $i$  at any future instant. This probability is larger than zero when  $i < j$  since the walk is transient. Given the bp index  $i$  the probabilities  $p_j^{(i)}$ 's fulfill the recursion relation

$$p_j^{(i)} = q_j p_{j-1}^{(i)} + (1 - q_j) p_{j+1}^{(i)} \quad (\text{F1})$$

where, in analogy with definition (65) for a repeated sequence,

$$q_j = \frac{e^{g_s(f)}}{e^{g_s(f)} + e^{g_0(b_j, b_{j+1})}} \quad (\text{F2})$$

is the probability that the next base visited by the fork in  $j$  is  $j - 1$ . Equation (F1) is complemented by the boundary  $p_i^{(i)} = 0$  and  $p_N^{(i)} = 1$ . Mathematically speaking the probability of not reaching  $i$  from  $N$  is not equal to unity since a random walk on a finite sequence is recurrent. However this approximation is quantitatively excellent for the long sequences considered here. Defining

$$E_j^{(i)} = \frac{p_j^{(i)}}{p_{j+1}^{(i)}} \quad (\text{F3})$$

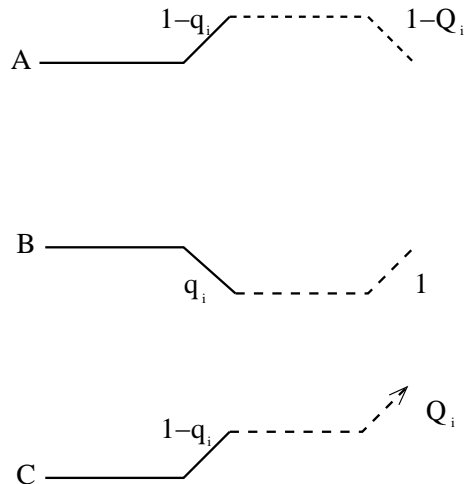


FIG. 29: Patterns A, B, and C present in the transition trace around base pair  $i$ . See text for definition.

we obtain the Riccati recursion relation

$$E_j^{(i)} = 0; \quad E_{j+1}^{(i)} = \frac{1 - q_{j+1}}{1 - q_{j+1} E_j^{(i)}} \quad \text{for } j \geq i. \quad (\text{F4})$$

We have solved equation (F4) numerically for the  $\lambda$ -phage sequence. The escape probability from  $i$  is then obtained from (F3) and (67),

$$E_i = \frac{1}{p_{i+1}^{(i)}} = \prod_{j \geq i+1} E_j^{(i)}. \quad (\text{F5})$$

## APPENDIX G: AVERAGE ERROR $\epsilon$ AT FINITE FORCE

### 1. Case of one-way unzippings

In this appendix we calculate the average prediction error after one unzipping over the distribution of the unzipping time traces. For a given time-trace, the prediction error depends only on the observed set  $\{t_i, u_i, d_i\}$  of times  $t_i$  spent on each base, and numbers of opening ( $u_i$ ) or closing ( $d_i$ ) of each base. To make the average we have to calculate the distribution  $P_1(\{t_i, u_i, d_i\})$  of such sets on all the time traces.  $P_1(\{t_i, u_i, d_i\})$  is therefore the product of the probability to observe a set of  $\{t_i, u_i, d_i\}$  in a given time trace (given in equation 5) time the multiplicity of such a set  $\{t_i, u_i, d_i\}$  on all the possible time traces.

Let us start by calculate the distribution  $P_1(u_i, d_i)$  ignoring for a while the time  $t_i$  spent on this base. Let us focus on base  $i$ ; the sequence of opening and closing transitions around this base, hereafter referred to as transition trace, can be decomposed into three kinds of elementary patterns schematized in Fig 29, and labeled with letters A, B and C:

- Pattern A (Fig 29A) corresponds to staying on base  $i$  for some time, moving forward ( $i \rightarrow i + 1$ , probability  $1 - q_i$ ), then coming back to  $i$  after a random walk throughout the upper part of the sequence ( $i + 1 \dots N$ ) with probability  $1 - E_i$ . The probability of pattern A is thus  $P_A = (1 - q_i) \times (1 - E_i)$ .
- Pattern B (Fig 29B) corresponds to staying on base  $i$  for some time, moving backward ( $i \rightarrow i - 1$ , probability  $q_i$ ), then coming back to  $i$  after a random walk throughout the lower part of the sequence ( $i + 1 \dots N$ ) with probability 1. The probability of pattern A is thus  $P_B = q_i$ .
- Finally, pattern C (Fig 29C) corresponds to staying on base  $i$  for some time, moving forward ( $i \rightarrow i + 1$ , probability  $1 - q_i$ ), without ever coming back to this base later on (probability  $E_i$ ). This final pattern has probability  $P_C = (1 - q_i) \times E_i$ .

The number of closing transitions in a transition trace,  $d_i$ , is simply equal to the number of B patterns around base  $i$ . Similarly, the number of opening transitions,  $u_i$ , is the sum of the numbers  $N_A$  and  $N_C$  of A and C patterns respectively. As  $N_C = 1$  by definition, we have  $N_A = u_i - 1$ . A and B patterns can be randomly located in the transition trace and are followed by one C pattern, the distribution  $P_1(u_i, d_i)$  on the ensemble of transition trace is therefore:

$$P_1(u_i, d_i) = \binom{u_i - 1 + d_i}{d_i} q_i^{d_i} [(1 - q_i)(1 - E_i)]^{u_i - 1} E_i (1 - q_i). \quad (u_i \geq 1, d_i \geq 0). \quad (\text{G1})$$

Let us now focus on the total time  $t_i$  spent on base  $i$ . It is the sum of  $u_i + d_i$  times each exponentially distributed with average sojourn time

$$\langle t_i \rangle = \frac{1}{r(e^{g_0(b_i^L, b_{i+1}^L)} + e^{g_s(f)})} \quad (\text{G2})$$

Thus,  $\tau_i = t_i / \langle t_i \rangle$  is a stochastic variable obeying distribution  $P_R$  (27) where  $R = u_i + d_i$  plays the role of a fictitious number of unzippings. We obtain the joint probability of time  $\tau_i$ , opening and closing moves  $u_i$  and  $d_i$ ,

$$P_1(\tau_i, u_i, d_i) = \frac{q_i^{d_i} [(1 - q_i)(1 - E_i)]^{u_i - 1} E_i (1 - q_i) e^{-\tau_i} \tau_i^{d_i + u_i - 1}}{d_i! (u_i - 1)!}. \quad (\text{G3})$$

Summation over all values for  $d_i$  lead to the (single base) probability for unzipping data

$$P_1(\tau_i, u_i) = \frac{E_i (1 - q_i)}{(u_i - 1)!} [(1 - q_i)(1 - E_i) \tau_i]^{u_i - 1} e^{-\tau_i (1 - q_i)}. \quad (\text{G4})$$

Neglecting stacking effects between bases, the content  $b_i$  of base  $i$  is chosen to maximize the probability

$$P(b_i | \tau_i, u_i) = \frac{\exp\left(g_0(b_i) u_i - r e^{g_0(b_i)} \langle t_i \rangle \tau_i\right)}{\exp\left(g_0(W) u_i - r e^{g_0(W)} \langle t_i \rangle \tau_i\right) + \exp\left(g_0(S) u_i - r e^{g_0(S)} \langle t_i \rangle \tau_i\right)}, \quad (\text{G5})$$

where the average sojourn time is given by eqn (G2). This maximization can be done along the lines of Section IV A 1 devoted to the case of infinite force. We find the average fraction of mispredicted bases at force  $f$ ,

$$\epsilon_{f,1}^W = \sum_{u_i \geq 1} \int_0^{\frac{u_i \tau_i^W}{1 - q_i}} d\tau_i P_1(\tau_i, u_i) \quad , \quad \epsilon_{f,1}^S = \sum_{u_i \geq 1} \int_{\frac{u_i \tau_i^S}{1 - q_i}}^{\infty} d\tau_i P_1(\tau_i, u_i), \quad (\text{G6})$$

with definition (G4) for  $P_1$ . Hence eqn (69).

## 2. Case of two-way unzippings

We now suppose that the sequence is opened in both ways, and denote by  $\sigma = +$  the left-to-right and  $\sigma = -$  the right-to-left openings respectively. Let  $u_i^\sigma, \tau_i^\sigma$  denote the number of openings of bp  $i$  and the time spent by the fork on  $i$  for both directions ( $\sigma = \pm$ ). The joint distribution of  $u_i^\sigma, \tau_i^\sigma$  is  $P_1$  (G4) with  $q_i, E_i$  replaced with, respectively,  $q_i^\sigma$ , the probability to close back bp  $i$  when the fork is in  $i$ , and  $E_i^\sigma$ , the escape probability from base  $i$  in the  $\sigma$  direction.  $q_i^+$  and  $E_i^+$  are simply given by (F2) and (F5) respectively. In addition  $q_i^- = q_{N-i+1}^+$ , and  $E_i^-$  can be obtained along the lines of Appendix F.

As the unzippings in both directions produce statistically uncorrelated data the joint distributions of  $u_i^+, \tau_i^+$  and  $u_i^-, \tau_i^-$  factorize. The Bayesian probability that base  $i$  is of type  $b_i$  is simply given by (G5) with  $u_i = u_i^+ + u_i^-$ ,  $\tau_i = \tau_i^+ + \tau_i^-$ . In the framework of Maximum Likelihood Prediction we maximize this quantity to obtain the error on base  $i$ ,

$$\epsilon_{f,1}^{b_i} = \sum_{u_i^+, u_i^- \geq 1} \rho_1(u_i^+) \rho_1(u_i^-) \epsilon_{u_i^+, u_i^-}^{b_i}, \quad (\text{G7})$$

where

$$\begin{aligned}\epsilon_{u_i^+, u_i^-}^W &= \int_0^{+\infty} dx dy \theta(x + y - \tau^w(\hat{u}_i^+ + \hat{u}_i^-)) \frac{e^{-x} x^{(u_i^+ - 1)}}{(u_i^+ - 1)!} \frac{e^{-y} y^{(u_i^- - 1)}}{(u_i^- - 1)!} \\ \epsilon_{u_i^+, u_i^-}^S &= \int_0^{+\infty} dx dy \theta(\tau^s(\hat{u}_i^+ + \hat{u}_i^-) - x - y) \frac{e^{-x} x^{(u_i^+ - 1)}}{(u_i^+ - 1)!} \frac{e^{-y} y^{(u_i^- - 1)}}{(u_i^- - 1)!}\end{aligned}\quad (\text{G8})$$

and  $\rho_1$  is defined in (66) (beware of the dependence of  $E_i^\sigma$  on the unzipping direction  $\sigma$ ).

The generalization to the case of  $R/2$  unzippings in each direction is done along the lines of Section G 1, with the immediate result

$$\epsilon_{f,R}^{b_i} = \sum_{u_i^+, u_i^-} \rho_{R/2}(u_i^+) \rho_{R/2}(u_i^-) \epsilon_{u_i^+, u_i^-}^{b_i}, \quad (\text{G9})$$

where  $\rho_{R/2}$  is the  $(R/2)^{th}$  convolution power of the probability  $\rho_1$ , see eqn (68).

## APPENDIX H: CALCULATION OF $R_c$ IN PRESENCE OF DNA STRANDS FLUCTUATIONS

Let  $\hat{T}_i^r$  be the number of measures where the fork is really at location  $i = 0, 1, \dots, N$ . These integer numbers are stochastic and distributed according to, given the sequence  $B$ ,

$$\text{Proba}[\{\hat{T}_i^r\}|B] = \prod_i e^{-\Delta t r_o(b_i) \hat{T}_i^r} \left(1 - e^{-\Delta t r_o(b_i)}\right). \quad (\text{H1})$$

The number of times the fork is apparently at position  $j$ ,  $\hat{T}_j^a$ , given the set of  $\hat{T}_i^r$ , is stochastic too. Their probability is given by

$$\text{Proba}[\{\hat{T}_j^a\}|\{\hat{T}_i^r\}] = \sum_{\{f_{ij}=0,1,2,\dots\}} \prod_i \left\{ \frac{\hat{T}_i^r!}{\prod_j f_{ij}!} \prod_j [A_{j,i}]^{f_{ij}} \delta(\hat{T}_j^a, \sum_i f_{ij}) \right\} \delta(\hat{T}_i^r, \sum_j f_{ij}) \quad (\text{H2})$$

where  $\delta(a, b) = 1$  if  $a = b$ , 0 otherwise is the Kronecker function, and the fluctuation matrix  $A$  is defined in (89). It is convenient to work with the generating function of the  $\{\hat{T}_i^a\}$ ,

$$G_1(\{y_j\}|B) = \sum_{\{\hat{T}_j^a\}, \{\hat{T}_i^r\}} \text{Proba}[\{\hat{T}_j^a\}|\{\hat{T}_i^r\}] \text{Proba}[\{\hat{T}_i^r\}|B] \prod_j e^{y_j \hat{T}_j^a} = \prod_i \left( \frac{1 - e^{-\Delta t r_o(b_i)}}{1 - e^{-\Delta t r_o(b_i)} \sum_j A_{j,i} e^{y_j}} \right). \quad (\text{H3})$$

The generating function of the probability distribution of the apparent times  $t_j^a = \hat{T}_j^a \times \Delta t$  is simply  $G_1(\{y_j \Delta t\}|B)$ .

The above expression for  $G_1$  holds for one unzipping. For  $R$  unzippings the generating function  $G_R$  is simply given by the  $R^{th}$  power of  $G_1$ . In the large  $R$  limit we obtain

$$G_R(\{y_j \Delta t\}|B) = \exp \left[ -R \sum_i \ln(1 + \chi_i(\{y_j\})) \right] \quad (\text{H4})$$

where, to the first order in  $\Delta t$ ,

$$\chi_i(\{y_j\}) = \sum_j A_{j,i} y_j \left( \frac{\Delta t}{2} - r_o(b_i)^{-1} \right) - \frac{\Delta t}{2} \sum_j A_{j,i} y_j^2 r_o(b_i)^{-1}. \quad (\text{H5})$$

Assume now that the true sequence  $B^L$  is a repeated sequence of  $S$  bases with a  $W$  base at location  $n$ ; we call  $B^S$  the sequence made of  $S$  bases only. We furthermore assume that the free energy difference  $\Delta$  is small which makes inference harder. Using  $\rho$  defined in (91) and introducing  $s_j = y_j/r_o(S)$ , we obtain

$$G_R(\{s_j\}|B^L) = \exp [R \gamma(\{s_j\}|B^L)] \quad (\text{H6})$$

where

$$\gamma(\{s_j\}|B^L) = -\sum_j s_j h_j(B^L) - \frac{1}{2} \sum_{j,k} s_j \beta_{j,k} s_k + O(s_j^3) \quad \text{with} \quad h_j(B^L) = 1 - \frac{\rho}{2} + \Delta A_{j,n} \quad (\text{H7})$$

and matrix  $\beta$  defined in (92). Notice that the expressions for  $h$  and  $\beta$  were obtained using the approximation  $\sum_i A_{i,k} = 1$  for any  $k$ , and in the limit of small  $\rho, \Delta$ . The expression for  $\gamma(\{s_j\}|B^S)$  is obtained from (H7) when  $\Delta \rightarrow 0$ .

We obtain the large deviation expression for the distribution  $P_R$  of the apparent times through the Legendre transform of  $\gamma$ ,

$$P_R(\{t_j^a = \tau_j^a/r_o(S)\}|B) = \exp(-R \omega(\{\tau_i^a\}|B)) \quad \text{with} \quad \omega(\{\tau_i^a\}|B) = -\max_{\{s_j\}} \left[ \gamma(\{s_j\}|B) + \sum_j s_j \tau_j \right] \quad (\text{H8})$$

for the two sequences  $B = B^L, B^S$ . When  $\Delta$  is small we expect the distribution of apparent times for the two sequences to be very close and thus the set of times  $\{\tau_j^a\}^*$  for which they are equal will be close to the most likely apparent times with both distribution. This justifies the second order expansion in  $s$  in (H7). The exponent  $\omega^* = \omega(\{\tau_i^a\}^*|B^L) = \omega(\{\tau_i^a\}^*|B^S)$  of the probability of this crossing time  $\{\tau_j^a\}^*$  is equal, in the large  $R$  limit, to the inverse of  $R_c(n)$ . This statement can be graphically understood from the Figure 2 in [34]. A more detailed explanation will be given in [42]. The calculation of  $\omega^*$  is immediate from (H8) and leads to (92). For  $\rho = 0$  the value for  $R_c(n)$  coincide with its expression (32) in the absence of ssDNA fluctuation.

We now turn to the analysis of the Viterbi algorithm in the limit  $\Delta t = 0$ . The Laplace transform of the probability distribution  $P_R^{(i)}$  of the deconvoluted time  $\tau_i^d = t_i^d r_o(W)$  on base  $i$  is obtained from  $G_R$  by applying the deconvolution kernel  $D$ , with the result

$$\int_0^\infty dt_i^d P_R^{(i)}(t_i^d) e^{-y_i t_i^d} = \prod_{j=0}^N \frac{1}{(1 + y_i C_{ij})^R} \quad (\text{H9})$$

where  $C_{ij}$  is defined in (95). The error in predicting base  $i$  is then given by the integral of  $P_R^{(i)}$  over  $\tau_i^d > \tau^W$  since  $b_i^L = W$ , see (23,28),

$$\epsilon_{R,n}^W = \int_{\frac{R\Delta}{1-e^{-\Delta}}}^\infty dt_i^d P_R^{(i)}(t_i^d) = \int_0^\infty \frac{dx}{R} \int_{-\infty}^{+\infty} \frac{ds}{2i\pi} e^{Rf(x,s)} \quad (\text{H10})$$

where

$$f(x, s) = \left( x + \frac{\Delta}{1 - e^{-\Delta}} \right) s - \sum_i \ln(1 + s C_{n,i}) . \quad (\text{H11})$$

The result for  $R = 1$  unzipping is given by (94). In the large  $R$  limit we obtain expression (96) through a saddle-point calculation and a small  $s$  expansion (valid for small  $\Delta$ ). The saddle-point value for  $x$  can be located in 0, or in a strictly positive real value. This corresponds to the two cases listed in (96).

## APPENDIX I: ON SAMPLING AND LARGE DEVIATIONS OF THE ERROR $\epsilon$

We have calculated in Section (IV D 2) the average fraction of mispredicted bases within the hypothesis of exact sampling of the distribution of the number  $u_i$  of openings. Let us turn to the more realistic case of a finite number of samples,  $M$ . As  $M$  decreases, the values of  $u_i$  with exponentially small-in- $R$  probabilities are less and less likely to be sampled, leading to large deviations corrections. Let us fix on a base pair dropping the base index  $i$  to shorten the notation. The values of  $u$  which can be found in a sample of size  $M$  are the ones such that

$$\rho_R(u) \times M \gg 1 , \quad (\text{I1})$$

where  $\rho_R$  is given in eqn (68). Assume that we keep fix  $R$  and scale the number of samples according to  $M \sim e^{R\mu}$ . Upon introduction of the rate function for  $\hat{u} = u/R$ ,

$$\omega(\hat{u}) = \lim_{R \rightarrow \infty} \frac{1}{R} \ln \rho_R(R\hat{u}) = (1 - \hat{u}) \ln \left( \frac{\hat{u} - 1}{\langle \hat{u} \rangle - 1} \right) + \hat{u} \ln \left( \frac{\hat{u}}{\langle \hat{u} \rangle} \right) , \quad (\text{I2})$$

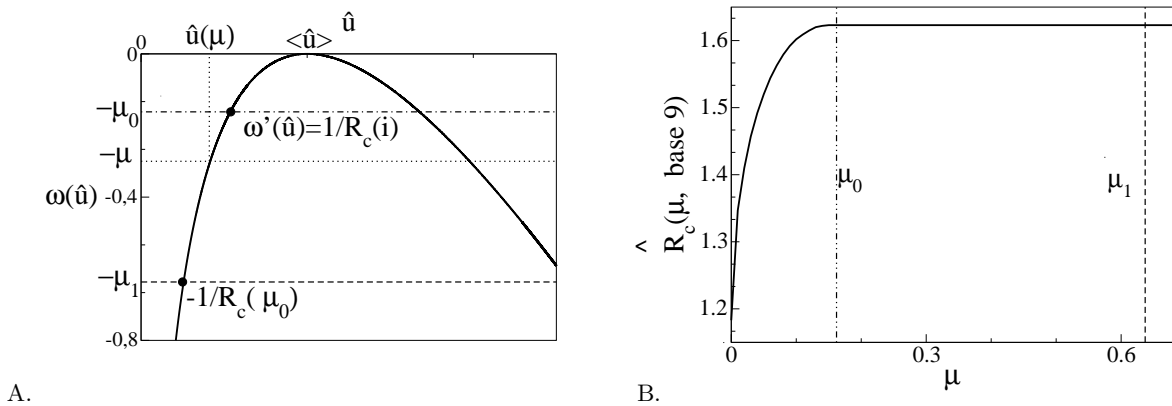


FIG. 30: **A.** rate function  $\omega(\hat{u})$  governing the large deviations of the number  $\hat{u}$  of openings of a base per unzipping.  $\omega$  vanishes when  $\hat{u}$  equals its average value,  $\langle \hat{u} \rangle$ , and is strictly negative otherwise. **B.**  $R_c$  vs. logarithm of the number of samples,  $\mu = \ln M/R$ , for the 9<sup>th</sup> base of the  $\lambda$ -phage sequence.

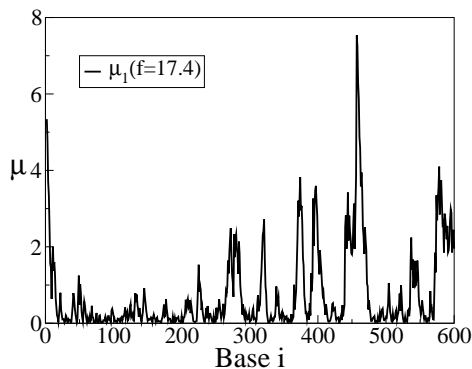


FIG. 31: Logarithm  $\mu_1$  of the number of samples (divided by  $R$ ) to obtain a good estimate of  $R_c(i)$  vs base pair index  $i$ .  $\mu_1$  strongly depends on the base index  $i$  e.g. we need to sample over  $M \sim e^{8 \times R}$  to accurately estimate  $R_c$  for all bases. The force is  $f = 17.4$  pN.

we rewrite condition (I1) into

$$\omega(\hat{u}) \geq -\mu, \quad (\text{I3})$$

This condition is graphically solved in Fig 30A. At fixed  $\mu$  a compact range of available values for  $\hat{u}$  is obtained, centered around the average number  $\langle \hat{u} \rangle$  of openings of a bp per unzipping. For instance, the smallest accessible value,  $\hat{u}(\mu)$ , is obtained when solving condition (I3) as an equality (Fig. 30A).

For each sample  $m = 1, \dots, M$  the measured error  $\epsilon_R^m$  takes value  $v = 0$  (if the base is correctly predicted) and 1 otherwise, with probabilities

$$P_v = \int_{\omega(\hat{u}) \geq -\mu} d\hat{u} e^{R \omega(\hat{u})} \left[ \left(1 - e^{-R\hat{u}/R_c}\right) \delta_{v,0} + e^{-R\hat{u}/R_c} \delta_{v,1} \right]. \quad (\text{I4})$$

We evaluate this probability through a saddle-point approximation,

$$P_v = \left(1 - e^{-R/R_c(\mu)}\right) \delta_{v,0} + e^{-R/R_c(\mu)} \delta_{v,1}, \quad (\text{I5})$$

where

$$R_c(\mu) = \max_{\hat{u} \geq \hat{u}(\mu)} \left[ \frac{R_c}{\hat{u} - R_c \omega(\hat{u})} \right]. \quad (\text{I6})$$

Let us call  $\mu_0 = \omega(\hat{u}_0)$  where  $\hat{u}_0$  is the root of  $\omega'(\hat{u}_0) = \frac{1}{R_c}$ , and  $\mu_1 = \frac{1}{R_c(\mu_0)} = \mu_0 + \frac{\hat{u}(\mu_0)}{R_c} > \mu_0$ . As  $\omega$  depends on the bp  $i$  so do  $\mu_0, \mu_1$ . Then,

- when  $\mu < \mu_0$  the maximum on the r.h.s. of (I6) is reached in  $\hat{u}(\mu)$  fulfilling the equality (I3), and is an increasing function of  $\mu$  (Fig 30B).
- when  $\mu_0 \leq \mu \leq \mu_1$   $R_c(\mu) = R_c(\mu_0) = \hat{R}_c$  does not depend on  $\mu$  anymore (Fig 30B). The average number of erroneous samples reads

$$M_{err} = M e^{-R/R_c(\mu)} = e^{R(\mu-1/R_c(\mu_0))} \quad (I7)$$

and is exponentially small in  $R$  by the very definition  $\mu_1$ , Hence no erroneous sample is detected and no estimate of  $R_c$  can be made.

- when  $\mu > \mu_1$   $M_{err}$  is exponentially large (I7), and the decay constant of the error can be safely measured and estimated to be  $R_c(\mu_0)$ .

Figure 31 shows  $\mu_1$ , the logarithm (divided by  $R$ ) of the number of samples needed to accurately estimate  $R_c$ , as a function of the base index  $i$ . We observe that  $\mu_1$  varies a lot from base to base.