The mechanical opening of DNA and the sequence content

S. Cocco [1], R. Monasson [2]
[1] *CNRS-Laboratoire de Physique Statistique de l'ENS, 24 rue Lhomond, 75005 Paris, France*
[2] *CNRS-Laboratoire de Physique Théorique de l'ENS, 24 rue Lhomond, 75005 Paris, France*

Separation of the strands of the DNA molecule play a fundamental role in biology. We here review different experimental and theoretical approaches to DNA opening in vitro. We then analyze the relationship between the sequence content and the opening signal. We focus on a theoretical study of the performances of Bayesian inference to predict the sequence of DNA molecules from fixed-force unzipping experiments.

I. INTRODUCTION

In all living cells long DNA molecules carry genetic information in their base-sequence. This information is read and processed to make proteins during transcription, to duplicate DNA during replication before cellular division and to repair DNA when it is damaged [1–3].

The DNA double helix is made of two polymeric chains, wound around one another to form a regular, right-handed helix. Each chain is a series of chemical units called nucleotides, joined together by single covalent bonds. The four types of nucleotides - containing the 'bases' adenine (A), thymine (T), guanine (G) and cytosine (C) - can be chained together in any order; the genetic information is carried precisely by the sequence of bases on a chain. The double-helix structure is only stable if the two chains carry complementary sequences, meaning juxtaposed A-T and G-C base pairs. The double helices found in cells obey this constraint (apart from damaged nucleotide) and thus can be reconstructed from only one strand, simply by synthesis of a complementary strand.

The two oppositely directed chains are bound together by hydrogen bonds and other physical interactions which generate only about k_BT of net cohesive free energy per base pair under conditions found in the cell (aqueous solution at room temperature with about 0.1 M salt and pH near 7.5). This means that double helices of more than about 30 base pairs in length are stable (they require thermal fluctuations $> 30k_BT$ to fall apart, which are rare enough to be ignored). At the same time, the two strands can easily taken apart without breaking the covalent backbone bonds. This strand separations takes place, indeed, if the solution condition are changed by increasing the temperature, increasing the pH, or decreasing the ionic condition [2].

The two DNA strands can also be separated mechanically. Micromechanical experiments allows one to gradually separate, or unzip, the two strands of long DNAs by the application of a force [4–8]. Finally strand separation can be also performed by enzymes, as is done in the cell; single molecule experiments follow a single enzyme working on DNA. [9–14].

Hereafter we will review some DNA strands separation or hybridization by in vitro experiments, from denaturation experiments (section II A) to single molecule experiments (section II B) . Among single molecule experiments [11] we will review the detection of single strand DNA hybridization [16], the DNA strand separation under translocation through a nano-pore [17, 18], the observation of the sequence-dependent activity of an exo-nuclease [9, 10], and the unzipping of DNA under a mechanical action at a constant velocity [4, 5, 7] or force [6, 8].

We will in particularly focus on how information on the DNA sequence can be extracted from experimental data. The work to separate the DNA strands depends indeed on the sequence, not only because the base pairs GC are most stable than the base pairs AT but also because there are base-dependent stacking interactions between neighboring base pairs in the double helical structure. The knowledge of the exact base content of DNA's sequence is of huge importance from both biological and medical points of view. Biochemical sequencing methods, described in section III, are very powerful and have been recently massively used for sequencing whole genomes, in particular the human one. Moreover looking for alternative way of sequencing DNA is an active field of research. In this regards DNA strand separation experiments on single molecules of DNA could be interesting. So far the single molecule experiments experiments have however not been combined with a concrete method for data analysis capable of extracting precise information on the sequence of DNA molecules. In section IV B we will review some theoretical studies on the prediction of a sequence from fixed force unzipping experiments.

	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: Pairing free energies $g_0(b, b')$ from the Mfold server [19] and Santa Lucia data [20] at temperature $T = 25^oC$, 150 mM NaCl and neutral pH. The columns refers to b', the lines to b in the $5' \to 3'$ direction along the molecule. The pairing free energies are given by the hydrogen bonds between the two bases of a same pair and stacking interactions between neighboring bases, see section IV A for more detail.

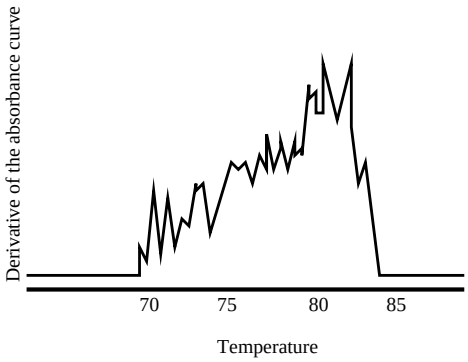

FIG. 1: Schematic representation of the differential melting curve for the λ phage sequence (see [2] for the exact plot)

II. STRAND SEPARATION OR HYBRIDIZATION

A. Thermal Denaturation of a DNA solution

The melting transition is, in classical biochemistry experiments, observed by measuring the optical properties of a solution containing DNA molecules, as the ultraviolet absorbance or the chromatic dichroism which are sensitive to the amount of double stranded regions in the molecule [2]. Alternatively DNA can be melted in calorimeters. In this set-up the melting transition is detected from the peaks in the heat capacity.

When raising the temperature the change in optical properties or in the heat capacity occur over a narrow temperature range (a few tenth of degrees), revealing that denaturation is a cooperative process. The melting temperature, defined as the temperature value at half the transition is then quite well defined. For a homogeneous sequence the melting transition is very sharp. The melting temperature ranges between 60^oC and 110^oC [2]; it depends on the sequence, on the length of the molecule and the chemical properties (pH, salt concentration) of the solution. For a fixed molecular length and chemical composition of the solution the base pairing free energy along the repeated sequence and the melting temperature are directly related. The pairing energies of the 16 possible combinations of two consecutive base pairs have been indeed precisely quantified from the experimental melting temperature, obtained in calorimeters, of artificial repeated sequences [19, 20] (see Table 1).

In the denaturation of heterogeneous sequences the melting temperature gives only information on the average sequence. The derivative of the absorbance melting curve as well as the peaks in the heat capacity give a more detailed information on the sequence. It reveals specific regions in the DNA that melt cooperatively over a narrow temperature range. AT-rich regions melt first, at lower temperature, and the GC-rich regions melt at higher temperature (see sketch in Fig. 1). Analysis of differential melting data can reveal specific characteristics of the sequence such as the presence of regions of special stability (rich in GC or in AT) or with special sensitivity to specific ions, or also the presence of mismatches. Melting data can thus be used as a rapid method to identify homology or differences among DNAs. However this method does not allow one to localize the identified regions into the DNA sequence.

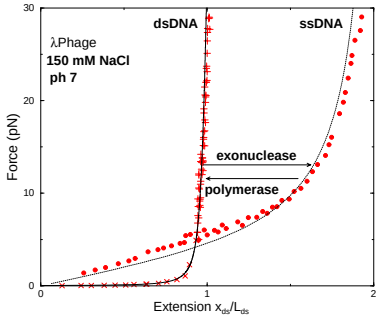

FIG. 2: Stretching curve of the λ phage DNA molecule, in its double strand conformation (dsDNA) and single strand conformation (ssDNA). L_{ds} is the crystallographic length of the ds-λ DNA. The ssDNA is shorter than the dsDNA for forces smaller than $f \simeq 5$ pN and longer for larger forces. The enzymes exonuclease convert dsDNA into ssDNA while the polymerases make the opposite.

B. Single molecule experiments

Strand separation and hybridization has recently been observed by mechanical micromanipulation of single DNA molecules [4–14]. By micromanipulation experiments a single molecule can be stretched by a force while measuring its extension. Experiments and theory have allowed to characterize the elasticity of a single strand DNA molecule (ssDNA) as well as the one of double strand DNA molecule (dsDNA) [21, 22]. The force extension curves at room temperature and standard ionic conditions are shown in Figure 2. At small forces the ssDNA and dsDNA look like random coils, but the ssDNA is more flexible than the dsDNA. Up to forces of 5 pN the ssDNA is indeed shorter than the dsDNA . On the contrary at forces larger than 5 pN both the ssDNA and the dsDNA are essentially straight up by the force; the length of the molecule is therefore the number of nucleotides times the distance between consecutive bases, of about 7 Å for the ssDNA backbone distance and 3.4 Å for the dsDNA axial distance. Therefore, in the range of forces of 5-60 pN, ssDNA has about twice the length of dsDNA.

The detection of a single DNA hybridization based on the different lengths of ssDNA and dsDNA has been experimentally performed by Singh-Zocchi and collaborators [16]. A small ssDNA fragment is attached between a glass slides and a bead that can be pulled by a force, while the complementary strand is put in the surrounding solution (Fig. 3). When the two strand hybridize in dsDNA the length of the molecule change. The change in length is detected by a fluorescence resonance energy transfer method. A second method that allows one to measure in a similar way the opening time of small DNAs is nanopore unzipping [17, 18]. In this experiment (Fig. 5) a DNA hairpin with a ssDNA tail at one extremity is first catched in a micropore by a voltage that is applied on the opposite extremity of the channel. Because the nanopore is very thin, the DNA hairpin has to open to pass through the nanopore. The presence of the DNA in the channel can be detected by the reduction of the current passing through the channel. The time that the DNA hairpin takes to open can be therefore measured. These experiments detect the overall hybridization or opening times which depend on the sequence, but, similarly to the thermal denaturation, allow one only to have a global information on the sequence.

Let us now describe method that give local information on the sequence content. A method recently developed is based on the single molecule detection of the polymerization or depolarization of a single DNA by enzymes. Enzymes called DNA polymerase synthesize dsDNA from ssDNA and a solution containing free nucleotides. Enzymes called exonucleases make the opposite: they remove one single strand from the double helix by cleaving one nucleotide after the other. Thanks to single molecule experiments, a single ssDNA or dsDNA molecule can be stretched by a micromanipulator and the activity of such enzymes can be monitored by the change in molecular length consequent to their action [11–13]. Interestingly, the experiments have shown that the rate of depolymerization by an exonuclease depends on the sequence. This rate dependence on the sequence demonstrate that enzymes feel the (different) binding energies of nucleotides while processing the sequence [9]. Moreover long pauses occur in correspondence with sequence-specific motif *eg.* GGCGA on the λ-phage sequence. These experiments therefore give the information on the sequence content and position on it. These methods are however, at first sight, quite involved to use as sequencing methods because enzyme are complicated and stochastic machines which for example undergo conformational changes during their activity. To obtain information on the sequence the sequence dependent fluctuations in the processing rate have to been separated from fluctuations independent on the sequence and attributed to conformational changes [9]. In the following we will focus on the unzipping of the DNA molecule by a force and not an enzyme.

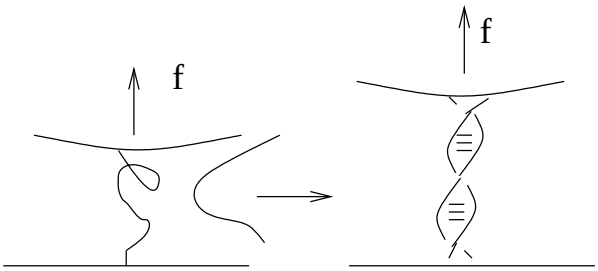

FIG. 3: Single molecule hybridization of a short DNA, attached from one hand to a bead and by the other hand to a surface stretched by a force $f < 5$ pN (at which the ssDNA is shorter than dsDNA as shown in Fig.2). The conformational change due to the hybridization with a free oligomers is optically detected [16].

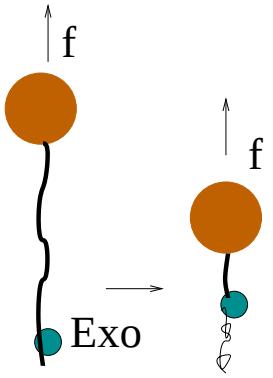

FIG. 4: The activity of an exonuclease on a single stretched DNA can be detected by the shortening of the molecule [9, 10] stretched by a force $f < 5$ pN.

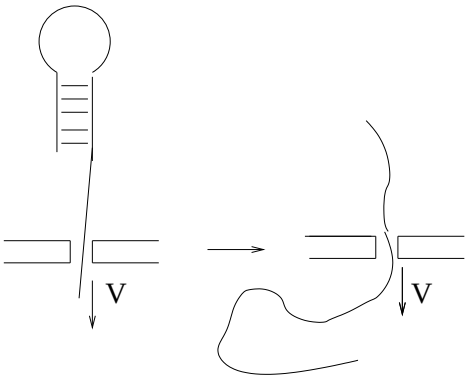

FIG. 5: Nanopore unzipping: a small dsDNA hairpin to which is attached a ssDNA link is "fished" into a channel in a membrane by applying a Voltage through the channel. The dsDNA cannot pass through the pore and therefore it goes through only after having been opened by thermal fluctuations.

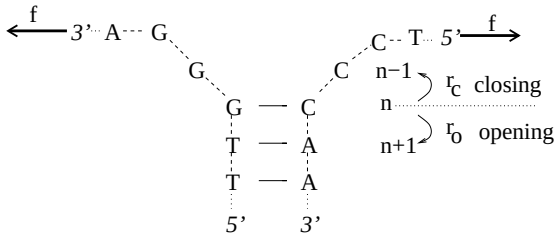

FIG. 6: Sketch of the fixed force opening: the fork in position n (number of open bases) move forward or backward with rates (probabilities per time) dependent on the free energy potential $G(n, f)$, see text and Figure 7.

III. DNA SEQUENCING METHODS

Natural DNA are usually too long to be sequenced directly. They are first fragmented into pieces of maximal length of about 1000 nucleotides which are then sequenced. The position of fragments within the large DNA are then reconstructed by mapping the regions with overlapping sequences between fragments. To be sequenced a DNA fragment is first cloned into many copies, usually by a PCR reaction. The DNA sequence of the fragments are then determined by chemical or enzymatic methods. The Maxam-Gilbert chemical method depends on base specific chemical reactions to break polynucleotide chains specifically at A,G, C or T. A ssDNA fragment is first labeled at either its 5'-or-3' end, then treated with a base specific reagent (*e.g.* specific for G), so that less than 1% of the G's react. Further chemical treatment causes strand breakage at each modified G. The breakage produces a set of end-labeled fragments whose lengths reveals the positions of G's within the sequence, the length can be determined with very high precision by gel electrophoresis. Many other fragments result but only the end-labeled ones are analyzed. Use of four base specific reactions in separate experiments provides the position of each base relative to the end label, *i.e.* the sequence. The principle of the enzymatic Sanger method is the same as that of Maxam and Gilbert: the length of an end-labeled oligonucleotide that terminates at one type of base reveals the position of that base. In the Sanger methods a complementary strands is synthesized. A small fraction of a single type of didexyonucleoside triphosphate *e.g.* didexyguanosine 5' triphosphate is used to cause chain termination during synthesis. Gel electrophoresis is then used to determine the chain lengths of the set of end labeled oligonucleotides.

These techniques are very powerful and reliable, with a 99.9% fraction of correctly predicted bases. However they are costly in terms of time and workforce. In addition some regions in the genome may be very difficult to clone, and obtaining information on their sequence is uneasy.

IV. UNZIPPING:EXPERIMENTS AND THEORY

In 1997, U. Bockelmann and F. Heslot demonstrated a new way to obtain information on the DNA base sequence, based on single molecule micromanipulation and force measurements [4]. The idea consists in unzipping the molecule by pulling apart the two complementary strands of the double helix (Fig. 6). If one opens the double helix by separating the extremities of the two single strands at constant velocity for instance, one can measure the force as a function of time, *i.e.* as a function of the number of opened base pairs. If one pulls slowly the unzipping force is of about 15 pN. The positive or negative excursions of the force around this average value reproduce the base content of the sequence, rich or poor in GC base pairs, respectively [4, 5]. The binding energy of the GC pair is higher than the one of the AT pair, hence a higher force is needed to break a GC pair.

With a more recent experimental setup Bockelmann and Heslot were able to observe the GC content of the sequence of long molecules (about 10,000 base pairs) with a sensitivity of about 10 DNA base pairs. Remarkably, a comparison between measured and calculated force signals (using the known DNA sequence) indicates that the signal could be sensitive to certain point mutations [5].

DNA unzipping appears as a mechanical way to read the sequence. Moreover the possibility to change the distance between the ssDNA extremities of the molecule, allows one to read the bases one after the other. Two information are collected together during unzipping: the distance between the ssDNA extremities of the molecule (Fig. 6) that is related to the position of the opening base and the force related to the type of the opening base.

Several limitations however do not allow one to directly read the sequence and lead to the feeling that sequencing by unzipping is not a achievable tool [23]. The first problem is that unzipping is stochastic: the work to open a base pair is only few k_BT so a base pair can open and close by thermal fluctuations. Two repetitions of an unzipping experiment give different data. The second problem is that unzipping traces do not give access directly to the position

of the opening fork $n(t)$, but to the distance $x(t)$ between the extremities of the molecule (Fig. 6). The unzipped single strands that link the extremities of the molecule and the opening fork are not rigid, but fluctuate. The value of $x(t)$ for a given $n(t)$ is therefore stochastic due to the thermal fluctuations of the unzipped single strands. Finally there are additional sources of noises from the experimental apparatus e.g. drift of the apparatus, and temporal and spatial resolutions.

Yet unzipping is a potentially powerful tool, since long molecules can be opened by unzipping in a fraction of seconds, and one single molecule can be opened and closed several times, to collect more opening signal. Our aim is to theoretically understand if extracting information on the sequence is possible despite this different sources of noise. We would like to design optimal protocols and methods of data analysis capable of providing us with the most accurate information possible about the sequence.

In the following we will focus on computer-generated data, for which the dynamical model and the source of noise are perfectly well defined.

To start with we have considered the fixed force unzipping [6, 8], in which a force is kept constant at the extremities of the molecule while the distance between them is measured.

As a first step we have focus only on the thermal movement of the fork position $n(t)$. In Section IV A we will introduce a simple model for the unzipping at constant force able to reproduce the experimental data. In Section IV B we will discuss the prediction of the sequence from unzipping data generated from this model by a Monte Carlo simulation.

Finally in conclusion we discuss how to tackle the more realistic case with several sources of noise at the end of the paper.

A. A model for fixed force unzipping

For artificial repeated sequences, the unzipping at fixed force f is easy to describe: the free energy difference $\Delta g(f)$ between two open bases and the close base pair is simply the difference between the elastic free energy $w_{ss}(f)$ of the two unzipped DNA bases submitted to a force f, and the binding energy g_0 of the base pair,

$$\Delta g(f) = 2w_{ss}(f) - g_0. \tag{1}$$

The critical force, at which the molecule unzip is therefore the value f^* such that $\Delta g(f^*) = 0$. The free energy $w_{ss}(f)$ is well known from ssDNA stretching experiments [21, 22]; it is the integral of the force-extension curve of Fig.2. The knowledge of f^* allows us to predict the binding free energy g_0 for repeated sequences [15]. Unzipping of repeated sequence has been performed by Rief *et al.* They have found $f^* = 20 \pm 3$ pN for poly(dG-dC) and $f^* = 9 \pm 3$ pN for poly(dA-dT). These forces give, $g_0 = 3.5\ k_BT$, $g_0 = 1.1\ k_bT$ respectively for the poly(dG-dC) and poly(dA-dT) sequence [22], in good agreement for the data of Table 1 obtained from thermal denaturation. For a heterogeneous sequence $\mathcal{S} = \{b_1 \dots b_N\}$, the free energy to unzip base pair i,

$$\Delta g(f, i) = 2w_{ss}(f) - g_0\,(b_i, b_{i+1}) \tag{2}$$

depends on the base and its neighbors due to stacking effects. The difference in free energy of the molecule when its first n bases are unzipped and when it is fully zipped is then

$$G(n, f) = \sum_{i=1,n} \Delta g(f, i). \tag{3}$$

The free energy landscape $G(n, f)$ for the first 450 bases of λ phage DNA is plotted in Fig. 7 C, at a force of 16.4 pN; the overall slope is negative, meaning that at this force the molecule opens. The free energy landscape is characterized by sequence-dependent barriers and minima between them. A simple modeling able to reproduce the unzipping dynamics (Fig. 2) describes the dynamic of the border (hereafter referred to as the fork) between the open and closed regions of the DNA molecule [5, 22, 24–28]. The motion of the fork, whose position corresponds to the number $n(t)$ of open base pairs at time t, is modeled as a one-dimensional random walk in the potential $G(n, f)$. Fig. 7 A shows pseudo unzipping data (in silico experiments) obtained from Monte Carlo simulation. A time-trace is the sequence $\mathcal{N} = \{n_i\}$ of the fork positions at discrete times $t = i \times \Delta$, where Δ is the delay between two measures (inverse of the temporal bandwidth). Fig. 7B and Fig. 7C show that the plateaus in the opening signal are in correspondence with the deepest minima of the free energy $G(n, f)$. Our model implicitly defines the probability $\mathcal{P}(\mathcal{N}|S)$ to measure a time-trace $\mathcal{N}$ given the sequence S of the molecule.

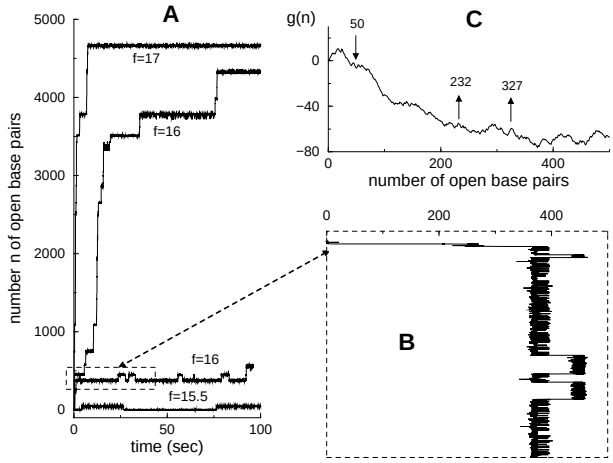

FIG. 7: Fixed force unzipping of the λ phage DNA obtained from our theoretical model. A. number n of open bases as a function of time t for forces f ranging from 15.5 to 17 pN. B. magnification of the boxed region in Figure 3A after left hand rotation by 90 degrees. C. potential V(n), equal to the free energy the λ phage DNA with the first n base pairs open (V=0 corresponds to the fully zipped molecule) for n¡450 and force f=16.4 pN. The up and down arrows indicate, respectively, a local minimum in n=50 and two maxima in n=232 and n=327.

B. Data treatment in fixed force-unzipping: from the signal to the sequence

Up to now we have considered the direct problem: from a given sequence $\mathcal{S} = \{b_i\}$ *e.g.* the λ-phage we have defined a model that gives un unzipping trace reproducing the experimental data. As we have seen the unzipping trace is just a unidimensional random walk in the free energy landscape $G(n, f)$ (function of the sequence $\mathcal{S}$)

The inverse problem can be stated in general terms as follows: from the trace $\mathcal{N} = \{n(t)\}$ of a discrete random walk is it possible to reconstruct the free energy landscape $G(n, f)$? This question translates, in our case, as follows: from the temporal trace $\mathcal{N}$ of the opening force can we deduce the sequence $\mathcal{S}$? We have studied this inverse problem in the Bayesian theory framework [31]. The probability to have a sequence $\mathcal{S}$ given a time-trace $\mathcal{N}$ can be written as,

$$\mathcal{P}(\mathcal{S}|\mathcal{N}) = \frac{\mathcal{P}(\mathcal{N}|\mathcal{S})\ \mathcal{P}_0(\mathcal{S})}{\mathcal{P}(\mathcal{N})}\ . \tag{4}$$

A prediction for a sequence is obtained by maximizing this probability over all possible sequences $\mathcal{S}$ (for a given $\mathcal{N}$). When no a priori information on the sequence is available, $\mathcal{P}_0(\mathcal{S})$ is uniform and the maximization of $\mathcal{P}(\mathcal{S}|\mathcal{N})$ over $\mathcal{S}$ reduces to that of $\mathcal{P}(\mathcal{N}|\mathcal{S})$. When, on the contrary, one does not aim to predict the content of a unknown molecule but to determine a mutation on a known molecule, $\mathcal{P}_0(\mathcal{S})$ contains the available a priori information on this sequence. $\mathcal{P}(\mathcal{N}|\mathcal{S})$ is defined from our dynamical model of the unzipping, and depends on the interval Δ between two measures.

We have first studied the unrealistic case of a very large bandwidth which allows us to follow all the base pair openings [30]. In this case the interval Δ between two measures is small with respect to the minimal base pair opening time (of about $1\mu s$ for an AT base pair). This value for Δ is not compatible with experimental limitations (e.g. due to the presence of the bead), indeed, nowadays, a realistic experimental value is $\Delta = 1ms$; however this limiting case can be studied in great details and some of the conclusions we obtained are generic in that they apply to the case of finite bandwidth. Once in silico unzipping data are generated with a Monte Carlo simulation a second program which ignores the phage sequence processes these data and makes a prediction for the sequence. This program is based on the Viterbi algorithm, widespread in information theory and in error correcting codes [31] In Figure 8 we show the probability ϵ_n (full line, $R = 1$) that base number n (along the 5' to 3' strand) is wrongly predicted, plotted for the first 450 bases of the λ phage at a force of 16.4 pN. ϵ_n varies a lot from base to base with values ranging from 0 (perfect prediction) to 0.75 (random choice of a base value among A,T,C,G). Comparison with the potential $G(n, f = 16.4)$ shows that ϵ_n is small in the flat regions of the potential ($350 < n < 450$), or in the local minima, for example the base $n = 50$ that is preceded by 4 weak bases and followed by 4 strong bases (...TTTA-A-GGCG...). On the contrary, the bases that are not well predicted correspond to local maxima of the potential e.g. bases n=327 and 328 located between 7 strong bases and 7 weak bases (...GCCGCCG-TC-ATAAAAT...). For a force of 16.4 pN, 67% of bases are correctly predicted. The error increases with the force because, keeping the duration of the experiment fixed, the time spent by the fork on each base decreases. The fractions of mispredicted bases are 47% and 20% for

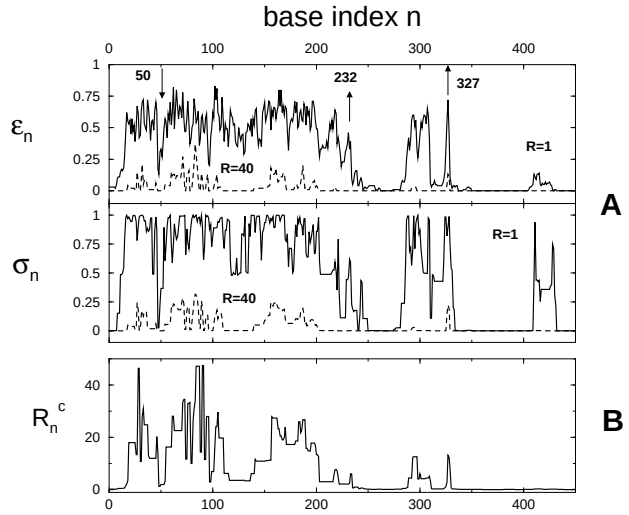

FIG. 8: A. Probability ϵ_n that base n is not correctly predicted (top) and entropy σ_n (middle) for the first 450 bases of the λ phage DNA unzipped under a force of 16.4 pN. The full lines correspond to the predictions from data coming from one unzipping, the dotted lines from combined data from R=40 unzippings. B. Theoretical values for the decay constants $R_c(n)$ of the errors ϵ_n. As an example, the base 232 (arrow) is characterized by $R_c(232) = 6$, and is badly (respectively, well) predicted from the data from 1 (respectively, 40) unzippings.

forces equal to, respectively, 17 and 15.5 pN.

The quality of prediction can be enormously improved by collecting the data from repeated unzipping of the same molecule, and processing them altogether. Indeed by opening and closing the molecule several times the fork will surely go over the same portion of the sequence several times, and the signal over noise ratio increases. Figure 4A shows that the error sharply decreases when the number R of unzippings increases from 1 to 40. We have confirmed this numerical result with theoretical tools borrowed from the theory of disordered systems, of biased random walks and of large deviations in probability theory [31]. The error in predicting base n decreases very rapidly with the number R of unzippings,

$$\epsilon_n \approx e^{-R/R_c(n,f)} \tag{5}$$

The typical decay constant $R_c(n, f)$ for the base n and at a force f can be calculated as the ratio of the decay constant at large force $R_c(i, f \geq 40)$ over the average number of openings of the base during a single unzipping, $\langle u_n \rangle$. $R_c(n, f \geq 40)$ depends on the sequence content around base n via the binding free energies $g_0(i)$ (2) and can be exactly computed for any given sequence [30]. Stacking interactions generate first-neighboring coupling between bases and give rise to block of bases with strongly correlated value for $R_c(n, f =\geq 40)$. The average number of openings of a base, $\langle u_n \rangle$, depends on the free energy landscape of the molecule (3), determined by the force and the sequence content, and can be computed for a given sequence [29, 30] . Our theoretical values for $R_c(n, f = 16.4\text{pN})$are shown in figure 8B, and vary between 0.1 to 45 depending on the base pair index n. The agreement with the decay of ϵ_n from $R = 1$ to $R = 40$ is very good (Fig. 8 A).

C. Toward more realistic studies

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. In particular we have calculated the decay constant R_c of the prediction error with the number of collected data with a finite time interval Δ between two measures. We have also implemented an inference algorithm in the case of a small bandwidth or equivalently a large interval between two measures $\Delta \sim 10^{-3}s$. This algorithm, at difference from the Viterbi algorithm, is not guaranteed to find the maximum of (4), but correctly

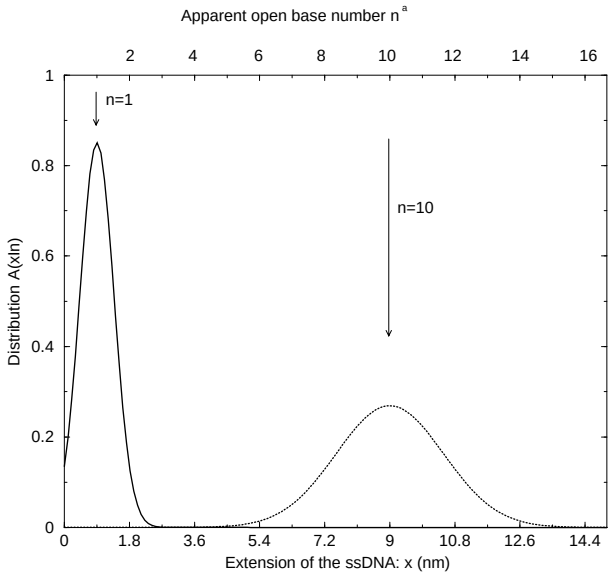

FIG. 9: Distribution $A(x|n)$ of the extension x of the open ssDNA at fixed position n of the opening fork , for $n = 1$ and $n = 10$. The variance of the distribution (at a force of 16 pN) increase as $0.22\ nm^2 \times n$. The apparent value of the number of opened bases is n^a = closest integer to x/x_o, with $x_o = 0.9\,nm$, as shown on the top label of the x axis.

find back a sequence of 30 bases when 100,000 measures are collected (which corresponds with $\Delta \sim 10^{-3}$ s to some minutes of experiments. This result is in agreement with the theoretical value $R_c = 20000$ we found. Moreover in the Monte Carlo simulations the dynamics of the opening fork is directly monitored, while in real experiments the distance between the opened extremities is measured, and the ssDNA are not rigid linkers. We have extended our theoretical calculation by including the fluctuations of the ssDNA. As illustrated in Fig.9 due to the ssDNA fluctuations for each position of the opening fork n we have a distribution $A(x|n)$ of the ssDNA extremities distance x of the molecule. Moreover the variance of the distribution A increases with n because longer the ssDNA is flopper it becomes. Our finding is that even for a large bandwidth, $\Delta \simeq 10^{-6}s$, corresponding to the opening time of the base AT, the reconstruction procedure would imply a pre-treatment of the signal through a deconvolution of the measured n^a with the pseudo-inverse of A; moreover a number of unzippings increasing as $\sqrt{n}$ is necessary to reach the same inference performance as in the absence of ssDNA fluctuation.

D. Conclusion

Micromanipulation experiments on single molecules of DNA are interesting approaches to extract information on DNA sequences. Among these experiments we have reviewed the unzipping of DNA under a mechanical action [4–8] or under translocation through a nano-pore [17, 18], and the observation of the sequence-dependent activity of an exo-nuclease [9, 10] and of a polymerase. Even if these experiments cannot be nowaday competitive with the standard sequencing method, they could be useful for applications that require for example partial information on the sequence and where a complete sequencing is a waste of time and effort while a rapid single molecule screening would be more appropriate. Let us cite the analysis of the genetic variability of the genome, that is, the ability to locate in the genome of an individual the differences with respect to the average sequence (mutations, SNIPs, displacements of some portions of the sequence, ...) available in data bank e.g. the human genome bank. Two applications can be distinguished here. First the detection of known mutations. Secondly the search for genetic signatures corresponding to a given phenotype, signaling for instance a predisposition or resistance to some disease, or a particular reaction to some drug, ... This application should ultimately provide a better understanding of genetic and tumoral diseases, and lead to the development of personalized medicine.

So far these promising experiments have however not been combined with a concrete method for data analysis capable of extracting precise information on the sequence of DNA molecules. To make the single-molecule sequencing successful optimal experimental protocols have to be designed as well as inference methods capable of providing us with the most accurate information possible about the sequence. The main difficulty comes from the intrinsically noisy nature of the measured force signal. Extracting information on the sequence from this signal will not be possible unless a detailed and deep comprehension of the noise is reached. In particular it is very important to have a good

understanding of the changes in the extension of the unzipped DNA strands due to thermal fluctuations.

---

[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] S. Cocco, J.F. Marko, *Physics World* **16**, 37 (2003)
[4] B. Essevaz-Roulet, U. Bockelmann, F. Heslot, *Proc. Natl. Acad. Sci. (USA)* **94**, 11935 (1997)
[5] U. Bockelmann, P. Thomen, B. Essevaz-Roulet, V. Viasnoff, F. Heslot, *Biophys. J.* **82**, 1537 (2002)
[6] J. Liphardt *et al. Science* **297**, 733 (2001).
[7] S. Harlepp *et al. Eur. Phys. J. E* **12**, 605 (2003).
[8] C. Danilowitcz *et al. Proc. Natl. Acad. Sci. (USA)* **100**, 1694 (2003).
[9] A.M. Van Oijen, P.C. Blainey, D.J.Crampton, C.C. Richardson, T. Ellemberg, X. Sunney Xie, *Science* **301**, 123 (2003)
[10] T.T. Perkins, R.V. Dalal, P.G. Mitsis, S.M. Block *Science* **301**, 1914 (2003).
[11] C. Bustamante, Z. Bryant and S. B. Smith *Nature* **421**, 423 (2003).
[12] GC Wuite, S.B. Smith, M. Young, D Keller, Bustamante C *Nature* **404**, 103 (2000).
[13] B. Maier, D. Bensimon, V. Croquette *Proc. Natl. Acad. Sci. (USA)* **97**, 12002 (2000).
[14] M.J. Levene, J Korlach J, SW Turner, M Foquet, HG Craighead, WW Webb *et al. Science* **299**, 682 (2003).
[15] M. Rief,H Clausen-Schaumann, H Gaub Nat Struct Biol. 1999 Apr;6(4):346-9.
[16] M. Singh-Zocchi,S. Dixit, V. Ivanov, G. Zocchi *Proc. Natl. Acad. Sci. (USA)* **100**, 7605 (2003).
[17] A.F. Sauer-Budge, J.A. Nyamwanda, D.K. Lubensky, D. Branton *Phys. Rev. Lett.* **90**, 238101 (2003)
[18] J. Mathé, H.Visram, V. Viasnoff, Y Rabin, A. Meller *Biophys. J.* **87**, 3205 (2004).
[19] M. Zuker *Curr. Opin. Struct. Biol.* **10**, 303 (2000).
[20] Santa Lucia Jr *Proc. Natl. Aca. Sci. (USA)* **95**, 1460 (1998),
[21] S.B. Smith, Y. Cui, C. Bustamante, *Science* **271**, 795 (1996)
[22] S. Cocco, R. Monasson, J. Marko. *C.R. Physique* **3**, 569 (2002).
[23] R.E. Thompson, E.D. Siggia. *Europhys. Lett.* **31**, 335 (1995).
[24] S. Cocco, R. Monasson, J. Marko. *Eur. Phys. J. E* **10**, 153 (2003).
[25] D.K. Lubensky, D.R. Nelson. *Phys. Rev. Lett.* **85**, 1572 (2000); *Phys. Rev. E* **65**, 031917 (2002).
[26] U. Gerland, R. Bundschuh, T. Hwa. *Biophys. J.* **81**, 1324 (2001).
[27] M. Manosas, F. Ritort, *Biophys. J.* **88**, 3224 (2004).
[28] D. Marenduzzo *et al. Phys. Rev. Lett.* **88**, 028102 (2002).
[29] V. Baldazzi *et al.*, *Phys. Rev. Lett.* **96**, 128102 (2006)
[30] V. Baldazzi *et al.*, Submitted to *Phys. Rev E* (2006).
[31] D.J.C. McKay, Information Theory, Inference and Learning Algorithms, Cambridge University Press (2003).