Journal of Biological Physics 31: 273–301, 2005.
DOI: 10.1007/s10867-005-3244-3 

# Can We Model DNA at the Mesoscale?

S. CUESTA-LÓPEZ[1,2,3], J. ERRAMI[2], F. FALO[1,3] and M. PEYRARD[2,*]
[1]*Dept. Física de la Materia Condensada, Universidad de Zaragoza, c/Pedro Cerbuna s/n 50009 Zaragoza, Spain;* [2]*Laboratoire de Physique, Ecole Normale Supérieure de Lyon, 46 allée d'Iltalie, 69364 Lyon cedex 07, France;* [3]*Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Spain*
*(*Author for correspondence, e-mail: michel.peyrard@ens-lyon.fr)*

**Abstract.** Modelling DNA is useful for understanding its properties better but it is also challenging because many of these properties involve hundreds of base pairs or more, or time scales which are much longer than the time scales accessible to molecular dynamics. It is therefore necessary to develop models at a mesoscale, which include enough details to describe the properties of interest, for instance the biological sequence, while staying sufficiently simple and realistic.

We discuss here two examples: a dynamical model to study the mechanical denaturation of DNA, which probes the sequence on various scales, and a model for the self assembly of DNA which describes the formation of hairpins and allows us to study its kinetics.

## 1. Introduction

When they proposed the famous structure of DNA, Watson and Crick supported their analysis by building a model of the molecule. One may wonder whether this approach, which has been so fruitful for the static case can also be used to investigate the *dynamical* properties of DNA, so important in many biological processes.

The dynamical properties of DNA can be investigated by computer simulations, but, although molecular dynamics and computers have made tremendous progress, all-atom simulations of DNA are still severely limited. To study most of the biological processes it is necessary to model DNA at a scale of hundreds or thousands of base pairs, which can only be done with a description that does not attempt to reach the atomic resolution. While such models may look simpler because they do not carry all the fine details of the molecule, their design raises difficult challenges: how to select the degrees of freedom that must be included, how to determine the interaction potentials between the components of these models at a mesoscale, larger than the atomic scale but still small enough to describe the biologically relevant properties of DNA?

Besides the possibility for carrying out calculations that would not be possible on an all-atom model, the development of simpler models is also useful to determine

what are the degrees of freedom which are actually controlling some specific properties of DNA. If a model is able to properly describe these properties, it indicates that it catches the phenomena that govern them.

To establish a DNA model, the first step is the selection of the appropriate scale. The answer depends on the properties of interest. For instance, to analyse the force-extension curves of DNA, which have been measured in single molecule experiments, the model can ignore all the internal details of the molecule and describe it as a flexible rod or a flexible polymer chain. Here we are interested in questions which are of biological relevance so that the model cannot ignore the base pair sequence which contains the genetic code. Hence the scale of the model must not be larger than the base pair. It may be tempting to select an atomic scale which would allow us to study very precisely the dynamics of all the atoms that make up the molecule. This level of modelling, generally referred to as "molecular dynamics", which has been extensively used for DNA [1], requires huge computing facilities to investigate the dynamics of the molecule on a scale larger than a few tens of base pairs and time scales of the order of nanoseconds. Our goal here is different because we are interested in properties of DNA which involve tens or hundreds of base pairs such as the mechanical denaturation of the double helix, or may be very slow at the molecular time scale such as the closing of a DNA hairpin. All atom-simulations would be unpractical.

The goal of this paper is to show that some properties of DNA can be properly described at a scale intermediate between the micro-scale of the elastic string and the atomic scale. The "meso-scale" that we consider is the scale of the base pair, where a single degree of freedom is used per base pair. This is a drastic simplification with respect to the atomic scale, but the description is however detailed enough to include the genetic code and recent results suggest that the biological relevance of such models is quickly growing as the models improve [2].

In order to discuss the validity of this "mesoscale" approach we shall focus our attention on two cases for which experimental results are available, and compare the output of the modelling with these results. We have chosen the dynamics of the mechanical opening of DNA [3, 4] and the fluctuations of DNA hairpins [5].

## 2. Mechanical Unzipping of DNA

The mechanical opening of DNA was first proposed by Viovy *et al.* [6] as a possible approach to determine the base sequence. It appears that it is not possible to proceed a base at a time, so that only some information on the sequence averaged on a scale of hundreds of bases can be obtained. These experiments raise questions concerning possible dynamical effects such as the role of the fluctuational opening of DNA on the results [7] because they involve out-of-equilibrium properties of the molecule. Current experiments are slow so that out-of-equilibrium effects can probably be neglected in most of them, but such a fundamental problem should however

not be forgotten, particularly if faster experiments on smaller DNA fragments are performed by pulling with an atomic force microscope.

Two different approaches can be used, and they are not probing the same thermodynamic ensemble [8]: one probes the constant-extension ensemble, while the other one probes the constant-force ensemble. The first experiments were performed by pulling at constant velocity [9] and while the opening of the two strands moved along the molecule, the variation of the force was recorded. In an experiment that opened about 10000 base pairs, a good correlation was observed between the magnitude of the force required for the opening, and the content of Guanine-Cytosine (GC) with respect to the Adenine-Thymine (AT) base pairs, computed by an average over 100 base pairs. This makes sense because the GC base pairs are linked by 3 hydrogen bonds, while the AT are linked by only two. The constant force experiments [4] are even more spectacular because they exhibit multiple metastable intermediates. The opening is characterised by rapid jumps and very long pauses, which can last for minutes or more. These experiments are explained by the existence of very deep local minima in the macroscopic barrier to unzipping which are found by coarse-graining thermodynamic data deduced from the fitting of experimental melting curves.

These results raise several questions:

– although observations are made on single molecules, the resolution of the experiments is of the order of hundreds or thousands of base pairs. Thus they probe DNA on a large scale. However experiments show that local effects on DNA thermal denaturation curves can be very significant since the melting curves of DNAs which differ by one base pair out of several hundred can be distinguished [10]. As mechanical denaturation can be expected to show similar properties, an analysis at a smaller scale would be useful.
– understanding the properties of DNA from basic principles rather than through empirical parameters would be very interesting, although it might be a challenge.

Our goal in this section is to show that a simple model of DNA can, at least partially, answer these questions.

### 2.1. A SIMPLE MODEL FOR NONLINEAR DNA DYNAMICS

Since we intend to model DNA from "first principles" we have to start from a description in terms of the elements that make up the molecule, and their interaction potentials. As discussed above we do not want to choose elements as small as the atoms, but rather describe the molecule at the scale of a base pair. We can start from a very simple dynamical model of DNA, which allows a qualitatively correct description of the thermal denaturation of DNA [11, 12] and turns out to provide results for the study of DNA thermal denaturation of short DNA sequences in good agreement with experiments [13].

The model is an extension of the Ising models which describe a base pair as a two-state system that can either be closed or open. It uses a real variable $y_n$

to describe the stretching of the $n$th base pair, which can increase continuously to infinity if the two bases separate completely as in DNA denaturation or take negative values, corresponding to a compression of the bond linking the bases with respect to its equilibrium length. Large negative values are however forbidden by steric hindrance introduced in the model by the potential linking the bases in a pair.

Of course such a description of the state of a base pair by a single variable cannot claim to describe the geometry of the molecule, and it is not our aim here. The model should be viewed as a truly minimal model. Including a variable which describes a displacement instead of a two-state system is essential for two reasons: we want to study the dynamics of the opening and thus it is necessary to model intermediate states between the open and closed base pair, and we want to establish the model on "first principles", which, in this case means that we would like to be able to express interaction potentials evaluated from our knowledge of the physical interactions in DNA, even though giving them quantitative values for a mesoscopic model is not trivial, as discussed below.

The model is shown on Figure 1 and it is defined by its Hamiltonian

$$H = \sum_n \frac{p_n^2}{2m} + W(y_n, y_{n-1}) + V(y_n), \quad \text{with } p_n = m\frac{dy_n}{dt}, \tag{1}$$

where $m$ is the reduced mass of the bases.

The potential $V(y)$ describes the interaction between the two bases in a pair. We use a Morse potential

$$V(y) = D(e^{-\alpha y} - 1)^2, \tag{2}$$

where $D$ is the dissociation energy of the pair and $\alpha$ a parameter, homogeneous to the inverse of a length, which sets the spatial scale of the potential. This expression has

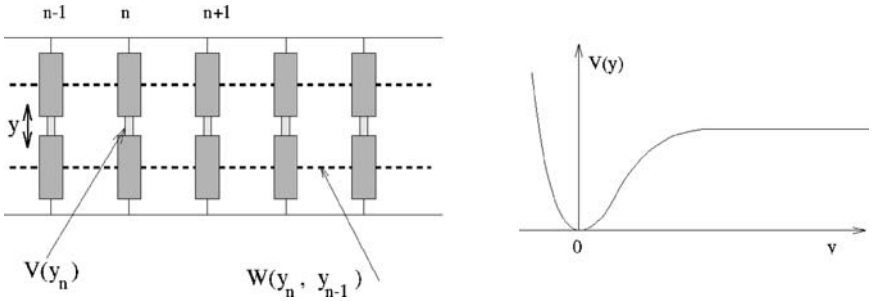


*Figure 1.* The simple dynamical model for DNA nonlinear dynamics, described by Hamiltonian (1).

been chosen because it is a standard expression for chemical bonds and, moreover, it has the *appropriate qualitative shape*.

- it includes a strong repulsive part for $y < 0$, corresponding to the sterical hindrance mentioned above,
- it has a minimum at the equilibrium position $y = 0$,
- it becomes flat for large $y$, giving a force between the bases that tends to vanish, as expected when the bases are very far apart; this feature allows a complete dissociation of the base pair, which would be forbidden if we had chosen a simple harmonic potential.

Other shapes may however be considered to get more quantitative results on the dynamics of the fluctuations of the bases. There is actually a potential barrier for reclosing because, when an open base comes back in the stack in the closed state, it has to force its way within the stack of the neighbouring bases. This effect can be described by a modification of the potential $V(y)$ [17].

The potential $W(y_n, y_{n-1})$ describes the interaction between adjacent bases along the DNA molecule. It has several physical origins:

- the presence of the sugar-phosphate strand, which is rather rigid and connects the bases. Pulling a base out of the stack in a translational motion tends to pull the neighbours due to this link. One should notice however that we have not specified the three dimensional motion of the bases in this simple model. An increase of the base pair stretching could also be obtained by rotating the bases out of the stack, around an axis parallel to the helix axis and passing through the attachment point between a base and the sugar-phosphate strand. Such a motion would not couple the bases through the stretching of the strands but their torsional rigidity would be involved. The potential $W(y_n, y_{n-1})$ is an effective potential which can be viewed as averaging over the different possibilities to displace the bases.
- the direct interaction between the base pair plateaux, which is due to an overlap of the $\pi$-electron orbitals of the organic rings that make up the bases.

The choice of $W(y_n, y_{n-1})$ is crucial for the validity of the model. A simple harmonic potential $W(y_n, y_{n-1}) = \frac{1}{2}K(y_n - y_{n-1})^2$ is not acceptable. Such a potential is sufficient to lead to a true phase transition between double-stranded and single-stranded DNA, i.e. a true phase transition in one dimension. This is interesting from the statistical physics point of view [11, 14], but it yields a very smooth denaturation, occurring over several tens of degrees, which does not agree at all with the observations of a sharp denaturation. A more elaborate potential has to be chosen, and it turns out that the potential

$$W(y_n, y_{n-1}) = \frac{1}{2}K\left(1 + \rho e^{-\delta(y_n + y_{n-1})}\right)(y_n - y_{n-1})^2 \tag{3}$$

meets the required conditions [12, 15]. When both interacting base pairs are closed ($y_n$ and $y_{n-1}$ small), the potential behaves like a harmonic potential with the coupling constant $K(1 + \rho)$. If either one of the base pairs is open ($y_n$ or $y_{n-1}$ large)

the effective coupling constant drops to $K$. Thus, with this anharmonic potential, the stacking interaction is significantly reduced when the base pairs open. There are several reasons which explain this behaviour of the coupling along the DNA molecule:

– The overlap of the $\pi$ electrons on the plateaus formed by the bases is an important part of the stacking energy, which is modified because the breaking of the hydrogen bonds leads to a redistribution of the $\pi$ electrons.
– In the open state the plateaus of the bases are displaced with respect to the regular stacking of double-stranded DNA, and the average overlap of consecutive bases is reduced.
– In the double helix, the rigidity of the strands which contributes to the collective behaviour of DNA is high because the hydrogen bonds between the bases prevent many rotational motions around the single bonds that make up the stand. As soon as the interaction is broken, rotations become possible and the DNA strands start to behave like RNA, which is very flexible.

Although the analytical form that we choose for $W(y_n, y_{n-1})$ is certainly not crucial, the physics behind the nonlinearity of $W$, which makes it become weaker when the bases pairs break, is essential to the validity of the model. This illustrates the importance of setting up appropriate interactions when a mesoscale model is designed. The nonlinear stacking allows a very large increase in the fluctuations of the strands when base pairs open, thereby increasing the entropic effects. This is what makes the DNA thermal denaturation so sharp [15].

### 2.2. Sequence and model parameters

In order to test the properties of the model to detect specific features of the DNA sequence, we studied an artificial sequence obtained by the juxtaposition of two characteristic sequences extracted from the DNA of bacteriophage T7, a termination sequence, i.e. a sequence which signals the end of a gene, and a promoter sequence, which, on the contrary, corresponds to the beginning of the gene. The promoter is a place where DNA opens to allow the beginning of the transcription of the gene, and thus it can be expected to have specific properties regarding opening. The sequence is repeated twice in our test model, giving a set of 154 base pairs. It allows us to test the reproducibility of the results and provides a longer DNA chain for the simulation of the mechanical opening. The promoter sits between sites 56 and 77, and between sites 133 and 154. Figure 2 shows the sequence used in our simulations.

The choice of the potential parameters is difficult because the potentials entering in the model are effective potentials, which combine many actual interactions. For instance $V(y)$ includes the hydrogen bonds between the bases but also the repulsion between the charged phosphate groups, which is partly screened by the ions in solution. However, physics can give an estimate for a typical set of parameters, which can then be refined by comparison with experiments. For $V(y)$ this leads to the following estimate: $D = 0.03$ eV, which is slightly above $k_B T$ at room

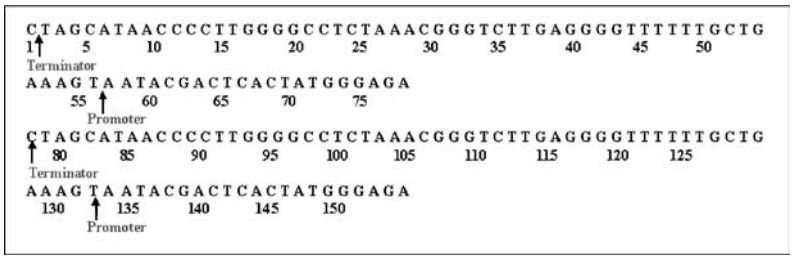


*Figure 2.* Sequence used in the simulation of DNA mechanical denaturation. A terminator (bases 1–55 and 78–132) and a promoter sequence (bases 56–77 and 133–154) of bacteriophage T7 are repeated twice to give a sequence of 154 base pairs.

temperature ($k_B$ being the Boltzmann constant) and $\alpha = 4.5$ Å$^{-1}$. For a stretching of the base pair distance of 0.1 Å, these parameters give a variation of energy of 0.006 eV, which is consistent with the values that can be expected for hydrogen bonds. A typical upper value for $K$ is $K = 0.06$ eV/Å$^2$, which corresponds to a weak coupling between the bases, as attested by the experimental results showing that proton-deuterium exchange can occur on one base pair without affecting the neighbours. The average mass of the nucleotides is 300 atomic mass units. For our simulations we work with a system of units adapted to the scale of the problem: lengths in units of $\ell = 1$ Å, energies in units of $e = 1$ eV, mass in units of $m_0 = 1$ atomic mass unit. This defines a natural time unit $t_0$ through $e = m_0 \ell^2 t_0^{-2}$, which is equal to $t_0 = 1.018 \times 10^{-14}$ s. This is of the order of magnitude of the period of the vibrational motions of the base pairs.

However the parameter values that we listed above are only averaged values for a hypothetical DNA homopolymer. To describe the base-pair sequence, they must be refined to make the difference between the two types of base pairs, AT linked by 2 hydrogen bonds, and GC linked by 3 hydrogen bonds. This has been done by Campa and Giansanti [13] who compared experimental curves for the denaturation of short DNA sequences with the denaturation curves which are given by the model. In their study the stacking interaction parameters have been considered as independent of the sequence, so that the sequence only enters into the parameters of the potential $V(y)$. Their optimisation leads to $D_{\mathrm{AT}} = 0.05$ eV, $D_{\mathrm{GC}} = 0.075$ eV (corresponding to a ratio 3/2 corresponding to the ratio of the number of hydrogen bonds), and $\alpha_{\mathrm{AT}} = 4.2$ Å$^{-1}$ $\alpha_{\mathrm{GC}} = 6.9$ A$^{-1}$. In the following we shall use these parameters for the potential $V$. For the stacking interaction, ref. [13] proposes $K = 0.025$ eV Å$^{-2}$, $\rho = 2$ and $\delta = 0.35$ Å$^{-1}$, however these values are harder to confirm from physical arguments due to the complicated nature of the stacking interaction. This is why we carried out some tests by varying the stacking.

### 2.3. MELTING OF AN INHOMOGENEOUS DNA SEQUENCE

In order to test the ability of the model to simulate an actual DNA sequence, we first examined the behaviour of the model at constant temperature without any

unzipping force. The simulation is carried with a multi-thermostat Nosé method [16], which we found efficient to reach a good thermalization. We use a time step $\Delta t = 0.05 t_0$, i.e. $\Delta t \approx 0.5 \times 10^{-15}$ s. We start from an initial condition with a random velocity distribution corresponding to an initial temperature of 100 K, which is first equilibrated for $2.5 \times 10^5$ time steps, then heated to the working temperature with a temperature ramp lasting $2.5 \times 10^5$ time steps and again equilibrated for $2.5 \times 10^5$ time steps before any recording is made. A simulation involves 10 runs with different initial conditions, and the data are collected during 1 ns for each run. In these studies of the melting, the model is not subjected to any external force. Its right end ($n = 154$) is fixed and its left end ($n = 1$) is free.

One issue is the appropriate choice of the nonlinear coupling parameters which are hard to determine *a priori*. Figure 3 shows the statistics of opening for the base pairs along the chain for several sets of $K$ and $\rho$ parameters which give the same effective coupling constant $K' = K(1 + \rho) = 0.06$ eV Å$^{-2}$ when both base pairs are closed. The opening events are counted by probing each base every 0.5 ps, and counting a base pair as open when its stretching $y_n$ exceeds 2 Å, a value corresponding to the plateau of the Morse potential. Although this value is arbitrary, the results are not qualitatively changed by changing it, provided we select a value which is at least beyond $2y_{\text{inflex}}$, where $y_{\text{inflex}}$ is the value of $y$ corresponding to the inflexion point of the Morse potential. For small values of $\rho$ the coupling constant does not change very much when DNA is open or closed, while for $\rho = 2$, the coupling constant drops from 0.06 to 0.02 eV Å$^{-2}$ when either of the two interacting bases is open.

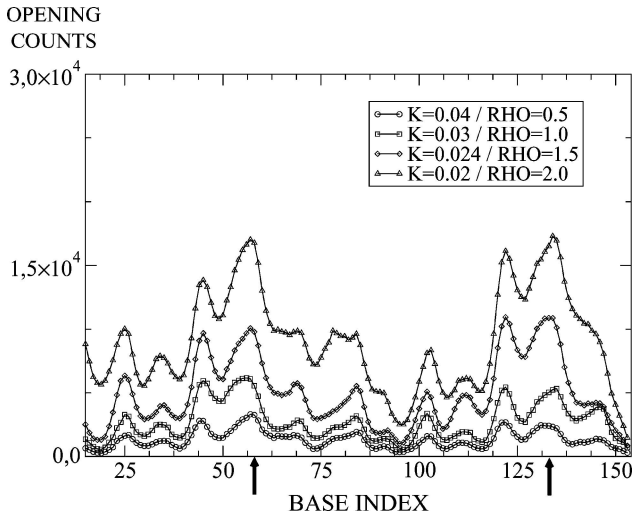


*Figure 3.* Statistics of the opening of the base pairs in an inhomogeneous DNA for various values of the nonlinear coupling parameters. The effective coupling constant $K' = K(1 + \rho)$ is kept constant and equal to 0.06 eV Å$^{-2}$. The arrows point to the promoter regions in the sequence.

The first point that should be noticed is that the results on the opening appear to be robust: the two halves of the model, which have the same base-pair sequence in our model (see Figure 2), show very similar patterns of opening. The results cannot be expected to be strictly identical on both halves because for several reasons: (i) the boundary conditions on both ends of the chain are not the same, (ii) due to the limited time of the simulation we do not get exactly equilibrium properties, and dynamical effects may play some role. However our results show that the reproducibility of the results for a given sequence is good. Figure 3 shows the importance of the nonlinear coupling. For small values of $\rho$ the fluctuational opening of the molecule varies only slightly along the sequence, while for $\rho = 2$ very large variations are observed. It is interesting to notice that the promoter region is the region where the largest number of opening events is found. This result is consistent with some recent findings using the same mesoscopic model of DNA to look for promoters in actual DNA sequences [2] but different diagnostics because these studies were explicitly looking for open bubbles. They found that promoters have a higher probability of opening than other regions of the sequence, in agreement with experimental observations. The simulations raise interesting questions about the collective effects which control the probability of opening. A simple view would conclude that any series of AT base pairs, which are weaker than the GC pairs, would lead to a high probability of opening. This is of course partly true because examining the regions of high opening probability on Figure 3, one indeed finds that they are in AT rich regions, but the actual opening probability relies on more subtle effects because the promoter region between indices 56 and 60, with 5 AT pairs shows a slightly higher opening that the region 42–47 of the terminator which has 6 adjacent AT pairs. Understanding these subtleties in DNA opening is still an unsolved problem, which would be of high practical interest.

### 2.4. SIMULATING THE MECHANICAL UNZIPPING

Studying the constant force mechanical unzipping requires an extension of the model. Figure 4 shows the configuration that we have chosen to mimic this experiment. As the model only has one degree of freedom per base pair, one strand can be viewed as fixed, while the other one is pulled by a constant force $F$ applied to a bead of mass $M$ and coordinate $y_M$, attached to DNA by a molecular linker, which is simply represented by a harmonic spring of constant $K_0$, connected to the first base pair. The DNA model is still thermalized by a numerical thermostat which can be either a Nosé thermostat as discussed above, or a Langevin thermostat in some calculations to allow us to explore another regime of coupling between the outside medium, i.e. the solution, and the DNA molecule. The bead, being much more massive ($M = 3000$ atomic mass unit) and bigger than a DNA nucleotide, experiences a macroscopic friction force, which is described in our calculation by an extra term $-\gamma_M dy_M/dt$ in its equation of motion. The damping has been set to $\gamma_M = 100t_0^{-1}$ which is 5 times the damping at which a harmonic oscillator made by

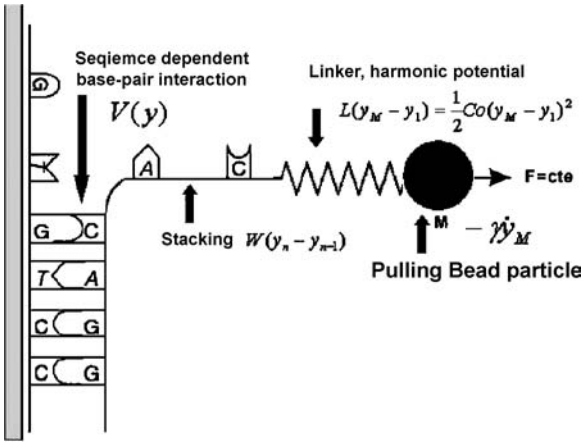


*Figure 4.* Model chosen for the simulation of the unzipping. A harmonic spring is added to the DNA chain to simulate the molecular linker which connects DNA to the bead on which the force $F$ is applied. Note that the model is drawn to look similar to the experimental setting for clarity, but everything is one-dimensional, and all displacements, whether they are stretching $y_n$ of the bases of displacement $y_M$ of the pulling bead, are along the same direction.

the bead and a spring of coupling constant $K = 0.04\,\text{eV/Å}^{-2}$ becomes overdamped. This corresponds to the physical situation where the motion of the bead in the fluid is highly damped by the hydrodynamic flow. On our numerical experiments, the applied forces have never been smaller than 50 pN although experiments may use smaller forces because, in spite of the simplicity of the model, the time range that can be studied in a simulation is limited. Small forces lead to an unzipping too slow for the numerical study. For $F = 50$ pN, the mechanical opening of the DNA model occurs typically in 300 ns.

Figure 5 shows the propagation of the opening when we apply a force $F = 51$ pN, measured on one hand by the number of broken base pairs, defined as for the melting by $y_n > 2$ Å, and, on the other hand, by the displacement of the pulling bead versus time. Large fluctuations of the number of open pairs are observed. They are not statistical fluctuations associated to different realisations with different samples because Figure 5 displays the results of a single numerical simulation, but they come from dynamical effects. With a force of $F \approx 50$ pN the opening is slow and the strand which is already denaturated is not under strong tension. Its fluctuations allow temporary re-closings of some segments of the molecule. Such fluctuations involve 10 to 30 bases so that they are well below the resolution of the experiments.

At the scale of our simulations, the propagation of the opening shows periods of fast progress separated by pauses where the unzipping stops or slows down very much. Constant force unzipping experiments [4] find similar phenomena, but, in the experiments they appear at a much larger scale of hundreds or thousands of base pairs. The opening pattern is clearly related to the sequence: series of AT base pairs lead to fast unzipping while GC rich regions lead to pauses. Even a sequence

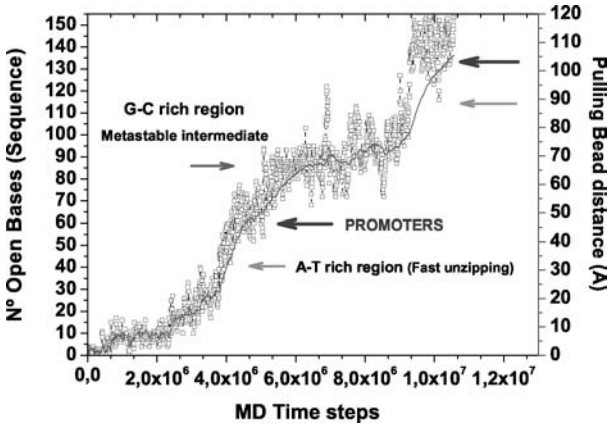


*Figure 5.* Propagation of the unzipping in the sequence of Figure 2 driven by a constant force $F = 51$ pN. The squares show the number of open bases versus time (left scale) and the full line shows the displacement of the pulling bead (right scale).

of 2 GC base pairs in a series of AT leaves a detectable signature in the motion of the pulling bead, as seen for instance at the level of the promoter.

Although the timing of the opening may vary significantly from one numerical experiment to another carried with the same parameters but a different pattern of fluctuations, the shape of the pattern is rather well preserved, showing pauses or fast opening in the same regions of the sequence. This indicates that the unzipping pattern is actually probing the sequence and not simply the thermal fluctuations of the sample. Figure 6 shows that the details of opening pattern are sensitive to the

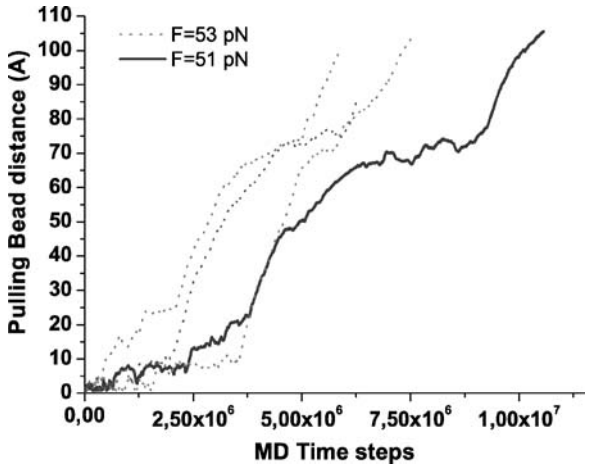


*Figure 6.* Propagation of the unzipping for different values of the pulling force. For each force ($F = 51$ pN, thick lines and $F = 53$ pN, thin lines) the figure shows the result of two different numerical experiments carried with Nosé thermostats using different realisations of the fluctuations.

exact value of the force since $F = 51$ pN and $F = 53$ pN give distinguishable patterns, but the general features are preserved. What appears to be more important is a change in the environment of the molecule, which, in our simulations is modelled by the thermostat. The Nosé thermostat describes an outside medium which is weakly coupled to the molecule. The results that it yields are similar to results obtained by a Langevin simulation using a small damping ($\gamma < 0.1 t_0^{-1}$, for a damping force in the equation of motion of the $n$th base pair written as $-m\gamma(dy_n/dt)$. In the simulations a stronger coupling to the fluid surrounding the molecule can be described by a Langevin thermostat using a larger damping, such as $\gamma = 1$. It modifies drastically the patterns, which become closer to the experimental patterns: long pauses (of the order of $5 \times 10^6$ time steps for instance) are observed, followed by a quick unzipping of about 60 base pairs. This points out the important role of the environment, which could perhaps be probed in experiments too, by varying the viscosity of the solution as it has been done for some studies of protein dynamics. This also raises the question of the validity of numerical simulations of a thermalized biomolecule at a mesoscale. In this case, the thermostat plays a more crucial role than in a microscopic simulation, where all the water molecules are present. For a mesoscopic model the choice of an optimal thermostat may be difficult. When the bases are stacked they are only weakly exposed to the solvent and a small value of $\gamma$ (or a simulation with the Nosé thermostat) is probably appropriate, but when the bases are unstacked a significantly larger damping should be used [17].

### 2.5. Analysing the mechanical unzipping

In the previous section, mechanical unzipping was viewed as a dynamical process, but it can also be viewed as an out-of-equilibrium transition between a closed and an open state. Figure 7 illustrates this point of view by showing the phase diagram of the DNA sequence of Figure 2 in the $(F, T)$ plane for two values of the interaction parameter $\rho$.

As expected the transition temperature decreases when $F$ increases. It can even drop to 0 for a sufficiently large force. In this case the denaturation is purely mechanical and can be expected to happen when the effective potential in the presence of the force, $V_{\text{eff}}(y) = D[\exp(-\alpha y) - 1]^2 - Fy$ looses its minimum, i.e. for $F_c = \alpha D/2$. For an inhomogeneous model things are more complex, but if we use the potential parameters of an AT base pair, we get $F_c = 168$ pN, while a GC pair would open mechanically for $F_c = 414$ pN. The frontier separating the double-stranded region from the melted region shows a sharp increase of the force required to open the molecule at low $T$, which is not surprising because, in the limit of vanishing temperature, a single GC pair would be enough to prevent a full separation of the strands for any force below 414 pN. What is noticeable is that, although the transition temperature under force depends on the value of $\rho$, this effect is weak. The global behaviour of the system appears to be less affected by the nonlinearity of the coupling than the properties concerning local opening.

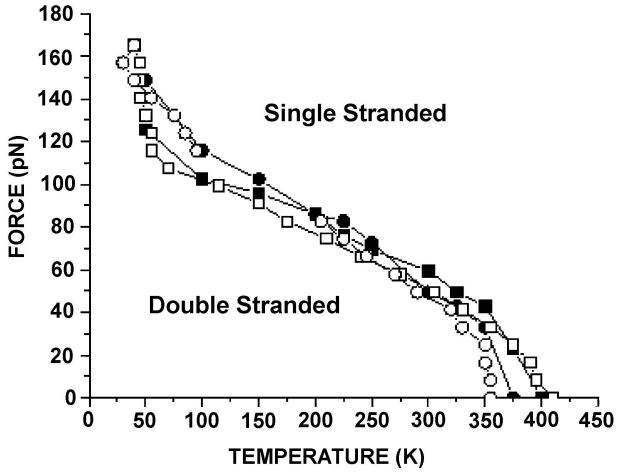


*Figure 7.* Phase diagram of DNA under traction as a function of the pulling force $F$ and temperature $T$. Below the lines (low $F$, low $T$ domain), DNA is in the double helix state, while above the two strands are separated. The sequence is the sequence shown in Figure 2. The potentials $V(y)$ have the values listed for *AT* and *GC* base pairs above. The coupling constant is $K = 0.04$ eV Å$^{-2}$ and $\rho = 0.5$ (square symbols) or and $\rho = 2$ (circles). The open symbols correspond to simulations carried at constant temperature and increasing force, until denaturation was found. The filled symbols correspond to simulations at constant force $F$, and increasing temperature until the denaturation was observed. The lines are not based on a model calculation, and should be viewed as a guide for the eye only.

In order to understand how the denaturation moves along the DNA molecule, it would be useful to evaluate the energy cost for the opening of a given region of the molecule. A crude estimate would be to sum up the Morse potential energies of the broken bases, but this is a local analysis which does not consider any collective effect in the opening. Another possibility, which is still approximate but turns out to be fruitful, is to examine what the equations of motion of the model can tell us about the opening. For a homogeneous DNA chain it is possible to analytically compute the shape of the "domain wall" which separates a closed part of the molecule from the part which has been opened by pulling. For this calculation it is convenient to move to a dimensionless set of variables by introducing the dimensionless stretching $Y = ay$ and measuring the energies in units of the depth $D$ of the Morse potential so that the Hamiltonian becomes $H' = H/D$. We can also introduce a dimensionless coupling parameter $S = K/(Da^2)$ and a dimensionless time $\tau = \sqrt{Da^2/m}t$. For the homogeneous chains where the parameters do not depend on the site, all these quantities are defined without ambiguity and the dimensionless Hamiltonian becomes

$$H' = \sum_n \frac{1}{2}P_n^2 + \frac{2}{S}(Y_n - Y_{n-1})^2 + (e^{-Y_n} - 1)^2 \quad \text{with } P_n = \frac{dY_n}{d\tau}. \tag{4}$$

This Hamiltonian is a simplified form of Hamiltonian (1), in which the anharmonic interaction potential $W$ has been replaced by its harmonic approximation to allow analytical calculations. If $\rho \neq 0$, the value of the coupling constant $K$ entering in $S$ should actually be $K' = K(1 + \rho)$ to get the best approximation. From this Hamiltonian one can derive dimensionless equations of motion, which depend on a single parameter $S$.

In the continuum limit approximations, these equations become

$$\frac{\partial^2 Y}{\partial t^2} - S\frac{\partial^2 Y}{\partial x^2} + \frac{\partial V(Y)}{\partial Y} = 0 \qquad \text{with } V(Y) = (e^{-Y} - 1)^2. \tag{5}$$

This equation has the exact static solution

$$Y(x) = \ln\left[1 + e^{\sqrt{2/S}(x - x_0)}\right], \tag{6}$$

where $x_0$ is an integration constant that determines the position of the solution. It describes a configuration where one part of the molecule ($x < x_0$) is closed, while for $x \gg x_0$ the base pair separation grows linearly with space and the molecule is fully denaturated. It corresponds to a *domain wall* between two states of the DNA molecule. Let us evaluate the energy of this solution for a finite chain of $N$ base pairs. Sites with an index smaller than $x_0$ are such that $Y \simeq 0$. The Morse potential and the coupling energy between adjacent sites vanish. For sites with an index larger than $x_0$, $Y \gg 1$ so that the Morse potential takes the value $+1$ while $dY/dx \simeq \sqrt{2/S}$ corresponds to a coupling energy $\frac{1}{2}S(\sqrt{2/S})^2 = 1$. Therefore each site with an index larger than $x_0$ contributes to the energy by $e = 2$. As a result the dimensionless energy of the domain wall is

$$E_P^+ = 2(N - x_0) + \mathcal{O}(N^0), \tag{7}$$

where the term $\mathcal{O}(N^0)$ corresponds to the core of the wall ($x \simeq x_0$) where $Y$ evolves smoothly from the bottom of the Morse potential towards the plateau. In the limit $N \to \infty$ the energy of the domain wall becomes infinite. For finite $N$, one can notice that the energy of the wall gets smaller if $x_0$ increases, i.e. if the closed region of the molecule extends. The solution (6) is thus unstable. It tends to move in the direction that closes the base pairs. This is not surprising because, if thermal effects are not taken into account, the stable state of DNA is the closed double helix. This domain wall solution can be used to compute the thermal denaturation temperature of DNA [14, 18], but it can also be used to analyse the mechanical opening. An external force on the $N$th site can stabilise the domain wall in the open position, and a force exceeding this equilibrating force will move the domain wall towards opening.

The domain wall solution can be used to scan the energy barrier for opening as follows. We *assume* a given domain wall shape, according to Eq. (6) and we move it within the sequence. For each position its energy is calculated. The energy

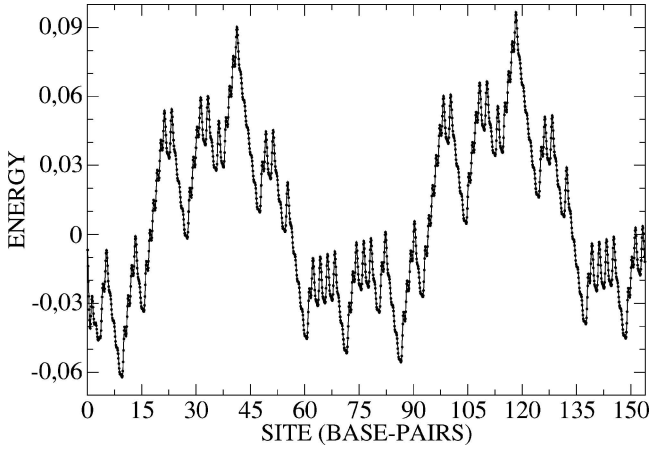


*Figure 8.* Variation of the energy of a domain wall given by Eq. (6) with $S = 0.5$ when it is moved along the sequence shown in Figure 2. The average slope (Eq. (7)) has been removed.

shows a general trend according to Eq. (7), i.e. it increases as the size of the open region increases. But, in addition, it varies from site to site because the sequence is not homogeneous. Removing the average trend, we get thus a pattern of the energy required to open each part of the molecule. Figure 8 shows the pattern that we get by selecting $S = 0.5$.

This pattern is remarkably correlated to the thermal opening pattern of this DNA sequence. It appears as a "negative" of the thermal opening pattern: the regions which open easilly thermally are also the regions where the motion of the domain wall requires less energy. In particular the promoter region appears as a minimum when we scan the sequence with the domain wall. This approach appears as a method which could be used to quickly probe a sequence. Basically it provides an averaging process to evaluate the opening probability, but instead of being based on some arbitrary weighting, it uses a physical argument to determine the weighting. This weighting is still approximate because it uses a solution valid for a homopolymer to probe a molecule with an actual sequence. Moreover the analytical expression of the domain wall is obtained in the case of a harmonic coupling. The simulations are performed with the anharmonic coupling but, as the bases are not wide open in the core of the wall, the effective coupling constant $K'$ can be used in the definition of $S$. In spite of these approximations, the domain wall gives the correct qualitative shape of the opening. It should however be noticed that we used a value for $S$ which is significantly higher than the one that would be derived from the model parameters: for a pure AT sequence $S_{\mathrm{AT}} = K'/(D_{\mathrm{AT}}\alpha_{\mathrm{AT}}) = 0.068$ and for a GC sequence $S_{\mathrm{GC}} = K'/(D_{\mathrm{GC}}\alpha_{\mathrm{GC}}) = 0.016$. When the calculation is performed with such very small values of $S$ the landscape of energy obtained by moving the domain wall over the sequence is extremely spiky, showing many small

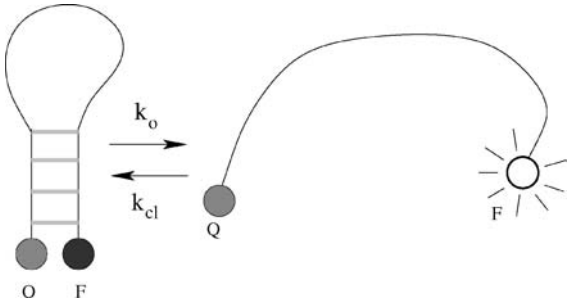


*Figure 9.* Schematic view of the opening-closing fluctuations of a DNA hairpin. A fluorophore (F) and a quencher (Q) can be used to monitor these two conformations. In the open conformation the fluorophore is far enough from the quencher to be actually fluorescent.

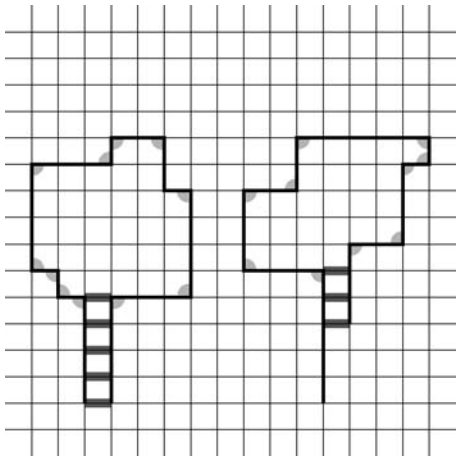

*Figure 10.* Two configurations of the hairpin model of a lattice. The DNA strand is indicated by the thick line on the lattice. The hydrogen bonds are marked by the thick bonds connecting two points of the stand, and the shaded corners represent the bending energy contributions. The left case corresponds to the perfect closing, while the right figure shows an example of a mismatched partial closing.

peaks. A larger value of $S$ corresponds to a broader wall and it can be viewed as a phenomenological way to take into account the thermal fluctuations which broaden the interface between the closed and open region (Figures 9 and 10).

This calculation amounts to computing an effective potential for the domain wall, in the spirit of the collective coordinate approach which has been used in soliton models of DNA [19]. Of course the calculation can be improved if we rely on a numerical solution in the inhomogeneous DNA. For each position of the domain wall it is possible to numerically relax its shape before computing its energy, and moreover the thermal fluctuations can be taken into account in the process, but such an approach loses part of the interest of the method because it involves heavy computations.

## 3. Self Assembly of DNA Hairpins

### 3.1. MODEL

DNA hairpin are short nucleotide chains which have, in their two terminating regions, complementary bases which can therefore self assemble to form a short double helix called the stem of the hairpin. They can exist in two states, the open and the closed state, and fluctuate between the two, being mostly closed at low temperature and mostly open at high temperature [5]. Using DNA chains which carry a fluorophore at one end and a quencher at the other end, it is possible to detect the state of the hairpin, which is only fluorescent in its open state. Hairpins are interesting both for the physicist and the biologists. For the physicist they provide a simple system to study the self assembly of DNA with two pieces of strand which are maintained in the vicinity of each other by the loop of the hairpin, so that they can easily find each other for the assembly. For the biologists they may provide very sensitive probes for short DNA sequences: a loop which is complementary to a sequence to recognise can self assemble with it. This prevents the hairpin from closing and it is detected by fluorescence.

Modelling the fluctuations of a hairpin is more challenging than modelling the mechanical denaturation of DNA for two reasons:

– the self assembly of a structure is not simply the reverse process of its opening because the elements must find each other is space and then orient properly with respect to each other, before actually assembling in a final stage which is the only stage of the process which can be viewed as the reverse of the breaking;
– the time scales for the assembly can be very long (hundreds of $\mu$s for instance), i.e. many orders of magnitude longer than the typical time scale of the microscopic dynamics of a macromolecule.

For these reasons, even the simple DNA model that we introduced in the previous section cannot be used for this investigation. We shall introduce a model which is even simpler, and examine to what extent it is valid to study such a problem.

Our hairpin model is inspired by the lattice models which have been used to study protein folding. It is a lattice model so that only discrete motions are allowed, thus it cannot describe the true dynamics of the hairpin. Instead we use a Monte-Carlo dynamics where the moves are discrete and determined by their probability at the temperature of the simulation, depending on their energy cost or gain. To carry such a calculation we only have to specify the energy of the model in each configuration. As a first approach to this problem we decided to choose the simplest underlying lattice, a planar square lattice. The interest is that it restricts the number of accessible states with respect to a more complex three-dimensional lattice, but, as discussed below, this introduces of course some restrictions on the ability of the model to describe actual hairpins.

The energy of the DNA strand is assumed to depend on two terms only, a bending energy which appears when two consecutive segments are at some angle, and the energy of the base pairs which can form in the stem. The total number of nucleotides

in the DNA strand is denoted by $N$. The number of nucleotides which can form the stem is denoted by $n_s$. In order to specify the kind of pairing allowed in the stem, each nucleotide of the stem, denoted by index $j$ is affected of a "type" $t_j$. Only two nucleotides having the same "type" are allowed to form a base pair by hydrogen bonding. Thus, rather that actually specifying the type of a base ($A$, $T$, $G$, $C$) we specify the type of pairing that it can form. The energy of the model is written as

$$E = n_A E_A + \frac{1}{2} \sum_{j=1}^{n_s} \sum_{j'=1}^{n_s} e(j, j') \tag{8}$$

$$e(j, j') = \delta(t_j - t_{j'})\delta(d_{jj'} - 1)a(j)a(j')E_{HB}(t_j) \tag{9}$$

where

- $n_A$ is the number of angles in the DNA strand on the lattice, and $E_A$ is a positive model parameter giving the energy costs of a bent. In some calculations, $E_A$ may be different for a bent in the stem or in the loop.
- $e(j, j')$ is the pairing energy between nucleotides $j$ and $j'$ of the stem. The factor $\delta(t_j - t_{j'})$ enforces the condition that the two nucleotides should be of the same "type", $\delta(d_{jj'} - 1)$ indicates that the pairing is only possible if the two nucleotides are adjacent on the lattice. The factors $a(j)$ and $a(j')$ are equal to 1 only if the nucleotide is available for pairing, i.e. if it is not already involved in another pair. Otherwise the pairing is not formed and they are set to 0. They are necessary because some geometries of the chain could put a nucleotide in a position adjacent to two sites occupied by nucleotides of the same type. Finally $E_{HB}(t_j)$ is the pairing energy for nucleotides of type $t_j$. It is a negative quantity, which means that the pairing is favourable because it lowers the energy of the hairpin.

We studied this model using Monte Carlo simulations in the same spirit as the studies performed on lattice models of proteins, i.e. we generate a random walk of the DNA chain on the lattice with the condition that the system should be in thermal equilibrium at temperature $T$. A configuration of energy $E$ must therefore have a probability proportional to $\exp(-E/T)$, where $T$ is measured in units of energy. If the moves are selected in order to stay as close as possible to the actual motion of a polymer in a fluid, the method can even be used to study dynamical effects with a fictitious time scale which is simply given by the number of Monte Carlo steps [20]. For this reason we selected only local motions of the chain. On the two-dimensional square lattice, there are only three such motions: the change of the angle between the two segments at one end of the chain, the flipping of a corner of a lattice cell with respect to the diagonal of the cell and a crank mechanism. If it does not lead to a clash with another part of the chain, an attempted motion is accepted with probability $P = \min[\exp(-\Delta E/T), 1]$, where $\Delta E = E_2 - E_1$ is the difference between the energy after and before the move, using a standard Metropolis algorithm.

## 3.2. EQUILIBRIUM PROPERTIES OF THE OPENING-CLOSING TRANSITION

### 3.2.1. *The Transition in the Absence of Mismatch*

Let us consider first the *equilibrium* properties of DNA hairpins in the simple case when they can only close with a correct matching of the bases in the stem. This would be the case if the base sequence in the stem forbids any mismatch. In order to compare with experimental results [21] we considered the case of a stem having 5 base pairs ($n_s = 5$). Since there are only 4 types of bases, at least one has to appear twice in the stem. Thus the Watson-Crick pairing rules allow at least one mismatched pairing, but it may be very unfavourable because, if it occurred, the other bases of the stem would not be paired and may even experience some steric hindrance. In the model it is easy to strictly forbid any mismatched closing by using a sequence $t_i = \{1, 2, 3, 4, 5\}$ where all base pairs have different types. Besides this condition, in our calculations we gave same energy $E_{\mathrm{HB}} = -1$ to all types of base pairs. This value sets the energy scale, and thus the temperature scale. With these parameters, the model does not attempt to mimic any real DNA hairpin, but it is designed to stay as simple as possible in order to exhibit the basic mechanisms that govern the hairpin properties.

Figure 11 shows the variation of the number of hydrogen-bonded base pairs versus temperature for chains having different numbers $N$ of nucleotides. The number of nucleotides in the loop is $N - 10$ since the stem is always made of two segments of 5 nucleotides. In these calculations, the bending energy $E_A$ has been set to $E_A = 0.02$, and it has the same value along the whole DNA strand. The results have been obtained with different initial conditions: we start either from a closed hairpin or a random coil. Each point in the figure is an average of 100 calculations with different sets of random numbers to generate the initial conditions and the stochastic motions of the chains on the lattice, each calculation involving between

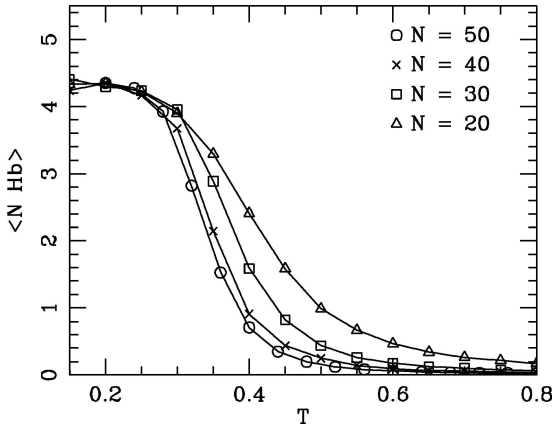


*Figure 11.* Variation versus temperature of the number of hydrogen-bonded pairs in the stem for hairpins of different lengths $N$, in the absence of mismatches.

$4 \times 10^8$ and $8 \times 10^8$ Monte Carlo steps (depending on temperature and chain length). The first $2 \times 10^7$ steps are discarded in the analysis to allow the model to equilibrate to the selected temperature. For $T \geq 0.15$ a good equilibration is achieved, while results at lower temperatures show some dependence on the initial conditions because an equilibrium state has not been reached. This is why they are not shown in Figure 11.

As expected, when temperature increases we observe a fairly sharp decrease of the number of hydrogen-bonded base pairs. It corresponds to the opening of the hairpin, which occurs over a temperature range of about 0.2 energy units, around the so-called "melting temperature" $T_m \approx 0.35$, which is well below the the temperature $T = 1$ corresponding to the binding energy of a base pair. This indicates that the entropy gain provided by the opening of the hairpin contributes to lower the free energy barrier for opening. Increasing the length of the loop lowers $T_m$, in agreement with the experiments [21]. It also makes the transition sharper, which is not observed in the experiments.

The role of the rigidity of the loop can be tested by changing the value of the bending energy $E_A$ for all the bends in the loop, without changing its value in the stem. Figure 12 shows that a more rigid loop leads to an opening at lower temperature, in agreement with the experimental observations [21]. However the variation of $T_m$ given by the model appears to very small, and moveover, as discussed below, the effect of the rigidity of the loop on the thermodynamics of the hairpin is not

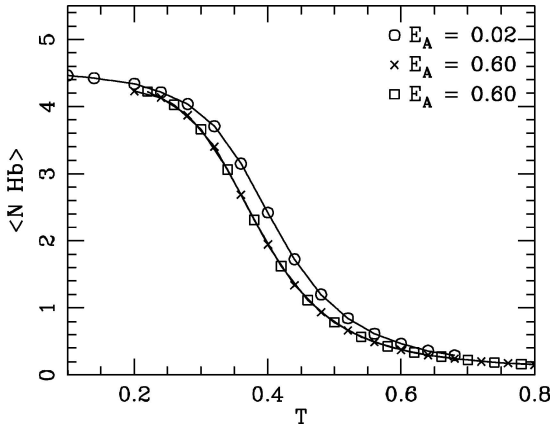


*Figure 12.* Effect of the rigidity of the loop on the opening of the hairpin: variation versus temperature of the number of hydrogen-bonded pairs in the stem for loops with different bending energies $E_A = 0.02$ and 0.60, in the absence of mismatches. In the stem the bending energy has been set to $E_A = 0.02$ for both calculations. The two sets of points for $E_A = 0.6$ (crosses and squares) have been obtained in two independent calculations, with different sets of temperatures and different initial conditions. The crosses show results obtained with a closed hairpin initial condition, while the squares have been obtained with random initial conditions. Each point on this figure is an average over 100 sets of initial conditions and random numbers.

correctly described in our model. This points out some limitations of the simplified model, although a quantitative comparison with the experiments is difficult because, in the experiments, the rigidity was varied by changing the bases from *T* to *A*. The larger purine bases *A* are assumed to give a higher rigidity to the strand but this could only be related to the variation of $E_A$ by extensive all-atom numerical simulations [22]. Moreover, the role of base stacking in the loop is certainly more complex than the simple change of the rigidity of the chain that our simplified model can describe.

### 3.2.2. *Role of the Mismatches*

One feature of DNA hairpins is that, unless they have a specifically designed sequence, they may close with a wrong pairing in the stem (see Figure 10). These imperfect, mismatched, closings have a higher energy that the perfectly closed hairpin, but they can be very long-lived.

They affect the opening-closing transition as shown in Figure 13 which compares the melting curves in the presence and in the absence of mismatches. In order to allow mismatches, the sequence of bases of the stem has been set to $t_i = \{1, 1, 1, 1, 1\}$, i.e. all base pairs are of the same type so that many mismatched pairings are possible, with 1, 2, 3, 4 hydrogen-bonded base pairs. In this case we show the mean value $\langle d \rangle$ of the distance between the first and last nucleotide of the chain rather than the number of hydrogen-bonded stem base pairs because $\langle d \rangle$ provides a more complete picture of the configuration of the hairpin.

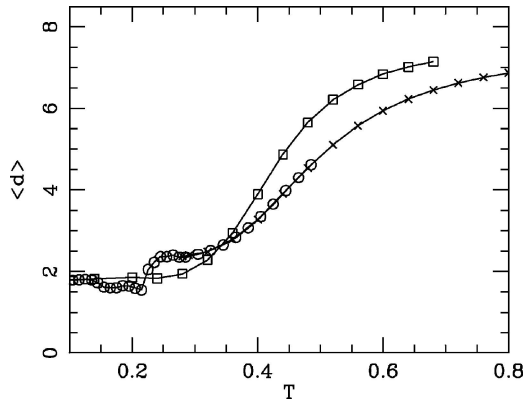


*Figure 13.* Comparison of melting curves with and without mismatches. The mean value $\langle d \rangle$ of the distance between the first and last nucleotide is plotted versus temperature. The chain has $N = 20$ nucleotides, with $E_{HB} = -1$ for all base pairs of the stem, $E_a = 0.02$. The squares show data without mismatch ($t_i = \{1, 2, 3, 4, 5\}$), while the circles and crosses show data with mismatches ($t_i = \{1, 1, 1, 1, 1\}$). In this case two sets of calculations have been performed. The circles have been obtained with $8 \times 10^8$ Monte Carlo steps, while the crosses involve only $4 \times 10^8$ Monte Carlo steps. For $T > 0.25$ the two sets give identical results, but, at low *T*, a smaller number of Monte-Carlo steps slightly affects the results.

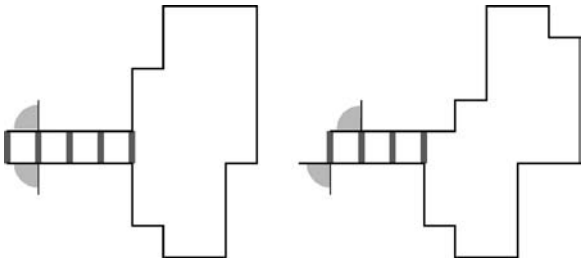

*Figure 14.* Schematic plot of the fluctuations of the free end of the chain in a perfectly closed state (left) and in a mismatched state (right).

In Figure 13, the case without mismatch shows a smooth melting curve, similar to the results of Figure 11. In the low temperature domain where the hairpin is closed, $\langle d \rangle$ is larger than the value $\langle d \rangle = 1$ that could be expected from a static image of the closed hairpin because there are fluctuations. They are particularly important at the free end of the stem, as schematised on Figure 14.

When mismatches are allowed, the curve $\langle d(T) \rangle$ shows a fairly sharp kink around $T = 0.215$, and then an increase, qualitatively similar to cases without mismatch, but occurring however more more smoothly and at higher temperature. The kink, which corresponds to a jump of $\langle d \rangle$ of about one unit, is due to the formation of a mismatched closing where only 4 base pairs of the stem are formed (Figure 14, right part). As temperature is raised further, the number of paired bases in the stem keeps decreasing, but, as there are many more possibilities for binding than in the no-mismatch case, the opening of the hairpin is more gradual.

### 3.3. KINETICS OF THE OPENING AND CLOSING

Up to now we spoke of the opening transition of the hairpin as if the hairpin should be closed at low $T$ and open at high $T$. It is actually more complex because, in a small system like the hairpin, a phase transition between two states does not exist. Actually we always have a equilibrium between the open form $O$ and the closed form $C$

$$C \underset{k_o}{\overset{k_{cl}}{\rightleftarrows}} O, \tag{10}$$

which can be studied like a chemical equilibrium rather than a phase transition. At low $T$ the equilibrium is displaced towards closing and at high $T$ it is displaced towards opening.

This suggests that the methods of chemical kinetics can be used to analyse the dynamics of the fluctuations of the hairpin. Let us consider that the hairpin is a two-state system. This is obviously an approximation which becomes very crude when mismatches are allowed since, in this case, the hairpin can also exist in some

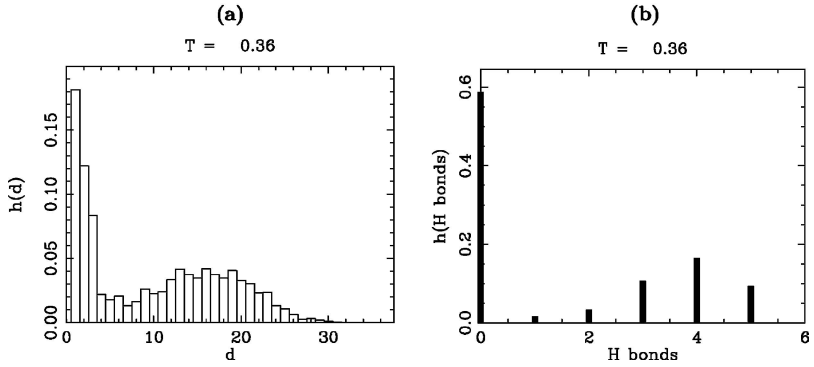


*Figure 15.* Normalised histograms of the distance $d$ between the two ends of the chain (a), and number of hydrogen bonds (b) for a hairpin with $N = 50$ and no mismatches, at temperature $T = 0.36$. This temperature is close to the opening temperature $T_m$ of this hairpin. Model parameters $E_{\mathrm{HB}} = -1$, $E_a = 0.02$. The histograms show the coexistence of two populations: one population of completely open hairpins (large values of $d$ and 0 hydrogen bonds) and a population of closed hairpins in which some of the hydrogen bonds are formed, the highest probability being with 4 hydrogen bonds formed.

intermediate states where it is incompletely closed. In the absence of mismatch, the two-state picture is a satisfactory approximation, as shown in Figure 15. This figure shows the histogram of the distance $d$ between the two ends of the chains, and the histogram of the number of hydrogen-bonded base pairs at temperature $T = 0.36$ for a model without mismatch with $N = 50$. This temperature is close to the melting temperature $T_m$ for this model, and the histograms clearly show the coexistence of two populations of states: (i) an open state, where there are no hydrogen-bonded pairs in the stem, which corresponds to the hump for $d > 5$ on Figure 15(a), (ii) a closed state corresponding to the sharp maximum for $d < 4$ in Figure 15(a) and to the existence of 2 to 5 hydrogen-bonded base pairs in Figure 15(b) (with a maximum at 4, due to the opening fluctuations at the end of the stem as discussed above and schematised in Figure 14, left).

The two-state picture allows us to write standard kinetic equations for the populations $[C]$ and $[O]$ of the closed and open states as

$$\frac{d\,[C]}{dt} = -k_o[C] + k_{cl}[O] \tag{11}$$

$$\frac{d\,[O]}{dt} = +k_o[C] - k_{cl}[O], \tag{12}$$

where $k_0$ and $k_{cl}$ are the kinetic constants for the opening and closing events respectively. This system has the solution

$$[C](t) = \frac{C_0 k_o}{k_o + k_{cl}} e^{-(k_o + k_{cl})t} + \frac{C_0 k_{cl}}{k_o + k_{cl}}, \tag{13}$$

where $C_0$ is the value of $[C]$ at time $t = 0$. This shows that, if we start from a population of closed hairpins, we expect it to decay exponentially with a characteristic time $\tau = 1/(k_o + k_{cl})$ until an equilibrium is reached with

$$\frac{[O]}{[C]} = \frac{k_o}{k_{cl}} = K_e, \tag{14}$$

where $K_e$ is the equilibrium constant.

Therefore, if we follow the evolution of the population of closed hairpins in a Monte Carlo simulation which starts from $C_0$ closed configurations, we can determine separately $\tau$ (from the decay of the closed population) and $K_e$ from the final equilibrium state, so that we can determine the kinetic constants for opening and closing, given by

$$k_o = \frac{1}{\tau}\frac{1}{1+K_e} \quad k_{cl} = \frac{1}{\tau}\frac{K_e}{1+K_e} \tag{15}$$

Figure 16 shows the results of such an analysis for a case without mismatches. The open/closed state of the chain was measured with two different criteria: from the distance $d$ between the two ends (a value $d > 4$ is considered as an open state) or from the number of hydrogen-bonded base pairs (an open state must not have any bound base pair). Both give very similar results, in agreement with the above discussion of Figure 15 which shows that both criteria can be used to separate between the open and closed states. When they are plotted in logarithmic scale

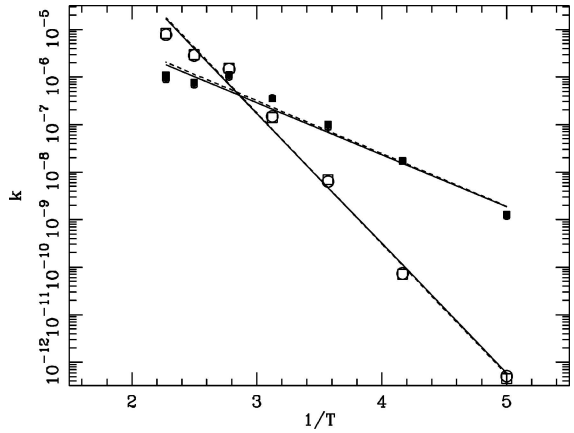


*Figure 16.* Arrhenius plot of the kinetic constants $k_{op}$ (open symbols) and $k_{cl}$ (closed symbols) versus $1/T$ for a model without mismatch, $N = 50$, $E_{\mathrm{HB}} = -1$, $E_a = 0.02$. The time unit is a Monte-Carlo step. The lines are least square fits of the points (full lines for opening state defined by $d > 4$, and dashed lines for opening defined by the absence of hydrogen bonded base pairs).

versus $1/T$, the kinetic constants are well fitted by straight lines, which allows us to define activation energies $E_o$ and $E_{cl}$ for the opening and closing events by

$$k_o = K_o e^{-E_o/T} \quad k_{cl} = K_{cl} e^{-E_{cl}/T} \tag{16}$$

The fits of Figure 16 give $E_o = 6.3$ and $E_{cl} = 2.5$. Figure 16 is very similar to the figures showing $k_0$ and $k_{cl}$ which can be obtained experimentally [5]. The experiments also find an opening activation energy much larger than the closing energy. The experimental ratio $E_o/E_{cl}$ is even larger than the ratio that we derive from our model. Owing to the simplicity of the model, it would be meaningless to try to adjust parameters to get the experimental ratio. What is more interesting is the meaning of this result $E_o \gg E_{cl}$, which can be related to the need to break the hydrogen bonds linking the base pairs to open the hairpin, while the kinetic of the closing is dominated by entropic effects because it occurs when the two sides of the stem managed to reach the correct spatial position after a random walk in the configuration space.

Experiments show that the opening kinetics is almost insensitive to the length of the loop, while the closing slows down significantly when the length of the loop increases ($k_{cl}$ decreases) while its activation energy does not depend on the length of the loop. The model confirms that the activation energies do not vary when we change $N$, but it only finds a very small variation of $k_{cl}$ as a function of $N$, contrary to the experiments. This points out one of its severe limitations: the entropy of the loop is not sufficiently well described when its motions are constrained on a two-dimensional square lattice. This limitation also appears when we study the effect of the rigidity of the loop. As noticed above, the effect is very small and to obtain some noticeable influence of the rigidity, we have to increase the bending energy very significantly, for instance up to $E_A = 0.6$ (Figure 12). In this case the activations energies become $E_o = 5.5$ and $E_{cl} = 2.5$, i.e. the opening activation energy is reduced by about 12% and the closing energy is only weakly affected, while the experiments found a large increase of the closing activation energy and almost no change for $E_o$. This shows that, for this study, our model does not correctly describe the experiment. Besides an incorrect description of entropic effects in the model, that we already mentioned above, other phenomena could enter, and particularly a possible role of the mismatches in the experimental sequence. While the model strictly forbids mismatches, in the experiments, changing the bases in the loop from $A$ to $T$ modifies the possible mismatches.

As one could expect, the kinetics of the hairpin fluctuations is strongly affected by the presence of mismatches. The two-state approach is no longer valid. Mismatched states are open if we define them in terms of the distance between the ends but still show many hydrogen-bonded base pairs. Although the time evolution of the closed states is no longer a simple exponential decay, an approximate fit by an exponential gives the order of magnitude of the characteristic time $\tau$. Figure 17 shows the values of $\tau$ determined with two definitions of an open state: (i) a state

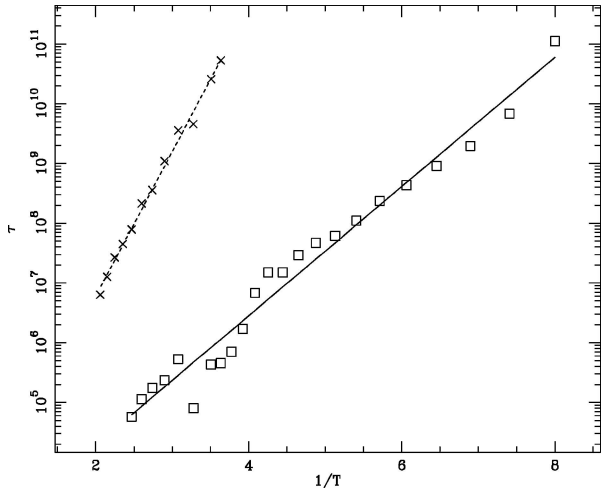


*Figure 17.* Logarithmic plot of the characteristic time for opening $\tau$ versus $1/T$ for a case with mismatches. The squares (fitted by the full line) correspond to a definition of the opening from the distance of the two ends ($d > 2$) and the crosses (fitted by the dashed line) define opening by the absence of any hydrogen-bonded base pair. The time unit is a Monte-Carlo step.

where the distance of the two ends of the chain is $d > 2$, (ii) a state where all the hydrogen bonds linking the bases in the stem have been broken. Figure 17 shows that the lifetime of closed hairpins defined according to these criteria vary by several orders of magnitude. This is not surprising because a hairpin which is closed in a mismatced state may be counted for open for the first criterion ($d > 2$) but closed with respect to the second one since some of its base pairs are hydrogen bonded. In this case the above analysis to calculate $k_o$ and $k_{cl}$ loses its meaning.

The role of the mismatches in the experimental studies of molecular beacons [5] has not been investigated so that we cannot compare the results of the model with experimental data. Although the sequence used in [21] could in principle allow wrong closing, there were certainly much less likely than in our study where all base pairs of the stem are the same. Moreover, studies using a fluorophore and a quencher are only probing the distance $d$ between the ends of the chain, so that they are not sensible to wrong closings. For such a study the hairpin is still a two-state system.

## 4. Conclusion

We examined two approaches to study DNA at the scale of a nucleotide rather than the atomic scale considered in molecular dynamics simulations.

The model of Section 2 is similar to molecular dynamics because it solves the equations of motions of a system of units interacting through given potential terms. The difference lies in the units, which are nucleotides instead of atoms, but also

in the interaction potentials which are specifically designed to properly describe the interactions between complex objects. This is particularly noticeable for the stacking potential. It is not expressed, as usual, in terms of the distance between the units since it includes the differences of their coordinates ($(y_n - y_{n-1})$, which is the one-dimensional distance), but also the sum $y_n + y_{n-1}$ which is necessary to express that, if any of the two bases which interact belongs to a broken pair, the interaction decreases. Establishing such potentials is crucial to the success of a mesoscale approach, and the comparison with experiments is essential to determine their validity. For the dynamical DNA model that we discussed in Section 2, studies of the thermal denaturation of DNA were very important to establish and validate the model.

The determination of appropriate parameters is one major difficulty of mesoscopic models because the interactions that they describe are effective interactions, actually involving many local interactions. For instance the potential $V(y)$ which links the bases in a pair is not only determined by the hydrogen bonds linking the bases. The repulsion between the negatively charged phosphate groups, screened by counter ions coming from the surrounding solution, is also essential, but it is hard to calculate from first principles. This is why single molecule experiments are very helpful to calibrate the models. They provide data which complement the thermodynamic studies, such as the investigation of DNA thermal denaturation, which have been available for decades and have been very useful in the derivation of parameters for the Ising-like models of DNA. Single molecule experiments are useful first because their results are not averaged by statistics over $10^{23}$ (or much more) molecules, and they may also give dynamical quantities. Another difficulty of mesoscale simulations is the proper description of thermal fluctuations with an appropriate thermostat. A simple Langevin or Nosé thermostat may be too crude to describe for instance what happens to a base when it is inside the DNA stack, i.e. weakly coupled to the surrounding fluid, or, on the contrary in an unstacked position where it should be strongly coupled to the surrounding.

We have shown that the model can give interesting results to study the mechanical denaturation of DNA. There are however many open questions, that complementary studies using experiments and modelling may answer. One of them is the correlation between the sequence and the opening probability. Although it is related to the ratio between the AT and GC pairs in the sequence because AT pairs are easier to break, the full picture is more complex. We have shown for instance that a region with 5 AT pairs may open more easily than another which has 6 consecutive AT pairs. Such effects, which are also observed in experimental and theoretical studies of the opening of DNA in the vicinity of promoters [2], are still not understood.

The problem of the self assembly of DNA hairpin is very different because, in this case, we are interested in events which occur on time scales of nanoseconds to microseconds or even longer. In this case we gave up dynamical simulations and used a Monte Carlo simulation which amounts to a stochastic exploration of the phase space. The two-dimensional lattice model that we introduced is extremely

simple, but we showed that it nevertheless captures some important aspects of the fluctuations of DNA hairpins such as the variation of the melting temperature with the length of the loop and its rigidity, and can even study some kinetic properties and detect the large difference between the closing and opening activation energies. This tells us that these properties of the hairpin are not due to peculiarities of the DNA structures, but are on the contrary general properties of a polymer chain that can form a hairpin by establishing bonds between two terminal regions.

However this model suffers from an incorrect evaluation of the entropy of the loop. This is a problem of the lattice model, but it is particularly acute for a two-dimensional square lattice.

In both cases, the interest of such models is not their ability to reproduce the reality, because this would not tell us very much about DNA. This ability should only be viewed as a test of the validity of the model, which can then be used to explore some properties which are hardly accessible to experiments. For instance the dynamical model of DNA can explore the details of the fluctuations of the molecule as a function of its sequence on a scale of a few tens to a few hundreds of base pairs. This may provide a tool to analyse the sequence which completes the static analysis, or the thermodynamics studies which have shown their interest for very long sequences (tens of thousands of base pairs) but do not have a sufficient resolutions to study a few tens of bases. Similarly the hairpin model could be used to study the role of the mismatches on the fluctuations for instance.

## Acknowledgments

Part of this work have been supported by the project BFM 2002-00113 DGES (Spain), and the Aragon Government (DGA – Group of Non Linear and Statistical Physics). S. Cuesta-López is supported by the Spanish Ministry of Science and Education (FPU-AP2002-3492).

## References

1. Lavery, R.: Modelling the DNA Double Helix: Techniques and Results, in M. Peyrard (ed.), *Nonlinear Exciations in Biomolecules*, Editions de Physique/Springer-Verlag, Les Ulis, 1995.
2. Choi, C.H., Kalosakas, G., Rasmussen, K.O., Hiromura, M., Bishop, A.R. and Usheva, A.: DNA Dynamically Directs Its Own Transcription Initiation, *Nucleic Acid Research* **32** (2004), 1584–1590.
3. Essevaz-Roulet, B., Bockelmann, U. and Heslot, F.: Mechanical Separation of the Complementary Strands of DNA, *Proc. Natl. Acad. Sci. USA* **94** (1997), 11935–11940.
4. Danilowicz, C., Coljee, V.W., Bouzigues, C., Lubensky, D.K., Nelson, D.R. and Prentiss, M.: DNA Unzipped under a Constant Force Exhibits Multiple Metastable Intermediates, *PNAS* **100** (2003), 1694–1699.
5. Bonnet, G., Krichevsky, O. and Libchaber, A.: Kinetics of Conformational Fluctuations in DNA Hairpin-Loops, *Proc. Natl. Acad. Sci. USA* **95** (1998), 8602–8606.

6. Viovy, J.-L., Heller, C., Caron, F., Cluzel, P. and Chatenay, D.: Séquençage de l'ADN par ouverure mécanique de la double hélice: une évaluation théorique, *C.R. Acad. Sci. Paris, Sciences de la Vie/Life sciences* **317** (1994), 795–800.
7. Peyrard, M.: Using DNA to Probe Nonlinear Localised Excitations? *Europhys. Lett.* **44** (1998), 271–277.
8. Lubensky, D.K. and Nelson, D.R.: Single Molecule Statistics and the Polynucleotide Unzipping Transition, *Phys. Rev. E* **65** (2002), 031917-1-25.
9. Essevaz-Roulet, B., Bockelmann, U. and Heslot, F.: Mechanical Separation of the Complementary Strands of DNA, *Proc. Natl. Acad. Sci. USA* **94** (1997), 11935–11940.
10. Wartell, R.M. and Benight, A.S.: Thermal Denaturation of DNA Molecules: A Comparison of Theory with Experiments, *Physics Reports* **126** (1985), 67.
11. Peyrard, M. and Bishop, A.R.: Statistical Mechanics of a Nonlinear Model for DNA Denaturation, *Physical Review Letters* **62** (1989), 2755–2758.
12. Dauxois, T., Peyrard, M. and Bishop, A.R.: Entropy Driven DNA Denaturation, *Physical Review E* **47** (1993), R44–R47.
13. Campa, A. and Giansanti, A.: Experimental Tests of the Peyrard-Bishop Model Applied to the Melting of Very Short DNA Chains, *Phys. Rev. E* **58** (1998), 3585–3588.
14. Dauxois, T., Theodorakopoulos, N. and Peyrard, M.: Thermodynamic Instabilities in One Dimension: Correlations, Scaling and Solitons, *J. Stat. Phys.* **107** (2002), 869–891.
15. Theodorakopoulos, N., Dauxois, T. and Peyrard, M.: Order of the Phase Transition in Models of DNA Thermal Denaturation, *Phys. Rev. Lett.* **85** (2000), 6–9.
16. Martyna, G.J., Klein, M.L. and Tuckerman, M.: Nosé-Hoover Chains: The Canonical Ensemble via Continuous Dynamics, *J. Chem. Phys.* **97** (1992), 2635–2643.
17. Cuesta-Lopez, S. and Peyrard, M. unpublished.
18. Theodorakopoulos, N. Peyrard, M. and MacKay, R.S.: Nonlinear Structures and Thermodynamic Instabilities in a One-Dimensional Lattice System, *Phys. Rev. Lett.* **93** (2004), 258101-1-4.
19. Cuenda, S. and Sanchez, A.: Nonlinear Excitations in DNA: Aperiodic Models Versus Actual Genome Sequences, *Phys. Rev. E* **70** (2004), 051903-1-8.
20. Landau, D.P. and Binder, K.: *Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, 2000.
21. Goddard, N.L., Bonnet, G., Krichevsky, O. and Libchaber, A.: Sequence Dependence Rigidity of Single-Stranded DNA, *Phys. Rev. Lett.* **85** (2000), 2400–2403.
22. Cuesta-López, S., Peyrard, M. and Graham, D.J.: Model for DNA Hairpin Denaturation, *Eur. Phys. J. E.* (to be published).