Journal of Statistical Physics, Vol. 120, Nos. 3/4, August 2005 (© 2005)
DOI: 10.1007/s10955-005-5962-0

A Probabilistic Study of DNA Denaturation

Clément Dombry[1]

Received June 7, 2004; accepted May 4, 2005

We consider Benham's model for strand separation in supercoiled circular DNA. This is a mean field model in external inhomogeneous field, conditioned to small values of the perimeter. Under some conditions on the external field, we prove a large deviations principle for the distribution of the magnetization under the Gibbs measure. The rate function strongly depends on the structure of the external field. It allows us to prove a law of large numbers and to study denaturation as a function of the temperature and the superhelical density.

KEY WORDS: Statistical mechanics; inhomogeneous field; large deviations principle; DNA denaturation.

1. INTRODUCTION

Fundamental biological mechanisms such as replication and transcription of DNA require the two strands of the DNA double helix to separate. The separation of the two strands – called denaturation – can be partial or total. Denaturation depends on several factors such as the temperature, the composition of the DNA sequence and the geometry of the DNA polymer. Benham proposes a mathematical model for the process of denaturation (see refs. 2–4). His model is based on statistical mechanics ideas and takes into account the temperature, the nature of the bases and the superhelicity of DNA. He also develops algorithms to locate the regions where the denaturation is highly susceptible to occur. In ref. 13, the author computes the thermodynamic limit of Benham's model in the homopolymer case, and for some special types of heteropolymer. Denaturation only occurs in a few regions, and then eventually expands. Therefore Mazza proposes a

[1]Institut Camille Jordan, Université Claude Bernard – Lyon 1, 50, avenue Tony Garnier, Bâtiment RECHERCHE [B], Domaine de Gerland 69367 Lyon Cedex 07, France; e-mail: dombry@univ-lyon1.fr

modification of Benham's model focusing on configurations having a small number of 'denaturation bubbles'. He computes the thermodynamic limit of this new model for some types of DNA sequence and shows that the model exhibits phenomena of phase transition. The present work is motivated by extending these computations to more general heteropolymers. Identifying the cases, where it is possible is a large part of the work, and is connected with the notion of the structure of the DNA sequence.

We introduce Benham's model in a somewhat formal way, that is in the framework of one-dimensional spin model in inhomogeneous external field. Consider a circular graph with N sites, labeled successively $i = 1, \ldots, N$. This graph stands for a circular DNA heteropolymer of length N. This polymer is a double helix, each strand of the helix consists in N nucleotides. Each vertex represents a pair of nucleotides. At each site there is a spin denoted by σ_i taking values $\{-1, +1\}$. For convenience, we define the variables

$$n_i = \frac{1+\sigma_i}{2} = \begin{cases} +1 \text{ if } \sigma_i = +1, \\ 0 \quad \text{if } \sigma_i = -1, \end{cases} \qquad i = 1, \ldots, N.$$

The meaning of $n_i = 0$ is that the bases of the double helix at site i are linked by an hydrogen bond, the link is closed, and $n_i = 1$ means that this bond is broken, the link is open. Let $\Omega_N = \{-1, +1\}^N$ be the configuration space. We define some macroscopic quantities on the configuration space. The magnetization of a configuration $\sigma \in \Omega_N$ is defined by

$$M_N(\sigma) = \frac{1}{N} \sum_{i=1}^{N} n_i \in [0, 1].$$

The magnetization stands for the proportion of denatured bonds and measures the denaturation: $M_N = 0$ means that all links are closed, the DNA polymer is not denatured, $M_N = 1$ means that all links are open, the polymer is totally denatured. The field B_N is given by a sequence of reals $(b_i^N)_{1 \leqslant i \leqslant N}$. The value of the field at site i represents the energy of the bond between the bases located at site i. As a nucleotide pair is either $A+T$ or $G+C$, the field takes only two values denoted by b_{AT} and b_{GC}. The AT-links are formed of two hydrogen bonds and the GC-links consist in three hydrogen bonds so that $b_{GC} > b_{AT}$. The interaction of the spin system with the external field is measured by the localized magnetization

$$M_{NB_N}(\sigma) = \frac{1}{N} \sum_{i=1}^{N} b_i^N n_i.$$

It is a measure of the energy needed to break all the bonds and get the configuration σ. The perimeter of a configuration $\sigma \in \Omega_N$ is defined by

$$R_N(\sigma) = \frac{1}{2} \sum_{i=1}^{N} \mathbb{1}_{\sigma_i \neq \sigma_{i+1}}.$$

We make use of the circular boundary conditions: the site N is regarded as being followed by site 1, so that σ_{N+1} has to be seen as σ_1. This comes from the circular structure of the graph and ensures that the system is invariant under translation. Because of the periodic boundary conditions, the perimeter is an integer. It represents the number of connected domains of denatured bonds or 'denaturation bubbles'.

The Hamiltonian introduced by Benham is

$$H_N(\sigma) = a R_N(\sigma) + N F(M_N(\sigma), M_{NB_N}(\sigma)), \tag{1}$$

where F is the function defined by

$$F(m, \tilde{m}) = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m} \left(\kappa + \frac{m}{A}\right)^2 + \tilde{m}.$$

In this formula, κ is the superhelicity of the DNA polymer and is considered as a parameter and the other terms are biological constants. In Sun *et al.*, the following values are given: at 0.01 mol Na^+ concentration and temperature $T = 310\,\mathrm{K}$, $a = 10.5\,\mathrm{kcal/mol}$, $b_{AT} = 0.258\,\mathrm{kcal/mol}$, $b_{GC} = 1.305\,\mathrm{kcal/mol}$, $C = 3.6\,\mathrm{kcal/rad^2}$, $A = 10.4$, $K_0 = 2350$ RT with $R = 8.3146\,\mathrm{J/K/mol}$. For a discussion about this Hamiltonian, the reader should refer to the original works of Benham[2–4] or to Clote and Backofen.[8] In this Hamiltonian, we take into account the energy of nucleation initiation (through the constant a), the energy of AT or GC separation (through the constants b_{AT} and b_{GC}), the torsional or rotational free energy (through the constant C) and the free energy associated to the residual linking number (through the constant K_0).

Let ρ_N be the uniform probability measure on Ω_N. The Gibbs measure is defined by

$$\pi_N(\sigma) = \frac{1}{Z_N} e^{-H_N(\sigma)} \rho_N(\sigma),$$

where the normalization factor Z_N is the partition function defined by

$$Z_N = \int_{\Omega_N} e^{-H_N(\sigma)} \rho_N(d\sigma).$$

What we are interested in is the asymptotic behavior of the Gibbs measure when it is conditioned on untypical small values of the perimeter. For $r_N \in \mathbf{N}$, define:

$$\rho_{N,r_N} = \rho_N(\,.\,\mid R_N \leqslant r_N),$$
$$\pi_{N,r_N} = \pi_N(\,.\,\mid R_N \leqslant r_N).$$

We suppose that the typical equilibrium state of the DNA complex is distributed according to π_{N,r_N}. In ref. 13, the author shows that the small perimeter assumption ensures the possibility of phase transitions, that is the possibility for the existence of a stable and robust denatured state when the superhelical density is small enough. Conditioning on small values of the perimeter makes the Ising nearest-neighbour structure to disappear in the limit, at least for the aspects considered in this paper. So, we can consider the small perimeter condition as a slight modification of Benham's model that allows to compute the thermodynamic limit effectively. The justification of this assumption relies on the principle of equivalence of ensembles (see refs. 11 and 12) and references herein). Roughly speaking, this principle states that in the thermodynamic limit, the microcanonical measures and the grand canonical measures are equivalent. Hence we believe that the thermodynamic limit of the conditioned measures is equivalent to the thermodynamic limit of the Benham's measures with the value a equal to infinity. This is relevant since the biological constants satisfy $a \gg b_{AT}, b_{GC}$. In this paper, we present an extension of Mazza's results for more general heteropolymer.

Question. What is the asymptotic behavior of the magnetization M_N and the localized magnetization M_{NB_N} under the conditioned Gibbs measure π_{N,r_N}, when $r_N \ll N$?

In order to observe an asymptotic behavior for the localized magnetization M_{NB_N}, we have to impose that the external fields B_N converge in some sense. The most interesting feature of this study is that the influence of the field on the magnetizations M_N and M_{NB_N} is explicited. We obtain indeed a law of large numbers for the pair (M_N, M_{NB_N}), i.e. the pair converges to a deterministic limit denoted by $(M_\infty, \widetilde{M}_\infty)$. This limit can be evaluated and strongly depends on the external fields B_N. The term

M_∞ represents the limit proportion of broken links, i.e. the limit denaturation. The term $\widetilde{M}_\infty$ stands for the amount of energy needed to break all the bounds, it gives information on the localization of denaturation. In order to prove the law of large numbers, we study the large deviations properties of the magnetization (M_N, M_{NB_N}).We firstly prove a large deviations principle (LDP) for the pair, and then deduce the law of large numbers.

Our paper is organized as follows. Section 2 is devoted to the exposition of the results: we state and discuss the hypotheses, give some examples, state the LDP for the distribution of $(M_N, M_{NB_N}, R_N/N)$ under the conditioned Gibbs measure π_{N,r_N}, and deduce a law of large numbers. In Section 3, the results are proved. Section 4 is devoted to applications: we study DNA denaturation as a function of the superhelical density κ, and give numerical computations (see Figs. 1 and 2).

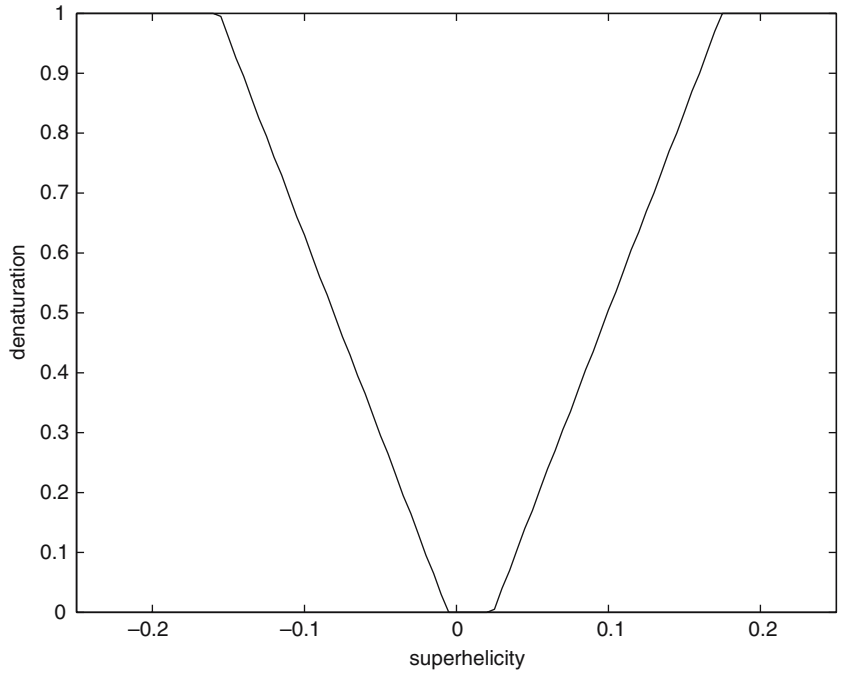

Fig. 1. Denaturation M_∞ as a function of the superhelicity κ in the homogeneous case.

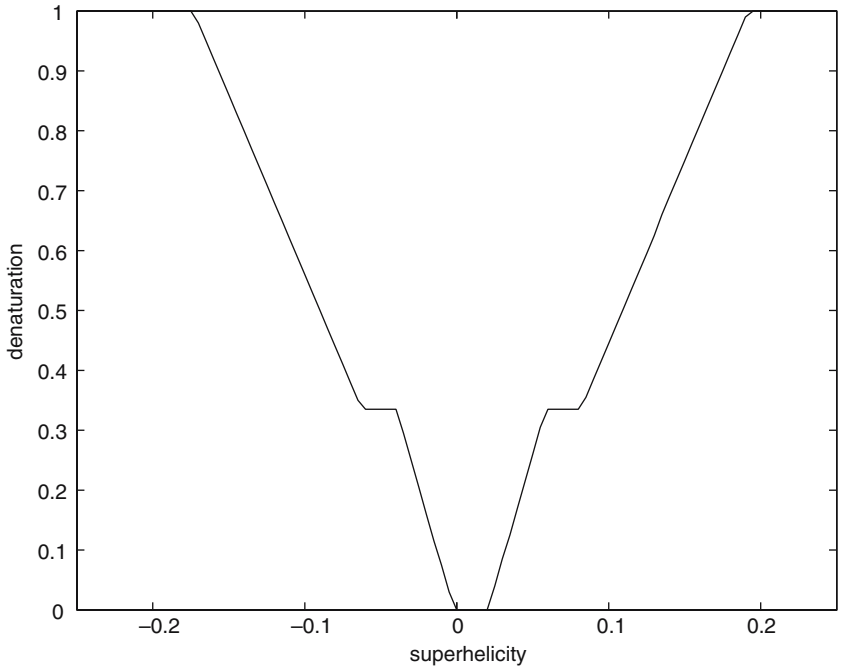

Fig. 2. Denaturation M_∞ as a function of the superhelicity κ in the very inhomogeneous case.

2. STATEMENT OF THE RESULTS

2.1. Hypotheses and Examples

In the sequel, for $N \geqslant 1$, let $B_N = (b_i^N)_{1 \leqslant i \leqslant N} \in \{b_{AT}, b_{GC}\}^N$ be external fields and $r_N \geqslant 1$ be integers. We suppose that the sequence of integers $r_N \geqslant 1$ satisfies the small perimeter assumption:

- $(\mathbf{H_0})$: $\lim_{N \to +\infty} \frac{r_N}{N} = 0.$

2.1.1. Hypotheses on the Sequence of External Fields

Let $(T_N)_{N \geqslant 1}$ be a sequence of integers. We define the average of the field B_N with scale T_N and we denote by $\overline{B}_N = (\bar{b}_i^N)_{1 \leqslant i \leqslant N} \in [b_{AT}, b_{GC}]^N$ the field defined by

$$\bar{b}_i^N = \frac{1}{T_N} \sum_{k=i}^{i+T_N-1} b_k^N, \quad 1 \leqslant i \leqslant N \tag{2}$$

The periodic boundary condition $b_{N+i}^N = b_i^N$ is used.

Roughly speaking, the first hypothesis (H_1) is that the averaged field $\overline{B}_N$ is almost constant with at most r_N values. Denote by L_{N,r_N} the set of fields $G \in \mathbf{R}^N$, which are constant on r_N subintervals: that is $G = (g_i)_{1 \leqslant i \leqslant N}$ belongs to L_{N,r_N} if and only if there exist integers $0 = l_0 < l_1 \cdots < l_{r_N} = N$ such that g_i is constant on each of the r_N sets of the form $l_k < k \leqslant l_{k+1}$. The first hypothesis is:

• $(\mathbf{H_1})$: For each $N \geqslant 1$, there exists a field $\widetilde{B}_N = (\tilde{b}_i^N)_{1 \leqslant i \leqslant N} \in L_{N,r_N}$ such that the distance $\delta_N = \frac{1}{N} \sum_{i=1}^N |\bar{b}_i^N - \tilde{b}_i^N|$ has limit zero as $N \to +\infty$.

Since $\bar{b}_i^N \in [b_{AT}, b_{GC}]$, we can also suppose that $\tilde{b}_i^N \in [b_{AT}, b_{GC}]$.
For each field $B_N = (b_i^N) \in \mathbf{R}^N$, we denote by $\mu(B_N)$ the probability measure

$$\mu(B_N) = \frac{1}{N} \sum_{i=1}^N \delta_{b_i^N}.$$

The sequence of fields B_N is said to converge in distribution to a probability measure μ if the sequence of measures $\mu(B_N)$ converges to μ as N goes to infinity. We suppose that there exists a probability measure μ on $\mathbf{R}$ such that:

• $(\mathbf{H_2})$: The sequence of fields $\overline{B}_N$ converges in distribution to μ as $N \to +\infty$.

Under assumption (H_1), this is equivalent to

• $(\mathbf{H_2'})$: The sequence of fields $\widetilde{B}_N$ converges in distribution to μ as $N \to +\infty$.

The last assumption states that the scale T_N used to define the averaged field $\overline{B}_N$ is small:

• $(\mathbf{H_3})$: $\lim_{N \to +\infty} \frac{r_N T_N}{N} = 0.$

We focus on configurations with at most r_N connected domains of denatured bonds, thus the mean length of a domain is of order N/r_N. Assumption (H_3) states that $T_N \ll N/r_N$. Averaging is done on a local

scale. Note that we have three length scales: the global scale is of order N, the intermediate scale is of order N/r_N, and the local scale is of order T_N. In several applications, the length scale T_N will be a constant independent of N, that is why we use the term local scale.

Remark. In this paper, we focus on fields having values in $\{b_{AT}, b_{GC}\}$ because the external field stands for the DNA sequence. It is worth noting that we could work with general fields $B_N \in \mathbf{R}^N$. The results and proofs extend to this more general case under analogous but somewhat stronger hypotheses (essentially, we have to impose the convergence of the measures $\mu(B_N)$ to the measure μ and also the convergence of the first moment of $\mu(B_N)$ to the first moment of μ which is not automatic in the general case.)

2.1.2. Examples

In order to explain the hypotheses, we exhibit some sequences of external fields B_N satisfying (H_1)–(H_3). The case of quasi-homogeneous fields and very-inhomogeneous fields were mentioned in ref. 13.

Quasi-homogeneous fields. A sequence of fields B_N is said to be quasi-homogeneous with intensity b if it satisfies hypotheses (H_1)–(H_3) with limit measure $\mu_1 = \delta_b$, for some $b \in \mathbf{R}$. Since the fields B_N only take values b_{AT} and b_{GC}, the parameter b must belong to $[b_{AT}, b_{GC}]$. This is a generalization of homogeneous fields: it is straightforward to check that if the fields are constant, for example with value b_{AT}, then hypotheses (H_1)–(H_3) hold with limit measure $\mu = \delta_{b_{AT}}$ (take $T_N = 1$). Another example is that of periodic fields. Let W be a word in the letters $\{b_{AT}, b_{GC}\}$ of length T. We construct the fields B_N by repeating the word W (to have N letters, repeat it $[N/T]$ times and eventually add a few letters of the beginning of W). The mean intensity of the fields is approximatively $b = 1/T \sum_{i \in W} b_i$. Take $T_N = T$. Because of periodicity, the averaged field $\overline{B}_N = (\bar{b}_i^N)_{1 \leqslant i \leqslant N}$ is constant and equal to b, at least for $1 \leqslant i \leqslant N - T$. Take $\widetilde{B}_N$ be the constant field with value b and length N. It is easy to check that hypotheses (H_1)–(H_3) are satisfied with the limit measure $\mu_1 = \delta_b$. We give a last example of quasi-homogeneous fields in the context of random external fields. Suppose that

• $(\mathbf{H}_0')$: There exists a sequence of integers T_N satisfying (H_3) such that for every $K > 0$, the series $\sum_{N \geqslant 1} N \exp(-K T_N)$ converges.

This is stronger than assumption (H_0). Note that Hypothesis (H_0') is satisfied if $r_N = N^\delta$ for some $\delta \in (0, 1)$, with $T_N = N^{(1-\delta)/2}$. Let $b \in$

$[b_{AT}, b_{GC}]$. Suppose that the external field $B_N = (b_i^N)_{1 \leqslant i \leqslant N}$ is random and corresponds to independent and identically distributed variables b_i^N, with values b_{AT} or b_{GC} and expectation b. Then almost every sequence of fields B_N is quasi-homogeneous with intensity b. We omit the proof since it is a consequence of Proposition 1, which will be proved in the sequel.

Very inhomogeneous fields. A sequence of fields B_N is said to be very inhomogeneous with intensity $b \in [b_{AT}, b_{GC}]$ if it satisfies hypotheses (H_1)–(H_3) with limit measure

$$\mu_2 = \frac{b_{GC} - b}{b_{GC} - b_{AT}} \delta_{b_{AT}} + \frac{b - b_{AT}}{b_{GC} - b_{AT}} \delta_{b_{GC}}.$$

This measure is the one of greatest variance among the probability measures on $[b_{AT}, b_{GC}]$ with first moment b. This explains the terminology 'very inhomogeneous field'. Suppose the fields B_N belong to L_{N,r_N} and that $\lim_{N \to \infty} 1/N \sum_{i=1}^{N} b_i^N = b$. There are large areas where the field B_N is constant (with values in $\{b_{AT}, b_{GC}\}$) and the mean intensity of the field is approximatively b. Then the sequence of fields B_N is very inhomogeneous with intensity b. To check it, take $T_N = 1$ and $\widetilde{B}_N = \overline{B}_N = B_N$. Hypothesis (H_1) is verified since B_N belongs to L_{N,r_N} and the convergence of the mean intensity to b ensures that hypothesis (H_2) holds with limit measure μ_2.

Random external fields. The random framework provides us an interesting family of examples.

Proposition 1. Suppose that assumption (H_0') holds. Let $\beta : [0, 1] \to [b_{AT}, b_{GC}]$ be a piecewise continuous function. For $N \geqslant 1$, define a random external field B_N such that the variables b_i^N, $1 \leqslant i \leqslant N$ are independent and distributed according to

$$\mathbf{P}(b_i^N = b_{GC}) = 1 - \mathbf{P}(b_i^N = b_{AT}) = \frac{\beta(i/N) - b_{AT}}{b_{GC} - b_{AT}}.$$

Then, $\mathbf{P}$-almost surely, the sequence of external fields (B_N) satisfies assumptions (H_1)–(H_3) with limit distribution μ, which is the distribution of $\beta(U)$ for U uniformly distributed on $[0, 1]$.

It is worth noting that a good choice of the function β allows us to obtain any limit measure μ on $[b_{AT}, b_{GC}]$.

Biological fields. What about biological DNA sequences? In order to apply our results on denaturation to real biological DNA sequences, can we claim that our assumptions are satisfied? The answer is far from being easy, and it is very striking to note that the question is of great biological

interest. The first difficulty is that we generally consider only one DNA sequence and not a sequence of DNA sequences. Thus the asymptotic has to be interpreted as an approximation, which can be thought as effective since the length of DNA sequences is really high (for example several thousands pairs of bases for the DNA sequence of a virus, several billions pairs of bases for the human genome). What is the meaning of our assumptions? The averaged field $\overline{B}_N$ is strongly connected to the *GC*-content along the sequence, which biologists plot with a moving window. Such plots are popular in genetics. Long domains with almost constant averaged field correspond to long parts of the sequence with homogeneous *GC* content. This strongly remind us to the notion of isochore in genetics. An isochore is a long portion of a DNA sequence with homogeneous *GC* content. This notion was first mentioned in ref. 5. Since then, several computational methods have been developed to exhibit the isochore structure in genome sequences (see for example the work of Olivier *et al.*[14] It appears that some sequences do exhibit such a structure and others do not. The notion of isochore is still a subject of research in genetics, and even of controversy, as show the most recent publications.[7,10]

2.2. The Large Deviations Principle

At this point, we present the main mathematical result of the paper:

Theorem 1. Assume that assumptions (H_0)–(H_3) hold.
Then the distribution of $(M_N, M_{NB_N}, R_N/N)$ under the measure π_{N,r_N} satisfies a large deviations principle with speed N and good rate function J_Δ defined by

$$J_\Delta(m,\tilde{m},r)=\begin{cases} F(m,\tilde{m})-\inf_\Delta F & \text{if } (m,\tilde{m})\in\Delta \text{ and } r=0,\\ +\infty & \text{otherwise,}\end{cases}$$

where Δ is the domain

$$\Delta=\left\{(m,\tilde{m})\in\mathbf{R}^2 \mid 0\leqslant m\leqslant 1\ ,\ \int_0^m F_\mu^{-1}(x)\,dx\leqslant\tilde{m}\leqslant\int_{1-m}^1 F_\mu^{-1}(x)\,dx\right\}. \tag{3}$$

The function F_μ^{-1} denotes the pseudo-inverse of the repartition function of the probability measure μ.

In this large deviations principle, the rate function J_Δ is defined in terms of the function F appearing in the definition of Benham's Hamiltonian, and in terms of the domain Δ, which catches all information about

the field B_N. This domain will be of main importance in the sequel. In the quasi-homogeneous case with intensity b, the limit measure is $\mu_1 = \delta_b$ and the corresponding domain Δ_1 is the segment defined by

$$\Delta_1 = \{(m, \tilde{m}) \in \mathbf{R}^2 \mid 0 \leqslant m \leqslant 1 \ , \tilde{m} = bm\}.$$

In the very inhomogeneous case with intensity b, the limit measure is μ_2 and straightforward calculations show that the associated domain Δ_2 is the parallelogram defined by

$$\begin{aligned}\Delta_2 = \{(m, \tilde{m}) \in \mathbf{R}^2 \mid 0 \leqslant m \leqslant 1, \max(b_{AT} m, b - b_{GC}(1-m)) \leqslant \tilde{m} \\ \leqslant \min(b_{GC} m, b - b_{AT}(1-m))\}.\end{aligned}$$

One diagonal of this parallelogram is the segment Δ_1. The two previous examples are extremal in the sense that they correspond to the most homogeneous and inhomogeneous fields respectively. The general shape of Δ satisfies the following:

Proposition 2. For any distribution μ on $[b_{AT}, b_{GC}]$ with mean b, the domain Δ is convex, compact and symmetric around the point $(1/2, b)$. Furthermore, it contains the segment Δ_1 corresponding to the quasi-homogeneous case and is contained in the parallelogram Δ_2 corresponding to the very inhomogeneous case.

2.3. The Law of Large Numbers for Denaturation

The asymptotic behavior of $(M_N, \widetilde{M}_{NB_N}, R_N/N)$ follows from the study of the good rate function J_Δ. We obtain the following:

Proposition 3. Let $\kappa \neq 4\pi^2 C/K_0 A$. When N goes to infinity, the distribution of $(M_N, \widetilde{M}_{NB_N}, R_N/N)$ under π_{N,r_N} converges in distribution to the Dirac measure at point $(M_\infty, \widetilde{M}_\infty, 0)$. The point $(M_\infty, \widetilde{M}_\infty)$ is the unique minimizer of the function F on Δ. It satisfies

$$\widetilde{M}_\infty = \int_0^{M_\infty} F_\mu^{-1}(x)\, dx. \tag{4}$$

Equation (4) means that $\widetilde{M}_\infty$ lies on the lower boundary of Δ. Denaturation is localized so as to minimize $\widetilde{M}_\infty$. It occurs in regions where the averaged field $\overline{B}_N$ is low, i.e. regions with high AT concentration (since $b_{AT} < b_{GC}$).

In the last section, we use Proposition 3 to study denaturation as a function of superhelicity and give numerical applications.

3. PROOF OF THE MAIN RESULTS

3.1. Proof of Theorem 1

The strategy of the proof is the following: Varadhan's integral Lemma allows us to derive the LDP for the magnetization under the Gibbs measure π_{N,r_N} from a LDP for the magnetization under the uniform probability ρ_{N,r_N} formulated in Proposition 4. Thanks to the small perimeter assumption, the study of the magnetization under the uniform probability is reduced to combinatorial considerations formulated in Proposition 5. We have to study the existence of configurations with small perimeter and prescribed magnetization. The scale hypothesis (H_3) ensures that replacing the field B_N can be replaced by the field $\overline{B}_N$. Hypothesis (H_1) allows us to work with the field $\widetilde{B}_N$ instead of $\overline{B}_N$. It is helpful since $\widetilde{B}_N$ belongs to L_{N,r_N}. Hypothesis (H_2) ensures the convergence of the distribution of the fields and hence the convergence of several associated quantities.

Throughout this section, we suppose that $(B_N)_{N\geqslant 1}$ is a sequence of fields, (r_N) and (T_N) are sequences of integers, and that assumptions $(H_0)-(H_3)$ are satisfied. We note $L=\max(|b_{AT}|, |b_{GC}|)$.

3.1.1. Reduction of the Proof

Proposition 4. The distribution of $(M_N, M_{NB_N}, R_N/N)$ under ρ_{N,r_N} satisfies a LDP with speed N and good rate function I_Δ defined by

$$I_\Delta(m,\tilde{m},r)=\begin{cases}0 & \text{if } (m,\tilde{m})\in\Delta \text{ and } r=0,\\ +\infty & \text{otherwise,}\end{cases}$$

where Δ is the domain defined by (3).

Let us prove that Proposition 4 implies Theorem 1.
The Hamiltonian H_N defined by Eq. (1) is a function of M_N, M_{NB_N} and R_N only. Thus the distribution of $(M_N, M_{NB_N}, R_N/N)$ under the Gibbs measure π_{N,r_N} and under the uniform measure ρ_{N,r_N} are linked by the relation

$$\begin{aligned}&\int_{\Omega_N}\theta(M_N, M_{NB_N}, R_N/N)d\pi_{N,r_N}\\ &=1/Z_{N,r_N}\int_{\Omega_N}\exp N[-aR_N/N-F(M_N, M_{NB_N})]\\ &\qquad\times\theta(M_N, M_{NB_N}, R_N/N)d\rho_{N,r_N},\end{aligned}$$

where $\theta : \mathbf{R}^3 \to \mathbf{R}$ is any bounded continuous function and Z_{N,r_N} is the partition function defined by

$$Z_{N,r_N} = \int_{\Omega_N} \exp N[-aR_N/N - F(M_N, M_{NB_N})] d\rho_{N,r_N}.$$

Since the function $(m, \tilde{m}, r) \mapsto -ar - F(m, \tilde{m})$ is bounded and continuous on $[0,1] \times [b_{AT}, b_{GC}] \times [0,1]$, we can apply Varadhan's integral Lemma to derive the LDP for the distribution of $(M_N, M_{NB_N}, R_N/N)$ under π_{N,r_N} from the LDP for the distribution of $(M_N, M_{NB_N}, R_N/N)$ under ρ_{N,r_N}. That is, Proposition 4 implies Theorem 1 thanks to Varadhan's integral Lemma.

Lemma 1. The following inequalities hold ρ_{N,r_N}-almost surely:

$$|M_{NB_N} - M_{N\overline{B}_N}| \leqslant 2Lr_N T_N/N, \tag{5}$$

$$|M_{N\overline{B}_N} - M_{N\widetilde{B}_N}| \leqslant \delta_N. \tag{6}$$

Thus the distribution of M_{NB_N}, $M_{N\overline{B}_N}$ and $M_{N\widetilde{B}_N}$ under ρ_{N,r_N} are exponentially equivalent.

Proof. Let $\sigma \in \Omega_N$ be such that $R_N(\sigma) \leqslant r_N$. The subset $U = \{i | \sigma_i = +1\}$ of $\{1, \ldots, N\}$ is the disjoint union of at most r_N 'connected components' $U_l = \{u_l, \ldots, v_l\}$. The difference is estimated by

$$|M_{NB_N}(\sigma) - M_{N\overline{B}_N}(\sigma)| \leqslant \frac{1}{N} \sum_{l=1}^{r_N} \left| \sum_{i \in U_l} b_i^N - \bar{b}_i^N \right|. \tag{7}$$

For the connected component U_l, we have

$$\sum_{i \in U_l} \bar{b}_i^N = \sum_{i=u_l}^{v_l} \frac{1}{T_N} \sum_{j=i}^{i+T_N-1} b_j = \sum_{i=u_l}^{v_l+T_N-1} \alpha_i b_i,$$

where $\alpha_i T_N$ is the cardinal of the set $\{i - T_N + 1, \ldots, i\} \cap \{u_l, \ldots, v_l\}$. If $u_l + T_N - 1 \leqslant i \leqslant v_l$, then $\alpha_i = 1$, otherwise $0 \leqslant \alpha_i \leqslant 1$. Therefore,

$$\left| \sum_{i \in U_l} b_i^N - \bar{b}_i^N \right| \leqslant \left| \sum_{i=u_l}^{u_l+T_N-1} (\alpha_i - 1) b_i + \sum_{i=v_l}^{v_l+T_N-1} \alpha_i b_i \right| \leqslant 2LT_N. \tag{8}$$

Equations (7) and (8) together yield Eq. (5).

Equation (6) is a straightforward consequence of assumption (H_1):

$$|M_{N\overline{B}_N}(\sigma) - M_{N\widetilde{B}_N}(\sigma)| \leqslant \frac{1}{N}\sum_{i=1}^{N}\left|\bar{b}_i^N - \tilde{b}_i^N\right| = \delta_N.$$

The notion of exponentially equivalent measures is defined in ref. 9, p. 130: the distributions of M_{NB_N}, $M_{N\overline{B}_N}$ under ρ_{N,r_N} are said to be exponentially equivalent if for every $\alpha > 0$,

$$\limsup_{N\to\infty}\frac{1}{N}\log\ \rho_{N,r_N}\left(|M_{NB_N} - M_{N\overline{B}_N}| > \alpha\right) = -\infty. \tag{9}$$

Inequality (5) and hypothesis (H_3) imply that for large N,

$$\rho_{N,r_N}\left(|M_{NB_N} - M_{N\overline{B}_N}| > \alpha\right) = 0.$$

This implies Eq (9). In the same way, inequality (6) and Hypothesis (H_1) imply that the distributions of $M_{N\overline{B}_N}$ and $M_{N\widetilde{B}_N}$ under ρ_{N,r_N} are exponentially equivalent. ▮

Two sequences of exponentially equivalent measures have the same large deviations properties, i.e. if a LDP hold for one sequence of measures, the same LDP will hold for any sequence of exponentially equivalent measures – see Theorem 4.2.13 in ref. 9. Thus Lemma 1 allows us to work with $M_{N\widetilde{B}_N}$, instead of M_{NB_N}. It is equivalent to replace the field B_N by the field $\widetilde{B}_N$. It is helpful since $\widetilde{B}_N$ belongs to L_{N,r_N} and verifies Hypothesis (H_2').

It will be convenient to use simpler notations: in the sequel, we use the notation $\widetilde{M}_N$ instead of $M_{N\widetilde{B}_N}$.

Lemma 2. Let $\omega_N \subset \Omega_N$ be a sequence of events such that for large N, $\rho_{N,r_N}(\omega_N) > 0$. Then

$$\lim_{N\to+\infty}\frac{1}{N}\log\left[\rho_{N,r_N}(\omega_N)\right] = 0.$$

Proof. The probability measure ρ_{N,r_N} is the uniform probability on the set $\{\sigma \in \Omega_N | R_N(\sigma) \leqslant r_N\}$. As a configuration σ is uniquely determined

by the value σ_1 and the position of the $2R_N(\sigma)$ sites i such that $\sigma_i \neq \sigma_{i+1}$, we have

$$\text{card}\{\sigma \in \Omega_N | R_N(\sigma) \leqslant r_N\} = \sum_{r=0}^{r_N} 2\binom{N}{2r} \leqslant 2(r_N+1)\binom{N}{2r_N}.$$

The last inequality holds for large N, because if $4r_N \leqslant N$ and $0 \leqslant r \leqslant r_N$,

$$\binom{N}{2r} \leqslant \binom{N}{2r_N}.$$

If $\rho_{N,r_N}(\omega_N) > 0$, the following inequalities hold:

$$\left(2(r_N+1)\binom{N}{2r_N}\right)^{-1} \leqslant \rho_{N,r_N}(\omega_N) \leqslant 1.$$

We then use Stirling's formula to estimate the binomial coefficient and Hypothesis (H_0) to estimate the limit and the lemma is proved. ▮

Lemmas 1 and 2 allow us to reduce Proposition 4 to the following one:

Proposition 5.

• $(\mathbf{A_1})$: For every closed set $A \subset \mathbf{R}^2$ disjoint from Δ, there exists $N_0 \in \mathbf{N}$ such that for every $N \geqslant N_0$, $\rho_{N,r_N}((M_N, \widetilde{M}_N) \in A) = 0$.
• $(\mathbf{A_2})$: For every open set $A \subset \mathbf{R}^2$ intersecting Δ, there exists $N_0 \in \mathbf{N}$ such that for every $N \geqslant N_0$, $\rho_{N,r_N}((M_N, \widetilde{M}_N) \in A) > 0$.

Let us prove that Proposition 5 implies Proposition 4 and hence Theorem 1. In view of Lemma 1, the LDP of Proposition 4 is equivalent to a LDP for the distribution of $(M_N, \widetilde{M}_N, R_N/N)$ under ρ_{N,r_N}, with good rate function I_Δ. We have to show that for every open set $O \subset \mathbf{R}^3$,

$$-\inf_{x \in O} I_\Delta(x) \leqslant \liminf_{N \to +\infty} \frac{1}{N} \log\ \rho_{N,r_N}\left((M_N, \widetilde{M}_N, R_N/N) \in O\right) \tag{10}$$

and that for every closed set $C \subset \mathbf{R}^3$,

$$\limsup_{N \to +\infty} \frac{1}{N} \log\ \rho_{N,r_N}\left((M_N, \widetilde{M}_N, R_N/N) \in C\right) \leqslant -\inf_{x \in C} I_\Delta(x). \tag{11}$$

As I_Δ equals 0 on the set $\Delta \times \{0\}$ and $+\infty$ outside of this set, inequalities (10) and (11) are trivial unless O is open and intersect $\Delta \times \{0\}$ or C is closed and disjoint from $\Delta \times \{0\}$. Let $O \in \mathbf{R}^3$ be an open set intersecting $\Delta \times \{0\}$. There exists an open set $A \subset \mathbf{R}^2$ intersecting Δ and $\varepsilon > 0$ such that $A \times]-\varepsilon, \varepsilon[\subset O$. Hypothesis (H_0) implies that for large N, $r_N/N < \varepsilon$ and thus

$$\rho_{N,r_N}((M_N, \widetilde{M}_N, R_N/N) \in O) \geqslant \rho_{N,r_N}((M_N, \widetilde{M}_N) \in A).$$

According to assertion (A_2), this probability is strictly positive for large N. Apply then Lemma 2 with $\omega_N = \{(M_N, \widetilde{M}_N) \in A\}$. This yields Eq. (10).

Let $C \in \mathbf{R}^3$ be a closed set disjoint from $\Delta \times \{0\}$. There exists a closed set $A \subset \mathbf{R}^2$ disjoint from Δ and $\varepsilon > 0$ such that $F \cap \left(\mathbf{R}^2 \times [-\varepsilon, \varepsilon]\right) \subset A \times [-\varepsilon, \varepsilon]$. Hypothesis (H_0) implies that for large N, $r_N/N < \varepsilon$ and thus

$$\rho_{N,r_N}((M_N, \widetilde{M}_N, R_N/N) \in C) \leqslant \rho_{N,r_N}((M_N, \widetilde{M}_N) \in C).$$

According to assertion (A_1), this probability is equal to zero for large N. Hence, the lim sup in Eq. (11) is equal to minus infinity, and this equation holds. ∎

3.1.2. Proof of Proposition 5

The following lemma investigates the behavior of $\widetilde{M}_N$ conditionally to M_N.

Let $m_N \in [0, 1]$ be such that $Nm_N \in \mathbf{N}$. Define the set

$$\mathcal{V}_{N,m_N} = \{\widetilde{M}_N(\sigma) | R_N(\sigma) \leqslant r_N, M_N(\sigma) = m_N\}.$$

Lemma 3. Then,

$$\mathcal{V}_{N,m_N} = \{\widetilde{M}_N(\sigma) | M_N(\sigma) = m_N\} \tag{12}$$

Let $\tilde{m}$ be such that $\min \mathcal{V}_{N,m_N} \leqslant \tilde{m} \leqslant \max \mathcal{V}_{N,m_N}$. There exists $\tilde{m}' \in \mathcal{V}_{N,m_N}$ such that

$$|\tilde{m}' - \tilde{m}| \leqslant L/N. \tag{13}$$

Suppose that $m_N \to m$ as $N \to +\infty$. Then:

$$\lim_{N\to\infty} \min \mathcal{V}_{N,m_N} = \int_0^m F_\mu^{-1}(x)\,dx, \tag{14}$$

$$\lim_{N\to\infty} \max \mathcal{V}_{N,m_N} = \int_{1-m}^1 F_\mu^{-1}(x)\,dx. \tag{15}$$

Proof. Equation (12) states that we can forget the constraint $R_N \leqslant r_N$ in the definition of $\mathcal{V}_{N,m_N}$. This is true because the field $\widetilde{B}_N$ belongs to L_{N,r_N} – i.e. is constant on r_N subintervals. Divide $\{1,\ldots,N\}$ in r_N subintervals $U_1,\ldots,U_{r_N}$, where the field $\widetilde{B}_N$ is constant. Let $\sigma \in \Omega_N$ be a configuration such that $M_N(\sigma) = m_N$ and let J be the set of the Nm_N indexes, where $\sigma = +1$. The magnetization $\widetilde{M}_N(\sigma)$ only depends on the cardinality of the sets $J \cap U_1,\ldots,J \cap U_{r_N}$. We can modify the configuration σ in the following way: choose two indexes i and $j \in U_l$ such that $\sigma_i = 1$ and $\sigma_j = 0$ and modify the configuration σ by setting $\sigma_i = 0$ and $\sigma_j = 1$. This modification does not change the values $M_N(\sigma)$ and $\widetilde{M}_N(\sigma)$. It is possible to perform several modifications in such a way that all the elements of $J \cap U_i$ come on the left or on the right side of U_i. We obtain a configuration $\sigma' \in \Omega_N$ such that $M_N(\sigma') = M_N(\sigma) = m_N$, $\widetilde{M}_N(\sigma') = \widetilde{M}_N(\sigma)$ and $R_N(\sigma') \leqslant r_N$. This proves Eq. (12).

Let σ^- (resp. σ^+) be a configuration in $\{\sigma \in \Omega_N | M_N(\sigma) = m_N\}$ such that $\widetilde{M}_N$ is minimal (resp. maximal). We can find a path $\sigma_0 = \sigma^-, \ldots, \sigma_k = \sigma^+$ such that two successive configurations differ only by switching two spins, one from 0 to 1 and the other one from 1 to 0. Each configuration verifies $M_N(\sigma) = m_N$ and two successive values in the sequence $\widetilde{M}_N(\sigma_0), \ldots, \widetilde{M}_N(\sigma_l)$ differ by at most $2L/N$. This explains Eq. (13).

Let $\tilde{\mu}_N = \mu(\widetilde{B}_N) = 1/N \sum_{i=1}^N \delta_{\tilde{b}_i^N}$ be the distribution of the external field $\widetilde{B}_N$. We denote by $F_{\tilde{\mu}_N}^{-1}$ the pseudo inverse of the repartition function of the measure $\tilde{\mu}_N$. It is a step function defined on $(0,1)$, which value on $]k-1/N, k/N]$ is the kth smallest value among the $\tilde{b}_i^N$. Thus,

$$\min \mathcal{V}_{N,m_N} = \frac{1}{N}\left(F_{\tilde{\mu}_N}^{-1}\left(\frac{1}{N}\right) + \cdots + F_{\tilde{\mu}_N}^{-1}\left(\frac{Nm_N}{N}\right)\right) = \int_0^{m_N} F_{\tilde{\mu}_N}^{-1}(x)\,dx. \tag{16}$$

Hypothesis (H_2') states that the sequence of measures $\tilde{\mu}_N$ converge to μ. This implies the convergence almost everywhere of the inverse repartition

functions $F^{-1}_{\tilde{\mu}_N}$ to F^{-1}_{μ}. We also have the inequality $|F^{-1}_{\tilde{\mu}_N}(x)| \leqslant L$. Since $m_N \to m$ as $N \to +\infty$, Lebesgue's theorem implies that

$$\lim_{N\to+\infty} \int_0^{m_N} F^{-1}_{\tilde{\mu}_N}(x)\,dx = \int_0^{m} F^{-1}_{\mu}(x)\,dx. \tag{17}$$

Equations (16) and (17) together yield Eq. (14). Equation (15) is proved in the same way. ▮

Proof of Assertion (A_1). Suppose that the assertion does not hold: there exists a closed set A not intersecting Δ such that

$$\exists N_i \to +\infty,\ \exists \sigma_{N_i} \in \Omega_{N_i, r_{N_i}} \quad \text{such that} \quad (M_{N_i}(\sigma_{N_i}), \tilde{M}_{N_i}(\sigma_{N_i})) \in A.$$

To simplify the notations, we omit the index i and write N instead of N_i. Since the magnetization $(M_N(\sigma), \widetilde{M}_N(\sigma))$ belongs to the compact set $[0,1] \times [b_{AT}, b_{GC}]$, we can assume that

$$(M_N(\sigma_N), \tilde{M}_N(\sigma_N)) \underset{N\to\infty}{\to} (m, \tilde{m}).$$

As A is a closed set, $(m, \tilde{m}) \in A$. Furthermore we have for each N,

$$\min \mathcal{V}_{N, M_N(\sigma_N)} \leqslant \widetilde{M}_N(\sigma_N) \leqslant \max \mathcal{V}_{N, M_N(\sigma_N)}.$$

Let N go to infinity, and use equations (14) and (15), this yields

$$\int_0^m F^{-1}_{\mu}(x)\,dx \leqslant \tilde{m} \leqslant \int_{1-m}^1 F^{-1}_{\mu}(x)\,dx.$$

It means that $(m, \tilde{m}) \in \Delta$. But $(m, \tilde{m})$ belongs to A and A doesn't intersect Δ. There is a contradiction and assertion (A_1) must hold. ▮

Proof of Assertion (A_2). Let A be an open set intersecting Δ and let $(m, \tilde{m}) \in \Delta \cap A$. We exhibit a sequence $\sigma_N \in \Omega_N$ such that $R_N(\sigma_N) \leqslant r_N$ and

$$(M_N(\sigma_N), \widetilde{M}_N(\sigma_N)) \underset{N\to+\infty}{\longrightarrow} (m, \tilde{m}).$$

This will imply that for large N, $(M_N(\sigma_N), \widetilde{M}_N(\sigma_N))$ belongs to A and $\rho_{N, r_N}((M_N, \widetilde{M}_N) \in A) > 0$, proving assertion (A_2).

Let $m_N = [mN]/N$. This sequence verifies $m_N \underset{N\to+\infty}{\longrightarrow} m$.
If $\tilde{m}$ is such that $\min \mathcal{V}_{N,m_N} \leqslant \tilde{m} \leqslant \max \mathcal{V}_{N,m_N}$, then Eq. (13) implies that there exists $\sigma_N \in \Omega_N$ such that $R_N(\sigma_N) \leqslant r_N$, $M_N(\sigma_N) = m_N$ and $|\widetilde{M}_N(\sigma_N) - \tilde{m}| \leqslant L/N = \varepsilon_N^{(1)}$.

If $\tilde{m} < \min \mathcal{V}_{N,m_N}$, we choose σ_N such that $R_N(\sigma_N) \leqslant r_N$, $M_N(\sigma) = m_N$ and $\widetilde{M}_N(\sigma_N) = \min \mathcal{V}_{N,m_N}$. As $(m, \tilde{m}) \in \Delta$, we have the following inequalities

$$|\widetilde{M}_N(\sigma_N) - \tilde{m}| = \min \mathcal{V}_{N,m_N} - \tilde{m} \leqslant \min \mathcal{V}_{N,m_N} - \int_0^m F_\mu^{-1}(x)\,dx = \varepsilon_N^{(2)}.$$

In the same way, if $\tilde{m} > \max \mathcal{V}_{N,m_N}$, we prove a similar result with

$$\varepsilon_N^{(3)} = \int_{1-m}^1 F_\mu^{-1}(x)\,dx - \max \mathcal{V}_{N,m_N}.$$

The three cases together ensure that there exists $\sigma_N \in \Omega_N$ such that $R_N(\sigma_N) \leqslant r_N$, $M_N(\sigma_N) = m_N$ and $|\widetilde{M}_N(\sigma_N) - \tilde{m}\ | \leqslant \varepsilon_N$ with $\varepsilon_N = \max(\varepsilon_N^{(1)}, \varepsilon_N^{(2)}, \varepsilon_N^{(3)})$. Equations (14) and (15) imply that ε_N converges to 0 as N goes to infinity. As a consequence, $\widetilde{M}_N(\sigma_N) \underset{N\to+\infty}{\longrightarrow} \tilde{m}$ and assertion (A_2) is proved. ∎

3.2. Proof of Proposition 2

This proposition describes the general shape of Δ.

Proof. The function F_μ^{-1} is nondecreasing and bounded. Thus the function $m \mapsto \int_0^m F_\mu^{-1}(x)\,dx$ is continuous, convex, bounded and $m \mapsto \int_{1-m}^1 F_\mu^{-1}(x)\,dx$ is continuous, concave and bounded. This proves that Δ is compact and convex. The relation

$$\int_0^{1-m} F_\mu^{-1}(x)\,dx = \int_0^1 F_\mu^{-1}(x)\,dx - \int_{1-m}^1 F_\mu^{-1}(x)\,dx = b - \int_{1-m}^1 F_\mu^{-1}(x)\,dx,$$

shows that Δ is symmetric around the point $(1/2, b)$. Since the points $(0, 0)$ and $(1, b)$ belong to the convex domain Δ, the segment Δ_1 between these points is contained in Δ. Since F_μ^{-1} takes its values in $[b_{AT}, b_{GC}]$,

$$mb_{AT} \leqslant \int_0^m F_\mu^{-1}(x)\,dx \leqslant \int_{1-m}^1 F_\mu^{-1}(x)\,dx \leqslant mb_{GC}. \tag{18}$$

As $\int_0^1 F_\mu^{-1}(y)dy = b$,

$$b - b_{GC}(1-m) \leqslant \int_0^m F_\mu^{-1}(x)\,dx \leqslant \int_{1-m}^1 F_\mu^{-1}(x)\,dx \leqslant b - b_{AT}(1-m). \tag{19}$$

Equations (18) and (19) imply the inclusion $\Delta \subseteq \Delta_2$. ∎

3.3. Proof of Proposition 3

In this section, we show some applications of the large deviations principle stated in Theorem 1: the magnetizations obey a law of large numbers, that is to say converge to a deterministic limit as the system size N goes to infinity. This result is derived from the study of the minima of the good rate function J_Δ. We assume that the hypotheses of Theorem 1 are satisfied with the limit measure μ. The Hamiltonian is defined in Eq. (1), it depends on physical constants a, b_{AT}, b_{GC}, C, K_0, A, and on the superhelicity κ.

Proof. The good rate function J_Δ has value $+\infty$ outside of $\Delta \times \{0\}$. On $\Delta \times \{0\}$, it is defined by

$$\begin{aligned} J_\Delta(m, \tilde{m}, 0) &= F(m, \tilde{m}) - \inf_\Delta F = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m}(\kappa + \frac{m}{A})^2 + \tilde{m} - \inf_\Delta F \\ &= G(m) + \tilde{m}, \end{aligned}$$

where G denote the function defined on $[0, 1]$ by

$$G(m) = \frac{2\pi^2 C K_0}{4\pi^2 C + K_0 m}(\kappa + \frac{m}{A})^2 - \inf_\Delta F.$$

In order to minimize $J_\Delta(m, \tilde{m}, 0)$, we choose the lowest $\tilde{m}$, that is $\tilde{m} = \int_0^m F_\mu^{-1}(x)\,dx$, and minimize on $[0, 1]$ the function ϕ defined by

$$\phi(m) = G(m) + \int_0^m F_\mu^{-1}(x)dx.$$

The function G has a second derivative given by

$$G''(m) = \frac{4\pi^2 C K_0 (K_0 \kappa A - 4\pi^2 C)^2}{(4\pi^2 C + K_0 m)^3 A^2}.$$

Hence for $\kappa \neq 4\pi^2 C/K_0 A$, the function G is strictly convex. As F_μ^{-1} is non decreasing, its primitive is convex. Hence, for $\kappa \neq 4\pi^2 C/K_0 A$, the function ϕ is strictly convex on $[0,1]$, and achieves its minimum at a unique point M_∞. Let $\tilde{M}_\infty = \int_0^{M_\infty} F_\mu^{-1}(x)\,dx$. The rate function J_Δ achieves its minimum at a unique point which is $(M_\infty, \tilde{M}_\infty, 0)$. The large deviations principle stated in Theorem 1 implies the convergence of the distribution of $(M_N, \tilde{M}_N, R_N/N)$ under π_{N,r_N} to the Dirac measure at point $(M_\infty, \tilde{M}_\infty, 0)$ with an exponential speed. ∎

Remark. If $\kappa = 4\pi^2 C/K_0 A$, the good rate function J_Δ reduces on $\Delta \times \{0\}$ to the linear function

$$J_\Delta(m, \tilde{m}, 0) = \frac{\kappa}{2A}\left(\kappa + \frac{m}{A}\right) + \tilde{m} - \inf_\Delta F.$$

The uniqueness of a minimizer depends on the shape of Δ. In the case $0 < b_{AT} < b_{GC}$ (more important in applications), uniqueness always holds.

3.4. Proof of Proposition 1

In this section, $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space. For $N \geqslant 1$, the field B_N is a random variable from Ω to $\{b_{AT}, b_{GC}\}^N$. We denote by $B_N(\omega)$ a realization of the random variable B_N.

Let $\beta \to [b_{AT}, b_{GC}]$ be a piecewise continuous function. We suppose that for $N \geqslant 1$, the distribution of the random field B_N is such that the variables b_i^N, $1 \leqslant i \leqslant N$ are independent and distributed according to

$$\mathbf{P}(b_i^N = b_{GC}) = 1 - \mathbf{P}(b_i^N = b_{AT}) = \frac{\beta(i/N) - b_{AT}}{b_{GC} - b_{AT}}.$$

The variable b_i^N is $\{b_{AT}, b_{GC}\}$-valued, with expectation $\beta(i/N)$.

Suppose furthermore that the sequence of integers r_N satisfies assumption (H_0'). Proposition 1 states that **P**-almost surely, the sequence of fields B_N satisfies assumptions (H_1)–(H_3) with limit distribution μ, which is the distribution of $\beta(U)$ for U uniformly distributed on $[0,1]$.

The proof of Proposition 1 relies on a lemma of uniform exponential concentration for the empirical mean of Bernoulli variables:

Lemma 4. Let $X_1, \ldots, X_n$ be independent variables, X_i being $\{0,1\}$-valued with expectation x_i. Let $\bar{X}$ be the mean of the X_i's, and $\bar{x}$ be the mean of the x_i's. For every $\varepsilon > 0$, there exists $K(\varepsilon) > 0$ depending on ε only, such that

$$\mathbf{P}(|\bar{X} - \bar{x}| > \varepsilon) \leqslant 2e^{-K(\varepsilon)n}.$$

Proof. The proof of this lemma is a straightforward application of the standard concentration inequality for bounded martingales differences – see for example ref. 9 Section 2.4.1. ▮

Proof of Proposition 1. Choose T_N given by assumption (H_0'). For $1 \leqslant i \leqslant N$, the variable $\bar{b}_i^N$ is the empirical mean of T_N independent $\{b_{AT}, b_{GC}\}$-valued variables. The expectation of b_i^N is $\bar{\beta}_N(i/N)$, where $\bar{\beta}_N$ is the function defined by

$$\bar{\beta}_N(x) = \frac{1}{T_N} \sum_{k=i}^{i+T_N-1} \beta(x + k/N)$$

with the periodic boundary condition $\beta(x+1) = \beta(x)$.
We apply Lemma 4: for every $\varepsilon > 0$,

$$\mathbf{P}(|\bar{b}_i^N - \bar{\beta}_N(i/N)| > \varepsilon) \leqslant 2e^{-T_N K'(\varepsilon)}$$

with $K'(\varepsilon) = K(\varepsilon/(b_{GC} - b_{AT}))$. We need this normalization because the b_i^N are not in $\{0, 1\}$ but in $\{b_{AT}, b_{GC}\}$. As a consequence, we have the following estimation

$$\mathbf{P}\left(\max_{1 \leqslant i \leqslant N} |\bar{b}_i^N - \bar{\beta}_N(i/N)| > \varepsilon \right) \leqslant 2Ne^{-T_N K'(\varepsilon)}.$$

Since the series $\sum_{N \geqslant 1} 2Ne^{-T_N K'(\varepsilon)}$ is finite, Borel Cantelli's Lemma implies that **P**-almost surely,

$$\max_{1 \leqslant i \leqslant N} |\bar{b}_i^N - \bar{\beta}_N(i/N)| \underset{N \to +\infty}{\longrightarrow} 0. \tag{20}$$

We now study the behavior of the deterministic functions $\bar{\beta}_N$ as N goes to infinity. If β is continuous on $[0, 1]$ and $\beta(0) = \beta(1)$, then β is uniformly continuous on $[0, 1]$ (with the boundary condition). As a consequence, the sequence of functions $\bar{\beta}_N$ uniformly converges on $[0, 1]$ to β as N goes to infinity. If β is piecewise continuous on $[0, 1]$, the uniform convergence holds on any compact where β is continuous, and the sequence of functions $\bar{\beta}_N$ converges to β in $L^1([0, 1])$. This implies the convergence of the measures

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{\bar{\beta}_N(i/N)} \underset{N \to +\infty}{\longrightarrow} \beta(U) \tag{21}$$

with U uniformly distributed on $[0, 1]$.

Equations (20) and (21) implies that **P**-almost surely

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\tilde{b}_i^N} \underset{N\to+\infty}{\longrightarrow} \beta(U),$$

i.e. assumption (H_2) holds almost surely with limit measure μ equal to the distribution of the random variable $\beta(U)$.

As the piecewise continuous function β can be approximated by a sequence of piecewise constant functions, there exists $\widetilde{B}_N = (\tilde{b}_i^N)_{1\leqslant i\leqslant N} \in L_{N,r_N}$ such that

$$\frac{1}{N}\sum_{i=1}^{N} |\tilde{b}_i^N - \beta(i/N)| \underset{N\to+\infty}{\longrightarrow} 0. \tag{22}$$

Equations (22), (20) and the convergence in $L^1([0, 1])$ of the sequence $\bar{\beta}_N$ to β implies that **P**-almost surely

$$\frac{1}{N}\sum_{i=1}^{N} |\tilde{b}_i^N - \bar{b}_i^N| \underset{N\to+\infty}{\longrightarrow} 0.$$

It means that Hypothesis (H_1) holds **P**-almost surely. ▮

4. APPLICATION: DENATURATION AS A FUNCTION OF THE SUPERHELICITY

We study denaturation as a function of superhelicity κ and present numerical computations. Note that similar computations can be done to study the denaturation as a function of the temperature. The values taken for the physical constants $b_{AT}, b_{GC}, \ldots$ are those given by Clote and Backofen[8] recalled in the introduction. We plot the function $\kappa \mapsto M_\infty(\kappa)$ in the quasi-homogeneous case (Fig. 1) and in the very inhomogeneous case (Fig. 2), both with mean intensity $(b_{AT} + 2b_{GC})/3$. In both cases, we represent the denaturation M_∞ of a DNA polymer (on the y-axis) as a function of its superhelicity κ (on the x-axis). Both polymers consist in a large number of nucleotides, with a concentration in AT of 33.3%, and in GC of 66.7%. The difference between the two polymers lies in the repartition of the nucleotides. The first one has an homogeneous repartition of the nucleotides along the sequence whereas the second one contains large

areas with only A and T, and large areas with G and C. Note that for the computation, we don't need the DNA sequence, but only the limit distribution μ. The computations are based on Proposition 3: we compute the minimizer of the good rate function J_Δ on the domain $\Delta \times \{0\}$ in two different cases. In the homogeneous case, the limit distribution is $\mu_1 = \delta_{0.333 b_{AT} + 0.667 b_{GC}}$ and the domain Δ_1 is a segment. In the very inhomogeneous case, the limit distribution is $\mu_2 = 0.333\delta_{b_{AT}} + 0.667\delta_{b_{GC}}$ and the domain Δ_2 is a parallelogram.

We now comment upon these figures. In both cases, for $\kappa = 0$, the DNA polymer is not denatured – $M_\infty = 0$. This nondenatured state is stable: for $\kappa \approx 0$, we still have $M_\infty = 0$. On the other side, for large absolute value of the superhelicity $|\kappa|$, the DNA polymer is totally denatured – $M_\infty = 1$.

In the homogeneous case (Fig. 1), the nondenatured state $M_\infty = 0$ is obtained for a superhelicity κ between the critical values $\kappa_1^- \approx -0.005$ and $\kappa_1^+ \approx 0.024$. If the superhelicity overcrosses the critical value κ_1^+, partial denaturation occurs – $M_\infty > 0$. The denaturation increases with the superhelicity, until it reaches the critical value $\kappa_2^+ \approx 0.175$, where the denaturation is total – $M_\infty = 1$. This totally denatured state is stable: for $\kappa \geqslant \kappa_2^+$, we still have $M_\infty = 1$. For negative values of the superhelicity, $\kappa_1^- \leqslant \kappa \leqslant 0$ correspond to the stable nondenatured state, $\kappa_2^- < \kappa < \kappa_1^-$ to partial denaturation and $\kappa \leqslant \kappa_2^-$ to the stable totally denatured state, with the critical value $\kappa_2^- \approx -0.156$. Note that $|\kappa_1^-| < \kappa_1^+$ and $|\kappa_2^-| < \kappa_2^+$: this reflects the fact that for a fixed amount of absolute superhelicity, the denaturation is larger for negative superhelicity than for positive superhelicity. In other terms, negative supercoiling enhances denaturation.

In the very inhomogeneous case (Fig. 2), the stable nondenatured state corresponds to a superhelicity between the critical values $\kappa_1^- \approx -0.001$ and $\kappa_1^+ \approx 0.020$. The stable totally denatured state occurs for $\kappa \geqslant \kappa_2^+$ or $\kappa \leqslant \kappa_2^-$, with the critical values $\kappa_2^- \approx -0.173$ and $\kappa_2^+ \approx 0.192$. The intermediate values of the superhelicity $\kappa_2^- < \kappa < \kappa_1^-$ and $\kappa_1^+ < \kappa < \kappa_2^+$ yield partial denaturation. In this case, a new stable state appears corresponding to $M_\infty = 0.333$: this state corresponds to the complete denaturation of the AT domain and the non-denaturation of the GC domain. It occurs for $\kappa_3^+ \leqslant \kappa \leqslant \kappa_4^+$ and $\kappa_4^- \leqslant \kappa \leqslant \kappa_3^-$, with the critical values $\kappa_3^+ \approx 0.059$, $\kappa_3^- \approx -0.040$, $\kappa_4^+ \approx 0.081$ and $\kappa_4^- \approx -0.062$. For positive supercoiling, $0 \leqslant \kappa \leqslant \kappa_1^+$ correspond to the stable nondenatured state, for $\kappa_1^+ < \kappa < \kappa_3^+$ the AT domain is partially denaturated, for $\kappa_3^+ \leqslant \kappa \leqslant \kappa_4^+$ the stable state corresponding to the total denaturation of the AT domain is reached, for $\kappa_4^+ < \kappa < \kappa_2^+$ the GC domain is partially denatured and for $\kappa \geqslant \kappa_2^+$, the totally denatured state is reached. The same behavior holds for negative super-

coiling. Once again, note that the critical values corresponding to negative supercoiling are smaller: for $i = 1, \ldots, 4$, $|\kappa_i^-| < \kappa_i^+$. This shows that negative supercoiling enhance denaturation.

ACKNOWLEDGMENTS

We gratefully acknowledge A. Le Ny for encouraging and fruitful discussions, L. Gueguen for his competence in bioinformatics and the referees for useful remarks and advices.

REFERENCES

1. R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, Reprint of the 1982 original, (1989).
2. C. Benham, Torsional stress and local denaturation in supercoiled DNA, *Proc. Natl. Sci.* **76**(8):3870–3874 (1979).
3. C. Benham, Theoretical analysis of heteropolymeric transitions in superhelical DNA at high temperature, *J. Chem. Phys.* **92**(10):6294–6305 (1990).
4. C. Benham, Theoretical of the helix-coil transition in positively superhelical DNA at high temperature, *Phys. Rev. E* **53**(3):2984–2987 (1996).
5. G. Bernardi, The isochore organisation of the human genome, *Ann. Rev. Gen.* **23**:637–661 (1989).
6. P. Billingsley, *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, 2nd ed. (John Wiley & Sons Inc., New York, 1986).
7. O. Clay and G. Bernardi, Isochores: dream or reality? *Trends Biotechno.* **20**(6):237 (2002).
8. P. Clote, and R. Backofen, *Computational Molecular Biology: An Introduction* (Wiley, New York) 2000.
9. A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Applications of Mathematics, Vol. 38 2nd ed. (Springer-Verlag, New York, 1998).
10. W. Li, Are isochore sequence homogeneous? *Gene* **300**(1–2):129–139 (2002).
11. J. T. Lewis, C-E. Pfister and W. G. Sullivan, The equivalence of ensembles for lattice systems: some examples and a counterexample. *J. Statis. Phy.* **77**(1–2):397–419.
12. J. T. Lewis, C-E. Pfister and W. G. Sullivan, Entropy, concentration of probability and conditional limit theorems, *Markov Proces Related Fields* **1**(3):319–386 (1995).
13. C. Mazza, Strand separation in negatively supercoiled DNA. Available at the URL xxx.lanl.gov/cond-mat/0306476 (2002).
14. J. L. Olivier, P. Carena, M. Heisenberg and P. Barnaul-GAlvan, Islander: computational prediction of isochores in genome sequences. *Nucleid Acids Res.* **32**:W287–W292 (2004).
15. C. T. Zhang and R. Zhang, Isochore structures in the mouse genome, *Genomics* **83**(3):384–394 (2004).