DNA unzipping via stopped birth and death processes with random probability transition

Pierre Andreoletti, Roland Diel

DNA unzipping via stopped birth and death processes with random probability transition

P. Andreoletti, R. Diel *

January 3, 2011

Abstract

In this paper we provide an alternative approach to the works of the physicists S. Cocco and R. Monasson about a model of DNA molecules. The aim is to predict the sequence of bases by mechanical stimulations. The model described by the physicists is a stopped birth and death process with random probabilities of transition. We consider two models, a discrete in time and a continuous in time, as general as possible. We show that explicit formula can be obtained for the probability to be wrong for a given estimator, also we add some generalizations comparing to the initial model allowing us to answer some questions asked by the physicists.

1 Introduction

1.1 The physical approach

In this introduction we first summarize some ideas and results of the works of V. Baldazzi, S. Cocco, E. Marinari and R. Monasson ([3], [4]), and S. Cocco and R. Monasson [7] who are interested in a method for the sequencing of DNA molecules. They study a mechanical way, described below, instead of traditional bio-chemical or gel electrophoresis technics. These experiments for mechanical unzipping were first realized by Bockelmann, Helsot and

*Laboratoire MAPMO - C.N.R.S. UMR 6628 - Fédération Denis-Poisson, Université d'Orléans, (Orléans France).

MSC 2000 62P10 ; 82D30.

Key words : DNA unzipping, birth and death processes, random environment, maximum of likelihood

coworkers [6] and [5]. The principle is based on the fact that the force which links the two bases of a given pair depends on whether it is a $C \equiv G$ or a $A - T$ (see Figure 1). Indeed the links $A - T$ is weaker for biochemical

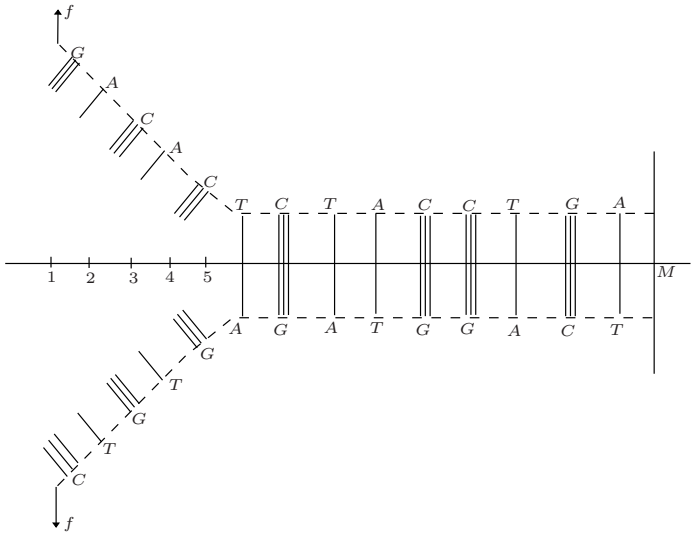

Figure 1: $X_{\cdot} = 5$, $b_1 : C \equiv G$, $b_2 : A - T$

reasons than the link between $C \equiv G$. Moreover there is also some stacking effects between adjacent bases, that is to say, the force needed to break, for example, the link $C \equiv G$ is different if the C is following by a A, or a T. This last factor is not negligible (see the table below) and therefore must be taken into account if we want the model to be as sharp as possible.

We now give a brief description of the experiment (for more details see [3]), the extremities of the DNA molecule are stretched apart under a force f. The force f is chosen in such way that it is large enough so the molecule

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

Figure 2: Binding free energies (units of $k_B T$)

can be totally unzipped. However f is also not too strong so that naturally the molecule rebuild itself. Though there is back and force movement of the number of open pair bases, this back and force movement generates a signal which can be measured by biologists. This signal can be modelised by a birth and death process with random probabilities of transition.

1.2 The model

We denote by M the length of the DNA chain and by $(b_1, b_2, \cdots, b_M)$ the sequence of bases of one of the strand of the molecule. So b_i is the i^{th} base which can be either a A, a T, a C or a G and the corresponding base of the other strand can be deduced. We will consider both a discrete and continuous time-sequence of the number of open base pairs, the first one is denoted X, the second one Y. We now make the link between X (and Y) and b. For this, we need some physical notions: the free energy g when the first x base pairs of the molecule are open is

$$g(x) := \sum_{i=1}^{x} g_0(b_i, b_{i+1}) - x g_1(f).$$

There are two different parts: first, $g_0(b_i, b_{i+1})$ is the binding energy of the pair i. Notice that stacking effects are taken into account: g_0 depends on the base content b_i and on the next pair b_{i+1}. The second contribution $g_1(f)$ is the work to stretch under a force f the open part of the two strands when one more base pair is opened, especially g_1 increases when f does. Note that g_1 is known, whereas $\sum_{i=1}^{x} g_0(b_i, b_{i+1})$ is random as we are looking for the b_i's. A typical trajectory of g is given in [4] page 7, it looks like Figure 3.

The number of opened base pairs evolves randomly with a probability directly connected to the difference of free energy g between two consecutive base pairs. Therefore it can be represented by a random walk in random environment:

The discrete case is defined as follows, assume that the random sequence $g_0 := (g_0(b_x, b_{x+1}), 1 \leq x \leq M-1)$ is fixed, then the probabilities of transition of the number of open pairs, are given by, for all $2 \leq x \leq M-1$

$$p_x = \mathbb{P}(X_{\cdot+1} = x+1 | X_{\cdot} = x, g_0) := \frac{1}{1 + \exp\left(\beta(g(x) - g(x-1))\right)}, \qquad (1)$$

where β is a constant parameter which is proportional to the inverse of the temperature, also we assume $p_1 = 1$. Note that this definition is such that

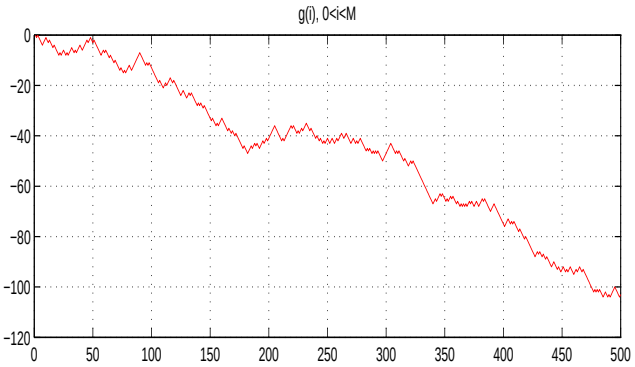

Figure 3: A typical trajectory of g, $M = 500$

the larger is f the greater is the probability to open a new pair. We easily get a simple expression for this probability which is

$$\mathbb{P}(X_{.+1} = x+1 | X_{.} = x,\ g_0) = \frac{1}{1+\exp(\beta \Delta g(b_x, b_{x+1}))}, \tag{2}$$

where we denote

$$\Delta g(b_x, b_{x+1}) := g_0(b_x, b_{x+1}) - g_1(f). \tag{3}$$

Formula (2) shows that we only need to have local information on the sequence b to get the probability of transition at site $x+1$. Though in the discrete case $X_{.}$ can only move forward with probability p_x or backward with probability $1-p_x$. We will discuss about some results on this well known model in the next section. A typical trajectory of X looks like Figure 4.

For *the continuous time model*, the physicists also take into account the time it takes X to go from a site to another. Thus we introduce a second time continuous model Y. Given the g_0, when Y is at the site x, it jumps in $x+1$ with rate $re^{-\beta g_0(b_x, b_{x+1})}$ and in $x-1$ with rate $re^{-\beta g_1(f)}$ where r is a constant which value depends on biological parameters. That is, given the DNA sequence b, Y is a Markov process with finite state space $\{1, \dots, M\}$ killed when it hits M whose transition rates are for $x > 1$,

$$p(x,y) = \begin{cases} re^{-\beta g_0(b_x, b_{y+1})} & \text{if } y = x+1, \\ re^{-\beta g_1(f)} & \text{if } y = x-1, \\ -r\big(e^{-\beta g_0(b_x, b_{x+1})} + e^{\beta g_1(f)}\big) & \text{if } y = x, \\ 0 & \text{otherwise.} \end{cases}$$

and for $x = 1$,

$$p(1,y) = \begin{cases} re^{-\beta g_0(b_1,b_2)} & \text{if } y = 2, \\ -re^{-\beta g_0(b_1,b_2)} & \text{if } y = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The process Y can be represented as the couple (X,T) where X is the sequence of the discrete jumps and has the same law as in (1) and T is the sequence of the successive times spent in each site between two jumps.

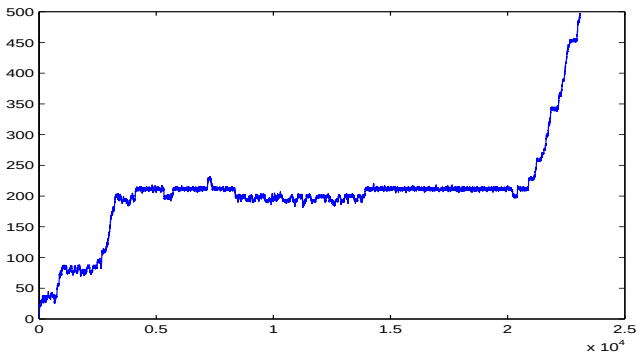

Figure 4: A typical trajectory of the number of unzipping pairs, $M = 500$.

We describe now briefly some results obtained by the physicists in the continuous time case

1.3 Some results obtained by the physicists

In their paper [3], [4], [7], they assume first that the model is without stacking effect, considering that g_0 is only a function of b_x and that $(g_0(b_x), x)$ is a sequence of independent and identically distributed random variables. In this case they compute the maximum likelihood estimator for b_x. For a better accuracy they consider several total unzipping instead of a single one, that is to say they look at a sequence of R independent trajectories $(Y_\cdot^{(l)}, l \leq R)$. In a second step they study the decreasing of the probability that this estimator gives a wrong sequence, and they show that this probability decreases exponentially; forall $i \leq M$,

$$\mathbb{P}(b_i \neq \hat{b}_i) \leq \exp(-R/R_c(i)).$$

The constant R_c is also estimated numerically. For the general case (with stacking effects) they use Viterbi algorithm [11] to compute the maximum of likelihood. Then they estimate the probability to be wrong with both analytics and numerical methods, they get a similar result than for the independent case.
After some discussions with S. Cocco and R. Monasson some questions rise: is it possible to get a general and rigorous method which can be applied to all these cases ? how the choice of the force can be used in order to improve the results ? and what is the difference between the discrete and continuous time model ? We study all those questions in the present paper.

1.4 A mathematical point of view

First we would like to recall some basic facts for the discrete time model. If we forget, for the moment, that the state space is finite, $(X_k,\ k \in \mathbb{N})$ is a random walk on a random environment on $\mathbb{Z}$ as Solomon defined it in [10]. We know, for example that if the $g_0(b_x, b_{x+1}) = g_0(b_x)$ are i.i.d with mean zero and $g_1 = 0$, then X is almost surely recurrent and transient on the other case. For the recurrent case, X is a Sinai's walk [9], for the transient one, the first study is due to H. Kesten, M.V. Kozlov, F. Spizer [8]. Here we are interested on what a trajectory of the walk can say about the environment, this aspect has not been studied a lot, there is a paper of O. Adelman, N. Enriquez [1] and for the special case of Sinai's walk a paper of P. Andreoletti [2]. More precisely [2] shows that the $g(x)$ can be estimated from a single trajectory of the walk by studying the asymptotic (in time) of the local time which is the amount of time the walk spends in one site. However this approach can not be used to give informations on a particular site, typically on $g_0(x)$ for a given x.
To move from Solomon walks to the problem asked by the physicists we have to make a sacrifice, more especially we are no longer interested in asymptotic in time. Indeed if the time goes to infinity that means that either we have to wait a very long time to reach the end of the molecule, or that if it is reached it can move back to the beginning. This last case is not possible because when the end of the molecule is reached then the two separate strands will not be able to reform the molecule properly. In compensation, we only have to study the process X or Y until it reaches M, that is until time

$$\tau_M = \inf\{k > 0, X_k = M\}. \tag{4}$$

So we are interested in the discrete time process $(X_k,\ k \leq \tau_M)$ and the continuous one $Y = (X_k, T_k,\ k \leq \tau_M)$. Note also that M is the length

of the DNA molecule, in term of the number of pairs, which can be big but finite. The other good news is the fact that the DNA molecule can be unzipped a large number of times, we have called, this number R, and we will be looking at asymptotic in this variable. So we are looking at R independent trajectories denoted $(Z_{t_l}^{(l)},\ 1 \leq l \leq R,\ 0 \leq t_l \leq \tau_M^{(l)})$ of random walks on a same unknown environment b with $\tau_M^{(l)}$ the first time the walk l hits M (Z is either X or $Y = (X, T)$). We state most of our results without any assumptions on the distribution of the sequence b. However to simplify the expressions we assume sometimes that all molecules are equiprobable.
The method is based on the fact that, given the trajectory of a random walk (or R random walks) on an environment b, the probability that a given estimator $\hat{b}$ gives a good sequence (typically $\mathbb{P}(b = \hat{b})$) depends only on elementary functions of the trajectory of this random walk.
For the *discrete time*, the important quantities are the number of times X goes from x to $x+1$ or to $x-1$:

$$L_x^{+,(l)} := \sum_{k=0}^{\tau_M^{(l)}-1} \mathbb{1}_{X_k^{(l)}=x;X_{k+1}^{(l)}=x+1}, \quad L_x^{-,(l)} := \sum_{k=0}^{\tau_M^{(l)}-1} \mathbb{1}_{X_k^{(l)}=x;X_{k+1}^{(l)}=x-1},$$
$$L_x^{+,R} := \sum_{l=1}^{R} L_x^{+,(l)} \quad \text{and} \quad L_x^{-,R} := \sum_{l=1}^{R} L_x^{-,(l)}.$$

For the *continuous time*, we have also to consider the total time spent in each site until the instant $\tau_M^{(l)}$ (which is as in the discrete case the hitting time of M for the processes $X^{(l)}$): for any $x \in [1, M]$,

$$S_x^{(l)} = \sum_{i=0}^{\tau_M^{(l)}} T_x^{(l)} \mathbb{1}_{X_i^{(l)}=x} \quad \text{and} \quad S_x^R = \sum_{l=0}^{R} S_x^{(l)}.$$

We will denote by X^R (Y^R in the continuous case) the σ-field generated by the trajectories of the R independent random walks killed when they hit the coordinate M. $\mathbb{P}$ denotes the probability distribution of the whole system, whereas P^α is the probability distribution for a given sequence of nucleotides α. Also E^α (resp. Var^α for the variance) is the expectation associated to P^α.
In Section 2, we start by the estimation base by base, we define the *information* at site x for both cases and show that the expression of the probability to get a given base at a site x conditionally on the trajectories are a simple function of the information. Then we study the asymptotic (in R) of the

probability that the maximum likelihood estimator gives a wrong base, we define and study a typical number of unzipping R_c which measures the quality of our prediction. In a second time we are interested in the estimation of the whole molecule, we start with a general expression of the probability to get a specific sequence given the trajectories of R random walks. We show that the maximum likelihood estimator converges. Then we study the probability to make at least h mistakes by considering this estimator, and study the decreasing of the probability to be wrong. We focus on the continuous case, and just quote the differences with the discrete case.
In Section 3, we focus on some possible improvements. The first one consists on modifying locally the force in order to trap the system in a specific region. It has a direct effect on the time spent in this region and therefore on the quality of the prediction. The second one consists also in modifying the force, but this time it is function of the binding energies.

2 Bayes estimator, asymptotics in R and typical number of needed unzipping R_c

In this section we will always assume that f is constant. We start with the estimation site by site:

2.1 Prediction site by site

We start with a general proposition true both for continuous and discrete time cases, then we discuss the differences between the two cases. First we define the following function i_x, it is called local information at site x of the system, it differs for the two cases. Let $x \in \{2, \cdots, M-1\}$ and $\alpha_{x-1}, \alpha_x, \alpha_{x+1} \in \{A, T, C, G\}^3$.
For the discrete case, the information is defined by

$$\begin{aligned} i_x(\alpha_{x-1}, \alpha_x, \alpha_{x+1}) := \\ L_x^{+,R} \log(1 + e^{\beta \Delta g(\alpha_x, \alpha_{x+1})}) + L_x^{-,R} \log(1 + e^{-\beta \Delta g(\alpha_x, \alpha_{x+1})}) \\ + L_{x-1}^{+,R} \log(1 + e^{\beta \Delta g(\alpha_{x-1}, \alpha_x)}) + L_{x-1}^{-,R} \log(1 + e^{-\beta \Delta g(\alpha_{x-1}, \alpha_x)}). \end{aligned}$$

and *for the continuous case,*

$$\begin{aligned} i_x(\alpha_{x-1}, \alpha_x, \alpha_{x+1}) := \beta g_0(\alpha_x, \alpha_{x+1}) L_x^{+,R} + S_x^R r e^{-\beta g_0(\alpha_x, \alpha_{x+1})} \\ + \beta g_0(\alpha_{x-1}, \alpha_x) L_{x-1}^{+,R} + S_{x-1}^R r e^{-\beta g_0(\alpha_{x-1}, \alpha_x)}. \end{aligned}$$

We are now ready to state the

Proposition 2.1. *For all $x \in \{2, \cdots, M-1\}$, and for $\alpha_x \in \{A, T, C, G\}$, denoting $b^x = (b_1, b_2, \cdots, b_{x-1}, b_{x,+1}, \cdots, b_{M-1})$, we have*

$$\mathbb{P}\left(b_x = \alpha_x | Z^R,\ b^x\right) = \frac{\exp(-I_x(\alpha_x, b))}{\sum_{\bar{\alpha}_x} \exp(-I_x(\bar{\alpha}_x, b))} \tag{5}$$

where

$$I_x(u, b) = I_x(u, b)(Z^R) := i_x(b_{x-1}, u, b_{x+1}) - \log \mathbb{P}(b_x = u | b^x),$$

and Z^R is either X^R for the discrete case or Y^R for the continuous one. The maximum likelihood estimator $\hat{b}_x$ for b_x, is given by:

$$\hat{b}_x = \sum_{\alpha_x \in \{A,T,C,G\}} \alpha_x \mathbb{1}_{\{I_x(\alpha,b) = \min_{\bar{\alpha}} I_x(\bar{\alpha},b)\}}. \tag{6}$$

and we have for all R large enough

$$\mathbb{P}\left(\hat{b}_x \neq b_x | Z^R, b^x\right) = \exp\left(-R/R_c(x) + \epsilon_x(R)\right), \tag{7}$$

where $R_c(x)$ is called the typical number of random walks at site x, *it is defined by*

$$1/R_c(x) := \lim_{R \to +\infty} \frac{-\log \mathbb{P}\left(\hat{b}_x \neq b_x | Z^R, b^x\right)}{R}. \tag{8}$$

Moreover, $\epsilon_x(R) \lesssim (2R \log\log R)^{1/2}((Var^b L_x^{+,(1)})^{1/2} + (Var^b L_x^{-,(1)})^{1/2})$ in the discrete case and $\epsilon_x(R) \lesssim (2R \log\log R)^{1/2}((Var^b L_x^{+,(1)})^{1/2} + (Var^b S_x^{(1)})^{1/2})$ in the continuous one. We denote $a \lesssim b$ if there exists a strictly positive constant r such that $a \leq rb$.

We will give the proof of a more general result in Section 2.2 so we do not give any details here.

Notice that $1/R_c(x)$ is no more and no less the rate function in the large deviation theory. The above proposition is general and gives only few informations on the decreasing of the probability to be wrong, so we now separate the two cases, and discuss about R_c.

The discrete case. First define the function $G_a : \mathbb{R} \to \mathbb{R}_+$,

$$G_a(u) := \log\left(\frac{1 + e^{\beta u}}{1 + e^{\beta a}}\right) + e^{\beta a} \log\left(\frac{1 + e^{-\beta u}}{1 + e^{-\beta a}}\right),$$

so G_a is positive and $G_a(a) = 0$. In fact the law of large numbers yields that almost surely Recall that Δg is defined in (3).Thanks to the law of large numbers

$$1/R_c(x) \approx \frac{1}{\bar{p}_{x-1}} \Delta G^-(b_{x-1}) + \frac{1}{\bar{p}_x} \Delta G^+(b_x),$$

with:

$$\begin{aligned}
\Delta G^-(b_{x-1}) &:= \min_{\alpha_x \neq \hat{b}_x} (G_{\Delta g(b_{x-1}, \hat{b}_x)}(\Delta g(b_{x-1}, \alpha_x)), \\
\Delta G^+(b_{x+1}) &:= \min_{\alpha_x \neq \hat{b}_x} (G_{\Delta g(\hat{b}_x, b_{x+1})}(\Delta g(\alpha_x, b_{x+1})). \\
\frac{1}{\bar{p}_x} = \frac{1}{\bar{p}_x(b)} &:= \sum_{k=x+1}^{M-1} \exp\left(\beta \left(g(k) - g(x)\right)\right) + 1 = \frac{1}{P^b_{x+1}(\tau_x \leq \tau_M)}. \qquad (9)
\end{aligned}$$

Note that we want $\Delta G^-(b_{x-1})$ and $\Delta G^+(b_{x+1})$ as large as possible, both of them measure the difference between the correct information which involves $\hat{b}_x$, and the other informations. Unfortunately they are possibly very small:

$$G_a(u) = \beta^2 (u-a)^2 \frac{G''_a(a)}{\beta^2} + o(u-a)^2, \qquad (10)$$

thus

$$\Delta G^-(b_{x-1}) \approx \beta^2 \min_{\alpha_x \neq \hat{b}_x} (\Delta g(b_{x-1}, \alpha_x) - \Delta g(b_{x-1}, \hat{b}_x))^2$$

when the right energy is close to another one. However this is not the only case where ΔG^+ and ΔG^- can be small, indeed, assume $u < 0$ and $a < 0$, then for large β,

$$G_a(u) \approx \exp(\beta a)(\exp(\beta(u-a)) - 1 - \beta(u-a)). \qquad (11)$$

This situation may appear when f is large and the binding energy at site x of the molecule is weak. We will see in Section 3 a method to avoid this kind of situation. We also have the following inequality:

$$\begin{aligned}
\frac{1}{\bar{p}_x} &\geq \exp(\beta(\max_{x \leq l \leq M} (g(l) - g(x)))) \\
&= \exp\left(\beta M_x\right), \qquad (12)
\end{aligned}$$

with $M_x := \max_{x \leq l \leq M} \left\{ \sum_{k=x+1}^{l} g_0(b_k, b_{k+1}) - (l-x) g_1(f) \right\}$. So, as expected, the convergence is better if there are obstacles in the path from x to M. Finally, we have

$$1/R_c(x) \geq \exp\left(\beta M_{x-1}\right) \Delta G^- + \exp\left(\beta M_x\right) \Delta G^+, \qquad (13)$$

with

$$\Delta G^- := \min\{\Delta G^-(\gamma), \gamma \in \{A, T, C, G\}\}$$

and

$$\Delta G^+ := \min\{\Delta G^+(\gamma), \gamma \in \{A, T, C, G\}\}.$$

A few words about $\epsilon_x(R)$ which comes from the iterated logarithm law: typically $(Var^b(L_x^+))^{1/2}$ as well as $(Var^b(L_x^-))^{1/2}$ behaves like $1/\bar{p}_x$ (see Lemma 2.2). Then $\epsilon_x(R)$ is always negligible comparing to the main term.

Formula useful for the estimation. As we have seen above, $R_c(x)$ characterizes locally the environment. However, what is really important to control the quality of the estimation at a point x is not the number of walks R but the total number of passages at this point, L_x^R. That is why we define the *typical number of visits at site* x, $L_c(x)$ by

$$1/L_c(x) := \lim_{R\to+\infty} \frac{-\log \mathbb{P}\left(\hat{b}_x \neq b_x | Z^R, b^x\right)}{L_x^R}, \tag{14}$$

we get

$$\frac{1}{L_c(x)} \geq \Delta G^+ \wedge \Delta G^-.$$

Total amount of time to reach M. An other important factor is the time required to unzip totally R times the DNA molecule. It should not be too large. This time is given by:

$$\tau_M^R = \sum_{l=1}^{R} \tau_M^{(l)} = \sum_{l=1}^{R} \sum_{x=1}^{M-2} (L_{x-1}^{+,(l)} + L_x^{+,(l)} - 1).$$

so we have

$$\begin{aligned} R\exp(\beta \max_x M_x) \lesssim E^b\left[\tau_M^R\right] =& R\sum_{x=1}^{M-2}\left(\frac{1}{\bar{p}_{x-1}} + \frac{1}{\bar{p}_x} - 1\right) \\ \lesssim& RM\exp(\beta \max_x M_x). \end{aligned}$$

Here is a problem with β: indeed as we have seen in the previous paragraph (see (12)), large β can lead to a better prediction, however it slows down the system. Of course it is worse if there is an obstacle between x and M because in this case M_x is large too.

The continuous time case.
Like for the discrete case we first define a function $F : \mathbb{R} \to \mathbb{R}_+$ by

$$\begin{aligned}
&F(u) = e^{\beta u} - 1 - \beta u \text{ and} \\
&\Delta F^- = \min\left(F(g_0(\alpha, u) - g_0(\alpha, v)), \alpha, u, v \in \{S, W\}, u \neq v\right), \\
&\Delta F^+ = \min\left(F(g_0(u, \alpha) - g_0(v, \alpha)), \alpha, u, v \in \{S, W\}, u \neq v\right),
\end{aligned}$$

then a similar analysis than for the discrete case leads to

$$1/R_c(x) \geq \frac{\Delta F^+}{\bar{p}_x} + \frac{\Delta F^-}{\bar{p}_{x-1}} \text{ and } 1/L_c(x) \geq \Delta F^+ \wedge \Delta F^-.$$

Note that the bad case observed for the discrete case (see equation (11)) does not appear here, however when $u - a$ is small, $F(u - a)$ is as $G_a(u)$ of the order of $a - u$ but the constant is better.

In the next section we look at the entire molecule, we define global information and study the decreasing of the probability to make a mistake by using the maximum likelihood estimator.

2.2 Inferring the whole molecule

First we present the joint distribution of $L_x^{+,(1)}, L_x^{-,(1)} = L_{x-1}^{+,(1)} - 1$, in fact it is not more difficult to get the joint distribution of $(L_x^{+,(1)}, 1 \leq x \leq M-1)$ and as we have not found it in the literature, we first prove the following lemma for one random walk and a constant force f:

Lemma 2.2. *If we denote $k = (k_i\ ,\ i \in \{1, \ldots, M-2\})$, then the distribution of $L^+ := L^{+,(1)}$ is*

$$\begin{aligned}
P^b\left(L^+ = k\right) =& p_{M-1}(1 - p_{M-1})^{k_{M-2}-1} \prod_{i=2}^{M-2} \binom{k_i + k_{i-1} - 2}{k_i - 1} p_i^{k_i} (1 - p_i)^{k_{i-1}-1}, \\
=& \prod_{i=2}^{M-1} \binom{k_i + k_{i-1} - 2}{k_i - 1} p_i^{k_i} (1 - p_i)^{k_{i-1}-1}
\end{aligned}$$

with $k_{M-1} = 1$. In particular, for $x \in \{2, \cdots, M-1\}$,

$$\begin{aligned}
&P^b\left(L_x^+ = k_x, L_x^- = k_{x-1}\right) = P^b\left(L_x^+ = k_x, L_{x-1}^+ = k_{x-1} + 1\right) \\
=& \binom{k_x + k_{x-1} - 1}{k_x - 1} (1 - p_x)^{k_{x-1}} (p_x(1 - \bar{p}_x))^{k_x - 1} (p_x \bar{p}_x), \qquad (15)
\end{aligned}$$

where $\bar{p}_x$ is given by (9). Moreover

$$E^b\left(L_x^+\right) = \frac{1}{\bar{p}_x},\ E^b\left(L_x^-\right) = \frac{e^{\beta\Delta g(b_x,b_{x+1})}}{\bar{p}_x} \text{ and } E^b\left(S_x^{(1)}\right) = \frac{e^{\beta g_0(b_x,b_{x+1})}}{r\bar{p}_x},$$
$$Var^b\left(L_x^+\right) = \frac{1}{\bar{p}_x}\left(\frac{1}{\bar{p}_x} - 1\right) \text{ and } Var^b\left(S_x^{(1)}\right) = \frac{e^{2\beta g_0(b_x,b_{x+1})}p_x}{r^2\bar{p}_x}.$$

Proof. The equality (15) of the lemma can easily be obtained by using the Markov property of X given b, the mean and the variance of L_x^+ and $S_x^{(1)}$ are direct consequences. Therefore we just prove the expression of the joint distribution of L^+. Define now for $n \geq 1$, $A_n := \bigcap_{j=n}^{M-1}\{L_j^+ = k_j\}$ where $k_{M-1} = 1$ (there is always only one jump from $M-1$ to M),

$$\begin{aligned} P^b\left(L^+ = k\right) &= P^b\left(A_1\right) = P^b\left(L_1^+ = k_1 \middle| A_2\right) P^b\left(A_2\right) \\ &= P^b\left(L_1^+ = k_1 \middle| L_2^+ = k_2\right) P^b\left(A_2\right) \end{aligned}$$

where the second equality comes from the Markov property of the walk X given b. Equation (15) implies for any $x \in \{2, \cdots, M-1\}$,

$$P^b\left(L_{x-1}^+ = k_{x-1} \middle| L_x^+ = k_x\right) = \binom{k_x + k_{x-1} - 2}{k_x - 1}(1-p_x)^{k_{x-1}-1}p_x^{k_x}$$

Thus,

$$P^b\left(L^+ = k\right) = \binom{k_2 + k_1 - 2}{k_2 - 1}p_2^{k_2}(1-p_2)^{k_1-1}P^b\left(A_2\right)$$

and we get the result of Lemma 2.2 recursively. □

We now define the *global information* I of the whole molecule. Let $\alpha \in \{A,T,C,G\}^M$.
For the discrete case X^R,

$$I(\alpha) := -\log \mathbb{P}(b = \alpha) + \sum_{x=1}^{M-1} L_x^{+,R}\log(1 + e^{\beta\Delta g(\alpha_x,\alpha_{x+1})}) + L_x^{-,R}\log(1 + e^{-\beta\Delta g(\alpha_x,\alpha_{x+1})}).$$

and *for the continuous case $Y^R = (X^R, T^R)$,*

$$I(\alpha) = -\log \mathbb{P}(b = \alpha) + \sum_{x=1}^{M-1} \beta g_0(\alpha_x,\alpha_{x+1})L_x^{+,R} + re^{-\beta g_0(\alpha_x,\alpha_{x+1})}S_x^R \quad (16)$$

We now give a general result and its proof:

Theorem 2.3. *For any $\alpha \in \{A,T,C,G\}^M$, we have:*

$$\mathbb{P}\left(b=\alpha\middle|Z^R\right)=\frac{e^{-I(\alpha)}}{\sum_{\bar{\alpha}} e^{-I(\bar{\alpha})}}. \tag{17}$$

The maximum likelihood estimator is the element $\hat{b}$ of $\{A,T,C,G\}^M$ which minimizes the function I. If the function

$$G_0:\ \alpha \to (g_0(\alpha_x,\alpha_{x+1}),\ x\in\{1,\cdots,M-1\})$$

is injective, then the maximum likelihood estimator converges almost surely to the DNA chain b.

Proof. We only give the proof for the continuous case, the discrete one is simpler and uses the same ideas. For a realization of $Y^R=\left(X^{(1)},\cdots,\ X^{(R)},\right.$ $\left.T^{(1)},\cdots,\ T^{(R)}\right)$, Bayes Lemma gives :

$$\begin{aligned}\mathbb{P}\left(b=\alpha|Y^R=y\right)&=\mathbb{P}\left(b=\alpha\middle|\bigcap_{l=1}^{R}\left\{X^{(l)}=x^{(l)},\ T^{(l)}=t^{(l)}\right\}\right)\\&=\frac{\mathbb{P}\left(\bigcap_{l=1}^{R}\left\{X^{(l)}=x^{(l)},\ T^{(l)}=t^{(l)}\right\}\middle|b=\alpha\right)\mathbb{P}(b=\alpha)}{\mathbb{P}\left(\bigcap_{l=1}^{R}\left\{X^{(l)}=x^{(l)},\ T^{(l)}=t^{(l)}\right\}\right)}\end{aligned}$$

(We still use $\mathbb{P}$ to denote a probability density.) When $X_i^{(l)}=x_i^{(l)}$, and the environment α are given, $T_i^{(l)}$ is an exponential variable independent from the other $X^{(k)},\ T^{(k)}$, of parameter $r\left(e^{-\beta g_0(\alpha_{x_i},\alpha_{x_i+1})}+e^{-\beta g_1(f)}\right)$ if $x_i^{(l)}\in\{2,\cdots,M-1\}$ or $re^{-\beta g_0(\alpha_1,\alpha_2)}$ if $x_i^{(l)}=1$. Thus,

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{l=1}^{R}\left\{X^{(l)}=x^{(l)},\ T^{(l)}=t^{(l)}\right\}\middle|b=\alpha\right)\\=&\mathbb{P}\left(\bigcap_{l=1}^{R}X^{(l)}=x^{(l)}\middle|b=\alpha\right)\prod_{l=1}^{R}\prod_{i=1}^{\tau_M^{(l)}-1}\mathbb{P}(T_i^{(l)}=t_i^{(l)}|b=\alpha,\ X_i^{(l)}=x_i^{(l)})\\=&\mathbb{P}\left(\bigcap_{l=1}^{R}X^{(l)}=x^{(l)}\middle|b=\alpha\right)(re^{-\beta g_0(\alpha_1,\alpha_2)})^{l_1^R}e^{-s_1^R re^{-\beta g_0(\alpha_x,\alpha_{x+1})}}\\&\times\prod_{x=2}^{M-1}(r(e^{-\beta g_0(\alpha_x,\alpha_{x+1})}+e^{-\beta g_1(f)}))^{l_x^R}e^{-s_x^R r(e^{-\beta g_0(\alpha_x,\alpha_{x+1})}+e^{-\beta g_1(f)})}\end{aligned}$$

$$\text{where } l_i^R = \sum_{l=1}^{R} \sum_{k=1}^{\tau_M^{(l)}-1} \mathbb{1}_{x_k^{(l)}=i} \quad \text{and} \quad s_i^R = \sum_{l=1}^{R} \sum_{k=1}^{\tau_M^{(l)}-1} \tau_{x^{(l)}}^{(l)} \mathbb{1}_{x_k^{(l)}=i}.$$

Moreover

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{l=1}^{R} X^{(l)} = x^{(l)} \middle| b = \alpha\right) \\
&= \prod_{x=2}^{M-1} \left(\frac{e^{-\beta g_0(\alpha_x,\alpha_{x+1})}}{e^{-\beta g_0(\alpha_x,\alpha_{x+1})} + e^{-\beta g_1(f)}}\right)^{l_x^{+,R}} \left(\frac{e^{-\beta g_1(f)}}{e^{-\beta g_0(\alpha_x,\alpha_{x+1})} + e^{-\beta g_1(f)}}\right)^{l_{x-1}^{+,R}-1} \\
&= \prod_{x=2}^{M-1} \frac{e^{-l_x^{+,R}\beta g_0(\alpha_x,\alpha_{x+1}) - (l_{x-1}^{+,R}-1)\beta g_1(f)}}{\left(e^{-\beta g_0(\alpha_x,\alpha_{x+1})} + e^{-\beta g_1(f)}\right)^{l_x^R}}
\end{aligned}$$

where

$$l_i^{+,R} = \sum_{l=1}^{R} \sum_{k=1}^{\tau_M^{(l)}} \mathbb{1}_{x_k^{(l)}=i,\ x_{k+1}^{(l)}=i+1}.$$

Then we have the following equality

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{l=1}^{R} \left\{X^{(l)} = x^{(l)},\ T^{(l)} = t^{(l)}\right\} \middle| b = \alpha\right) \\
&= \prod_{x=2}^{M-1} r^{l_x^R} e^{-s_x^R r(e^{-\beta g_0(\alpha_x,\alpha_{x+1})} + e^{-\beta g_1(f)}) - l_x^{+,R}\beta g_0(\alpha_x,\alpha_{x+1}) - (l_{x-1}^{+,R}-1)\beta g_1(f)}. \\
&\quad \times r^{l_1^R} e^{-s_1^R r e^{-\beta g_0(\alpha_1,\alpha_2)} - l_1^{+,R}\beta g_0(\alpha_1,\alpha_2)}.
\end{aligned}$$

It is now easy to obtain the expression of $\mathbb{P}(b = \alpha|Y^R)$ of the theorem.

We now prove the convergence of the maximum likelihood estimator. According to Lemma 2.2, for any $x \in \{1, \cdots, M-1\}$, the strong law of large number and the central limit theorem give

$$L_x^{+,R} = \frac{R}{\bar{p}_x(b)} + \epsilon_x(R) \quad \text{and } S_x^R = \frac{R e^{\beta g_0(b_x,b_{x+1})}}{r\bar{p}_x(b)} + \epsilon_x(R)$$

Then for any $\alpha \in \{A, T, C, G\}^M$, $\mathbb{P}$-a.s. the information $I(\alpha)$ is equivalent to

$$I(\alpha) = R \sum_{x=1}^{M-1} (\beta g_0(\alpha_x, \alpha_{x+1}) + e^{-\beta(g_0(\alpha_x,\alpha_{x+1}) - g_0(b_x,b_{x+1}))}) \frac{1}{\bar{p}_x(b)} + o(R).$$

This quantity is minimal if and only if for each $x \in \{1, \cdots, M-1\}$,

$$g_0(\alpha_x, \alpha_{x+1}) = g_0(b_x, b_{x+1}).$$

So if the function G_0 is injective, then, almost surely, for R large enough, $I(\alpha)$ is minimal iff $\alpha = b$. □

2.3 Control of the quality of the estimation for the continuous time case

In this part, we suppose that the function G_0 defined in the previous theorem is injective. Fix an $x \in \{1, \cdots, M-1\}$. The probability to make a mistake at site x when you know the environment around the point x is

$$\mathbb{P}(b_x = \hat{b}_x | Y^R,\ b^x) = \frac{e^{-i_x(b_{x-1}, \hat{b}_x, b_{x+1})}}{\sum_{\alpha_x} e^{-i_x(b_{x-1}, \alpha_x, b_{x+1})}}$$

As $\hat{b}_{x-1}$ and $\hat{b}_{x+1}$ converges to b_{x-1} and b_{x+1} according to Theorem 2.3, it is possible to estimate this probability by substituting the couple of variables (b_{x-1}, b_{x+1}) by $(\hat{b}_{x-1}, \hat{b}_{x+1})$. However this estimate is good only if $\hat{b}_{x-1}$ and $\hat{b}_{x+1}$ are themselves good estimates of b_{x-1} and b_{x+1}. It is also possible to obtain simpler asymptotic estimates which does not depend on b. As said before, for any $x \in \{1, \cdots, M-1\}$,

$$\begin{aligned} L_x^{+,R} = \frac{R}{\bar{p}_x} + \epsilon_x(R) \ \text{ and } S_x^R &= \frac{R}{\bar{p}_x r e^{g_0(b_x, b_{x+1})}} + \epsilon_x(R) \\ &= \frac{L_x^{+,R}}{r e^{g_0(b_x, b_{x+1})}} + \epsilon_x(R) \end{aligned}$$

where

$$\begin{aligned} \epsilon_x(R) &\lesssim (2R \log\log R)^{1/2} ((Var^b L_x^{+,(1)})^{1/2} + (Var^b S_x^{(1)})^{1/2}) \\ &= o((R \log\log R)^{1/2}). \end{aligned}$$

Then for any $\alpha_x \in \{A, T, C, G\}$, $i_x(b_{x-1}, \alpha_x, b_{x+1})$ is asymptotically equivalent to

$$\begin{aligned} i_x(b_{x-1}, \alpha_x, b_{x+1}) = &(\beta g_0(\alpha_x, b_{x+1}) + e^{-\beta(g_0(\alpha_x, b_{x+1}) - g_0(b_x, b_{x+1}))}) L_x^{+,R} \\ &+ (\beta g_0(b_{x-1}, \alpha_x) + e^{-\beta(g_0(b_{x-1}, \alpha_x) - g_0(b_{x-1}, b_x))}) L_{x-1}^{+,R} + \epsilon_x(R). \end{aligned}$$

This quantity is minimal if and only if $\alpha_x = b_x$ then

$$\sum_{\alpha} e^{-i_x(b_{x-1}, \alpha_x, b_{x+1})} = e^{-(\beta g_0(b_x, b_{x+1}) + 1) L_x^{+,R} + (\beta g_0(b_{x-1}, b_x) + 1) L_{x-1}^{+,R} + \epsilon_x(R)}.$$

As $\mathbb{P}$-almost surely, for R large enough,

$$\hat{b}_{x-1} = b_{x-1},\ \hat{b}_x = b_x \text{ and } \hat{b}_{x+1} = b_{x+1},$$

we obtain the following asymptotic estimate for the probability to make a wrong prediction at site x:

$$\begin{aligned}
&-\log \mathbb{P}(b_x \neq \hat{b}_x | Y^R,\ b^x) \geq \\
&F\big(g_0(\hat{b}_x, \hat{b}_{x+1}) - g_0(u_x^+, \hat{b}_{x+1})\big) L_x^{+,R} + \\
&F\big(g_0(\hat{b}_{x-1}, \hat{b}_x) - g_0(\hat{b}_{x-1}, u_x^-)\big) L_{x-1}^{+,R} + \epsilon_x(R)
\end{aligned}$$

where

$$\begin{aligned}
u_x^+ &= \text{argmin}\{u \neq \hat{b}_x,\ F\big(g_0(\hat{b}_x, \hat{b}_{x+1}) - g_0(u, \hat{b}_{x+1})\big)\}, \\
u_x^- &= \text{argmin}\{u \neq \hat{b}_x,\ F\big(g_0(\hat{b}_{x-1}, \hat{b}_x) - \ g_0(\hat{b}_{x-1}, u)\big)\},
\end{aligned}$$

recall also that $F(u) = e^{\beta u} - 1 - \beta u$, $\forall u \in \mathbb{R}$. Therefore this estimate is still true when b^x is not given:

$$\begin{aligned}
-\log \mathbb{P}(b_x \neq \hat{b}_x | Y^R) \geq &F\big(g_0(\hat{b}_x, \hat{b}_{x+1}) - g_0(u_x^+, \hat{b}_{x+1})\big) L_x^{+,R} + \\
&F\big(g_0(\hat{b}_{x-1}, \hat{b}_x) - g_0(\hat{b}_{x-1}, u_x^-)\big) L_{x-1}^{+,R} + \epsilon_x(R)
\end{aligned}$$

Same kind of arguments give the probability to be wrong on a chain of length $h+1$ starting at x.

Corollary 2.4. *Let n_e be the number of errors done by estimating the DNA chain. Almost surely for R large enough,*

$$\begin{aligned}
-\log \mathbb{P}\left(n_e \geq h \ | Y^R\right) &\geq K M_h^R + o(R) \\
&\geq KhR + o(R).
\end{aligned}$$

where

$$M_h^R = \min\left(\sum_{x \in I_h} L_x^{+,R} \ / \ I_h \subset \{1, \ldots, M-1\}\ ,\ |I_h| = h\right).$$

Proof. We begin with the definition of the information for the sites x to $x+h$: for $\alpha_{x-1}, \cdots, \alpha_{x+h+1} \in \{A, T, C, G\}^{h+3}$,

$$\begin{aligned}
i_{x,x+h}(\alpha_{x-1}, \cdots, \alpha_{x+h+1}) = &-\log \mathbb{P}(b_x = \alpha_x, \ldots, b_{x+h} = \alpha_{x+h}) \\
&+ \sum_{y=x-1}^{x+h} \beta g_0(\alpha_y, \alpha_{y+1}) L_y^{+,R} + r e^{-\beta g_0(\alpha_y, \alpha_{y+1})} S_y^R.
\end{aligned}$$

We can now express the probability we are looking for. Denote by $b^{x,x+h}$ the vector $(b_1, \cdots, b_{x-1}, b_{x+h+1}, \cdots, b_M)$, then

$$\mathbb{P}\left(b_x \neq \hat{b}_x, \ldots,\ b_{x+h} \neq \hat{b}_{x+h} | Y^R,\ b^{x,\ldots,x+h}\right)$$
$$= \sum_{\alpha_x \neq \hat{b}_x, \ldots, \alpha_{x+h} \neq \hat{b}_{x+h}} \frac{e^{-i_{x,x+h}(b_{x-1}, \alpha_x, \cdots, \alpha_{x+h}, b_{x+h+1})}}{\sum_{\bar{\alpha}_x, \ldots, \bar{\alpha}_{x+h}} e^{-i_{x,x+h}(b_{x-1}, \bar{\alpha}_x, \cdots, \bar{\alpha}_{x+h}, b_{x+h+1})}}$$

Reasoning in the same way as in the previous proof, we obtain

$$\sum_{\bar{\alpha}_x, \ldots, \bar{\alpha}_{x+h}} e^{-i_{x,x+h}(b_{x-1}, \bar{\alpha}_x, \cdots, \bar{\alpha}_{x+h}, b_{x+h+1})} = e^{-\sum_{y=x-1}^{x+h} (\beta g_0(b_y, b_{y+1})+1) L_y^{+,R} + o(R)}.$$

and then

$$\begin{aligned}
&-\log \mathbb{P}\left(b_x \neq \hat{b}_x, \ldots,\ b_{x+h} \neq \hat{b}_{x+h} | Y^R\right) \geq \\
&F\big(g_0(\hat{b}_{x-1}, \hat{b}_x) - g_0(\hat{b}_{x-1}, v_{x-1})\big) L_{x-1}^{+,R} + \\
&\sum_{y=x}^{x+h-1} F\big(g_0(\hat{b}_y, \hat{b}_{y+1}) - g_0(u_y, v_y)\big) L_y^{+,R} \\
&+ F\big(g_0(\hat{b}_{x+h}, \hat{b}_{x+h+1}) - g_0(u_{x+h}, \hat{b}_{x+h+1})\big) L_{x+h}^{+,R} + o(R)
\end{aligned}$$

where

$$(u_y, v_y) = \operatorname{argmin}\{u \neq \hat{b}_y,\ v \neq \hat{b}_{y+1},\ F\big(\beta g_0(\hat{b}_y,\ \hat{b}_{y+1}) - \beta g_0(u, v)\big)\}.$$

We can have as before a lower bound which does not depend from the estimator:

$$\begin{aligned}
-\log \mathbb{P}\left(b_x \neq \hat{b}_x, \ldots,\ b_{x+h} \neq \hat{b}_{x+h} | Y^R\right) &\geq K \sum_{y=x-1}^{x+h} L_y^{+,R} + o(R) \\
&\geq K(h+2)R + o(R).
\end{aligned}$$

which now leads easily to the corollary. □

Corollary 2.4 shows in particular that the probability to make at least h mistakes decreases exponentially with h.

The discrete time case leads to very similar results, in fact the main difference is that ΔF^+ and ΔF^- are replaced by ΔG^+ and ΔG^-.

3 Possible improvements of the method

In this paragraph, we use the results of the previous sections to propose two simple extensions which can be used to improve the predictions. The idea of both of them is based on the fact that the force f can be modified and adapted to the context.

3.1 Forces depending on the site we are interested in

In this section, we will discuss about $\frac{1}{\bar{p}_x}$, which appears in the lower-bounds of $1/R_c(x)$. Like we have seen before, $\frac{1}{\bar{p}_x}$ can be large, depending on the sequence g_0 and the force at site x, we recall that

$$\frac{1}{\bar{p}_x} = \sum_{l=x+1}^{M-1} \exp\left(\beta\left\{\sum_{k=x+1}^{l} g_0(b_k, b_{k+1}) - (l-x)g_1(f)\right\}\right)$$

For example when the force f is not too large, typically when $g_1(f) << \max(g_0(\alpha_x, \alpha_{x+1}))$, then valleys, that is to say portions of the sequences b such that $\sum_{k=x+1}^{l} g_0(b_k, b_{k+1}) - (l-x)g_1(f) >> 1$, can appear. So the quality of our prediction is good (the decreasing of the probability to be wrong is in $e^{-\text{const}Re^{\beta M_x}}$) only in some specific regions of the molecule.
When the above conditions do not appear, the force can be modified in order to slow down locally the system[1]. Assume that we are interested in a specific region centered on the point y, $[y-A, y+A]$, $A > 0$ where the $(L^+_{y+x} + L^-_{y+x}, x \in [-A, A])$ or the $(S_{y+x},\ x \in [-A, A])$ are small. Then we can take for $x \in [-A, A]$,

$$g_1(f_{y+x}) = C(A-x), \tag{18}$$

and we get:

$$\frac{1}{\bar{p}_{y+x}} \geq \exp\left(\beta\left\{\sum_{k=y+x+1}^{A+y} g_0(b_k, b_{k+1}) - \frac{C}{2}(A-x)(A-x-1)\right\}\right)$$

especially for $x = 0$,

$$\frac{1}{\bar{p}_y} \geq \exp\left(\beta\left\{\sum_{k=y+1}^{A+y} g_0(b_k, b_{k+1}) - \frac{C}{2}A(A-1)\right\}\right)$$

[1] According to the physicists, it is possible.

Then if $\mathbb{E}(g_0(b_k, b_{k+1})) \geq \frac{C}{2}(A-1)$ and A is large enough, $\frac{1}{\bar{p}_y}$ will be quite large too. Once again this will work if the region we are looking at is quite far from the end of the molecule, that is to say, A is large. On the other case what could be a good idea is to unzip the molecule from the other end. We now move to another possible improvement.

3.2 The energy point of view: forces depending on the values of the environment

In this paragraph we do not try to find directly the sequence of bases but the associated binding energies. We denote $g_0(x)$ for $g_0(b_x, b_{x+1})$and we assume that they are K distinct values for $g_0(x)$, typically for the DNA they are given by Table 2. The random variables $(g_0(x), x)$ are not independent, for the DNA molecule the dependence is also given by Table 2. For example, the energy 1.06 can only be followed by 1.78, 1.55, 2.52 or 2.22. We also assume that the whole sequences g_0 are equiprobable.

First let us introduce some new notations. We will denote $\mu_1, \mu_2, \cdots, \mu_K$, the possible values of $g_0(.)$, they are ordered in such a way that $\mu_i > \mu_{i+1}$ for all i. We also assume that the force f can take $K+1$ different values $\{f_1, f_2, \cdots, f_{K-1}, f_K, f_{K+1} = 0\}$, and then, that $g_1(.)$ takes equally $K+1$ distinct values denoted $\{r_1, r_2, \cdots, r_{K-1}, r_K, r_{K+1} = 0\}$. We also suppose that $r_1 > r_2 > \cdots > r_{K-1}$, and

$$\begin{aligned}
&\mu_1 - r_1 < 0,\ \mu_1 - r_2 > 0,\ \forall i > 1\ \mu_i - r_2 < 0,\\
&\mu_2 - r_2 < 0,\ \mu_2 - r_3 > 0,\ \forall i > 2\ \mu_i - r_3 < 0,\\
&\cdots\\
&\mu_K - r_K < 0,\ \mu_K - r_{K+1} = \mu_K > 0.
\end{aligned}$$

Let us define

$$q_m^i := (1 + e^{\beta(\mu_m - r_i)})^{-1}, \tag{19}$$

which is the probability to go on the right if the force f_i is applied and if the value of the environment is equal to μ_m. Notice that if f_1 is applied then for all $x \leq M$,

$$p_x := (1 + \exp(\beta(g_0(x) - r_1)))^{-1} \geq q_1^1 > 1/2,$$

and we denote $\Gamma_1 := \{x \leq M, p_x = q_1^1\}$. Then if f_2 is applied, for all $x \leq M,\ x \notin \Gamma_1$,

$$p_x \geq q_2^2, \tag{20}$$

in the same way as before we denote Γ_2 the sites such the equality in (20) is satisfied. We get a partition $\{\Gamma_1, \Gamma_2, \cdots, \Gamma_K\}$ of $\{1, \cdots, M\}$. Finally we can look at a certain number of random walks for each values taken by the force. We denote by R_j the number of random walks we consider for the force with value f_j.
From now on we will only focus on the discrete time case because it is the one where the gain is the most important, however what we suggest can be applied to the continuous time model as well. We introduce the information at site x, if the force f_j is applied then

$$i_x^j(m) := L_x^{+,R_j} \log q_m^j + L_x^{-,R_j} \log(1 - q_m^j), \tag{21}$$

and the relative information at site x

$$i_x^j(m,l) := i_x^j(m) - i_x^j(l). \tag{22}$$

We define the function $H_a : \mathbb{R} \to \mathbb{R}_+$,

$$H_a(u) := \log(1 + \exp(\beta u)) + \exp(\beta a) \log(1 + \exp(-\beta u)).$$

Proposition 3.1. *Assume that the forces f_k and then f_{k+1} are applied (everywhere) then, for all x, all sequence g_0^x and all estimator $\hat{g}_0(x)$,*

$$\begin{aligned}
&\mathbb{P}\left(\hat{g}_0(x) = \mu_k,\ g_0(x) = \mu_k | X^{R_k}, X^{R_{k+1}},\ g_0^x\right) \\
&= \left(1 + \sum_{m=1, m\neq k}^{K} \exp\left(-i_x^k(m,k) - i_x^{k+1}(m,k)\right)\right)^{-1} \mathbb{1}_{\hat{g}_0(x)=\mu_k}.
\end{aligned}$$

Let us define the following estimator:

$$\hat{g}_0(x) = \inf\left\{k > 0,\ \frac{L_x^{-,R_k}}{L_x^{+,R_k}} < 1,\ \frac{L_x^{-,R_{k+1}}}{L_x^{+,R_{k+1}}} > 1\right\} \tag{23}$$

then

$$\begin{aligned}
\frac{1}{R_c^k(x)} &:= -\lim_{R_k = R_{k+1} = R \to \infty} \log \frac{1}{R} \left(\mathbb{P}\left(\hat{g}_0(x) = \mu_k,\ g_0(x) = \mu_k | X,\ g_0^x\right) - 1\right) \\
&\geq \frac{H^{(k)}}{\bar{p}_x^k} + \frac{H^{(k+1)}}{\bar{p}_x^{k+1}},
\end{aligned}$$

where

$$H^{(k)} := \min_{l\in\{k-1,k+1\}} H_{\mu_l - r_k}(\mu_l - r_k) - H_{\mu_k - r_k}(\mu_k - r_k),$$

$$H^{(k+1)} := \min_{l\in\{k-1,k+1\}} H_{\mu_l - r_k}(\mu_l - r_{k+1}) - H_{\mu_k - r_{k+1}}(\mu_k - r_{k+1}),\ \textit{and}$$

$$\frac{1}{\bar{p}^l_x} := \frac{1}{\bar{p}^l_x(g^x_0)}\ \textit{with, for all sequence } \gamma^x, \tag{24}$$

$$\frac{1}{\bar{p}^l_x(\gamma^x)} := \sum_{z=x+1}^{M-1} \exp\left(\sum_{y=x+1}^{z} \gamma(y) - r_l\right) + 1. \tag{25}$$

γ^x *is (like for* b^x*) the sequence* $(\gamma(1), \gamma(2), \cdots, \gamma(x-1), \gamma(x+1), \cdots, \gamma(M))$.

Here we avoid a bad situation seen in the first section (see (11)), indeed, the worst case under the force f_{k+1}, leads to: for $a = \mu_k - r_{k+1} > 0, u = \mu_{k+1} - r_k + 1 > 0$,

$$H_a(u) - H_a(a) \gtrsim (\exp(\beta(a-u)) - 1 - \beta(a-u)), \tag{26}$$

which increases with β. However we have to be careful with this method, indeed in order to catch the small values of the energies, f_k should be small and slows down the system, indeed recall that

$$\begin{aligned} E^b[\tau_M] =& R\sum_{x=1}^{M-2}\left(\frac{1}{\bar{p}^k_x} + \frac{1}{\bar{p}^k_{x-1}} - 1\right) + R\sum_{x=1}^{M-2}\left(\frac{1}{\bar{p}^{k-1}_x} + \frac{1}{\bar{p}^{k-1}_{x-1}} - 1\right) \\ \gtrsim& R\exp(\beta \max_x M^k_x), \end{aligned}$$

with $M^k_x := \max_{x\le l\le M}\left\{\sum_{l=x+1}^{l} g_0(b_l, b_{l+1}) - r_k\right\}$ large if k is large.
An alternative approach is first to apply a large force f_1, from 0 to $x-1$ in order to reach rapidly the region we are interested in, then apply all the forces in x and then, after $x+1$ apply a small force (for example f_K) in order to slow down the system and stay focus on x. For each value of the force $(f_i, i \le K)$ at site x, R random walks are used. More precisely f is, as before, function of the energy and now of the sites:

$$f_i(z) = f_1 \mathbb{1}_{1\le z\le x-1} + f_i \mathbb{1}_{z=x} + f_K \mathbb{1}_{z\ge x+1} \tag{27}$$

we get

$$\begin{aligned}\frac{1}{R_c(x)} &:= -\lim_{R\to\infty}\frac{1}{R}\log\left(\mathbb{P}\left(\hat{g}_0(x)\neq g_0(x)|X^R,\ g_0^x, f_i(.), i\leq K\right)\right)\\ &\geq \frac{1}{\bar{p}_x^K}(H^{\rightarrow}+H^{\leftarrow})\\ H^{\rightarrow} &:= \max_{k\leq K-1}\min_{l\in\{k-1,k+1\}}(H_{\mu_l-r_k}(\mu_l-r_k)-H_{\mu_k-r_k}(\mu_k-r_k))\\ H^{\leftarrow} &:= \max_{k\leq K-1}\min_{l\in\{k-1,k+1\}}(H_{\mu_l-r_k}(\mu_l-r_{k+1})-H_{\mu_k-r_{k+1}}(\mu_k-r_{k+1})).\end{aligned}$$

The main interest in the above result comparing to the preceding one is the fact that $\frac{1}{\bar{p}_x^K}$ is large but the time to reach x is small. Of course this also increases the amount of time to reach the end of the molecule, but it can be stopped. The proof to get these results are very close to the one of Section 2 so we do not give any details.

In the above proposition we assume that g_0^x is known which can be restrictive, indeed we do not know anything on the molecule, we can remove this assumption and get the following result. First we need to introduce another function of information $\bar{i}$: for fixed x, m and j

$$\bar{i}_x^j(\mu_m^x) := L_x^{+,R_j}\log(1-\bar{p}_x^j(\mu_m^x))+R(\log(\bar{p}_x^j(\mu_m^x))-\log(1-\bar{p}_x^j(\mu_m^x))).$$

recall that $\bar{p}_x^j()$ is given by (25), it is a function of energies here μ_m^x is the sequence (from $x+1$ to M) of energies compatible with the energy μ_m of site x.

Proposition 3.2. *For all x, all estimator $\hat{g}_0(x)$, assume that each of the forces $(f_i(.), i\leq K)$ are applied to respectively $(R_i, i\leq K)$ random walks, then*

$$\begin{aligned}&\mathbb{P}\left(\hat{g}_0(x)=g_0(x)|\, L_x^{+,R_i}, L_{x,}^{-,R_i}, f_i(.),\ i\leq K\right)\\ &=\left(1+\sum_{m=1,\mu_m\neq\hat{g}_0(x)}^{K} e^{\sum_{j=1}^K -(i_x^j(m)-i_x^j(0))}\frac{\sum_{\mu_m^x}e^{\bar{i}_x^K(\mu_m^x)}}{\sum_{\mu_x^0}e^{\bar{i}_x^K(\mu_0^x)}}\right)^{-1}\end{aligned}$$

where the sum over the μ_m^x is the sum over all the compatible energies (from $x+1$ to the M) with μ_m of site x. μ_0^x is the same, but the compatibility is with $\hat{g}_0(x)$ instead of μ_m. $i_x^l(m)$ is given by (21), and for $i_x^l(0)$ we replace μ_m in $i_x^l(m)$ by $\hat{g}_0(x)$.
Moreover if we take for the estimator at site x defined in (23) *and we assume*

that all the R_i are equal we get as before

$$\frac{1}{R_c(x)} := -\lim_{R\to\infty} \log \frac{1}{R}\left(\mathbb{P}\left(\hat{g}_0(x) \neq g_0(x) | L_x^{+,R_i}, L_{x,}^{-,R_i}, i \leq K\right)\right)$$
$$\geq \frac{1}{\bar{p}_x^K}(H^{\rightarrow} + H^{\leftarrow}).$$

Proof. Thanks to Lemma 2.2, we easily get the first part of the proposition,

$$\lim_{R\to+\infty} \frac{1}{R}\bar{i}_x^j(\mu_m^x) = \left(\frac{1}{\bar{p}_x^j(\mu_0^x)} - 1\right)\log(1-\bar{p}_x^j(\mu_m^x)) + \log(\bar{p}_x^j(\mu_m^x)).$$

Also we have that $H(x) := (1/a-1)\log(1-x) + \log x$ reaches its maximum in a, we deduce from that $\frac{\sum_{\mu_m^x} e^{\bar{i}_x^K(\mu_m^x)}}{\sum_{\mu_x^0} e^{\bar{i}_x^K(\mu_x^0)}} \leq 1$, so we get the lower bound for $1/R_c(x)$. □

We finish with a discussion about the link between the energies and the sequences of bases. First let us recall the table of the binding free energies for DNA at room temperature:

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

We notice that the largest free energies which correspond to the most stable links are on the bottom right end corner of the table, in fact the largest binding energy is obtained when a G is followed by a C. Notice also that $g_0(G,G) = g_0(C,C)$ so we can not distinguish this two different links by looking only at the free energy. In the same way the lowest free energy are made with bases with a T and A followed by the same letters, again $g_0(A,A) = g_0(T,T)$. For the rest of the table we have the equality $g_0(W,S) = g_0(\bar{S},\bar{W})$, where S is either a C or a G and W a A or a T, $\bar{S}$ the complementary of S, and the same for $\bar{W}$.

Moreover it is possible to reconstruct the DNA molecule from the compatible binding energies only if there is only one sequence of base pairs which corresponds to the sequence of energies (see Theorem 2.3). This is not always the case, for example when the molecule repeats the same scheme which is undetermined: the energy of $C-C-\cdots-C$ is equal to the energy of $G-G-\cdots-G$, in the same way $A-C-A-C-\cdots-A-C$ has the

same energy than $G-T-G-T-\cdots-G-T$. Notice that if this kind of sequence is broken only once in the molecule then we turn to a determined case.

Acknowledgments We would like to thank Nathanael Enriquez and the members of the ANR MEMEMO who enable us to meet Rémi Monasson. Also we would like to thank Rémi Monasson and Simona Cocco for introducing the subject, sharing several discussions and a kind invitation at the ENS of Physic.

References

[1] O. Adelman and N. Enriquez. Random walks in random environment: What a single trajectory tells. *Israel J. Math.*, 142:205–220, 2004.

[2] P. Andreoletti. On the estimation of the potential of Sinai's rwre. *To appear in Braz. Journal of Prob. and Stat.*, 2010.

[3] V. Baldazzi, S. Cocco, E. Marinari, and R. Monasson. Infering dna sequences from mechanical unzipping: an ideal-case study. *Physical Review Letters E*, **96**: 128102–1–4, 2006.

[4] V. Baldazzi, S. Cocco, E. Marinari, and R. Monasson. Infering dna sequences from mechanical unzipping data: the large-bandwith case. *Physical Review Letters E*, **75**: 011904–1–33, 2007.

[5] U. Bockelmann, B. Essevaz-Roulet, and F. Heslot. Molecular stick-slip motion revealed by opening dna with piconewton forces. *Phys. Rev. Let.*, **79**: 4489–4492, 1997.

[6] U. Bockelmann, B. Essevaz-Roulet, and F. Heslot. Dna strand separation studied by single molecule force measurements. *Phys. Rev. E*, **58**: 2386–2394, 1998.

[7] S. Cocco and R. Monasson. Reconstructing a random potential from its random walks. *epl*, **81**: 1–6, 2008.

[8] H. Kesten, M.V. Kozlov, and F. Spitzer. A limit law for random walk in a random environment. *Comp. Math.*, **30**: 145–168, 1975.

[9] Ya. G. Sinai. The limit behaviour of a one-dimensional random walk in a random medium. *Theory Probab. Appl.*, **27**(2): 256–268, 1982.

[10] F. Solomon. Random walks in random environment. *Ann. Probab.*, **3**(1): 1–31, 1975.

[11] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**(2):260–269, 1967.