



Journal of Nonlinear Mathematical Physics

ISSN (Online): 1776-0852

ISSN (Print): 1402-9251

Journal Home Page: <https://www.atlantis-press.com/journals/jnmp>

Bayesian Approach to the Determination of the Kinetic Parameters of DNA Hairpins Under Tension

Marco Ribezzi-Crivellari, Mario Wagner, Felix Ritort

To cite this article: Marco Ribezzi-Crivellari, Mario Wagner, Felix Ritort (2011) Bayesian Approach to the Determination of the Kinetic Parameters of DNA Hairpins Under Tension, Journal of Nonlinear Mathematical Physics 18:Supplement 2, 397–410, DOI: <https://doi.org/10.1142/S1402925111001593>

To link to this article: <https://doi.org/10.1142/S1402925111001593>

Published online: 04 January 2021

Journal of Nonlinear Mathematical Physics, Vol. 18, Suppl. 2 (2011) 397–410

© M. R. Crivellari, M. Wagner and F. Ritort

DOI: [10.1142/S1402925111001593](https://doi.org/10.1142/S1402925111001593)

BAYESIAN APPROACH TO THE DETERMINATION OF THE KINETIC PARAMETERS OF DNA HAIRPINS UNDER TENSION

MARCO RIBEZZI-CRIVELLARI^{*,†,§}, MARIO WAGNER[†]
and FELIX RITORT^{†,‡,¶}

^{*}*Dipartimento di Fisica, Università di Roma 3
Via della Vasca Navale 84, Roma 00146 Italy*

[†]*Small Biosystems Lab, Departament de Física Fonamental
Universitat de Barcelona, Av. Diagonal 647
08028 Barcelona, Spain*

[‡]*CIBER-BBN de Bioingeniería, Biomateriales y Nanomedicina,
Instituto de Sanidad Carlos III, Madrid, Spain*

[§]*marco.ribezzi@gmail.com*

[¶]*fritort@gmail.com*

Received 5 November 2010

Revised 28 February 2011

Accepted 2 March 2011

In this paper we propose a Bayesian scheme for the determination of the unfolding and refolding kinetic rates of DNA hairpins under tension. This method is based on the hypothesis that the unfolding-refolding dynamics is well described by a Markov Chain. The results from the Bayesian method are contrasted to widely used techniques and good agreement is found. This work can be seen as a validation of the standard techniques from a statistical point of view.

Keywords: Optical tweezers; Bayesian reasoning; DNA hairpins.

2000 Mathematics Subject Classification: 22E46, 53C35, 57S20

1. DNA Hairpins

The structure of nucleic acid filaments in solution is determined by the specific sequence of nucleotides by which the filament is composed, also called “primary structure”, and by ambient conditions such as temperature, pH and salt concentration. Given the ambient conditions, a nucleic acid filament will often assume a specific three dimensional structure. This structure is a compromise between the free energy gain from base pairing (hydrogen bond formation between complementary bases) and base stacking (Wan Der Waals interactions between nearest-neighbor bases) on one side and entropic costs on the other side. In particular the pattern of base pairing in nucleic acid structures is often referred to as secondary structure. A particular kind of secondary structure are DNA hairpins, formed by palindromic single-stranded DNA sequences. Palindromic sequences are such that the first

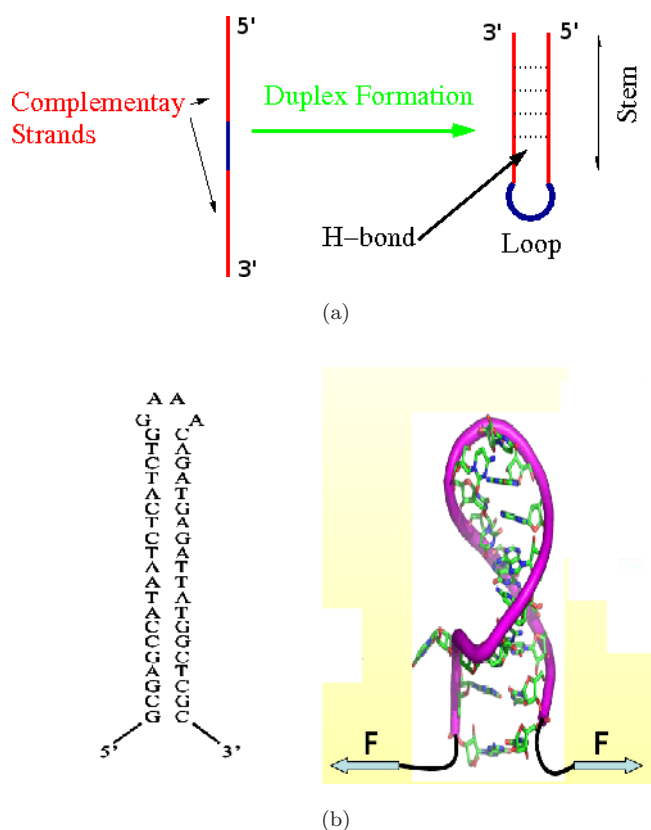


Fig. 1. Schematic description of the DNA hairpin structure. (a) Unfolded and folded hairpin structure, the red color denotes the complementary part of the hairpin, while blue denotes the loop region. (b) An example of hairpin forming sequence together with its three-dimensional structure. Force is applied at the beginning of the stem.

n bases near one end of the strand are complementary to the last n bases at the other end of the strand taken in reverse order (see Fig. 1). If the strand is N bases long, this means that the first base is complementary to the N th, the second to the $N - 1$ th and so on. In this way the single stranded filament can fold onto itself. The only exception are a few bases at the center of the sequence. These bases form the loop region. If the ambient conditions do favor base pairing, such single strand will assume the secondary structure shown in Fig. 1 which maximizes the free energy gain from pairing and stacking interactions [1]. The hairpin is a structural element that is present in DNA and RNA molecules in vivo as well as in vitro. RNA hairpins in vivo are known to play a key role in biological functions such as regulation of gene expression [2, 3]. But most importantly here, nucleic acid hairpins serve as model systems for secondary structure formation in DNA and RNA [4–10] (see Fig. 1). In the case of hairpins the two energetic factors involved in structure formation are the entropy cost for loop formation and the free energy gain from base pairing and stacking. In particular the formation of the loop can be considered as a nucleation step in the transition to the folded state and it is the rate limiting step in the folding kinetics. The kinetics and energetics of structure formation in DNA hairpins can be studied in detail by varying the sequence, the loop length and ambient conditions. Nowadays the native or minimum free energy structure can be predicted using the large amount of data obtained in the last decades from

bulk experiments and which were recently confirmed at the single molecule level by force unfolding experiments [1].

2. Single Molecule Experiments on DNA Filaments

Single molecule techniques have arisen as a new tool in the analysis of biological systems. In particular they have allowed to test for the effect of applied tensions on the structure of biological polymers. Among the force spectroscopy techniques, optical trapping is particularly fit to explore the properties of bio-molecules. As first proved by Ashkin in 1970 [11], highly focused laser beams can be used to trap and displace micron-sized dielectric beads, applying onto these objects forces in the range 1–100 pN. Moreover, by recollecting the light of the laser beam after its interaction with the trapped object it is also possible to accurately measure the applied force by linear momentum conservation [12]. Experimental setups designed for optical trapping are often called Optical Tweezers (OT). Indeed using OT it is possible to study the nonlinear elasticity of DNA filaments or unfold nucleic acid structures into an extended single strand by applying mechanical force in different physiological buffers and temperatures. In these experiments the nucleic acid sequence to be studied is chemically anchored to two micron sized dielectric beads. One of these beads is then captured in an optical trap formed by laser beams [1], while the other is trapped by air suction on the tip of a micropipette Fig. 2(a). This setup can be used to perform different kinds experiments.

In pulling experiments, for example, force is raised at a constant rate during the experiment. In this way it is possible to measure the elongation of a molecule as a function of the applied force, giving the so called Force-Distance Curve (FDC). The FDC is the molecular analogous of the volume pressure isotherm of classical thermodynamics. If the pulling rate is slow enough, a reversible FEC can be obtained, which gives, by integration, the free energy difference between the initial and final configuration. Nevertheless, in many practical cases, reversibility is never attained and the free energy difference can only be recovered from irreversible FECs by the use of nonequilibrium work relations such as the Jarzynski or Crooks relations [13, 14].

When the kinetics of DNA unfolding is to be measured it is convenient to keep the distance between the trap and the pipette fixed and to measure the lifetimes of the folded and unfolded states, the so-called Passive Mode (PM) [6, 7] hopping setup. The transition from the folded to the unfolded state will be detected as a drop in the applied force, since the longer end to end distance of the tethered molecule in the unfolded state allows for the bead to relax towards the center of the trap Fig. 2(a).

2.1. Hopping experiments on DNA hairpins

Some hairpins, when held at moderate tensions $\simeq 15pN$, will jump back and forth in between the folded and the unfolded states under the effect of thermal fluctuations. This allows a precise measurement of the folding kinetics and thermodynamics of the hairpin. In addition hopping experiments make possible the complete reconstruction of the folding free energy landscape along the molecular extension both in DNA hairpins [9, 10] and in proteins [15]. In this article we will consider hopping experiments performed on a DNA hairpin with a 20 bp stem and a five bp loop (see Fig. 3). In such a short hairpin the transition between

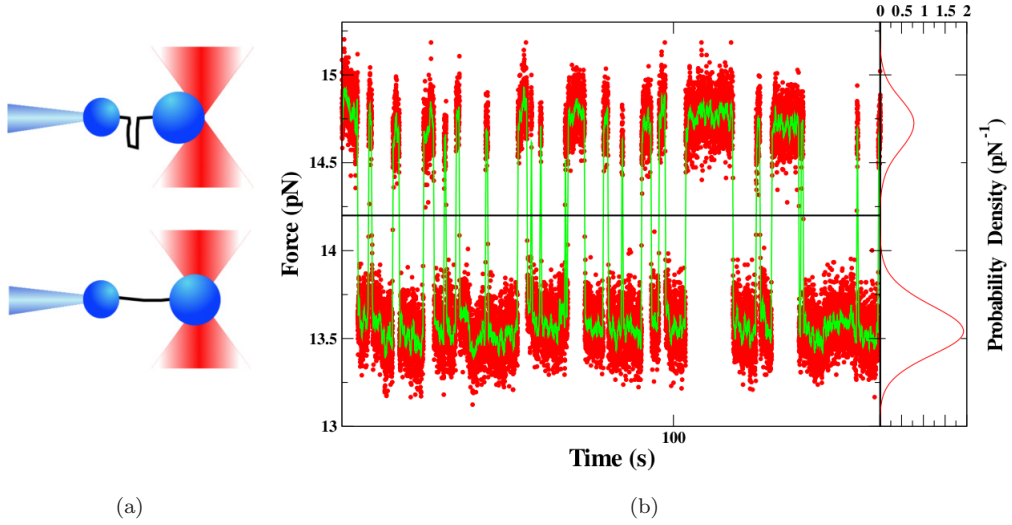


Fig. 2. (a) Schematic representation of the Passive Mode (PM) hopping setup [6, 7]. The DNA hairpin is linked to two dielectric beads. One bead is optically trapped, while the other is held by air suction on the tip of a micropipette. The distance between the micropipette and the bead is kept fixed. At moderate tensions the hairpin can hop back and forth between the folded and the unfolded state under the effect of thermal fluctuations. The transition from the folded to the unfolded state is detected as a drop in the applied force, as the longer contour length of the molecule in the unfolded state allows for a relaxation of the bead in the optical trap. (b) The force trace obtained from the DNA hairpin under tension. The molecule is folded when the force takes the higher value and unfolded when it takes the lower value. Some unfolding/refolding events are shown. The small circles show raw data, while the continuous line is a running average with 40 points window on the measured data. The continuous straight line is equal to \bar{f} , the median between the force in the folded state and the force in the unfolded state. On the right panel the probability density of the measured force is shown for the whole trace ($\simeq 300$ s), showing how no intermediate is detected and the two states are perfectly resolved.

the folded and the unfolded state happens abruptly without detectable intermediates (see Fig. 2(b)), all the stem unfolds at once.

The mean lifetime of the folded and the unfolded states are, in the case we will describe, on the order of the second and the high instrumental stability of optical tweezers allows for the measurement of hundreds of transitions. Each of these measurements leads to a force vs. time trace (Fig. 2(b)) in which the two states are clearly distinguishable. This is because the Signal to Noise Ratio (SNR), defined as the ratio between the force drop in the unfolding transition and the amplitude of thermal noise, is large enough. The data acquisition rate is finite, with a point being sampled every millisecond. The dots in Fig. 2(b) show the collected raw data, while the solid line shows an average of the measured data.

3. The Free Energy Landscape

The mechanical folding and unfolding of nucleic acid hairpins is commonly described in terms of a reaction coordinate and the corresponding free energy landscape. When subject to force, the end-to-end distance of the molecule along the force axis is an adequate reaction coordinate for the folding-unfolding reaction pathway. For a given applied force f it is common to consider only a single kinetic pathway for the unfolding and folding reactions, which is characterized by a single transition state (TS). The TS is the highest free energy

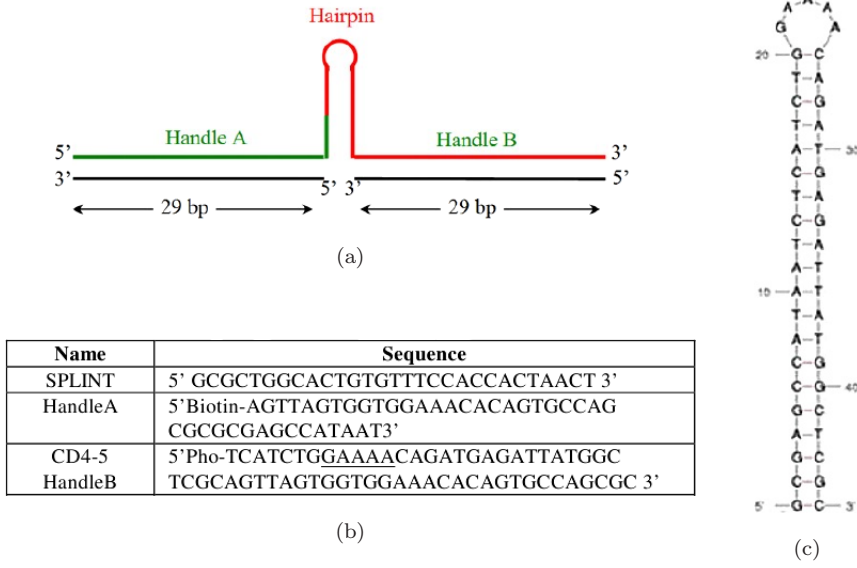


Fig. 3. Schematics of the hairpin synthesis. The actual construct used in the experiments involves the hairpin whose thermodynamics and kinetics are to be tested and two molecular “handles” which act as spacers to keep the hairpin away from the dielectric beads. (a) The synthesis uses three different oligos. One “Handle A” in the figure, encompasses the left handle and the first part of the hairpin sequence. The second oligo contains the second part of the oligo and the right handle. The third oligo completes both handles, exploiting their symmetry. (b) explicit sequence of the three oligos, the loop sequence is underlined. (c) The hairpin sequence. This synthesis procedure is described in detail in [16].

state along the reaction coordinate and determines the kinetics of the folding-unfolding reaction. In the following we shall denote by the variable σ the folding state of the hairpin, with $\sigma = 0$ corresponding to the folded state and $\sigma = 1$ corresponding to the unfolded state. The model we will use involves four parameters: the free energy difference between the two states $\sigma = 0$ and $\sigma = 1$ at zero force, $\Delta G_0 = G(0) - G(1)$, the height of the kinetic barrier B , defined as the free energy difference at force f between the TS and the folded $\sigma = 0$ state, and the distances X_0 and X_1 along the reaction coordinate that separates the transition state from states $\sigma = 0$ and $\sigma = 1$ respectively. The total distance along the reaction coordinate between $\sigma = 0$ and $\sigma = 1$ is defined as $X_m = X_0 + X_1$ (see Fig. 4). Under an applied force the free energy landscape is tilted along the reaction coordinate, changing the free energy difference ΔG and the barrier B . In a first approximation ΔG and B change linearly with the force whereas X_0 and X_1 are taken as constant: under these assumptions the reaction rates are given by:

$$k_{0 \rightarrow 1}(f) \equiv q(f) = k_0 e^{-\beta(B - G(0) - X_0 f)} = e^{z + x f} \quad (3.1)$$

$$k_{1 \rightarrow 0} \equiv r(f) = k_0 e^{-\beta(B - G(1) + X_1 f)} = e^{w - y f}, \quad (3.2)$$

with $\beta = 1/k_B T$ and where x, y, z, w are force independent parameters. The free energy difference at a given force is:

$$\Delta G(f) = -k_B T \log \left(\frac{q(f)}{r(f)} \right) = \Delta G_0 - X_m f. \quad (3.3)$$

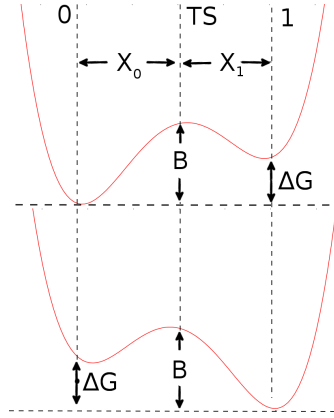


Fig. 4. Schematic picture of the two-state model. The free energy landscape of the hairpin molecule along the reaction coordinate axis at a given force has two minima corresponding to the two states Folded and Unfolded. When mechanical force is applied to the ends of the molecule the free energy landscape is tilted along x , decreasing the free energy of the Unfolded (1) state and the transition state (TS) relative to the Folded (0).

with $\Delta G_0 = G(0) - G(1)$. If the quantities z, w, x, y , entering in (3.1) and (3.2) are measured, the free energy landscape parameters can then be extracted as:

$$X_0 = \beta^{-1}x, \quad X_1 = \beta^{-1}y \quad (3.4)$$

$$\Delta G_0 = \beta^{-1}(w - z). \quad (3.5)$$

The standard method to extract such parameters is by measuring the kinetic rates, $q(f), r(f)$ at a given force as the inverse mean lifetimes of the corresponding states and then fitting a linear relation between the logarithm of the rates and the applied tension [7]. According to (3.1), (3.2) the slope and intercept of such fit should give estimates for z, w, x, y . We shall refer to this method as “the standard method” or “the histogram method”.

4. Analysis of a Single Hopping Trace

In order to perform a Bayesian analysis leading to the measurement of the kinetic rates of the hairpin, we map the force trace obtained from a PM hopping experiment to a dicotomic noise.

$$\sigma_i = \Theta(f(i\Delta t) - \bar{f}) \quad (4.1)$$

where $f(i\Delta t)$ is the force signal from a single hopping trace, sampled with finite acquisition time Δt ,

$$\begin{cases} \Theta(x) = 0, & \text{if } x \geq 0 \\ \Theta(x) = 1, & \text{if } x < 0. \end{cases} \quad (4.2)$$

and \bar{f} is the median between the force in the folded state and the force in the unfolded state (see Fig. 2). Since the force fluctuates in a PM hopping experiments we will consider the rates as functions of \bar{f} , which does not depend on the folding state but only on the distance between the trap and the pipette. The dicotomic trace is obtained in 4.1 takes only two

values, 0, 1 with 0 corresponding to the folded state and 1 corresponding to the unfolded state. The noise σ_i is then interpreted as a two-state Markov Chain, with a discrete time-step Δt corresponding to the data acquisition time (1 ms). The probability law associated with such a two-state Markov chain is determined by the transition probability from state j to state k , $T(j, k)$, which gives the probability for the molecule to be in state k at step $i + 1$ when it was in state j at the previous time i :

$$P(\sigma_{i+1} = k | \sigma_i = j) = T(j, k). \quad (4.3)$$

When the kinetic rates are much smaller than the data acquisition rate, the transition probability is well approximated by:

$$T(0, 0) = (1 - q\Delta t), \quad T(0, 1) = q\Delta t, \quad T(1, 1) = (1 - r\Delta t), \quad T(1, 0) = r\Delta t, \quad (4.4)$$

where $q \equiv q(\bar{f})$ is the unfolding kinetic rate and $r \equiv r(\bar{f})$ is the refolding kinetic rate. The transition rates also determine the stationary probability μ :

$$\mu(\sigma) = \delta_{\sigma,0} \frac{r}{q+r} + \delta_{\sigma,1} \frac{q}{q+r}. \quad (4.5)$$

The extraction of the kinetic rates of a DNA hairpin from a PM hopping experiment is then equivalent to the determination of the transition probabilities of the two-state Markov chain. Here we shall deal with this problem from a Bayesian viewpoint along the same lines as in [17] but deepening the analytical study of the posterior probability distribution and applying the techniques to real experimental data. In particular we will obtain explicit formulas for the maximum likelihood value of the kinetic parameters. Bayesian reasoning in data analysis [18] aims to the reconstruction of a posterior probability distribution for the parameters to be estimated $\{\theta\}$ (r and q in our case) from the experimental outcome $\{\sigma\}$. Using the definition of conditional probability we have:

$$P(\{\sigma\}|\{\theta\})P_0(\{\theta\}) = P(\{\theta\}|\{\sigma\})Q_0(\{\sigma\}), \quad (4.6)$$

or

$$P(\{\theta\}|\{\sigma\}) = \frac{P(\{\sigma\}|\{\theta\})P_0(\{\theta\})}{Q_0(\{\sigma\})}. \quad (4.7)$$

Here $P_0(\{\theta\})$ is the prior distribution on the values of the parameters, $P(\{\sigma\}|\{\theta\})$ is the likelihood function and $Q_0(\{\sigma\})$ may be seen as a normalization constant [18]. The evaluation of the likelihood function in the case of two-states Markov chains is straightforward, the normalized probability of a sequence $\{\sigma\}$, $i \in \{0, N\}$ is given by the probability of the initial condition, times the probability of each transition:

$$P(\{\sigma\}|\{\theta\}) = \mu_{\{\theta\}}(\sigma_0) \prod_{i=0}^{N-1} T_{\{\theta\}}(\sigma_i, \sigma_{i+1}). \quad (4.8)$$

The likelihood is completely determined from the initial condition and the total number of “bonds” n_{kj} from state k to state j along the trace.

$$P(\{\sigma\}|\{\theta\}) = \mu_{\{\theta\}}(\sigma_0) \prod_{k=0,1} \prod_{j=0,1} (T_{\{\theta\}}(k, j))^{n_{kj}}. \quad (4.9)$$

We will write the likelihood in an exponential form:

$$P(\{\sigma\}|\{\theta\}) = e^{\mathcal{N}S(\phi_{kj}) + \log(\mu_{\{\theta\}}(\sigma_0))} \simeq e^{\mathcal{N}S(\phi_{kj})}. \quad (4.10)$$

where $S(\phi_{kj}) = \sum_{k=0,1} \sum_{j=0,1} \phi_{kj} \log(T_{\{\theta\}}(k, j))$, $\mathcal{N} = \sum n_{kj}$, $\phi_{kj} = \frac{n_{kj}}{\mathcal{N}}$ and the error introduced in the last step is of the order of $1/\mathcal{N}$.

From the likelihood function (4.10) the posterior distribution can be obtained choosing a prior. The prior will be used to enforce the reasonable constraints that the rates must be positive and that they should be smaller than the data acquisition frequency Δt^{-1} . The maximally uninformative prior enforcing the two constraints is the uniform probability distribution on the set $\mathcal{P} : \{(r, q) \in [0, \Delta t^{-1}] \otimes [0, \Delta t^{-1}]\}$. With such prior the posterior probability is proportional to the likelihood on \mathcal{P} and zero outside of it:

$$P(\{\sigma\}|\{\theta\}) \propto P(\{\theta\}|\{\sigma\}), \quad \forall (q, r) \in \mathcal{P} \quad (4.11)$$

and Eq. (4.10) can be also read as a probability for the jump rates given the experimental outcome.

Using the explicit form for the transition probabilities of a two-state Markov chain we get:

$$S(q, r) = \phi_{00} \log(1 - q\Delta t) + \phi_{01} \log(q\Delta t) + \phi_{11} \log(1 - r\Delta t) + \phi_{10} \log(r\Delta t), \quad (4.12)$$

where S is now a function of (q, p) which has ϕ_{ik} as parameters. The most probable values for the parameters r and q , according to the posterior distribution, are found solving the following equations:

$$\partial_q S(q, r) = -\frac{\phi_{00}}{1 - q\Delta t} + \frac{\phi_{01}}{q\Delta t} = 0, \quad (4.13)$$

$$\partial_r S(q, r) = -\frac{\phi_{11}}{1 - r\Delta t} + \frac{\phi_{10}}{r\Delta t} = 0, \quad (4.14)$$

which yields:

$$\bar{q} = \frac{1}{\Delta t} \frac{\phi_{01}}{\phi_{00} + \phi_{01}} \quad \bar{r} = \frac{1}{\Delta t} \frac{\phi_{10}}{\phi_{11} + \phi_{10}}. \quad (4.15)$$

This is exactly the same result as in the histogram method, as the ratio of the number of transitions divided by the total number of points in one state is exactly the mean residence time in that state. Expanding S around its maximum, given by (\bar{q}, \bar{r}) we get:

$$S(q, r) \simeq S(\bar{q}, \bar{r}) - \frac{1}{2} \sigma_q^2 (q - \bar{q})^2 - \frac{1}{2} \sigma_r^2 (r - \bar{r})^2, \quad (4.16)$$

with

$$\sigma_q^2 = (\Delta t)^2 (\phi_{00} + \phi_{01})^2 \left(\frac{1}{\phi_{00}} + \frac{1}{\phi_{01}} \right) \simeq (\Delta t)^2 (\phi_{00} + \phi_{01})^2 \frac{1}{\phi_{01}} \quad (4.17)$$

$$\sigma_r^2 = (\Delta t)^2 (\phi_{11} + \phi_{10})^2 \left(\frac{1}{\phi_{11}} + \frac{1}{\phi_{10}} \right) \simeq (\Delta t)^2 (\phi_{11} + \phi_{10})^2 \frac{1}{\phi_{10}} \quad (4.18)$$

where the last step follows from the fact that $q\Delta t, r\Delta t \ll 1$ so that $\phi_{00} \gg \phi_{01}$ and $\phi_{11} \gg \phi_{10}$. Equations (4.16) and (4.17) show that the probability of a value of q such that $q - \bar{q} = \alpha \bar{q}$

is exponentially suppressed with a factor:

$$\exp(-\alpha^2 \mathcal{N} \Delta t \bar{q}). \quad (4.19)$$

The quantity $\mathcal{N} \Delta t \bar{q}$ is the ratio of the length of the trace (in seconds) to the mean lifetime of the folded state ($1/\bar{q}$) and can be seen as the number of expected transitions along the trace. So that what it really matters in the steepness of the probability distribution is the number of observed transitions. The posterior probability distribution is proportional to the likelihood obtained from Eqs. (4.10) and (4.12):

$$P(\{\theta\}|\{\sigma\}) \propto (1 - q\Delta t)^{n_{00}}(q\Delta t)^{n_{01}}(1 - r\Delta t)^{n_{10}}(r\Delta t)^{n_{11}}, \quad (4.20)$$

and must be properly normalized. To obtain the normalization constant we integrate (4.20) on \mathcal{P} :

$$\begin{aligned} Z &= \int_{\mathcal{P}} dq dr (1 - q\Delta t)^{n_{00}}(q\Delta t)^{n_{01}}(1 - r\Delta t)^{n_{10}}(r\Delta t)^{n_{11}} \\ &= \left(\frac{1}{\Delta t}\right)^2 \beta(n_{00} + 1, n_{01} + 1) \beta(n_{11} + 1, n_{10} + 1). \end{aligned} \quad (4.21)$$

Where

$$\beta(x, y) = \int_0^1 (1 - u)^{x-1} u^{y-1} du.$$

The posterior probability distribution is thus the product of two beta distributions [18]:

$$P(\{\theta\}|\{\sigma\}) dp dq = \frac{(1 - q\Delta t)^{n_{00}}(q\Delta t)^{n_{01}}(1 - r\Delta t)^{n_{10}}(r\Delta t)^{n_{11}}}{Z} dp dq. \quad (4.22)$$

In a Bayesian setting the best estimate for the model parameters are their averages with respect to the posterior distribution. Let $\mathbb{E}(\cdot)$ denote the such average. Using the special properties of the beta distribution [18], the expected kinetic rates are:

$$\mathbb{E}(q) = \frac{1}{\Delta t} \frac{\phi_{01} + (1/\mathcal{N})}{\phi_{00} + \phi_{01} + (2/\mathcal{N})} \quad (4.23)$$

$$\mathbb{E}(r) = \frac{1}{\Delta t} \frac{\phi_{10} + (1/\mathcal{N})}{\phi_{11} + \phi_{10} + (2/\mathcal{N})}. \quad (4.24)$$

In any practical situation $\mathcal{N}\phi_{01}, \mathcal{N}\phi_{00} \gg 1$, so that in the case of a single trajectory the most probable value of the transition rates is never very far from the mean value. Moreover also the variance goes to 0 as the length of the trace grows, meaning that the posterior probability distributions concentrates (see Fig. 5) around a point which is determined by the inverse mean lifetimes of the two different states. In this situation the information conveyed by the posterior distribution is the same as that obtained by the standard methods outlined at the end of Sec. 3.

5. Putting Different Traces Together

The theory introduced in the case of one single trace can be easily generalized to include different traces, measured at different forces, f_j , and from different molecules. All the force

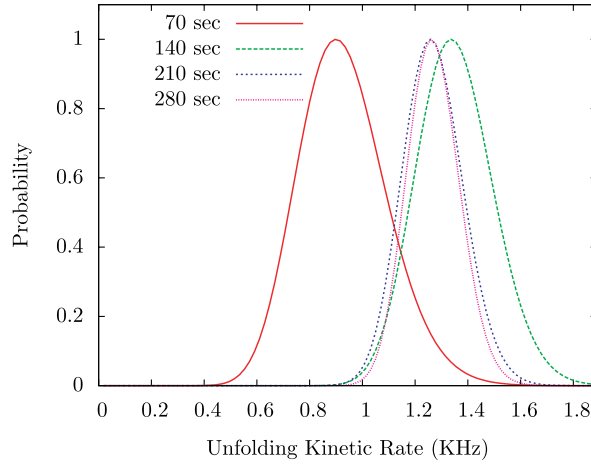


Fig. 5. The plot shows the evolution of the posterior probability distribution as the length of the force trace is increased. The posterior distribution has been plotted using one fourth, half, three fourths or the whole of a 280 s PM hopping trace. The different posterior distributions are normalized setting their maximum value to one. As the length of the trace increases the expected value of the transition rate fluctuates and then stabilizes. It is also possible to see how the measure assigned by the probability distribution gets more and more concentrated around the maximum. A third effect is the progressive disappearance of the initial asymmetry in the posterior distribution as the length of the trace is increased.

traces will be mapped to dicotomic noises as shown at the beginning of the last section. Since the different observations are to be considered independent, the probability of the whole collection of traces will be simply given by the product of the probability of each trace:

$$P(\sigma_i^1, \dots, \sigma_i^n | (q_1, r_1), \dots, (q_n, r_n)) = P(\sigma_i^1 | q_1, r_1) \dots P(\sigma_i^n | q_n, r_n), \quad (5.1)$$

where $q_j \equiv q(f_j)$ and $r_j \equiv r(f_j)$ so that, using (4.10) we obtain:

$$P(\sigma_i^1, \dots, \sigma_i^n | (q_1, r_1), \dots, (q_n, r_n)) = e^{\mathcal{N} \sum_j S^j(q_j, r_j)}, \quad (5.2)$$

where \mathcal{N} is now the total number of points in the collection of traces and

$$S^j(q_j, r_j) = \sum_{k,l} \psi_{kl}^j T^j(k, l), \quad \psi_{kl}^j = \frac{n_{kl}^j}{\mathcal{N}}. \quad (5.3)$$

According to the discussion in Sec. 3 the transition rates at different forces should be of the form:

$$q_j = e^{z+xf_j}, \quad r_j = e^{w-yf_j} \quad (5.4)$$

so that by Bayes theorem we can transform the probability distribution for the traces (5.2) into a probability distribution for the parameters z, x, w, y . As already noted in the previous section the folding and unfolding rate are independent random variables, we shall only discuss the case of the unfolding rate. Identical results do hold in the other case. The independence of the rates is due to the fact that the rate function $\sum_j S^j(q_j, r_j)$ can be

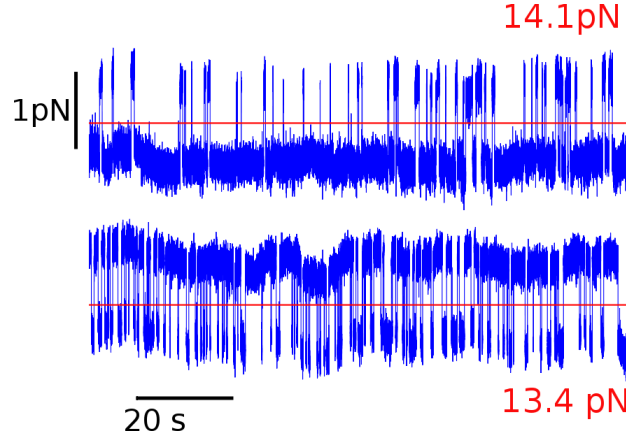


Fig. 6. Force traces obtained from hopping experiments on a single molecule at different forces (13.4 pN for the lower trace and 14.1 pN for the upper trace). The probability of the different states is different for different median force: in the lower trace, at a lower median force the folded state (higher force) is more probable than the unfolded state. At higher tension, upper curve, the unfolded state (lower force) is favoured. Force and time scales as shown in the plot. The low frequency oscillations of the force signal are due to instrumental drift effects.

written as a sum of two contributions, one depending only on the q_j s and the other only on the r_j s:

$$\sum_j S^j(q_j, r_j) = \sum_j \sum_{k=0,1} \psi_{0,k}^j T^j(0, k) + \sum_j \sum_{k=0,1} \psi_{1,k}^j T^j(1, k). \quad (5.5)$$

Since $T^j(0, k)$ does only depend on q_j (see (4.12)) and thus on z, x (5.4) and $T^j(1, k)$ does only depend on r_j and thus on w, y we can write:

$$\sum_j S^j(q_j, r_j) = S_0(z, x) + S_1(w, y), \quad (5.6)$$

with

$$S_0(z, x) = \sum_j n_{00}^j \log(1 - e^{z+xf_j}) + n_{01}^j (z + xf_j) \simeq \sum_j -n_{00}^j e^{z+xf_j} + n_{01}^j (z + xf_j)$$

$$S_1(w, y) = \sum_j n_{11}^j \log(1 - e^{w+yf_j}) + n_{10}^j (w + yf_j) \simeq \sum_j -n_{11}^j e^{w+yf_j} + n_{10}^j (w + yf_j).$$

In the following we shall restrict our analysis to the rate function $S_0(z, x)$. The equations for the maximum of the rate function (5.3) are:

$$\partial_z S_0 = \sum_j n_{01}^j - n_{00}^j e^{z+xf_j} = 0 \quad (5.7)$$

$$\partial_x S_0 = \sum_j f_j n_{01}^j - f_j n_{00}^j e^{z+xf_j} = 0, \quad (5.8)$$

whose solution is implicitly given by:

$$\frac{\sum_j f_j n_{00}^j e^{x f_j}}{\sum_j n_{00}^j e^{x f_j}} = \frac{\sum_j f_j n_{01}^j}{\sum_j n_{01}^j} \quad (5.9)$$

$$z = \log \left(\frac{\sum_j n_{01}^j}{\sum_j n_{00}^j e^{x f_j}} \right). \quad (5.10)$$

Note that rate function is convex, as it is a sum of convex functions, so that any maximum is a global maximum. Moreover it is easy to show that (5.9) does always have a solution for $x \in [-\infty, \infty]$: the two sides of the equation have the formal aspect of the “average” of f_j for different j with respect to the probability distributions $\mathcal{P}_l(j), \mathcal{P}_r(j)$:

$$\mathcal{P}_l(j) = \frac{n_{00}^j e^{x f_j}}{\sum_j n_{00}^j e^{x f_j}}, \quad \mathcal{P}_r(j) = \frac{n_{01}^j}{\sum_j n_{01}^j}.$$

Table 1. Comparison of the values for the Free Energy Landscape parameters as extracted by the two methods.

	Bayesian method	Standard method
X_0 (nm)	8.6 ± 0.6	9.0 ± 0.8
X_1 (nm)	6.0 ± 0.6	6.2 ± 0.6
X_m (nm)	14.6 ± 0.8	15.2 ± 1
ΔG_0 ($k_B T$)	51 ± 4	52 ± 3

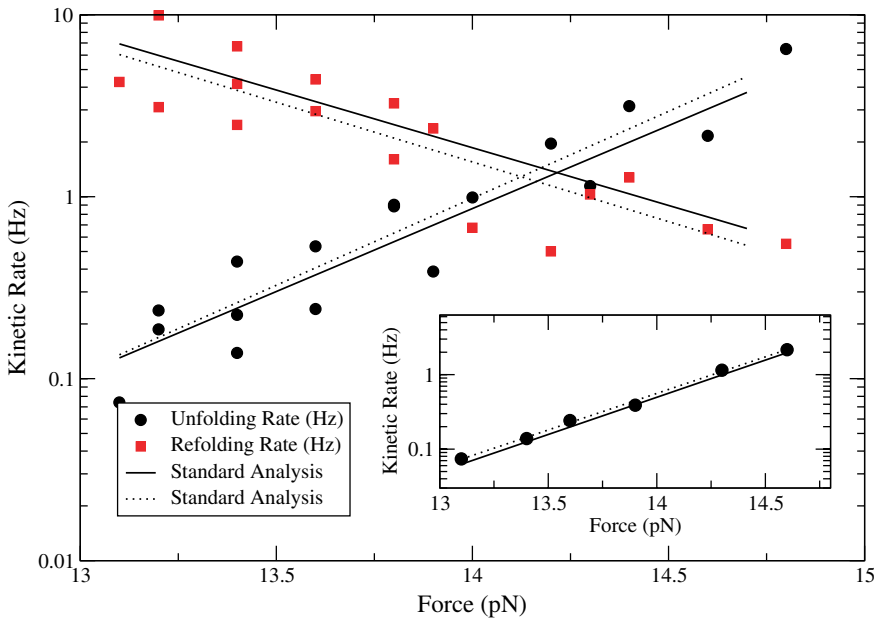


Fig. 7. Comparison of the Free Energy Landscape parameters as obtained from the Bayesian method (continuous lines) and the standard methods (dotted line). Inset: the two methods applied to data coming from a single molecule, perfect agreement between the two methods is shown. Main figure: The methods are applied to a collection of traces coming from different molecules. Points do represent the unfolding rates measured via the standard method and squares represent the refolding rates measured in the same way. The Bayesian method and the standard method show a low 5% discrepancy.

The right-hand side of (5.9) does always take value in the interval $[f_{\min}, f_{\max}]$, while, due to the term e^{xf_j} , the left hand side spans the same range when x varies from $-\infty$ to ∞ . The maximum likelihood analysis can be easily performed on data from PM hopping experiments and the results obtained compared with those obtained by standard methods. When analyzing the trace for a single molecule (Fig. 7 inset) the results from the two methods can coincide pretty well and even when traces from different molecules are analyzed together (Fig. 7 main figure) the results show only a moderate discrepancy.

6. Conclusions

Two of the salient experimental issues in single molecule biophysics are the inherently stochastic nature of the observed processes, which are usually strongly affected by thermal fluctuations, and the large heterogeneity between molecules of the same species. As a consequence of these two facts Bayesian reasoning, which turns probability distribution on the experimental outcomes into probability distributions for the system's parameters, might prove very fruitful when applied to single molecule experiments as it was done for example in [17, 19, 20]. We have derived a posterior distribution for the parameters describing the folding kinetics of a DNA hairpin under tension. This posterior distribution is based on the assumption that the unfolding/refolding dynamics can be described as a two-state Markov chain and that the kinetic rates obey a simplified Transition State model. Under these assumptions the posterior distribution for the rates obtained from a single force trace have the form of a beta distribution, whose variance decreases with the number of observed transitions. At the level of the single force trace the Bayesian approach and the standard method do agree pretty well. The Bayesian approach does also allow the extraction of a probability distribution for the parameters describing the free energy landscape of the hairpin from a set of force traces obtained at different forces and from different molecules. The equations for the maximum of such probability distribution were derived and solved. In the case in which data from many different molecules are analyzed together, the two methods of analysis can show moderate discrepancies, which are due to the fact that the Bayesian method does not take into account all the traces on an equal basis, but weights them according to their length. In the case of the data discussed in this paper the discrepancy proved to be below the experimental error. The introduction of Bayesian methods in single molecule biophysics seems thus useful to validate widely used data analysis techniques from a statistical point of view and maybe to discover possible improvements to the actual methods.

References

- [1] J. M. Huguet, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante and F. Ritort, Single-molecule derivation of salt dependent base-pair free energies in DNA, *Proc. Nat. Acad. Sci.* **107**(35) (2010) 15431.
- [2] G. Varani, Exceptionally stable nucleic acid hairpins, *Annu. Rev. Biophys. Biomolecular Structure* **24**(1) (1995) 379–404.
- [3] J. SantaLucia Jr and D. Hicks, The thermodynamics of DNA structural motifs, *Annu. Rev. Biophys. Biomolecular Structure* **33** (2004) 415.
- [4] A. Mossa, M. Manosas, N. Forns, J. M. Huguet and F. Ritort, Dynamic force spectroscopy of DNA hairpins: I. Force kinetics and free energy landscapes, *J. Stat. Mech. Theory Exp.* **2009** (2009) P02060+.

- [5] M. Manosas, A. Mossa, N. Forns, J. M. Huguet and F. Ritort, Dynamic force spectroscopy of DNA hairpins: II. Irreversibility and dissipation, *J. Stat. Mech. Theory Exp.* **2009** (2009) P02061+.
- [6] J. D. Wen, M. Manosas, P. T. X. Li, S. B. Smith, C. Bustamante, F. Ritort and I. Tinoco Jr, Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results, *Biophysical Journal* **92**(9) (2007) 2996–3009.
- [7] M. Manosas, J. D. Wen, P. T. X. Li, S. B. Smith, C. Bustamante, I. Tinoco Jr and F. Ritort, Force unfolding kinetics of RNA using optical tweezers. II. Modeling experiments, *Biophysical Journal* **92**(9) (2007) 3010–3021.
- [8] M. Manosas and F. Ritort, Thermodynamic and kinetic aspects of RNA pulling experiments, *Biophysical Journal* **88**(5) (2005) 3224–3242.
- [9] M. T. Woodside, P. C. Anthony, W. M. Behnke-Parks, L. Kevan, H. Daniel and S. M. Block, Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid, *Science* **314** (2006) 1001–1004.
- [10] M. T. Woodside, W. M. Behnke-Parks, K. Larizadeh, K. Travers, D. Herschlag and S. M. Block, Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins, in *Proc. Nat. Acad. Sci.* **103** (2006) 6190–6195.
- [11] A. Ashkin, Acceleration and trapping of particles by radiation pressure, *Phys. Rev. Lett.* **24**(4) (1970) 156–159.
- [12] S. B. Smith, Y. Cui and C. Bustamante, Optical-trap force transducer that operates by direct measurement of light momentum, in *Biophotonics, Part B*, eds. G. Marriotti and I. Parker, *Methods in Enzymology*, Vol. 361 (Academic Press, 2003) pp. 134–162.
- [13] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco Jr and C. Bustamante, Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski’s equality, *Science* **296** (2002) 1832–1835.
- [14] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco Jr and C. Bustamante, Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies, *Nature* **437**(7056) (2005) 231–234.
- [15] J. Gebhardt, T. Bornschl gl and M. Rief, Full distance-resolved folding energy landscape of one single protein molecule, *Proc. Nat. Acad. Sci.* **107**(5) (2010) 2013.
- [16] N. Forns, M. Manosas, K. Hayashi, S. de Lorenzo, J. M. Huguet and F. Ritort, Improving signal-to-noise resolution in single molecule experiments using molecular constructs with short handles, *Biophysical Journal* **100**(7) (2011) 1765–1774.
- [17] X. Xue, H. Tong, F. Liu and Z. Ou-Yang, Bayesian analysis of folding and unfolding time series of single-forced RNAs, *J. Phys. Chem. B* **112**(44) (2008) 13680–13683.
- [18] G. D’Agostini, *Bayesian Reasoning in Data Analysis: A Critical Introduction* (World Scientific Publishing Co., 2003).
- [19] P. Maragakis, F. Ritort, C. Bustamante, M. Karplus and G. E. Crooks, Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise, *J. Chem. Phys.* **129** (2008) 024102.
- [20] S. C. Kou, X. S. Xie and J. S. Liu, Bayesian analysis of single-molecule experimental data, *J. Roy. Statist. Soc. C* **54**(3) (2005) 469–506.