Reconstruction and Identification of DNA Sequence Landscapes from Unzipping Experiments at Equilibrium

Carlo Barbieri,[†] Simona Cocco,[†*] Thomas Jorg,[†‡] and Rémi Monasson[‡]
[†]Laboratoire de Physique Statistique and [‡]Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique and Université Pierre et Marie Curie, Paris, France

ABSTRACT Two methods for reconstructing the free-energy landscape of a DNA molecule from the knowledge of the equilibrium unzipping force versus extension signal are introduced: a simple and fast procedure, based on a parametric representation of the experimental force signal, and a maximum-likelihood inference of coarse-grained free-energy parameters. In addition, we propose a force alignment procedure to correct for the drift in the experimental measure of the opening position, a major source of error. For unzipping data obtained by Huguet et al., the reconstructed basepair (bp) free energies agree with the running average of the true free energies on a 20–50 bp scale, depending on the region in the sequence. Features of the landscape at a smaller scale (5–10 bp) could be recovered in favorable regions at the beginning of the molecule. Based on the analysis of synthetic data corresponding to the 16S rDNA gene of bacteria, we show that our approach could be used to identify specific DNA sequences among thousands of homologous sequences in a database.

INTRODUCTION

Single-molecule techniques make possible the unzipping of a single DNA or RNA molecule, that is, the separation of the two nucleotidic strands under a mechanical action, e.g., at fixed force (1–3), or at constant pulling rate (4–6). The output signal, e.g., the distance between the two ends of the open strands in a constant force experiment (1–3), or the force versus trap displacement in a constant pulling rate experiment (4–6), is known to reflect the basepairing free energies, which depend on the sequence of the biomolecule. A natural question is whether this signal can, in practice, be used to reconstruct the DNA or RNA sequence.

The development of second-generation, high-throughput DNA sequencing methods (7–11) has revolutionized molecular biology and medicine over the past decades. These methods, e.g., sequencing by synthesis, commercialized by Illumina (8,9) (San Diego, CA), sequencing by ligation, called SOLID, commercialized by Life Technologies (12) (Carlsbad, CA), or sequencing by hybridization of complementary DNA probes (13), achieve parallel sequencing of many short DNA fragments, which are then reassembled to obtain the whole genome. There is, however, still a need for improvement to achieve massive, cheap, accurate, fast and individual sequencing. In the third generation of sequencing techniques single-molecule experiments, which in principle avoid the amplification stage and the segmentation of the sequence in shorter subsequences (reads), hold promise for limiting sequencing errors. Two promising methods are sequencing in zero-mode waveguide developed by Pacific Bioscience (Menlo Park, CA), in which the incorporation of nucleotides during the synthesis of a new DNA is observed continuously in real time (14) and nanopore sequencing, based on the readout of the sequence-dependent electrical signal resulting from the passage of a DNA molecule through a nanopore (15). A recently developed method based on a combination of constant-force unzipping and hybridization of complementary probes allows for the accurate readout of the positions of the probe sequences on a single DNA molecule (16). Finally mechanical unzipping of a single-molecule has been shown to be effective to reconstruct small DNA sequences in constant-force experiments (3). Even if unzipping experiments will not be, in the immediate future, competitive with commercial sequencing technologies they can provide complementary techniques, which can be advantageous, as well as simpler and cheaper, for particular applications. Among these applications are the rapid classification of an individual sequence, e.g., to characterize which bacterium has infected a patient, and the detection of genetic variations responsible for diseases, such as variations in the copy number of repeated sequences, which are particularly difficult to quantify with current sequencing methods.

Apart from direct application to the development of sequencing technologies, unzipping experiments have become a good experimental system to test equilibrium and out-of-equilibrium theories in statistical mechanics. This is due both to the very high control of the experimental system and to the fact that unzipping is very well modeled by a one-dimensional random walk of the opening fork (the boundary between the open and closed portion of the DNA double helix) in a disordered potential caused by the DNA sequence (5,6,17–19). Theoretical works have, in particular, focused on the possibilities of reconstructing the features of the sequence-specific free-energy landscape through equilibrium (2,3,20) and out-of-equilibrium measurements (21–26).

Submitted June 20, 2013, and accepted for publication November 27, 2013.

*Correspondence: cocco@lps.ens.fr

Editor: David Rueda.

0006-3495/14/01/0430/10 $2.00

http://dx.doi.org/10.1016/j.bpj.2013.11.4496

An important issue in extracting information on the sequence from unzipping experiments is the limitations due to the experimental setup (6,20,27). Thompson and Siggia have stressed the difficulty of inferring the sequence because the position of the opening fork cannot be read out directly from the displacement of the optical trap (see Fig. 1), as the thermal fluctuations in the single-stranded (ss) DNA may exceed the average gain of ~1 nm consecutive to the opening of one basepair (27). Other limitations of the experimental system are the thermal drift of the optical trap, the precision over the measured force, and the limited spatial resolution. Subnanometric spatial resolution, and a precision of measured forces on the order of a fraction of a piconewton, can nowadays be routinely achieved. However, the thermal drift of the optical trap remains a major problem in unzipping data, limiting the ability to associate local features of the force signal with an absolute position in the sequence.

In this work, we explicitly take into account in the inference model thermal fluctuations coming from the single strands of DNA, the linkers, and the bead in the optical trap. We express the average force at a given position as a convolution of the force signal over the possible positions of the opening fork with these distance fluctuations (4,6). Two techniques for inferring the sequence-specific free-energy landscape from the force signal at equilibrium are introduced. The first method, called saddle-point (SP) approximation, is fast and simple, and requires very little computational effort. The second method, called Box approximation, relies on the maximum likelihood inference of the free-energy parameters, coarse-grained over an appropriate number of basepairs, and is more demanding from a computational point of view. In addition, we show how multiple force signals corresponding to the same molecule can be aligned using an extension of alignment algorithms developed in bioinformatics. This alignment procedure can be used to reduce the drift effect in the data.

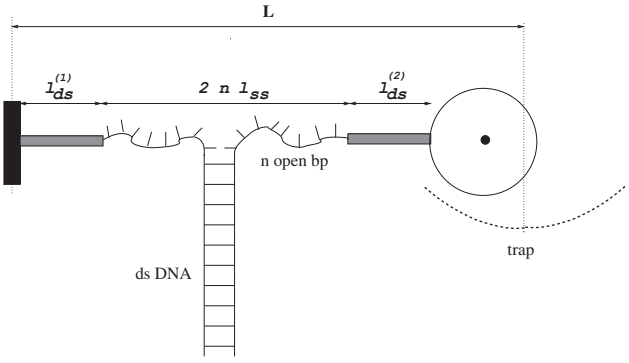

FIGURE 1 Sketch of the setup for the single DNA molecule unzipping experiment. The DNA molecule is attached to the surface (*left*) and a bead (*right*) held in an optical trap at distance L from the surface through rigid linkers of total length ℓ_{ds}. When the unzipping force is equal to its average value f_{av}, the single strands corresponding to the n unzipped basepairs of the molecule have extension $2n\,\ell_{ss}$, and the bead is displaced by ℓ_{av} from the center of the trap. When the force deviates from f_{av} the extension of the ssDNA strands and the displacement vary according to their stiffnesses, $K_{ss}(n)$ and K_{trap}.

First, we use our alignment and inference methods to reanalyze experimental data from a study of unzipping at constant and low velocity by Huguet et al. (4). We show how the sequence free-energy landscape can be reconstructed on the scale of several basepairs, and we discuss how this characteristic scale depends on the setup features. Based on those findings, we then show that our approach can be used to identify specific DNA sequences among a large database of homologous sequences. A proof of principle is given from synthetic force data corresponding to the 16S rDNA genes of 2076 bacteria. We show that our procedure is capable of matching a 16S gene with the same gene in the database, and to distinguish it from homologous genes with a few mismatches.

The article is organized as follows. In the Materials and Methods section, the model for DNA unzipping (6) is exposed and a local harmonic approximation of ssDNA elastic properties around the unzipping force is introduced. We then describe the two inference methods, called SP approximation and Box approximation, which allow us to reconstruct the free-energy landscape from the measured forces. We also explain how force data obtained from the same DNA molecule can be aligned to remove the drift. In the Results and Discussion section, we investigate the inference performances of the SP and Box approximations along the sequence for two repetitions of the unzipping experiment on the same sequence (4). The inference procedures are applied to synthetic force data generated from the 16S rDNA genes of a bacterial database, and we discuss their ability to identify one gene across a family of thousands of homologous sequences. Perspectives and open problems are discussed in the Conclusion.

MATERIAL AND METHODS

Model for DNA unzipping

Let $s_i = A, C, G, T$ be the bases in the DNA sequence along the 5′-to-3′ strand, with $i = 1, \ldots, N$ being the base index (Fig. 1). The free-energy cost for unzipping the first n basepairs of the double-stranded (ds) DNA molecule is given by

$$G_{ds}(n) = \sum_{i<n} g_0(s_i, s_{i+1}), \qquad (1)$$

where $g_0(s_i, s_{i+1})$ takes into account both pairing and stacking contributions between neighbor bases on the strand. The 10 independent values of $g_0(s_i, s_{i+1})$ are given as functions of the temperature and ionic condition (28,4), and are reported for the data we have analyzed in the Section I in the Supporting Material.

Each one of the two unzipped strands of the molecule are modeled as harmonic springs, with stiffness constant $K_{ss}(n)$, rest length $n\,\ell_{ss}$ and rest free energy $n\,g_{ss}$; $K_{ss}(n)$, ℓ_{ss} and g_{ss} are effective parameters obtained from a local harmonic approximation of the freely-jointed chain model,

known to be accurate for ssDNA elastic properties (29,30), around the average unzipping force, $f_{\rm av}$. In addition, the setup includes two very short dsDNA linkers with total length $\ell_{\rm ds}$, which we consider to be rigid, and the optical trap, with stiffness constant $K_{\rm trap}$. We denote by L the position of the trap (see Fig. 1).

After integrating out the degrees of freedom related to the unzipped strand extensions and the displacement of the bead in the trap, we obtained an effective free energy for the number of unzipped basepairs, n, as a function of the trap position, L, given by

$$G(n|L) = G_{\rm ds}(n) - 2\,n\,g_{\rm ss} + \frac{1}{2}K(n)(L - \ell_{\rm av} - \ell_{\rm ds} - 2\,n\,\ell_{\rm ss})^2, \tag{2}$$

where the effective spring constant is given by $1/K(n) = (2/K_{\rm ss}(n)) + (1/K_{\rm trap})$ and $\ell_{\rm av} = f_{\rm av}/K_{\rm trap}$ is the displacement of the bead in the trap at the average unzipping force. It is important to stress that the effective stiffness, $K(n)$, is dominated at the beginning of the opening by the trap stiffness. When the number, n, of open bases is such that $K_{\rm ss}(n)$ becomes small with respect to $K_{\rm trap}$, the stiffness is, conversely, dominated by the single-strand thermal fluctuations, and decreases with n. For the experimental setup in Huguet et al. (4), the crossover takes place for a few hundreds of open basepairs (see Fig. S1 in the Supporting Material). The effective stiffness does not vary significantly when, for fixed L, the number of unzipped basepairs changes around its average number. Hereafter we will therefore consider that it is a function of L only, denoted by $K(L)$. Details about the harmonic approximation, the value of $K(L)$, and the derivation of Eq. 2 can be found in Section II of the Supporting Material.

The free energy of the system is a function of the trap position, L, given by (in units of $k_{\rm B}T$)

$$G(L) \;=\; -\log Z(L), \;\text{with}\; Z(L) \;=\; \sum_{n=0}^{N} e^{-G(n|L)}. \tag{3}$$

Knowing the free energy, we can easily compute the value of the force at equilibrium for fixed L:

$$\langle f\rangle(L) \;= f_{\rm av} - \frac{dG}{dL}(L). \tag{4}$$

Hereafter, we refer to $G_{\rm ds}(n)$ (1) as the cumulative basepair free energy, and to the set of basepair free energies, $g_0(s_i, s_{i+1})$, versus basepair index, i, as basepair free energies. We now present two procedures to infer the basepair free energies from the knowledge of the experimental unzipped force, $f_{\rm exp}(L)$.

Inference of the basepair free-energy landscape: SP approximation

Given the position, L, of the trap, the most likely value of the number of unzipped basepairs is $n^{\rm SP}(L)$, minimizing $G(n|L)$. The SP approximation consists of approximating the sum over the values of n in Eq. 3 for $Z(L)$ by its dominant contribution, coming from $n = n^{\rm SP}(L)$. Within the SP approximation, the free energy simply corresponds to $G(L) = G(n^{\rm SP}(L)|L)$, and the equilibrium force is given by

$$\begin{aligned}\langle f\rangle(L) - f_{\rm av} &\simeq -\frac{\partial G}{\partial L}\left(n^{\rm SP}(L), L\right)\\ &\simeq K(L)\left(L - \ell_{\rm av} - \ell_{\rm ds} - 2n^{\rm SP}(L)\ell_{\rm ss}\right).\end{aligned} \tag{5}$$

As the unzipping proceeds, dsDNA pairing and stacking free energy is converted into ssDNA elastic free energy, equal to

$$g(L) \equiv 2g_{\rm ss} + 2(\langle f\rangle(L) - f_{\rm av})\ell_{\rm ss} \tag{6}$$

per unzipped basepair within the harmonic model of ssDNA outlined above. $g(L)$ is, at equilibrium, equal to the mean value of the basepair free energy, $g_0(s_n, s_{n+1})$, averaged over the distribution of the number, n, of unzipped basepairs at fixed L.

Upon replacement of the equilibrium force, $\langle f\rangle(L)$, with the experimental measure, $f_{\rm exp}(L)$, in Eqs. 5 and 6, we obtain the number of unzipped basepairs,

$$n^{\rm SP}(L) \;= \frac{1}{2\ell_{\rm ss}}\left(L - \ell_{\rm av} - \ell_{\rm ds} - \frac{f_{\rm exp}(L) - f_{\rm av}}{K(L)}\right), \tag{7}$$

and the corresponding equilibrium basepair free energy,

$$g^{\rm SP}(L) \;= 2g_{\rm ss} + 2\big(f_{\rm exp}(L) - f_{\rm av}\big)\ell_{\rm ss}, \tag{8}$$

respectively, at trap position L. The basepair free-energy landscape of the DNA molecule can then be parametrically plotted by representing $(n^{\rm SP}(L), g^{\rm SP}(L))$ for various values of L (see Results).

Inference of the basepair free-energy landscape: Box approximation

The SP approximation is fast and easy to implement, but neglects all the fluctuations of the number of unzipped basepairs around its most likely value, $n^{\rm SP}$. To take into account those fluctuations, we resort to another approximation scheme, where the average value of the force is computed exactly through the sum over all possible values of n as in Eq. 4, but where the cumulative basepair free energy, $G_{\rm ds}$ in Eq. 1, depends on a limited number of parameters, which can be optimized to reproduce the experimental force signal. To do so, we write the cumulative free energy as a sum of box functions of width b,

$$G_{\rm ds}^{\rm Box}(n) \;= b \sum_{k=0}^{\text{integer part of } n/b} g_k. \tag{9}$$

Parameter g_k represents the box average of the free energies over the basepairs in the interval $i = kb + 1, \ldots, (k+1)b$. The value of b can be chosen at convenience; the order of magnitude coincides with the typical fluctuations over the position of the bead in the optical trap at fixed position L (in units of ℓ_{ss}),

$$b_{\rm B}(L) \;= \sqrt{\frac{k_{\rm B}T}{4K(L)\ell_{\rm ss}^2}}. \tag{10}$$

We have chosen $b = b_B(L)/2$. In Section IVB of the Supporting Material, we indeed show that it is optimal to adapt b to the characteristic fluctuations of the apparatus or, equivalently, to choose the precision of the inference in accordance with the stiffness of the setup. Note that b varies with L due to the dependence of the effective stiffness, K, on the trap position (see Fig. S1).

The aim of the inverse problem is to infer the parameters $g_0, g_1, \ldots, g_{N/b-1}$ from the experimental unzipping curve, $f_{\rm exp}(L)$. As a result of thermal fluctuations of the number of unzipped basepairs at fixed L, we expect the force measures, $f_{\rm exp}(L)$ and $f_{\rm exp}(L')$, to be correlated (influenced by the same part of the sequence) as long as $|L' - L| < \sqrt{k_{\rm B}T/K} \sim b\,\ell_{\rm ss}$. To reduce redundancy in the data, we consider the set of measured forces, $f_{\rm exp}(L_k)$, at discrete positions $L_k = L_0 + k \times 2b\,\ell_{\rm ss}$, with integer-valued k; the offset position L_0 encompasses the linker length, $\ell_{\rm ds}$, and the average bead displacement, $\ell_{\rm av}$. We

further assume that the experimental error in measuring the force is a normal variable with zero mean and variance ϵ^2, with $\epsilon = 0.1$ pN. The logarithm of the probability of the set of measured forces at positions L_k is

$$\log P\left(\{f_{\exp}(L)\}\middle|\{g_k\}\right) = -\frac{1}{2\epsilon^2}\sum_{k=0}^{N/b-1}\left(f_{\exp}(L_k) - \langle f\rangle^{\mathrm{Box}}(L_k)\right)^2 - \frac{1}{2\Delta^2}\sum_{k=0}^{N/b-1}\left(g_k - \overline{g}\right)^2 \tag{11}$$

up to an additive constant independent of the g_k parameters. $\langle f\rangle^{\mathrm{Box}}(L_k)$ is the equilibrium force at trap position L_k, given by Eq. 4, with the true cumulative basepair free energy, G_{ds}, replaced with $G_{\mathrm{ds}}^{\mathrm{Box}}$. The second term in Eq. 11 represents the a priori contribution to the log probability; it regularizes the inference problem by imposing that the inferred free-energy parameters, g_k, should be around the typical value, $\overline{g} = 2.5$. The penalty parameter, $\Delta \simeq 1$ (in units of k_BT), corresponds to the expected deviation of g_k around $\overline{g}$.

In the spirit of maximum-likelihood inference, we maximize $\log P$ over the g_k coefficients using a gradient ascent procedure. The maximum does not seem to depend on the initial values of g_k. In addition we find that the optimum over g_k depends weakly on the parameter $\gamma = (\epsilon/\Delta)^2$ in the expected range $10^{-4} - 10^{-2}$ (see Section IVA in the Supporting Material).

Alignment of experimental unzipping forces

We have reanalyzed the data of Huguet and collaborators (4), in which a part of a λ-DNA molecule of 6800 base pairs, of known sequence, is unzipped at low velocity (10 nm/s), at 1 M monovalent salt concentration. The two complementary strands of the DNA molecule are attached, through two 29-nucleotide-long dsDNA handles, to a bead and to a micropipette. The force on the bead is measured through the displacement in the optical trap (see Fig. 1) and acquired at 1 kHz frequency. We have filtered this signal at a frequency of 1 Hz to obtain the average force at each position. We have analyzed two unzipping curves (see Fig. 4) corresponding to two molecules with the same sequence, hereafter called Molecules 1 and 2. Notice that the unzipping curves do not start at the beginning of the sequence, as the first recorded forces correspond to ~700 open basepairs for Molecule 1 and 950 open basepairs for Molecule 2.

An important source of experimental error is a low-frequency drift of the instrument, resulting from dilatations or contractions after local changes in temperature. The drift adds extra noise in the measurement of position L of the optical trap. Experimental data were preprocessed by Huguet et al. to reduce experimental drift (see Huguet et al. (4) and their Supplementary Information). Even after this preprocessing, however, the two experimental force curves for Molecules 1 and 2 were not perfectly superimposed (see Fig. 4, *upper*), and the two corresponding sets of 10 free-energy parameters, g_0, calculated in Huguet et al. (4), referred to as best sets, differed by 10% (see Fig. S19).

To align two force curves, we propose the following procedure, based on the celebrated Needleman-Wunsch alignment algorithm of bioinformatics (31). First, we compute the average unzipping forces, f_1 and f_2, for the two molecules, and apply a global shift, $f_2 - f_1 = 0.5$ pN, on the force curve of Molecule 2. This correction compensates a global error (offset) on the absolute force measure (typically of the order of some fractions of a piconewton). Second, we align the two force curves using the Matlab routine nwalign, which implements the Needleman-Wunsch algorithm. This routine aligns sequences of symbols (generally, bases or amino acids) according to a matrix of scores, expressing the similarities between pairs of symbols. Here, symbols are force values, and the score is a measure of how close two values are. In practice, we discretize trap positions L with a step of $\Delta L = 1$ nm and the force values, $f(L)$, in $N_f = 22$ increments, $\Delta f \sim 0.2$ pN. This choice allows us to cover the 4–5 pN total range of variations of the unzipping force along the molecules. Each force curve is therefore turned into a discretized sequence, $i(L)$, with $i = 1$ for the minimal value of the force and $i = N_f$ for the maximal value. The score for aligning two force increments i and j at the same position is given by

$$S_{ij} \equiv S(i-j) = -\frac{(i-j)^2}{2\sigma^2}. \tag{12}$$

Parameter σ is related to the experimental resolution of the force (of the order of 0.1–0.5 pN) divided by the discretization interval, Δf; hereafter, we choose $\sigma^2 = 5$. The minimal score, $S(N_f)$, corresponding to the ~ 4 pN maximal difference between two unzipping forces, is $S(N_f) \simeq -35$. In addition, gaps can be inserted in the alignment, with a fixed score of $S_{\mathrm{gap}} = -20$, about halfway between the scores $S(0) = 0$ and $S(N_f)$. Gaps are necessary to compensate for the drift of the trap position in one force signal relative to the other one. We have verified that the force signal alignments are weakly affected by the choice of another set of parameters or of another number of discretization intervals, N_f (Fig. S24).

In the Results section, we will compare the basepair free energies inferred using the data of Molecule 1 to the values computed from its best set and the λ-DNA sequence; results for Molecule 2 are reported in the Section V in the Supporting Material. Moreover, we will realign the force curves with the procedure described above and compare the two inferred basepair free-energy landscapes with the one obtained from the free energies given by the Mfold server (see Section VI in the Supporting Material). Finally, we will also generate synthetic force data by computing the equilibrium unzipping force as a function of the displacement, given the sequence and the Mfold free energies. These synthetic data allow us to infer the beginning of the sequence, which was lost in the experimental data. In addition, and of more importance, synthetic data are useful to estimate the performance of the inference method in ideal conditions (no drift, strict equilibrium).

Synthetic data on 16S bacterial genomes

A potential application of unzipping experiments is to identify an unknown DNA sequence from a database of reference sequences. To illustrate how DNA screening can be implemented, we focus on the 16S ribosomal RNA gene, of about $N = 1540$ bp. The 16S rDNA gene is widely used for phylogenetic classification of bacteria, and is relatively long to have a good statistics in sequence comparison (32). To better understand the resolution that could be achieved with unzipping analysis, we have downloaded from the NCBI RefSeq database (33) ~2500 well cured 16S bacterial sequences. The 16S rDNA sequence of one bacterium, hereafter called the test sequence, is chosen in the database. We then compute the theoretical unzipping force curves for all 16S bacterial sequences, infer the corresponding basepair free-energy landscapes, and compare them with the test landscape. We estimate their discrepancies, and whether they are large enough compared to the experimental uncertainty computed from the detailed analysis of the experimental data above.

RESULTS AND DISCUSSION

SP inference: reconstructed basepair free-energy landscape

The inference of the basepair free energy is easily done using the SP approximation, as Eqs. 7 and 8 provide a parametric representation of g versus the number of unzipped

basepairs, n, as a function of the trap position, L. The outcome for the data of Molecule 1 is shown in Fig. 2 and compared with its counterpart obtained from the sequence and the best free-energy parameters (Table S1), and averaged on a sliding window of 30 bp. The results of the SP inference for synthetic data generated from the model (4) are also plotted in Fig. 2. The difference between the basepair free energy inferred from the experimental data and that inferred from the synthetic data is small with respect to their common discrepancy with the true basepair free energies. This good agreement entails that our unzipping model based on the local harmonic approximation is accurate, and that out-of-equilibrium effects are weak: the unzipping velocity is low enough for the system to be effectively at equilibrium for each trap position L.

The performance of the SP procedure strongly depends on the local features of the free-energy landscape. The unzipping force signal is characterized by the so-called stick-slip phenomenon (34). When strong basepairs are followed by weaker base pairs the cumulative free energy, $G(n|L)$, Eq. 2, may have two local minima in n_1 and n_2, with $n_1<n_2$. The stick phase corresponds to the first minimum ($n = n_1$): the single strands and the bead in the trap are pulled and stretched without breaking strong bases. In the slip phase (the second minimum, in $n = n_2$), not only the strong basepairs but also the contiguous weak basepairs have opened. The stick-slip phenomenon gives rise to the characteristic sawtooth behavior of the unzipping force. Conversely, in regions where weak basepairs are followed by strong basepairs, the cumulative free energy, $G(n|L)$, has generally a unique minimum.

Fig. 2 shows that the SP approximation, which replaces the sum of the contributions associated with different n in Eq. 3 with a unique contribution from n^{SP} is accurate in the non-stick-slip regions (e.g., region $800<n<820$ in Fig. 2, *lower*), and less accurate in the stick-slip regions (e.g., region $820<n<860$ in Fig. 2, *lower*), where it cuts the true free-energy landscape. Detailed calculations presented in the Section III of the Supporting Material show how the error in reconstructing the landscape of stick-slip regions done by the SP approximation depends on the total stiffness of the apparatus, $K(L)$.

Box inference: reconstructed basepair free-energy landscape

In Fig. 3, the basepair free energies, $g(n)$, inferred with the Box approximation from the unzipping data of Molecule 1 and from the synthetic data are compared to their true counterparts, computed from the best free-energy parameters found in Huguet et al. (4), and averaged on a sliding window of 30 bp. Fig. 3 (*middle row*) shows how the Box inference allows us to better follow the variation of the basepair free-energy landscape along the sequence, whereas the SP inference tends to cut the free-energy barriers. The results for Molecule 2 are very similar (see Fig. S14). The strong similarity between the free-energy landscapes corresponding to experimental and synthetic data inferred with the Box approximation in the two regions magnified in the middle row of Fig. 3 (at the beginning and end of the data sets) provides further support for the validity of the model and the equilibrium assumption.

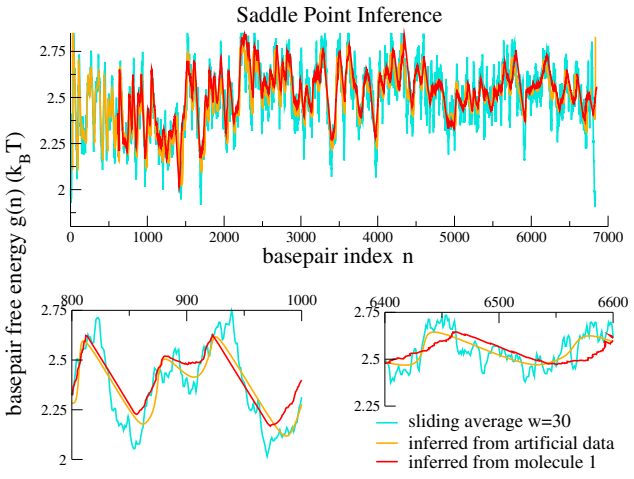

FIGURE 2 SP inference. The basepair free energies, $g(n)$, inferred from Molecule 1 force data (*red curve*) are compared to the true free energies (computed from Mfold, sliding average over $w = 30$ bp; *turquoise curve*) and to the free energies inferred from synthetic data generated from the model (*orange curve*). (*Upper*) Complete sequence. (*Lower*) Magnifications of two regions, one at the beginning of the molecule (*left*) and one at the end (*right*). To see this figure in color, go online.

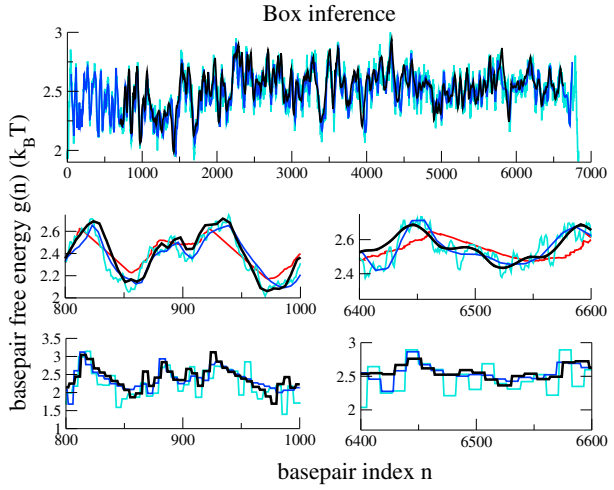

FIGURE 3 Box inference. Comparison of the basepair free energies, $g(n)$, inferred with the Box procedure from the force data of Molecule 1 ($b = 5$; *black curve*) and from the synthetic data ($b = 5$; *blue curve*) with the true free energies ($w = 30$; *turquoise curve*). (*Upper and middle*) Sliding averages with width $w = 30$ bp. Red curve (*middle*) shows the SP inference for Molecule 1, as in Fig. 2. (*Lower*) Box averages of widths 5 (*left*) and 10 bp (*right*). To see this figure in color, go online.

In Fig. 3, *lower*, we show the basepair free-energy landscape inferred with the Box approximation and the true basepair free-energy landscape with a box average with the same window size, b. The value of b (see Eq. 10) ranges from 5 bp at the beginning of the unzipping curve to 10 bp at the end of the sequence.

A detailed description of the reconstruction error along the sequence, i.e., of the difference between the true and the inferred basepair free-energy landscapes within the SP and Box approximations can be found in Section V of the Supporting Material. The reconstruction error does not show any systematic (monotonic) behavior with the number, n, of unzipped basepairs along the sequence. However, the error is larger in stick-slip regions (Fig. S10) and in regions for which the thermal drift of the optical trap has not been appropriately corrected, and the inferred basepair free-energy landscape is shifted with respect to the true landscape. This statement is corroborated by the fact that the inference error in the synthetic data set has much smaller peaks. The inference error in the real data are dominated by this drift problem, with the consequence that, apart from the very beginning of the sequence (700–1500 bp), it is not much lower with the Box approximation than with the SP approximation.

Alignment of unzipping force signals and experimental uncertainty in the inferred free-energy landscapes

To compare the inferred basepair free-energy landscapes of Molecules 1 and 2 with those obtained from the Mfold pairing parameters (Table S3 and Section VI in the Supporting Material), we aligned the two experimental force signals using the procedure described in Methods (see Fig. 4, *middle*). The agreement between the two force signals is much better than in the absence of alignment (Fig. 4, *upper*), though some differences are still visible, e.g., around $n = 4400$ bp. These differences allow us to quantify the experimental resolution of the force signal for two unzipping experiments with the same sequence in the setup used by Huguet and collaborators. We estimate the resolution of the SP landscape through the discrepancy between the free energies inferred for the two molecules after alignment by

$$\begin{aligned} \Delta g^{SP}_{av} &= \frac{1}{N}\sum_{n}\left|\Delta g^{SP}(n)\right|, \text{ with} \\ \Delta g^{SP}(n) &= g^{SP}_{1}(n) - g^{SP}_{2}(n), \end{aligned} \tag{13}$$

where subscripts 1 and 2 refer to Molecules 1 and 2, respectively. We find $\Delta g^{\mathrm{SP}}_{\mathrm{av}} \simeq 0.025$ (in units of $k_{\mathrm{B}}T$). Hence, the resolution per basepair is very small compared to the differences in free energy between different basepair types, which proves the efficiency of the alignment procedure.

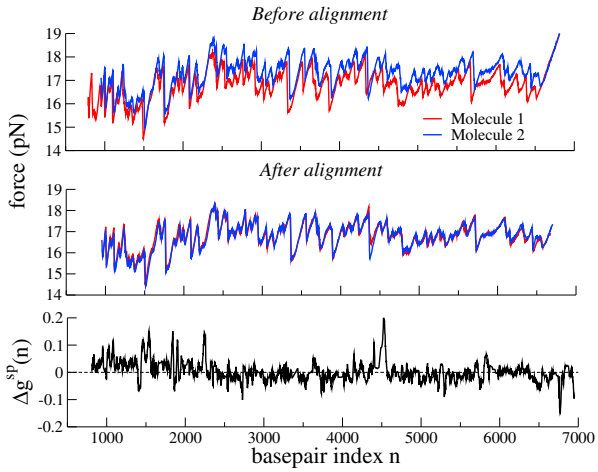

FIGURE 4 Force-curve alignment for Molecules 1 and 2. (*Upper*) Experimental force versus trap displacement for two unzipping experiments (Molecules 1 and 2) in Huguet et al. (4). (*Middle*) Coincidence between the force curves is strongly improved by our alignment procedure. We first globally shift the curve of Molecule 2 by $\simeq 0.5$ pN, then apply the Matlab routine for aligning the force curves along the basepair index axis, described in the main text. (*Lower*) Difference, $\Delta g^{SP}(n)$, between the SP free energies inferred from Molecules 1 and 2 after alignment. To see this figure in color, go online.

Fig. 4 (*lower*) shows that the discrepancy $\Delta g^{\mathrm{SP}}(n)$ is not uniform along the sequence and can reach values about 10 times higher than $\Delta g^{\mathrm{SP}}_{\mathrm{av}}$ values in some regions.

Sequence identification of the bacterial gene from synthetic force signal

We now compare the inferred free-energy landscapes between one 16S rDNA gene (the test sequence) and three other reference sequences based on the synthetic data. The test sequence, which we have chosen at random from the NCBI database (33), is a Brevibacterium of the *frigoritolerans* species (B-F); it is responsible for foot odor and is used for cheese fabrication. The reference sequences are a cyanobacterium, *Nostoc azollae* (N-A), another Brevibacterium, *B. halotolerans* (B-H), and the bacterium *Bacillus simplex* of the DSM 1321 strain (B-S). N-A and B-F have quite different 16S genes (329 mismatches), whereas B-H and B-F are more similar (102 mismatches); B-S, though not classified in the same family, is very similar to B-F (18 mismatches).

The first plot of Fig. 5 shows the theoretical unzipping force curves corresponding to the B-F and N-A genes and computed according to Eq. 4. As the two sequences differ widely in composition and length ($N = 1478$ and 1540 bp for N-A and B-F, respectively), the force curves are not well aligned, even if drift is not present in the synthetic data. We therefore align the two force curves using the force alignment procedure described in Methods (see Fig. 5, *second plot*). The SP free-energy landscapes of the two

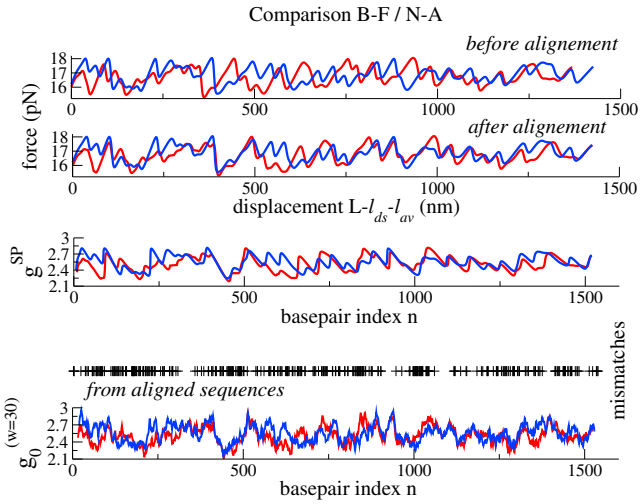

FIGURE 5 Comparison of the 16S gene of B-F and N-A bacteria from synthetic unzipping force signals. (*Upper two plots*) Equilibrium forces calculated from the 16S gene of the B-F (*blue*) and N-A (*red*) bacteria before (*upper*) and after (*lower*) alignment of the force signals. (*Middle*) Basepair free energies inferred from the aligned force signals with the SP procedure. (*Lower*) Basepair free energies computed from the aligned sequences and Mfold at 150 mM NaCl, averaged on a 30 bp sliding window; mismatch positions are shown with black crosses. To see this figure in color, go online.

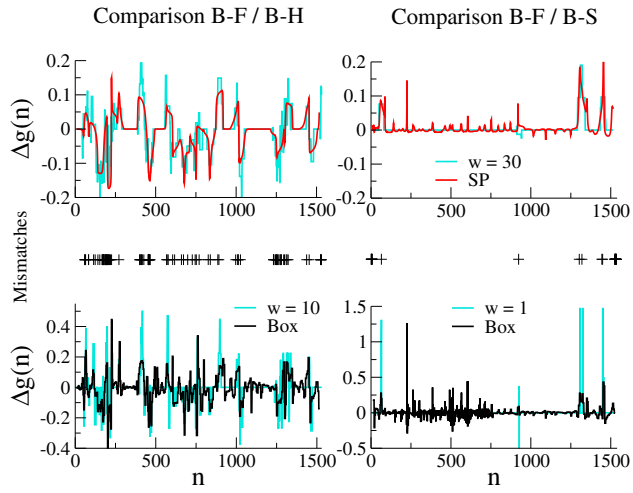

FIGURE 6 Comparison of the 16S gene of B-F bacterium with those of B-H (*left*) and B-S (*right*) bacteria from synthetic unzipping force signals. Differences, $\Delta g(n)$, between basepair free energies of the two bacteria after inferences with the SP (*upper*) and Box (*lower*) approximations. Sliding averages over $w = 30$ (*upper*), 10 (*lower left*), and 1 (*lower right*) of the differences between the Mfold free energies at 150 mM NaCl are shown for comparison (turquoise line). (*Middle*) Mismatches between B-F and B-H (*left*) and between B-F and B-S (*right*) sequences, obtained from direct alignment of the sequences; there are 102 mismatches between B-F and B-H, and 18 between B-F and B-S. To see this figure in color, go online.

sequences are then inferred (Fig. 5, *third plot*). Remarkably, the SP inferred free-energy landscapes are well aligned, as are the true free-energy landscapes obtained from the directly aligned sequences (Fig. 5, *bottom*). Moreover, the locations (basepair indices) of the discrepancies between the two inferred landscapes coincide with those between the true landscapes.

For the comparison of the test sequence (B-F) with the more similar reference sequences (B-H and B-S), we show in Fig. 6 the basepair free-energy differences, $\Delta g(n)$, between the landscapes inferred from the equilibrium force curves of the test and reference sequences using the SP (Fig. 6, *upper*) and Box (Fig. 6, *lower*) approximations; the corresponding free-energy landscapes can be found in Figs. S29–S34. Fig. 6 (*upper*) shows in addition the differences between the true landscapes, averaged over 30 bp. B-F and B-H can clearly be distinguished on this scale based on their SP free-energy landscapes. The difference between the inferred SP landscapes is $\Delta g^{\mathrm{SP}} = 59\ k_{\mathrm{B}}T$ in total or, equivalently, $\Delta g^{\mathrm{SP}}_{\mathrm{av}} \simeq 0.04\ k_{\mathrm{B}}T$/bp (see Eq. 13), which is larger than the resolution of $\simeq 0.02\ k_{\mathrm{B}}T$ estimated from the experiments on the λ-phage Molecules 1 and 2. We therefore conjecture that B-F and B-H could be distinguished using the SP procedure on unzipping data obtained with the setup of Huguet et al. (4).

The 16S genes of B-F and B-S differ by 18 mismatches only, 12 of which are located at the extremities of the molecules. Mutations can nevertheless be detected from the SP landscapes (Fig. 6, *upper right*) inferred from the synthetic data. Small peaks in the SP free-energy difference in mutation-free regions come from local errors in the force alignment, presumably due to the finite increment (≈ 0.2 pN) in the discretization of the force signal. The total difference between the inferred SP landscapes is $\Delta g^{\mathrm{SP}} = 12\ k_{\mathrm{B}}T$, that is, $\Delta g^{\mathrm{SP}}_{\mathrm{av}} \simeq 0.008\ k_{\mathrm{B}}T$/bp. This small value suggests that B-F and B-S probably could not be distinguished with unzipping data obtained with the setup of Huguet et al. (4)

With the Box procedure (Fig. 6, *lower*) the differences, $\Delta g(n)$, between the inferred landscapes agree with the differences between the true landscapes, averaged over $w = 10$ bp for B-F and B-H (Fig. 6, *lower left*) and over $w = 1$ bp for B-F and B-S (Fig. 6, *lower right*). In the latter case, all six internal mismatches coincide with peaks in the difference between the inferred basepair free energies, and can be detected; an additional peak, around basepair 250, is due to a local error in the force alignment. However, the Box approximation method is slower than the SP method (it takes several hours to fit the parameters for the $\simeq 1500$ bp sequence on a commercial Desktop computer with Mathematica). In addition, the Box procedure is more sensitive than SP to small errors in the force alignment procedure, e.g., errors around $n = 500$ in Fig. 6, *lower right*. The Box method should therefore be used as a refinement procedure to better quantify the differences in free-energy landscapes detected by the SP method.

Large-scale screening of bacterial database

We now carry out a large-scale screening of the bacterial database, keeping the test sequence (B-F) unchanged; the results of a similar large-scale screening where the test sequence is N-A are presented in Fig. S35. For each of the 2076 sequences in the database (see Section VIIF in the Supporting Material), we compute the synthetic equilibrium force curve, align it with the test force signal, and infer the SP free-energy landscape. The total free-energy difference with the test SP landscape is shown in Fig. 7 as a function of the number of mismatches in the pairwise sequence alignments (Fig. 7, *left*) and of the total difference between the true free-energy landscapes computed with Mfold (Fig. 7, *right*). The whole calculation for the 2076 sequences in the database, including the computation of the unzipping force curves, the alignments of the force signals, the SP inference, and the computation of free-energy differences is done with a Matlab code in ~15 min on a Macbook Pro computer.

Fig. 7 shows that there is a good correlation between the total SP free-energy difference and the number of mismatches. No simple linear relationship can be expected, as the difference in basepair free energy depends on the type of mismatch. In a similar way, the total differences between the inferred landscaped are strongly correlated to the total differences between the true landscapes (Fig. 7, *right*). As barriers are generally underestimated by the SP approximation, the former are generally smaller than the latter. We show in addition that the total SP free-energy differences increase when the synthetic data are generated with a fourfold-stiffer optical trap (Fig. 7, *light green dots*; see also Figs. S31 and S32).

The whole-database analysis with synthetic data and SP inference shows, as expected, that the only sequence with zero SP free-energy difference is the test sequence itself (B-F). The next most similar sequence is B-S. However, as discussed in the previous paragraph, the total difference in free energy is smaller than the experimental resolution between two identical molecules estimated from the data

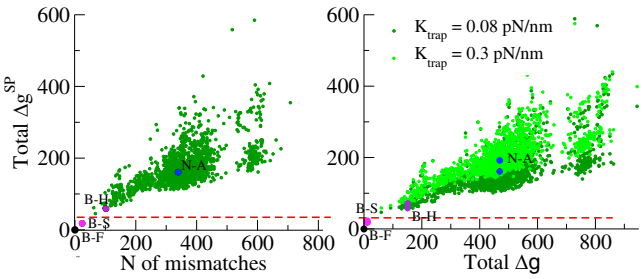

FIGURE 7 Large-scale comparison of the 16S gene in B-F with those in the other sequences of the database. (*Left*) SP free-energy differences versus the number of mismatches between B-F and the other sequences. (*Right*) SP versus Mfold (150 mM NaCl) free-energy differences between B-F and the other sequences. The straight line at $\Delta g = 35$ represents the experimental resolution estimated from Fig. 4. To see this figure in color, go online.

of Huguet and collaborators (4) and indicated in Fig. 7 by the red dashed line. It seems, however, that B-F can be distinguished from any other sequence in the database, including B-H, when experimental errors are taken into account. In Section VII of the Supporting Material, the 16S rDNA gene of N-A is compared to the other 2076 sequences in the database. The results are similar to what is shown in Fig. 7 for the B-F test sequence, with the difference that N-A could also be distinguished from its closest sequence with the estimated experimental resolution.

CONCLUSION

In this article, we have shown how the basepair free-energy landscape of a single DNA molecule with 6800 bases could be inferred from the unzipping data published in Huguet et al. (4) with a resolution of 30–40 basepairs. Sequencing techniques with low resolution but used on very long sequences could be interesting in practical applications, and could complement current techniques, which are limited to short reads.

The inference of the whole free-energy landscape is a difficult problem, as it requires determination of a large number of parameters, increasing in a linear fashion with the length of the molecule. We have proposed and compared two approximation approaches to solve this problem. The first approach, the SP method, is in practice a reparameterization of the force-extension curve and requires very little computational effort. The second procedure, called Box approximation, consists of approximating the free-energy landscape with a piecewise constant function on the scale of b bases and fitting the corresponding coarse-grained energetic parameters to match the equilibrium force computed from an unzipping model to the experimental signal. We find that the best value for b is about half the ratio of the length of thermal fluctuations of the bead in the optical trap over the typical length of two open basepairs. This choice allows us to adjust the procedure automatically to the precision of the experimental setup, and to avoid overfitting the data. As the size of ssDNA fluctuations increases with the number, n, of unzipped basepairs, so does the natural resolution, b, ranging from ~5 bases at the beginning of the molecule to 20 bases at the end of the molecule (see Fig. S1) in the setup of Huguet et al. (4). It is important to stress that the value of b at the beginning of the opening depends on the stiffness of the optical trap and of the dsDNA linkers (assumed to be rigid here, since they are very short) and could easily be made smaller in other experimental setups. Indeed, the optical trap stiffness in the setup of Huguet and co-workers (4), $K_{\text{trap}} = 0.08$ pN/nm, is relatively small. As a matter of comparison, consider the unzipping experiments of Woodside et al. (3), for which $K_{\text{trap}} = 0.3 - 0.4$ pN/nm. With the Box inference method, we expect to be able to resolve the free-energy landscape over the first 200 bp of the sequence with a resolution of

the order of $b \simeq 0.5\sqrt{k_B T/K 4\ell_{ss}^2} \approx 2$ bp (Fig. S26). As a consequence, it seems possible to drastically improve the reconstruction scale of the inference by changing the setup for small n, until K_{ss} becomes the smallest stiffness of the setup and ssDNA fluctuations are the dominant contribution to the bead fluctuations. Unfortunately, in the data sets analyzed, the unzipping signal starts at $n = 700 - 900$ open basepairs (for Molecules 1 and 2, respectively), and the part of the unzipping dominated by the trap stiffness is missing. In this range, as shown with the synthetic data, the resolution on the inferred free-energy landscape is expected to be of the order of 5–10 bases.

Comparison with the synthetic unzipping data obtained from the known sequence show that the major source of error is the drift of the apparatus. The presence of drift, the intensity of which could be reduced by the use of specific setups, e.g., double optical traps, considerably affects the accuracy of inferred free-energy landscapes. We have proposed an alignment procedure of force curves, which makes use of the celebrated Needleman-Wunsch algorithm for aligning nucleotidic or protein sequences. We have shown that the procedure is efficient for aligning two experimental force signals (Molecules 1 and 2) affected by drift and corresponding to the same DNA sequence. We expect that drift could be practically eliminated and that the inferred free energy could be assigned to unambiguous basepair indices, even in the absence of any a priori information on the sequence, by aligning a large number of unzipping curves corresponding to the same sequence. A systematic check of the efficiency of our alignment procedure on other experimental data, with several unzipping signals, would therefore be very useful.

In the second part of the article, we have given a proof of principle, with synthetic force data, that unzipping experiments combined with our inference approach could be used as a method of identifying one among thousands of 16S rRNA bacterial sequences. The standard method for detecting homologous sequences is DNA-DNA hybridization. DNA-DNA association kinetics is informative about the similarity between test and reference DNA sequences. However, this hybridization method is quite involved, as it is time-consuming, labor-intensive, and expensive to perform (32). Moreover, it gives only a global measure of the difference between the test and reference sequences. Unzipping-based methods could, in principle, also give local information on similarities or dissimilarities between the sequences.

The gene screening procedure proposed here allows us to find, within experimental limitations, the sequence corresponding to the test gene in the database. If this sequence is not present, the SP inference procedure identifies the sequence most similar to that of the test gene, partially reconstructs the sequence of the test gene in the matching zones, and gives the coarse-grained differences between the two free-energy landscapes in the nonmatching regions on a 10–50 bp scale, which depends on the experimental resolution and on the inference method. In particular, the SP method is robust and fast, and can be carried out with no extra computation cost with respect to the comparison of unzipping forces. Once the most similar sequence has been found, a more precise resolution over the differences of the free-energy landscapes can be obtained by the Box approximation. Note that we used our force alignment procedure to compare theoretical free-energy landscapes of homologous (but distinct) sequences, which are free of drift. When comparing an experimental force curve to one or more theoretical force curves computed from a sequence database, the force alignment procedure will, in addition, be helpful in removing the drift from the data.

We stress that differences between the inferred free-energy landscapes are more meaningful than differences between the true and inferred landscapes. Although there may be important differences between the SP free-energy landscape and the true one, e.g., due to the stick-slip characteristics of the unzipping signal, we have shown that homologous sequences, even a few mutations away from one another, could be distinguished by comparing their SP landscapes. SP comparison provides information not at the basepair level, but on larger scales. To achieve basepair accuracy, one could combine unzipping experiments with the hybridization of oligonucleotide probes (16), which could be engineered to bind to the part of the sequence where a different landscape has been detected. It would be interesting to test the hybridization of different probes with nucleotide contents compatible with the average free-energy difference inferred from the unzipping signal and the SP or Box approximations.

The study described in this article could be extended in several ways. We based our inference on the equilibrium force signal by filtering the force data at a resolution of 1 Hz. However, the temporal resolution of the data acquisition is much higher (here, 1 kHz). The data therefore contain in principle more information than the average force at each position. Thus, one way to expand on this study would be to exploit for the inference not only the average force but the distribution of forces at each position. A second, and very interesting way would be to incorporate in the model elements of the unzipping dynamics by taking into account the bead, single strand, and linker relaxation dynamics (35). In addition, although we focused here on the basepair free-energy landscape associated with the sequence, we did not attempt to infer the sequence itself. Thus, a third way to extend this study would be to look for the most likely sequence capable of generating the inferred Box-averaged free energy, g_k. It would be useful to introduce more complicated priors over the energetic parameters used in this work, in particular to constrain the basepair free energies to take values from a set of only 10 possible known values.

SUPPORTING MATERIAL

Four tables, 35 figures, and a detailed description of the model and theory are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(13)05761-5.

We are grateful to J. M. Huguet, M. Ribezzi, and F. Ritort for the communication of their data and for enlightening discussions. We thank V. Croquette for discussions and for having provided us with the 16S-rDNA database.

This work has benefited from the financial support of the Agence Nationale de la Recherche Jeunes Chercheurs grant ANR-06-JCJC-0051.

REFERENCES

1. Liphardt, J., B. Onoa, …, C. Bustamante. 2001. Reversible unfolding of single RNA molecules by mechanical force. *Science.* 292:733–737.
2. Danilowicz, C., V. W. Coljee, …, M. Prentiss. 2003. DNA unzipped under a constant force exhibits multiple metastable intermediates. *Proc. Natl. Acad. Sci. USA.* 100:1694–1699.
3. Woodside, M. T., W. M. Behnke-Parks, …, S. M. Block. 2006. Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins. *Proc. Natl. Acad. Sci. USA.* 103:6190–6195.
4. Huguet, J. M., C. V. Bizarro, …, F. Ritort. 2010. Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proc. Natl. Acad. Sci. USA.* 107:15431–15436.
5. Essevaz-Roulet, B., U. Bockelmann, and F. Heslot. 1997. Mechanical separation of the complementary strands of DNA. *Proc. Natl. Acad. Sci. USA.* 94:11935–11940.
6. Bockelmann, U., P. Thomen, …, F. Heslot. 2002. Unzipping DNA with optical tweezers: high sequence sensitivity and force flips. *Biophys. J.* 82:1537–1553.
7. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 74:5463–5467.
8. Harris, T. D., P. R. Buzby, …, Z. Xie. 2008. Single-molecule DNA sequencing of a viral genome. *Science.* 320:106–109.
9. Fuller, C. W., L. R. Middendorf, …, D. V. Vezenov. 2009. The challenges of sequencing by synthesis. *Nat. Biotechnol.* 27:1013–1023.
10. Metzker, M. L. 2010. Sequencing technologies: the next generation. *Nat. Rev. Genet.* 11:31–46.
11. Kircher, M., and J. Kelso. 2010. High-throughput DNA sequencing—concepts and limitations. *Bioessays.* 32:524–536.
12. Mir, K. U., H. Qi, Salata, …, G. Scozzafava. 2009. Sequencing by cyclic ligation and cleavage (CycLiC) directly on a microarray captured template. *Nucleic Acids Res.* 37:e5.
13. Pihlak, A., G. Baurén, …, S. Linnarsson. 2008. Rapid genome sequencing with short universal tiling probes. *Nat. Biotechnol.* 26:676–684.
14. Mardis, E. R. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9:387–402.
15. Stoddart, D., A. J. Heron, …, H. Bayley. 2009. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. USA.* 106:7702–7707.
16. Ding, F., M. Manosas, …, V. Croquette. 2012. Single-molecule mechanical identification and sequencing. *Nat. Methods.* 9:367–372.
17. Lubensky, D. K., and D. R. Nelson. 2000. Pulling pinned polymers and unzipping DNA. *Phys. Rev. Lett.* 85:1572–1575.
18. Gerland, U., R. Bundschuh, and T. Hwa. 2001. Force-induced denaturation of RNA. *Biophys. J.* 81:1324–1332.
19. Cocco, S., J. F. Marko, and R. Monasson. 2003. Slow nucleic acid unzipping kinetics from sequence-defined barriers. *Eur Phys J E Soft Matter.* 10:153–161.
20. Baldazzi, V., S. Cocco, …, R. Monasson. 2006. Inference of DNA sequences from mechanical unzipping: an ideal-case study. *Phys. Rev. Lett.* 96:128102–128106.
21. Collin, D., F. Ritort, …, C. Bustamante. 2005. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature.* 437:231–234.
22. Pavliotis, G. A., and A. M. Stuart. 2007. Parameter estimation for multiscale diffusions. *J. Stat. Phys.* 127:741–781.
23. Junier, I., A. Mossa, …, F. Ritort. 2009. Recovery of free energy branches in single molecule experiments. *Phys. Rev. Lett.* 102:070602.
24. Zhang, Q., J. Brujić, and E. Vanden-Eijnden. 2011. Reconstructing free-energy profiles from nonequilibrium relaxation trajectories. *J. Stat. Phys.* 144:344–366.
25. Crommelin, D. 2012. Estimation of space-dependent diffusions and potential landscapes from non-equilibrium data. *J. Stat. Phys.* 149:220–233.
26. Alemany, A., A. Mossa, …, F. Ritort. 2012. Experimental free-energy measurements of kinetic molecular states using fluctuation theorems. *Nat. Phys.* 8:688–694.
27. Thompson, E. R., and E. D. Siggia. 1995. Physical limits on the mechanical measurement of the secondary structure of bio- molecules. *Europhys. Lett.* 31:335–340.
28. Zuker, M. 2000. Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.* 10:303–310.
29. Smith, S. B., Y. Cui, and C. Bustamante. 1996. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science.* 271:795–799.
30. Cocco, S., J. F. Marko, and R. Monasson. 2002. Theoretical models for single-molecule DNA and RNA experiments: from elasticity to unzipping. *C. R. Physique.* 3:569–584.
31. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.
32. Janda, J. M., and S. L. Abbott. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45:2761–2764.
33. RefSeq Targeted Loci Project e, and 16S Bacterial Ribosomal RNA project: http://www.ncbi.nlm.nih.gov/genomes/static/refseqtarget.html.
34. Bockelmann, U., B. Essevaz-Roulet, and F. Heslot. 1997. Molecular stick-slip motion revealed by opening DNA with piconewton forces. *Phys. Rev. Lett.* 79:4489–4492.
35. Barbieri, C., S. Cocco, …, F. Zamponi. 2009. Dynamical modeling of molecular constructions and setups for DNA unzipping. *Phys. Biol.* 6:025003–025023.

Supporting Material

Reconstruction and identification of DNA sequence landscapes from unzipping experiments at equilibrium

C. Barbieri, S. Cocco, T. Jorg, R. Monasson

I. PARAMETERS FOR TRAP STIFFNESS, SS-DNA AND DS-DNA ELASTICITY, AND BASE PAIRING AND STACKING FREE ENERGIES

The condition of the unzipping experiment performed by Huguet and collaborators [1] are the following: temperature = 25^oC, ionic concentration of the solution= $1M$, pH= 7.5. The stiffness of the optical trap is $K_{trap} = 0.080$ pN/nm. As in [1] the ssDNA, released during unzipping, is modeled by a Freely-Jointed-Chain with Kuhn length $b_o = 1.15$ nm and interphosphate distance $d = 0.59$ nm between consecutive bases. The two dsDNA (handles) are modeled according to a Worm-Like-Chain with persistence length $l_p = 50$ nm and contour length $L_0 = 9.18$ nm. The free-energy parameters $g_0(s, s')$, which account for both pairing and stacking contributions, extracted from [1], are given in Table S1 and Table S2. These values correspond to the best energetic parameters, *i.e.* reproducing as close as possible the unzipping forces of Molecule 1 and of Molecule 2 respectively. In Table S3 we give the pairing parameters extracted from the MFold server for the experimental condition of [1] ($T = 25^oC$, $Na = 1M$) [2]. In Table S4 we give the pairing parameters extracted from the MFold server for the ionic concentration $Na = 150mM$ and $T = 25^oC$ [2].

II. MODEL FOR UNZIPPING

A. Derivation of the free energy $G(n|L)$ for n unzipped base pairs

The elastic free energy of the single strand (ss) of DNA at fixed force f is given by the modified freely jointed chain expression [1, 3]:

$$G_{ss}(n, f) = n\, g_{ss}(f) = n\, b_o\, \log\left[k_B T\, \frac{\sinh(d\, f/k_B T)}{d\, f}\right] \tag{1}$$

The parameter values $d = 0.59$ Å, $b_o = 1.15$ Å for 1M ionic conditions are extracted from [1].

The free energy of ssDNA at fixed distance x_{ss} between its two extremities is

$$G_{ss}(n, x_{ss}) = f(x_{ss})\, x_{ss} - n\, g_{ss}(f(x_{ss}))\ , \tag{2}$$

g_0	A	T	C	G
A	2.05	1.67	2.37	2.15
T	1.34	2.05	2.38	2.79
C	2.79	2.15	3.06	3.8
G	2.38	2.37	3.89	3.06

TABLE S1: Best binding free energies $g_0(s_i, s_{i+1})$ (units of k_BT) obtained for Molecule 1 in [1]. Base values s_i and s_{i+1} correspond to lines and columns respectively.

g_0	A	T	C	G
A	2.05	1.81	2.41	2.26
T	1.42	2.05	2.63	2.83
C	2.83	2.26	3.18	4.08
G	2.631	2.41	4.11	3.18

TABLE S2: Best binding free energies $g_0(s_i, s_{i+1})$ (units of k_BT) for Molecule 2 as computed in [1].

g_0	A	T	C	G
A	2.13	1.88	2.87	2.57
T	1.41	2.13	2.64	2.89
C	2.89	2.57	3.49	4.2
G	2.64	2.87	4.25	3.49

TABLE S3: Binding free energies $g_0(s_i, s_{i+1})$ (units of k_BT) obtained from the MFold server [2] for DNA at room temperature, pH=7.5, and ionic concentration of 1 M.

g_0	A	T	C	G
A	1.78	1.54	2.52	2.21
T	1.05	1.78	2.28	2.53
C	2.53	2.22	3.14	3.84
G	2.28	2.52	3.89	3.14

TABLE S4: Binding free energies $g_0(s_i, s_{i+1})$ (units of k_BT) obtained from the MFold server [2] for DNA at room temperature, pH=7.5, and ionic concentration of 150m M.

where $f(x_{ss})$ is the force required for a single strand with n unzipped base pairs to have extension x_{ss} at equilibrium, implicitly defined through

$$x_{ss} = \frac{\partial G_{ss}}{\partial f}(n, f) = n \, \frac{dg_{ss}}{df}(f) \; . \tag{3}$$

Hereafter we simplify the above expression for G_{ss} through an expansion around the average unzipping force f_{av}. This expansion, referred to as local harmonic approximation, is expected to be valid for small fluctuations of the force around f_{av}.

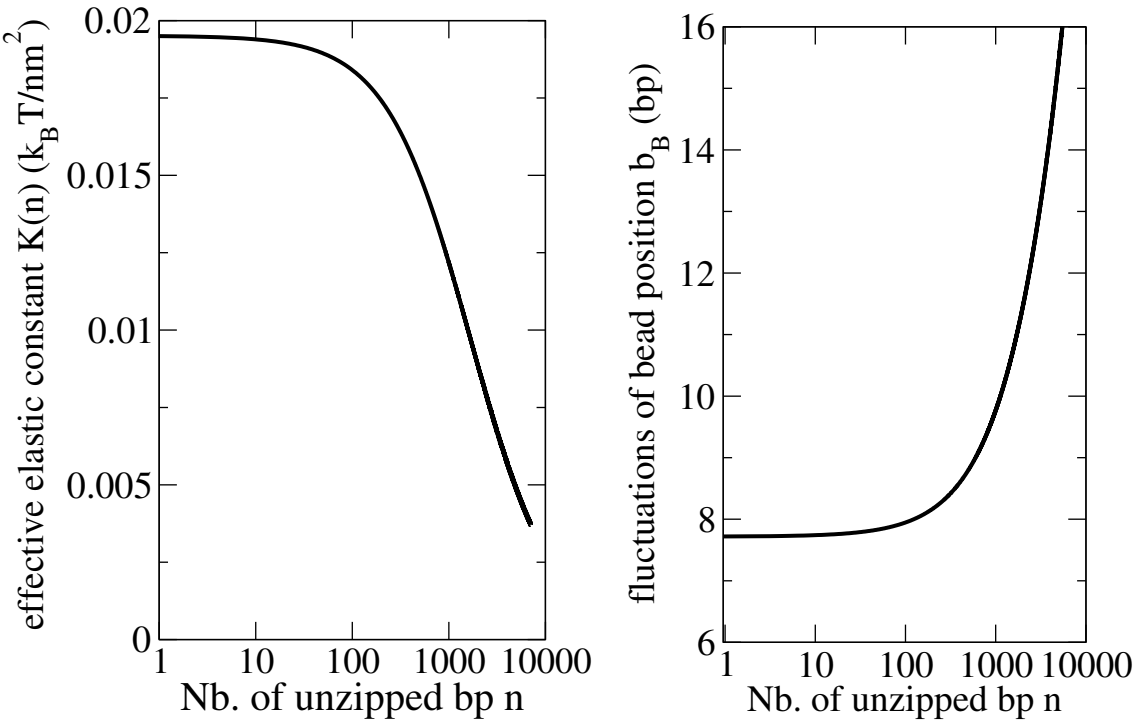

FIG. S1: Left: Stiffness of the experimental setup, $K(n)$, in units of k_BT/nm^2, as a function of the number of unzipped base pairs. Right: fluctuation $b_B(n)$ of bead in units of the extension ℓ_{ss} of an open bp as a function of the number of unzipped base pairs.

We start by choosing a reference value for the unzipping force f_{av}, and define the ssDNA extension per bp according to Eq. (3):

$$\ell_{ss} = \frac{dg_{ss}}{df}(f_{av}) \ . \tag{4}$$

A small deviation of x_{ss} from the equilibrium value $n\,\ell_{ss}$ corresponding to force $f_{av} = f(n\,\ell_{ss})$ will result in a small change of the force f applied on the ssDNA extremities. Linearizing Eq. (3) around $x_{ss} = n\,\ell_{ss}$ and $f = f_{av}$, we obtain

$$f - f_{av} \simeq K_{ss}(n)\ (x_{ss} - n\,\ell_{ss}) \ , \tag{5}$$

where the stiffness K_{ss} of the ssDNA is defined through

$$\frac{1}{K_{ss}(n)} = n\ \frac{d^2 g_{ss}}{df^2}(f_{av}) \ . \tag{6}$$

Notice that the effective stiffness for the ssDNA decreases with the number of unzipped base pairs.

The resulting expression for the free energy of the ssDNA at fixed extension is, within the local harmonic approximation corresponding to Eq. (5),

$$G_{ss}(n, x_{ss}) \simeq f_{av}\, x_{ss} + \frac{1}{2} K_{ss}\,(x_{ss} - n\,l_{ss})^2 - n\,g_{ss} \tag{7}$$

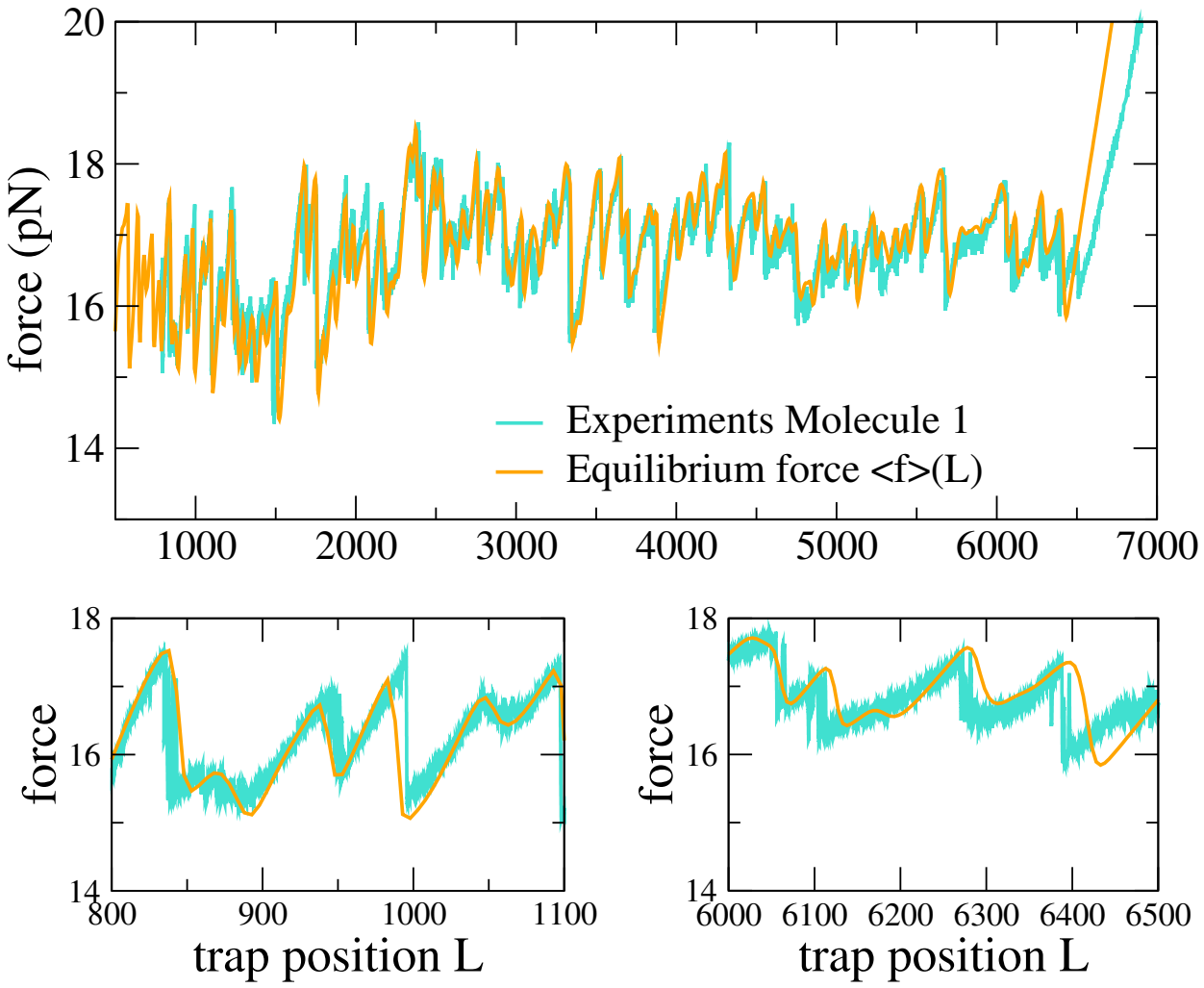

FIG. S2: Equilibrium force in the harmonic approximation compared to the experimental force. Top: Unzipping force, as a function of the trap position L (in nm). Turquoise line: experimental results for Molecule 1, Orange line: average force at equilibrium $\langle f \rangle(L)$. Bottom: magnification of two regions, at the beginning (left) and the end (right) of the sequence.

where $g_{ss} \equiv g_{ss}(f_{av})$.

The experimental setup includes, in addition to the ssDNA, the optical trap with stiffness constant K_{trap}, the small double strand (ds) DNA linkers which can be considered to be rigid for the force range $\simeq f_{av}$ considered here, and the dsDNA molecule which is unzipped (Fig. 1 of the main paper). We model the free energy cost for breaking apart the first n base pairs $(s_1, s_2, \ldots, s_n)$ of the molecule through

$$G_{ds}(n) = \sum_{i \leq n} g_0(s_i, s_{i+1}) \ , \tag{8}$$

where the energetic parameters $g_0(s_i, s_{i+1})$ are given in Tables S1, S2 & S3.

We may now write the total free energy $G(x_{ss}^{(1)}, x_{ss}^{(2)}, n|L)$ of the system as a function of the number n of unzipped base pairs, of the extensions $x_{ss}^{(1)}$ and $x_{ss}^{(2)}$, of the position L of the trap, and of the total extension ℓ_{ds} of the dsDNA linkers, see Fig. 1 of the main paper. Expressing the displacement of the bead with respect to the center of the trap as $L - x_{ss}^{(1)} - x_{ss}^{(2)} - \ell_{ds}$ we obtain

$$\begin{aligned} G(x_{ss}^{(1)}, x_{ss}^{(2)}, n|L) &= G_{ds}(n) + G_{ss}(n, x_{ss}^{(1)}) + G_{ss}(n, x_{ss}^{(2)}) + \frac{1}{2} K_{trap} (L - x_{ss}^{(1)} - x_{ss}^{(2)} - \ell_{ds})^2 \\ &= G_{ds}(n) - 2\, n\, g_{ss} + f_{av} (x_{ss}^{(1)} + x_{ss}^{(2)}) + \frac{1}{2} K_{ss} (x_{ss}^{(1)} - n\ell_{ss})^2 + \\ &\quad \frac{1}{2} K_{ss} (x_{ss}^{(2)} - n\ell_{ss})^2 + \frac{1}{2} K_{trap} (L - x_{ss}^{(1)} - x_{ss}^{(2)} - \ell_{ds})^2. \end{aligned} \tag{9}$$

All energetic parameters are expressed in units of $k_B T$.

The partition function for a fixed displacement L is

$$Z(L) = \sum_{n=0}^{N} \int_{-\infty}^{\infty} dx_{ss}^{(1)}\, dx_{ss}^{(2)}\, e^{-G(x_{ss}^{(1)}, x_{ss}^{(2)}, n|L)} \tag{10}$$

As a consequence of the local harmonic approximation the integration over the variables $x_{ss}^{(1)}, x_{ss}^{(2)}$ amounts to calculate two coupled Gaussian integrals, with the result

$$Z(L) = \sum_{n=0}^{N} e^{-G(n|L)} \tag{11}$$

where the effective free energy per unzipping n base pairs (at fixed L) is given by

$$G(n|L) = G_{ds}(n) - 2\, n\, g_{ss} + \frac{1}{2}\, K(n)\, (L - \ell_{av} - \ell_{ds} - 2\, n\, \ell_{ss})^2 \ . \tag{12}$$

The effective spring constant is

$$K(n) = \frac{K_{ss}(n)\, K_{trap}}{K_{ss}(n) + 2\, K_{trap}} \ . \tag{13}$$

We plot the effective stiffness $K(n)$ of the experimental setup as a function of the number n of unzipped bases in Fig. S1 (left); $K(n)$ is dominated by K_{trap} at small n and by $K_{ss}(n)$ at large n and, therefore, decreases as $1/n$ at large n.

Let us fix the displacement of the trap to some value L. As the fluctuations of n around its average value are small (see Section II D) compared to the inverse of the gradient of $K(n)$ we can in practice replace $K(n)$ with its value when the argument is equal to the average number of open base pairs,

$$\langle n \rangle (L) = \frac{1}{Z(L)} \sum_{n=0}^{N} n\, e^{-G(n|L)} \ . \tag{14}$$

The effective stiffness becomes a function of L, denoted by $K(L)$. Parameter $\ell_{av} = f_{av}/K(L)$ appearing in (12) is the displacement of the bead with respect to the trap center under the action of the average force f_{av}.

The standard deviation of the position of the bead at fixed L, $b_B(L)$ (see Fig. 1 in main text), expressed in units of the ssDNA extension ℓ_{ss} resulting from the opening of one bp has a simple expression in terms of the effective stiffness:

$$b_B(n) = \frac{1}{\sqrt{K(n)\,\ell_{ss}^2}}. \tag{15}$$

Figure S1 (right) shows the value of b_B as a function of n. Knowledge of b_B is useful to estimate the value of the box size b in the Box inference procedure, see Section IV B.

B. Parameters for the local harmonic approximation

The ss-DNA stretching free energy is expanded, in the local harmonic approximation, around the the force needed to unzip an uniform sequence with average base-pair free energy g_0. For Molecule 1 with the parameters given in [1] $g_0 = 2.5$ k$_B$T, giving $f_{av} = 16.6$ pN from the condition $2\,g_{ss}(f_{av}) = 2.5$ k$_B$T; at this force the extension of a ss-DNA base is $\ell_{ss} = 0.465$ nm, the extension of the two ds-DNA linker is $\ell_{ds} = 19.7$ nm and the displacement of the bead in the optical trap at the average unzipping force is $\ell_{av} = 208$ nm.

For Molecule 2 with the parameters given in [1] and Molecules 1 and 2 after simple alignment (see Section VI), we have used $f_{av} = 18$ pN, $\ell_{ss} = 0.946$ nm, $\ell_{ds} = 19.7$ nm, $\ell_{av} = 224.5$ nm. This unzipping force corresponds to the average free energy $g_0 = 2.8$ k$_B$T, obtained from the pairing parameters of MFold at 1M. We have verified that the outcome of the inference procedures does not depend much on the force f_{av} around which the ssDNA elasticity is expanded in the range of the unzipping force (14-18 pN).

C. Comparison of experimental and equilibrium forces with the local harmonic model

To validate the above model we show in Fig. S2 the unzipping force computed at equilibrium, $\langle f \rangle(L) = f_{av} + d \log Z(L)/dL$, compared to experimental data. The model fits quite well the data, even if slip events are steeper in experimental data than in the model. Note that at the end of the unzipping the theoretical curve and the experimental one are less well aligned, due to experimental drift.

Experimental data and model predictions differ in (at least) two important aspects:

- the force measured in experiments is averaged out over a 1 second time-window, and is not really sampled at equilibrium;

- the corresponding displacements of the trap (values of L) are averaged over on time intervals of 1 second, too.

On the contrary theory predicts the equilibrium value for the force for a fixed displacement, as we sum over all possible values for n, $x_{ss}^{(1)}$, $x_{ss}^{(2)}$. It would be interesting to take into account non equilibrium effects [4] in the theoretical calculations due to the changes in the displacement over the sliding window, and see if the comparison with the data is improved.

D. Number of open base pairs: average value and fluctuations

The average number of open base pairs is related to the displacement L and the average force $\langle f \rangle(L)$ at that displacement L by the equation, see Material and Methods Section,

$$\langle n \rangle(L) = \frac{L - \ell_{ds} - \ell_{av} - (\langle f \rangle(L) - f_{av})/K(L)}{2\ell_{ss}} \ . \tag{16}$$

Fluctuations around the average value are characterized by the standard deviation

$$\sigma_n(L) = \sqrt{\frac{1}{Z(L)} \sum_n (n - \langle n \rangle(L))^2 \, e^{-G(n|L)}} \ . \tag{17}$$

In Fig. S3 we show both the average number $\langle n \rangle(L)$ of unzipped pairs (top) and the standard deviation $\sigma_n(L)$ (bottom) as a function of the trap displacement L. For the sake of clarity we use the number of unzipped bases for the average force $f_{av} \simeq 16.65$ pN corresponding to a homogeneous sequence with uniform free energy $g_0 = 2\, g_{ss} = 2.5 \ k_B T$,

$$n_{av}(L) = \frac{L - \ell_{ds} - \ell_{av}}{2\ell_{ss}} \ , \tag{18}$$

as a dimensionless proxy for the trap position L. We observe that $\langle n \rangle(L)$ remains close to $n_{av}(L)$ as L increases, with positive or negative differences depending on whether the bp free energies are locally stronger or weaker than the average value g_0. Fluctuations at equilibrium, measured by $\sigma_n(L)$, can be of a few tens of bp. The standard deviations show strong heterogeneities with n but is, on the overall, larger at the end of the sequence than at the beginning, as expected from the fact that the setup stiffness decreases with n. Let us stress again that the effective stiffness $K(n)$ remains essentially unchanged when the bp number varies by $\sigma_n \sim$ few 10s bp; hence we are allowed to approximate $K(n)$ with a function of the trap position L only.

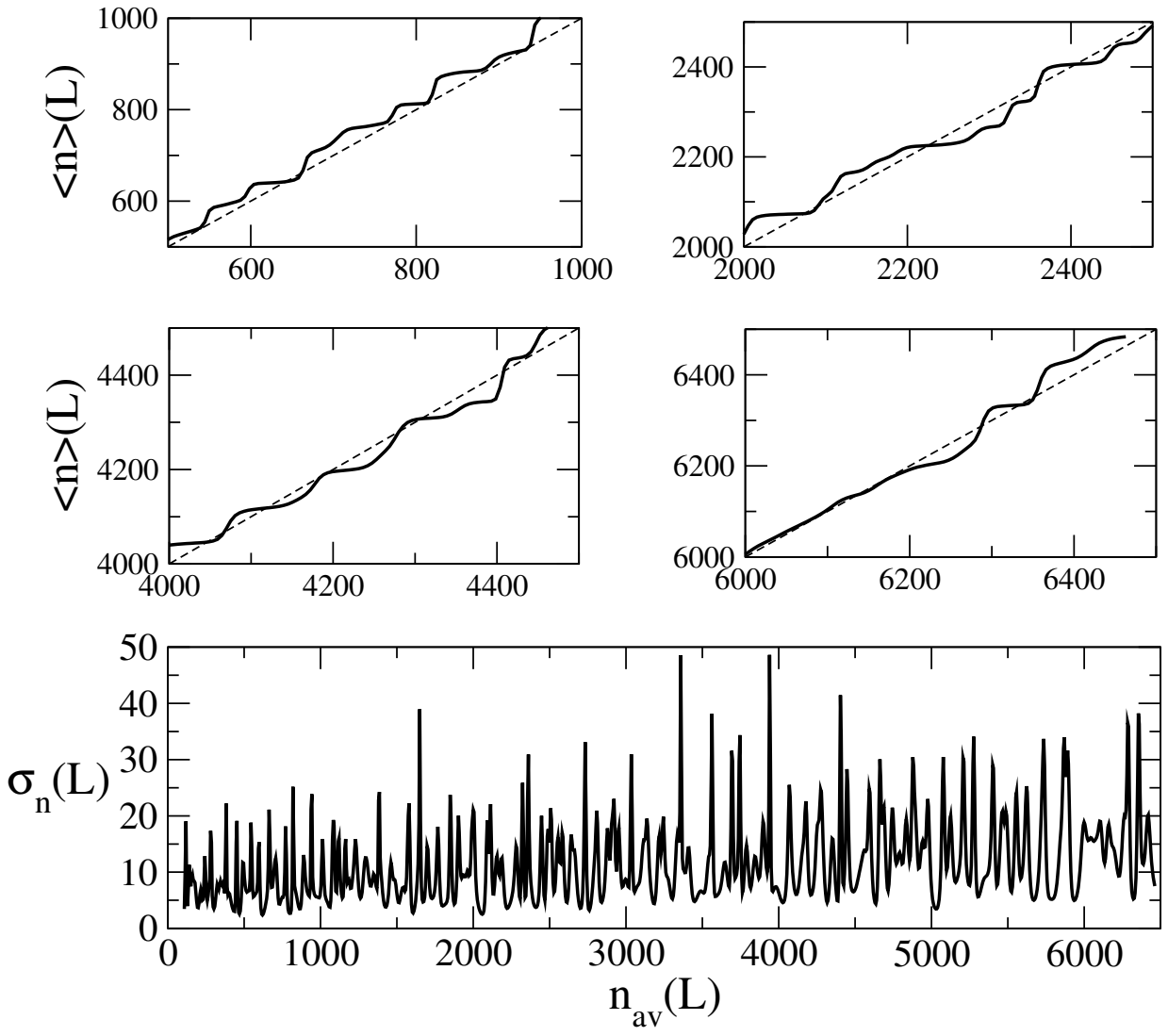

FIG. S3: Top and middle panels: Average number of open base pairs $\langle n \rangle(L)$ in the harmonic model (black) as a function of the its average sequence counterpart, $n_{av}(L)$, see Eq. (18). Dashed lines show the $n_{av} = \langle n \rangle$ curves. Bottom: Standard deviation of the number of open bp at equilibrium in the harmonic model, $\sigma_n(L)$, as a function of $n_{av}(L)$.

III. THEORETICAL STUDY OF THE INFERENCE ERROR IN THE SADDLE POINT APPROXIMATION

A. Deviations of the average number of open base pairs within the Saddle-Point approximation

We can check the self-consistency of the SP approximation by computing the difference $\Delta\langle n \rangle(L)$ between the average value of n at fixed L with the inferred sequence landscape, g_0^{SP}, and $n^{SP}(L)$.

If the SP approximation were exact this difference would vanish for all L. To lighten notations let us define $\ell = 2\ell_{ss}$ and rescale $L - 2\ell_{ds} - L_{av} \to L$. We write

$$\Delta\langle n\rangle(L) = \frac{1}{Z^{SP}(L)} \int_0^N dn\ n\ \exp\left(-G^{SP}(n) - \frac{K(L)}{2}(L - n\,\ell)^2\right) - n^{SP}(L) \tag{19}$$

with

$$Z^{SP}(L) = \int_0^N dn\ \exp\left(-G^{SP}(n) - \frac{K(L)}{2}(L - n\,\ell)^2\right)\ , \tag{20}$$

and

$$G^{SP}(n) = \int_0^n dn'\ (g_0^{SP}(n') - 2g_{ss}). \tag{21}$$

The result of the calculation for Molecule 1 is shown in Fig. S4. We observe that the deviations from the SP number of base pairs can reach substantial values, of a few tens of bases, comparable with the order of magnitude of the standard deviation σ_n.

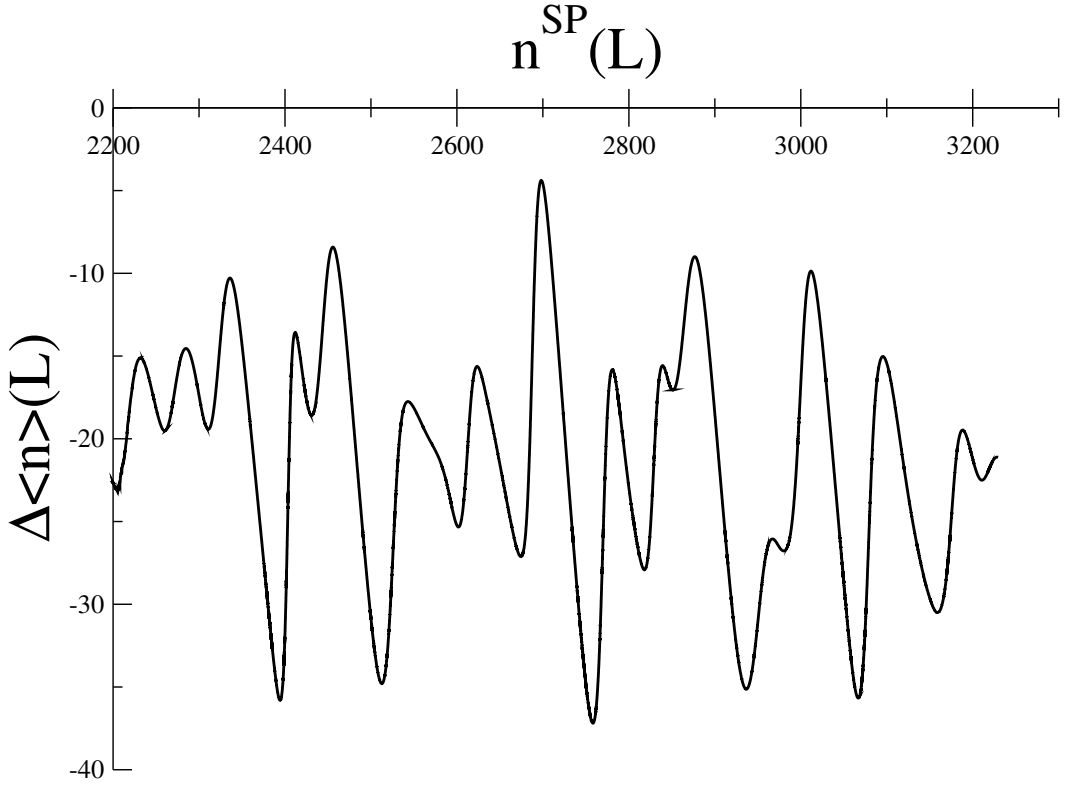

FIG. S4: Parametric representation of the deviation $\Delta\langle n\rangle(L)$ between the average number of open bp and the SP value vs. $n^{SP}(L)$ for a portion of the sequence landscape inferred with the SP approximation.

B. Theoretical curves for the SP inference for barriers

In this section we show that the SP approximation reproduces regions in in free energy landscape in the DNA sequence where weak bp are followed by stronger bp more faithfully than regions where strong bp are followed by weaker bp. The latter regions will be hereafter called Strong-Weak (S-W) barriers, and the former Weak-Strong (W-S).

Consider a barrier in the cumulative free-energy landscape, of height ΔG (with respect to the average free energy $2ng_{ss}$) and of width Δn. The barrier is of the Weak-Strong type if $\Delta G < 0$, and of the Strong-Weak type if $\Delta G > 0$. S-W barriers are responsible for the so-called stick-slip phenomenon [5]. The barrier can be locally approximated as a harmonic potential, whose stiffness is of the order of $-\Delta G/(\Delta n)^2$. This adds to the stiffness of the setup measured in terms of bp, $K(L)\ell_{ss}^2$, see Eq. (12). Two cases can be distinguished. For W-S barriers, both stiffnesses are positive, and the free energy has a unique minimum: we expect the SP approximation, which replaces the average value of n with is typical value n^{SP}, to be accurate. For S-W barriers, the two stiffnesses have opposite signs. There is a unique minimum if ΔG is smaller than $\Delta G_{c.o.} = K(L)(\Delta n\, \ell_{ss})^2$, and two separated minima if $\Delta G > \Delta G_{c.o.}$. We therefore expect the SP inference to be good at inferring W-S-barrier regions in the landscape, and to behave poorly for steep S-W barriers, *i.e.* such that ΔG exceeds the free energy $\Delta G_{c.o.}$. In this section we indeed show that the second derivative of the inferred cumulative free energy landscape, $\frac{d^2G^{sp}}{dn^2}$, is bounded from below by $-K(L)\ell_{ss}^2$, whatever the value of the large and negative second derivative of the true free energy G. To illustrate this statement we consider the following free-energy landscape:

$$\delta g_0(n) = -\Delta G\ \frac{n}{\Delta n^2}\ \exp\left(-\frac{n^2}{2\,\Delta n^2}\right). \tag{22}$$

Parameter Δn controls the width of the barrier. Here δg_0 represents the difference between g_0 and the reference value $2g_{ss}$. ΔG is equal to the extremal value of the cumulative free energy landscape $\delta G(n)$ at the center of the barrier $n = 0$:

$$\delta G(n) = \int_{-\infty}^{n} dn'\, \delta g_0(n') = \Delta G\ \exp\left(-\frac{n^2}{2\,\Delta n^2}\right). \tag{23}$$

The behaviors of the free energy per bp, $g_0(n)$, corresponding to, respectively, W-S ($\Delta G < 0$) and S-W ($\Delta G > 0$) barriers are shown in, respectively, Fig. S5 and Fig. S6.

Given the landscape defined in Eq. (23), and the stiffness constant K (which may depend on L) we calculate the average force $f(L)$ and use the SP inference formula to obtain $n^{SP}(L)$ and $g_0^{SP}(L)$. Results are shown in Figs. S5 and Fig. S6. We find two qualitatively different behaviors:

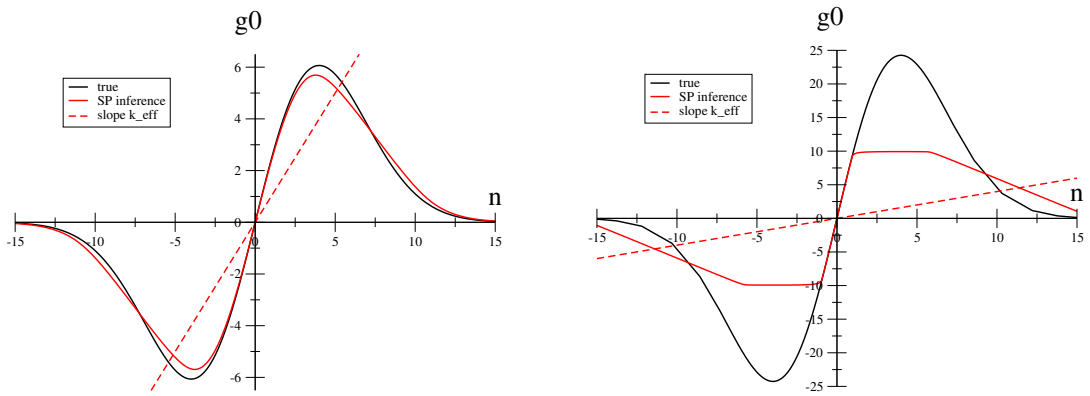

FIG. S5: Examples of W-S barriers. Left: $\Delta G = -10$, Right: $\Delta G = -40$. Other parameter are $\Delta n = 1$, $K_{eff} = K\,\ell^2 = 1$.

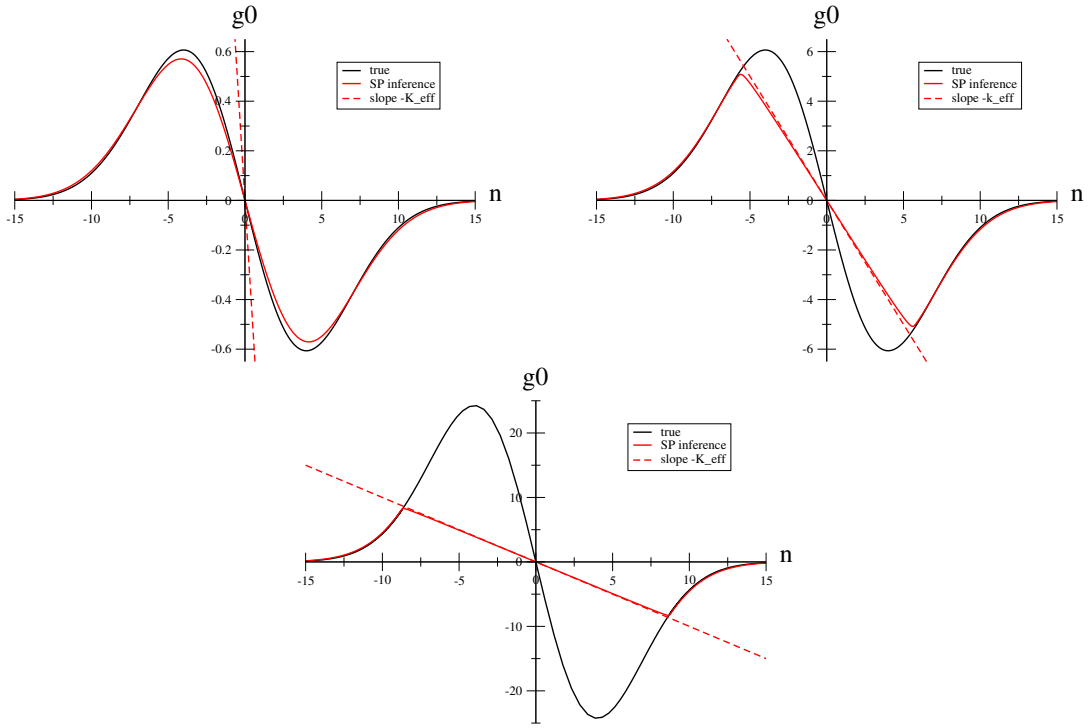

FIG. S6: Examples of S-W barriers. Top Left: $\Delta G = 1$, Top Right: $\Delta G = 10$, Bottom: $\Delta G = 40$. Other parameter are $\Delta n = 1$, $K_{eff} = K\ell^2 = 1$.

- For W-S barriers the inferred free energies are in good agreement in the central part of the barrier whatever the value of (negative) ΔG.

- For S-W barriers the slope of the inferred free energies at the origin is in good agreement with the true slope,

$$\frac{dg_0^{SP}}{dn^{SP}}(n^{SP} = 0) \simeq \frac{dg_0}{dn}(n = 0)\ . \tag{24}$$

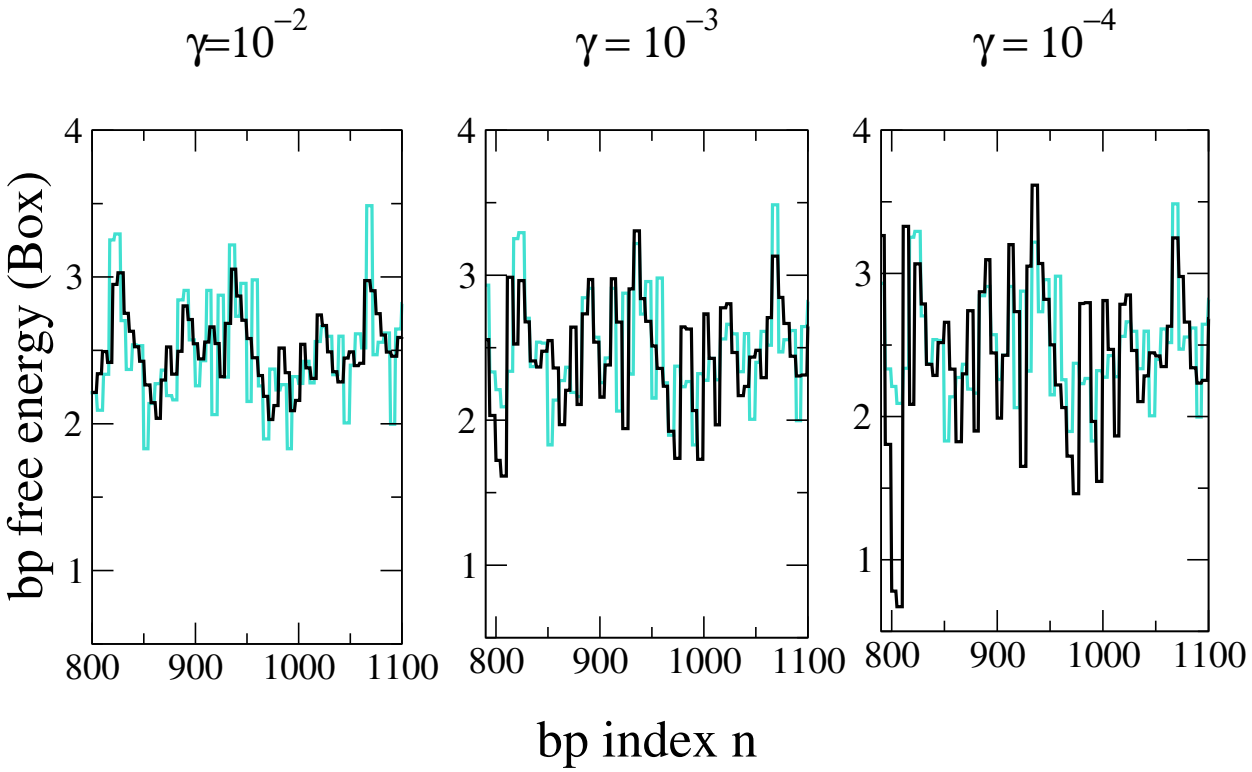

FIG. S7: Base pair free energies at the beginning of Molecule 2 (turquoise: true values, black: outcome of the Box inference procedure with $b = 5$). Left: penalty parameter $\gamma = 10^{-2}$. Middle: penalty parameter $\gamma = 10^{-3}$. Right: penalty parameter $\gamma = 10^{-4}$

for small ΔG only. Conversely the mean slope of the inferred barrier is much smaller (in absolute value) than the true one for large positive ΔG, and saturates to the value

$$\left|\frac{dg_0^{SP}}{dn^{SP}}\right| \simeq K\,\ell_{ss}^2 \ , \tag{25}$$

which depends on the dimensionless effective stiffness only. The cross-over between the two regime corresponds to a barrier height

$$\Delta G_{co} \simeq \Delta n^2\, K\,\ell_{ss}^2 \ . \tag{26}$$

IV. CHOICE OF THE PARAMETERS IN THE BOX APPROXIMATION

A. Penalty parameter

The Box approximation consists in maximizing the log-likelihood of the experimentally measured forces $f_{exp}(L_k)$ fo a set of positions L_k, see Material and Methods Section,

$$\log P(\{f_{exp}(L)\}|\{g_k\}) = -\frac{1}{2\epsilon^2}\sum_{k=0}^{N/b-1}\left(f_{exp}(L_k) - \langle f\rangle^{Box}(L_k)\right)^2 - \frac{1}{2\Delta^2}\sum_{k=0}^{N/b-1}(g_k-\bar{g})^2 \ , \qquad (27)$$

over the box free energies g_k. Given the experimental forces the outcome depends only on the dimensionless penalty parameter

$$\gamma = \left(\frac{\epsilon\,\ell_{ss}}{\Delta}\right)^2 \ , \qquad (28)$$

which is the squared ratio of the uncertainty over the work of the unzipping force and of the possible deviations of the free energy parameters around their mean $\bar{g}$. The inference is the result of a compromise between the reproduction of the force data (favored for small γ) and the pinning of the g_k around the average value $\bar{g}$ due to the prior probability (favored by large γ). Given the orders of magnitude of the uncertainty over the force, $\epsilon \sim 0.1$ pN, and of the fluctuations of g_k around $\bar{g}$, $\Delta \sim 1\ k_BT$, we expect γ to be comprised in the range $10^{-4} - 10^{-3}$.

In Fig. S7 we show the inferred free energy landscape with a penalty parameter $\gamma = 10^{-2}$, compared to the one inferred with $\gamma = 10^{-3}$ and $\gamma = 10^{-4}$. For most locations in the sequence the precise value of the penalty parameter does not have a large impact. For some bp, however, *e.g.* around $n = 800$, the regularization is helpful to prevent divergences in the inferred free energies, which very weakly affect the equilibrium value of the force, and are underconstrained by the data alone. In practice we find that $\gamma = 10^{-2}$ gives good predictions for the sequence free energies, when compared to the true values averaged over $w = 30$ bp. This value can be reduced to $10^{-3} - 10^{-4}$ when reconstructing the free energies at a better resolution (smaller scale) than 30 bp.

B. Width of the trial box

In Fig. S8 (right) we show the inference of the end of the sequence with a box-like trial function,

$$G_{ds}^{Box}(n) = b \sum_{k=0}^{\text{integer part of } n/b} g_k \ , \qquad (29)$$

with $b = 5$ bp; this value for b corresponds to roughly half the standard deviation of the nb of open bp at equilibrium, resulting from the ssDNA fluctuations with ≈ 6000 open bp. The inference is

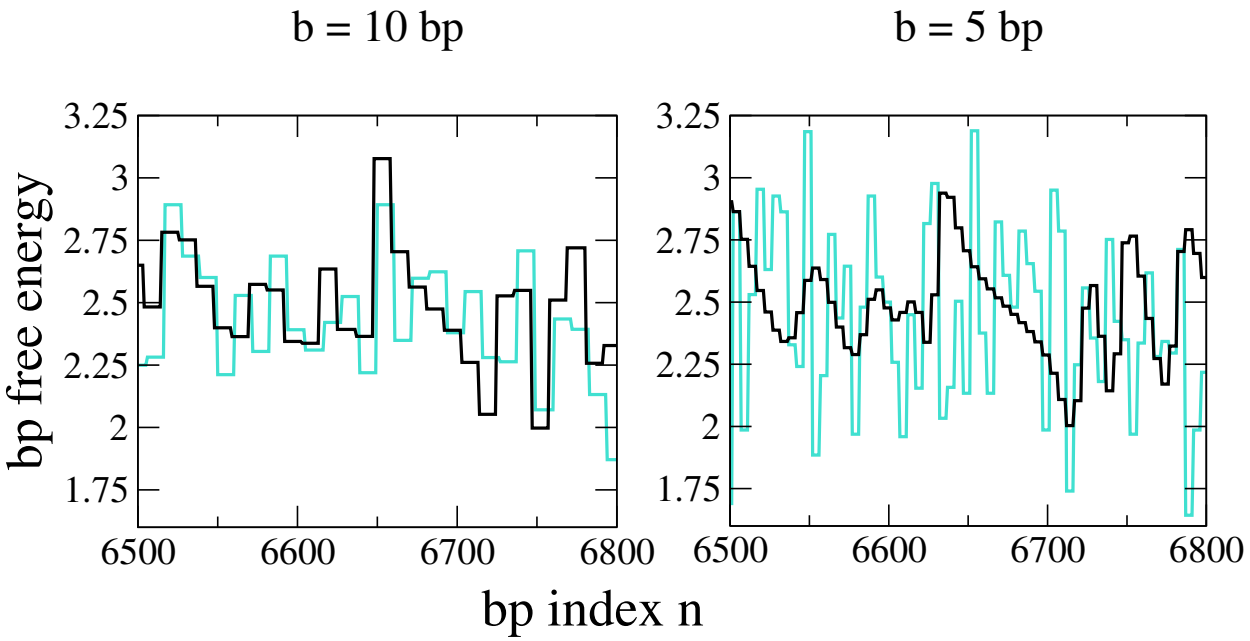

FIG. S8: Inference of the base pair energy at the end of Molecule 1. Left: trial function constant over $b = 10$ bp (black), comparison with the box average of the true free energy over 10 bp (turquoise). Right: trial function constant over $b = 5$ bp (black), comparison with the box average of the true free energy over 5 bases (turquoise).

compared to the one corresponding to $b = 10$ (Fig. S8, left panel). While the inference for $b = 5$ is twice more costly in terms of the number of parameters to infer, the inferred free energies are very similar to the ones obtained with a box trial functions over $b = 10$ bp. As expected it is useless to choose values for b smaller than half the standard deviation $b_B(L)$ defined in (15), see Fig S1.

C. Description of the optimization procedure

To find the local free energy parameter we minimize the difference between the experimental and theoretical forces with a regularization term as described in the Eq. (11) of the main text. We have implemented this minimization procedure in Mathematica 7. As running the minimization procedure on all 6800 base pairs is too slow we have defined unzipping zones of about 1200 base pairs, which overlap two by two over 100 base long regions. The 50 predicted bases at the beginning and at the end of each region are then discarded. This procedure is possible since the setup acts as a confining potential over the number n of open bp, and a base does not affect the average force to open a bp more than 100 base pairs away. In addition we introduce a cut-off in the sum over

n in Eq. (10) to estimate the average force; this cut-off limits the summation over a few hundred base pairs around the value $(L/\ell_{ss} - n)$, and is justified by the fact that the standard deviation of the number of open bases around this average value is of a few ten of bases at most. The small inference error on the synthetic data sets shows that this cut–and–paste procedure does not affect much the inference error along the sequence; However it could probably be improved by choosing carefully where to cut the data from the unzipping signal. The computation over 1200 base takes about 1 hour (for the values of b reported here) on an Intel Core 2 processor.

D. Theoretical unzippping forces from the inferred free-energy landscapes with the SP and Box approximation

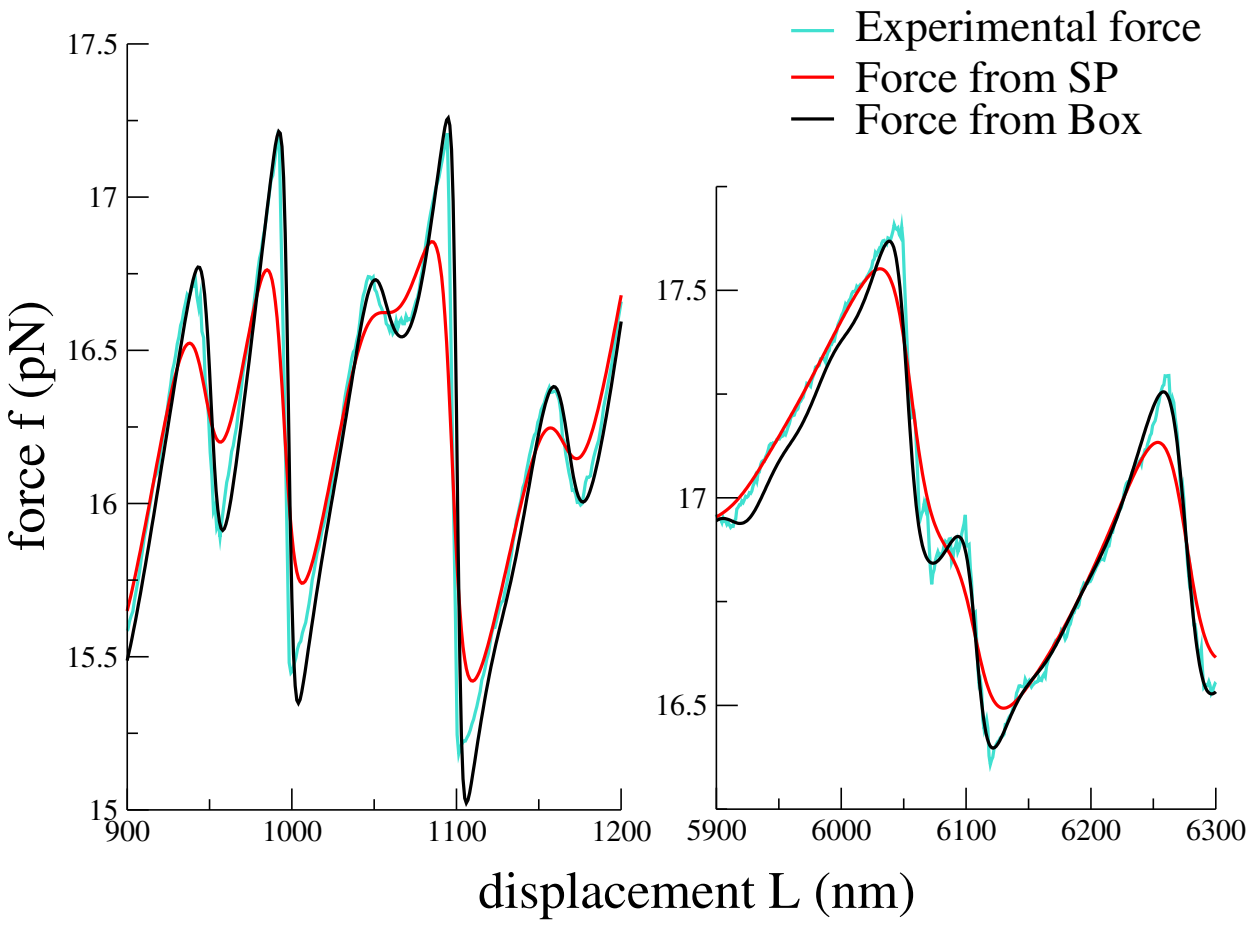

FIG. S9: Average force vs. trap position: Experimental values $f_{exp}(L)$ for Molecule 1 are shown in turquoise, while the equilibrium force $\langle f \rangle(L)$ computed the bp free energies inferred with the SP and Box approximations are shown in, respectively, red and black.

In Fig. S9 we show the equilibrium unzipping force obtained from Eq. (4) in the main text from the free-energy landscapes G inferred with the Box and SP inference. We observe that the unzipping force obtained from the Box approximation reproduces very closely the experimental force data, especially in stick-slip regions where the SP force show strong deviations.

V. RECONSTRUCTION ERRORS WITH RESPECT TO THE TRUE FREE-ENERGY LANDSCAPE

A. Definitions of the reconstruction errors

We denote by

$$g_0^{(w)}(n) = \frac{1}{w} \sum_{i=n-w/2}^{n+w/2} g_0(s_i, s_{i+1}) \tag{30}$$

the sliding average of the true bp free energy over a window of w bp. We want to estimate the error between the bp free energy $g(n)$ inferred with the SP or the Box approximations and $g_0^{(w)}(n)$. Of particular interest is the scale w which minimizes this difference, that is, on which the sequence free energies are better inferred.

The error

$$\epsilon_w(n_1, n_2) = \sqrt{\frac{1}{n_2 - n_1} \sum_{n_1 \leq n < n_2} \left(g(n) - g_0^{(w)}(n) \right)^2} \tag{31}$$

estimates the absolute discrepancy between g and $g_0^{(w)}$ in the portion of the sequence comprised between bp n_1 and n_2. To obtain a relative measure of this discrepancy we introduce the relative error

$$\rho_w(n_1, n_2) = \frac{\epsilon_w(n_1, n_2)}{\delta g_0(w)} \ , \tag{32}$$

where the denominator

$$\delta g_0(w) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\bar{g}_0 - g_0^{(w)}(i))^2} \quad \text{with} \quad \bar{g}_0 = \frac{1}{N} \sum_{i=1}^{N} g_0(s_i, s_{i+1}) \tag{33}$$

represents the characteristic fluctuations of the true free energy landscape (averaged on a window of w bp) with respect to its mean value. When w increases these fluctuations decrease because the sliding average includes more and more bases.

B. Dependence of the reconstruction error on the local landscape

In Fig. S10 we plot the position-dependent inference error, $\epsilon_w(n-50, n+50)$ defined in Eq. (31), averaged over one hundred bp and with a sliding average of the true sequence over $w = 30$ and $w = 50$ base pairs. We observe that the error fluctuates along the sequence, with no monotonic dependence on the number n of unzipped base pairs as would be expected from the increase of ssDNA fluctuations with n. The error mainly depends on the local free energy landscape and on the heights and the widths of barriers therein. The DNA molecule which has been unzipped is characterized by large variations in the free energy landscape especially at the beginning of the sequence, which make the error generally larger at the beginning than at the end of the sequence.

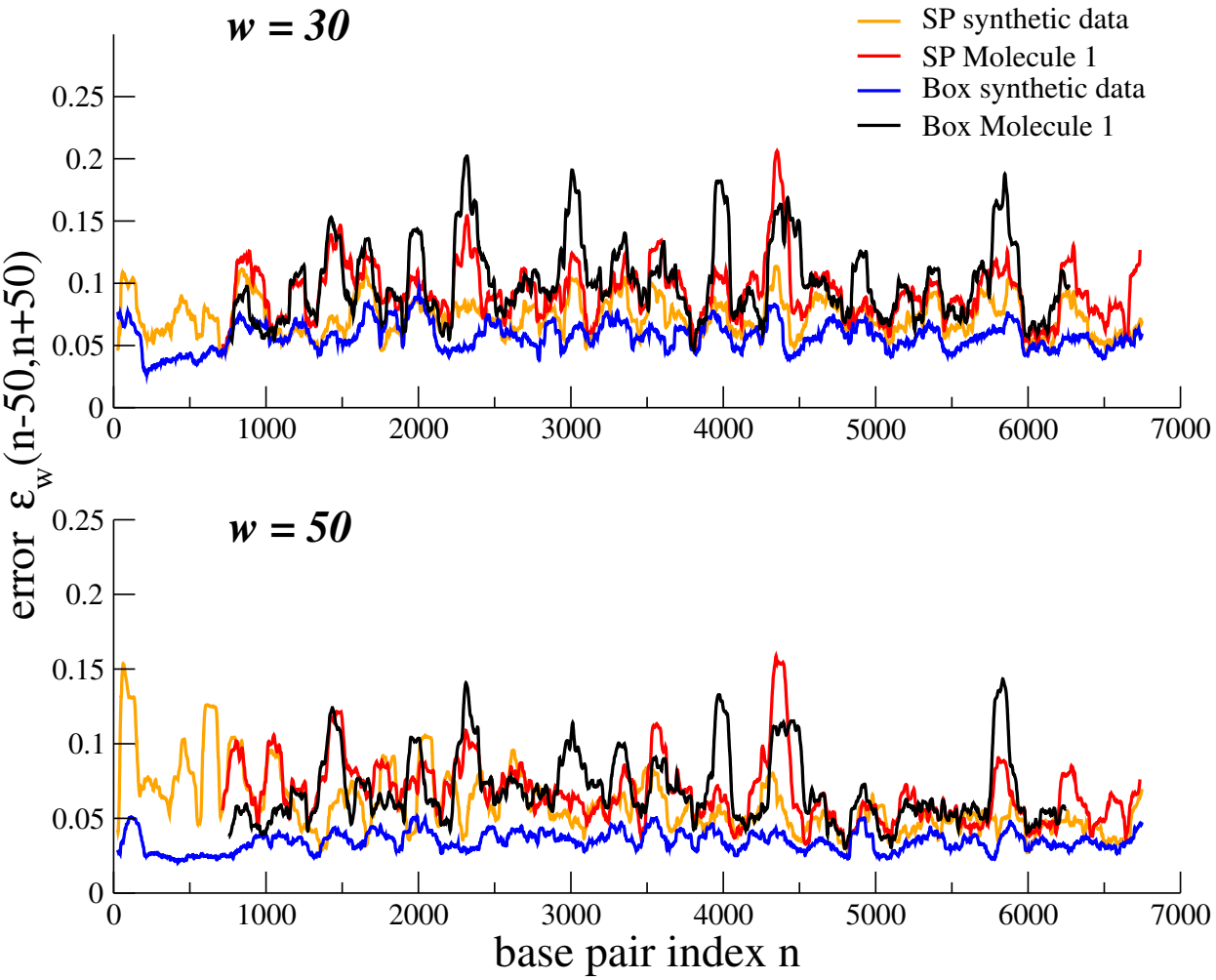

FIG. S10: Position-dependent errors in the inference. Error $\epsilon_w(n-50, n+50)$ on the inferred bp free energy with respect to the true value, averaged on a window $w = 30$ (top) and $w = 50$ (bottom), vs. bp index n.

As discussed in Section III the performance of the SP procedure strongly depends on the types

of the barriers in the free energy landscape, and on the stiffness of the experimental apparatus. In particular, as seen in Fig. 2 of the main text (bottom), the agreement between the true and inferred sequence free energies is worse on descending flanks (corresponding to S-W regions in the landscape) than on negative flanks (corresponding to W-S regions). We have checked that, in agreement with theoretical prediction, the SP inference is not capable of reproducing steep positive barriers in the free energy landscape. The asymmetry between W-S and S-W regions of the sequence landscape is visible from the analysis of the experimental force signal. Figure S11 shows the sliding average of $\frac{dg^{SP}}{dn^{SP}}$ over a 10-bp window as a function of n^{SP}. We see that this quantity, which coincides with the second derivative of G^{SP}, is bounded from below by minus the effective stiffness of the setup (lower blue curve), but is not bounded from above. Since the effective stiffness decreases with the number of open base pairs the limitation becomes stronger at large n.

As visible in Fig. 3 of the main text (bottom) the box approximation overcomes this limitation and is better able to reproduce S-W barriers (stick-slip regions) in the free energy landscape. We see in Fig. S11 that the second derivative of G^{Box} is not bounded from below by the stiffness of the setup, and that the agreement with the second derivative of the true free energy G is much better than in the case of SP inference. The Box approximation is therefore less sensitive to the curvature of the free energy landscape than the SP approximation. This statement is corroborated by the inference of synthetic data (Fig. S10): the error with the Box approximation shows less pronounced variations along the sequence than with the SP approximation.

To better understand how the reconstruction scale is affected by the drift of the apparatus we consider the inference error in the case of synthetic data, which are free of any drift effect. We find that the sequence free energies are in better agreement with their true counterparts than the free energies inferred from Molecule 1. In particular the bumps in the error in the region between base pairs 1300 and 1600 and around bp 2300 in Fig. S10 are not seen with the synthetic data, and seem to be due to a residual drift in the Molecule 1 data.

C. Characteristic reconstruction scale

To determine the scale on which the free energy landscape is better reproduced we plot in Fig. S12 the error ϵ_w over the whole sequence and over two portions of length 300 bp, located at the beginning and the end of the unzipping data. The SP inference error reaches a minimum at a scale of 30-40 bp at the beginning and at about 70 bp at the end of the sequence. The Box error is minimal when the sliding average of the true free energy landscape is compared to the sliding

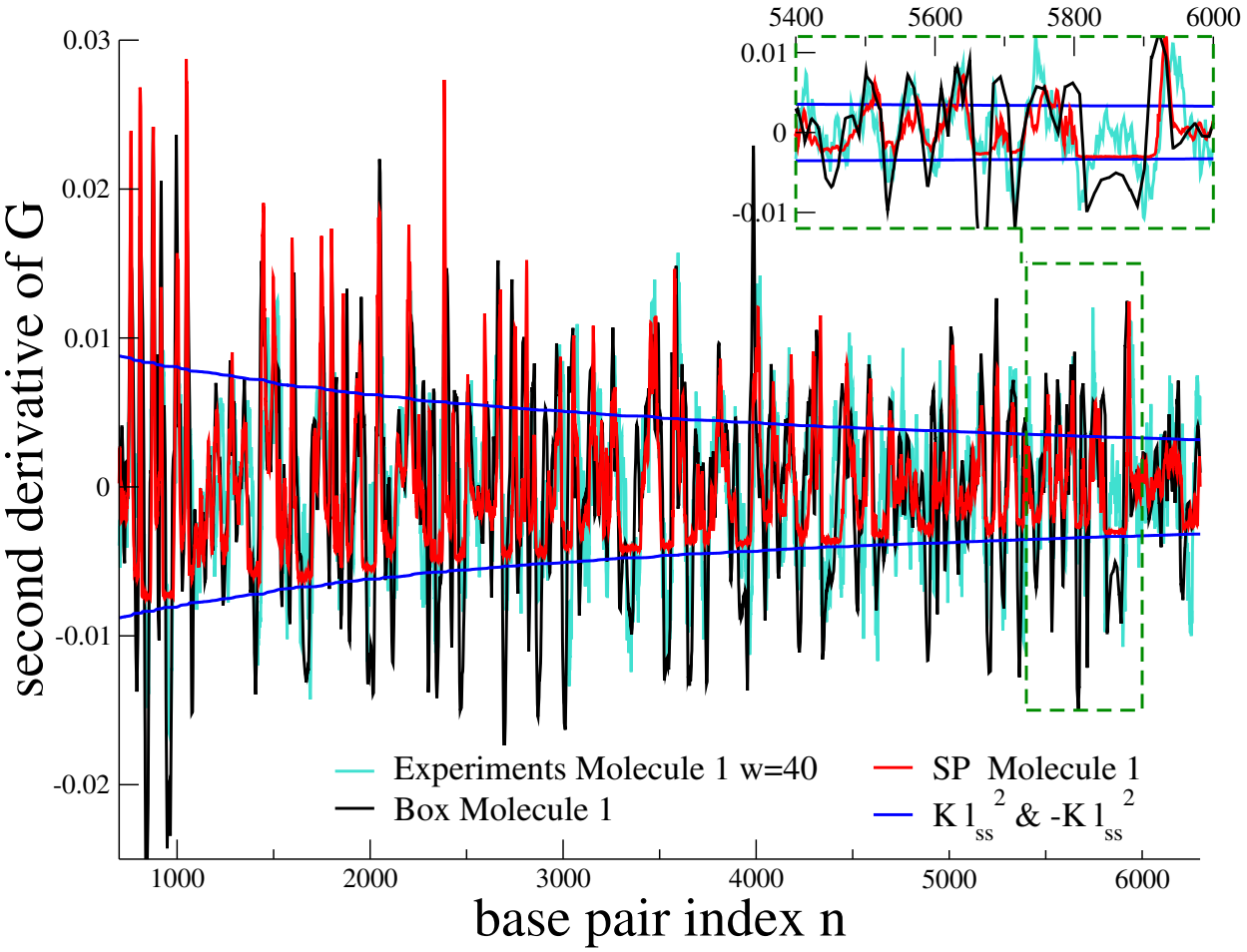

FIG. S11: Second derivative of the inferred cumulative free-energy landscape $G(n)$ (convoluted over 10 bp) plotted as a function of the bp index n for the SP (red curve) and Box (black curve) approximations, and for the experimental data (turquoise line, $w = 40$). Blue lines show the effective stiffness $K\ell_{ss}^2$ (upper curve) and its opposite (lower curve) in units k_BT. Inset: magnification of the $5400 < n < 6000$ region.

average of the inferred free energy landscape on the same window size, $w' = w$. The minimal error on the sliding-averaged free energies is of the order of 0.05 k_BT to 0.1 k_BT, depending on the approximation and on the location along the sequence. The relative error ρ_w ranges from 0.25 to 0.5 for scales ranging from 30 to 50 bp, meaning that the inference error ranges between one quarter and one half the standard deviation of the free energy g_0 around its mean along the sequence.

The reconstruction scale depends on the effective stiffness of the setup, which is larger at the beginning of the opening. As we do not have the force signal at the beginning of the unzipping in Molecules 1 and 2 we have resorted to the synthetic data to characterize this dependence. Figure S10 shows that the inference error in the region $n < 1000$ is smaller with the Box approximation than with the SP approximation; this effect is also visible on the few hundreds of bases in this

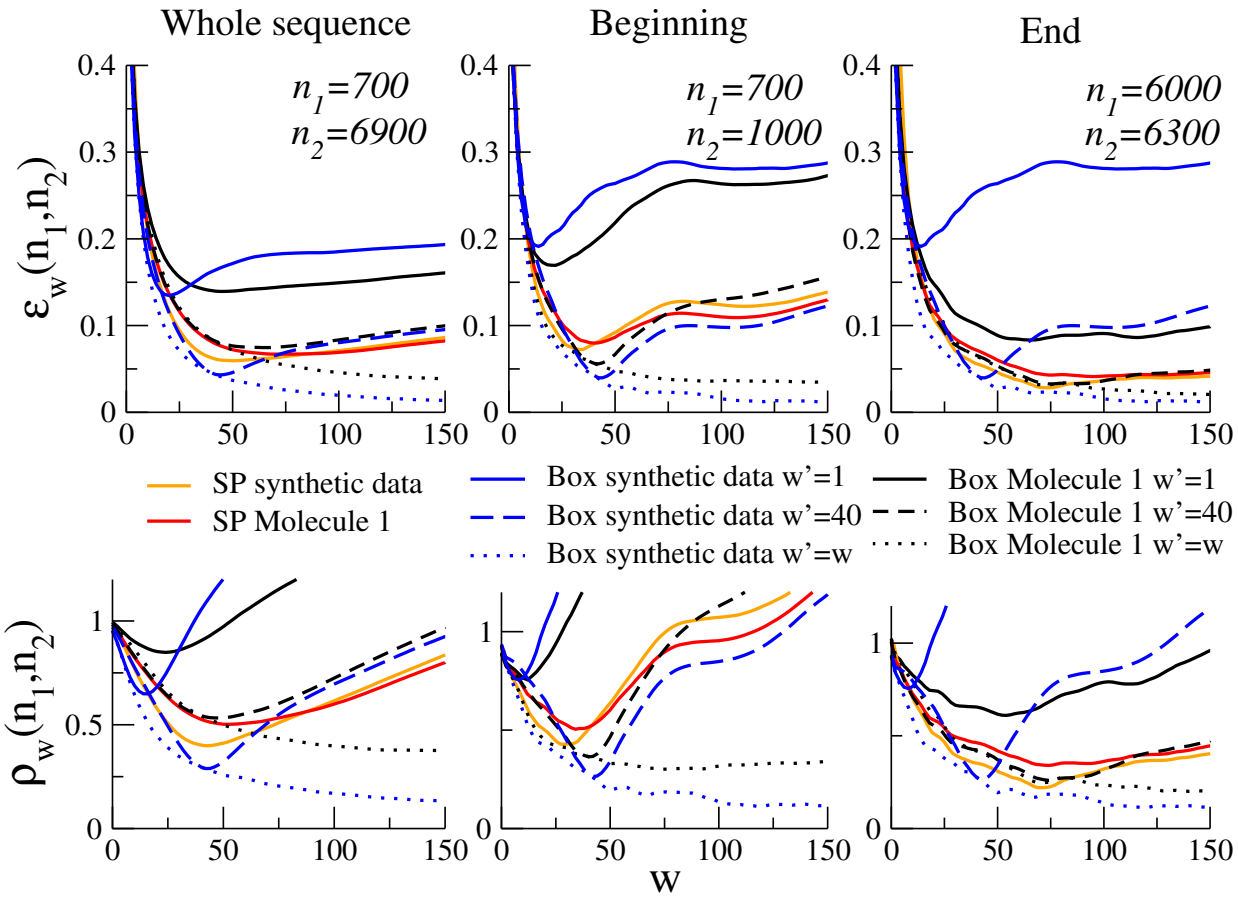

FIG. S12: Average errors in the inference: Average error over the inferred free energy ϵ_w (Top) and averaged relative error ρ_w (Bottom) as a function of the window size w of the running average on the true free energy landscape, and for different window size w' of the running average on the inferred landscape for Molecule 1 and synthetic data. Left: whole sequence, Middle: 300 bases at the beginning, Right: 300 bases at the end of the sequence.

region in the Molecule 1 data.

We show in Fig. S13 the inference of the free energies for the first 500 bases. The fluctuations of the displacement at the beginning of the sequence are $b = b_B/2 \simeq 4$ bp (see Fig. S1). As is shown in the figure, the free energies accurately inferred over a scale of about 20 base pairs. The inference error depends on the local features of the free energy landscape. It is very small in a favorable region $150 < n < 500$, and is larger at the beginning $n < 150$, where the free energy landscape has steeper variations.

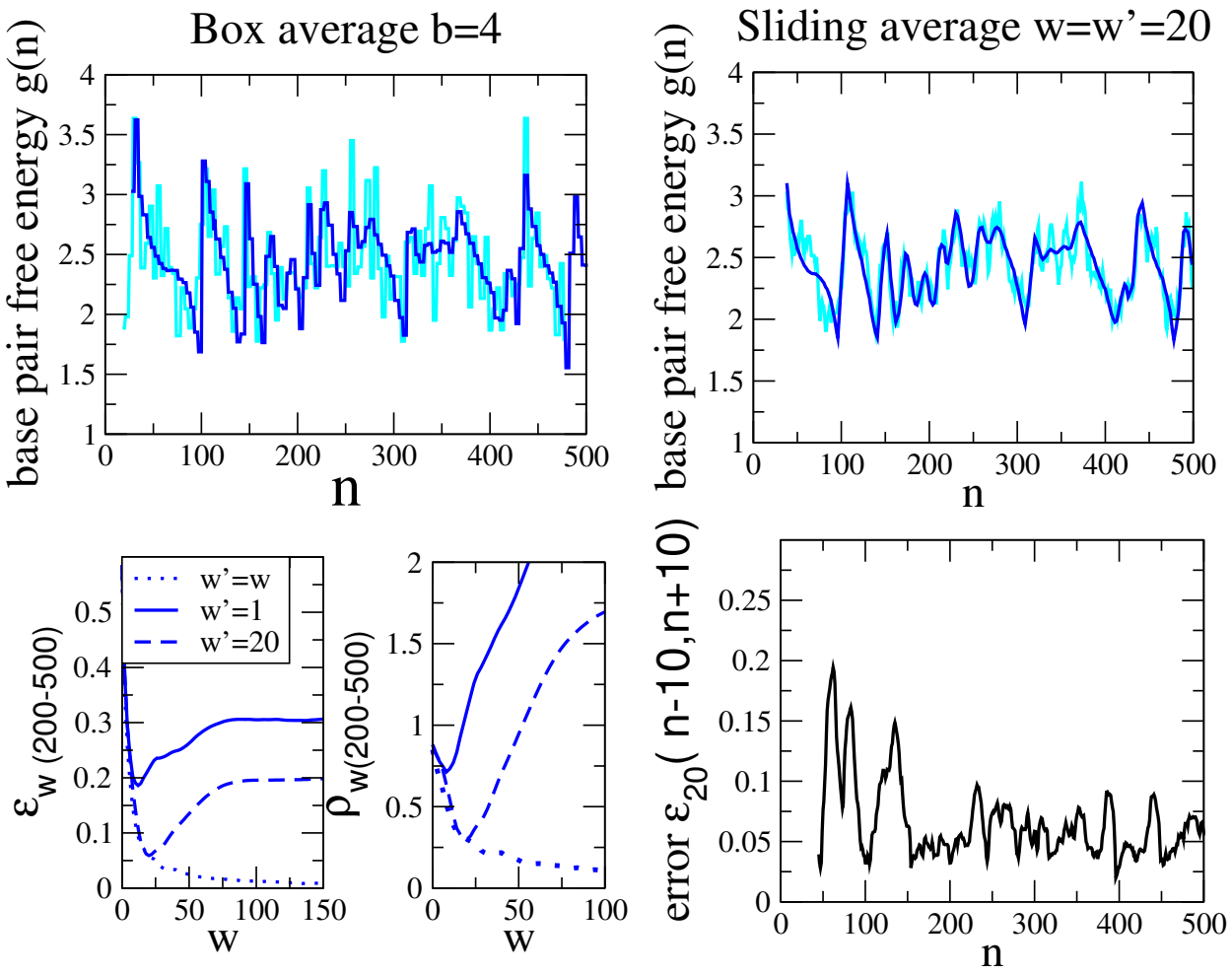

FIG. S13: Inference from the unzipping force for synthetic data. Top: Left: Free energy inferred for the first 500 base pairs (blue) compared with a box average of the true sequence with the same b (turquoise). Right: sliding average of the free energy over $w' = 20$ bp (turquoise) compared to the sliding average of the true sequence with $w = 20$ (blue). Bottom: Left: error ϵ_w and relative error ρ_w vs. window sizes w and w' for the sliding averages of, respectively, the true sequence and the inferred sequence. The error is $\epsilon_{20} = 0.05$ and $\rho_{20} = 0.3$ for $w' = w = 20$. Right: error $\epsilon_{20}(n-10, n+10)$ calculated as a sliding average over 20 bases along the sequence is shown for $w' = w = 20$.

D. Inference of the free-energy landscape for Molecule 2

We show in Fig. S14 the outcome of the SP and Box procedures to infer the sequence free energy of Molecule 2, using the best pairing parameters computed in [1]. The quality of the predictions, also shown in the position dependent error of Fig. S15 and in the average error of Fig. S16, is comparable to the ones for Molecule 1, shown in Fig. 2 and 3 of the main text.

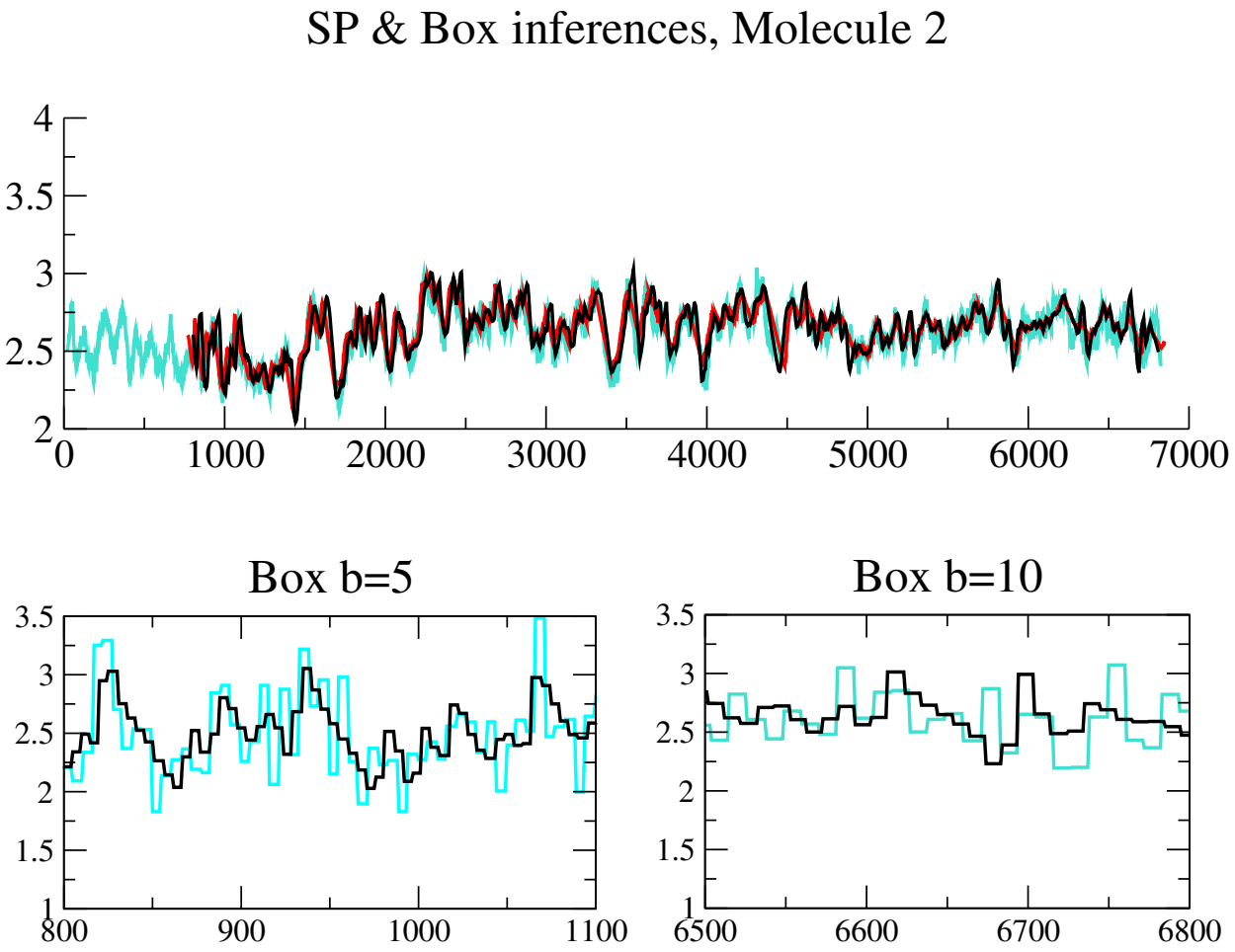

FIG. S14: Inference of the base pair energy from the unzipping force for molecule 2. Top: comparison of the SP (red line) and the Box (black line) landscapes with the true free energy (sliding average on 50 bp). Bottom: Box inference for the first (Left, $b = 5$) and the last (Right, $b = 10$) base pairs in the sequence.

E. Comparation of SP and Box inferences for Molecule 1 and Molecule 2 on the whole molecule

Figure S17 shows a magnification of the inferred free energies from Molecule 1 data with the SP approximation (see Fig. 2 of the main text), and with the Box approximation (see Fig. 3 of the main text), to allow for a better comparison with the true free energies along the whole molecule. In Fig. S18 the free energies inferred for the data of Molecule 2 with the SP approximation (see Fig. 2 of main text) and with the Box approximation are shown.

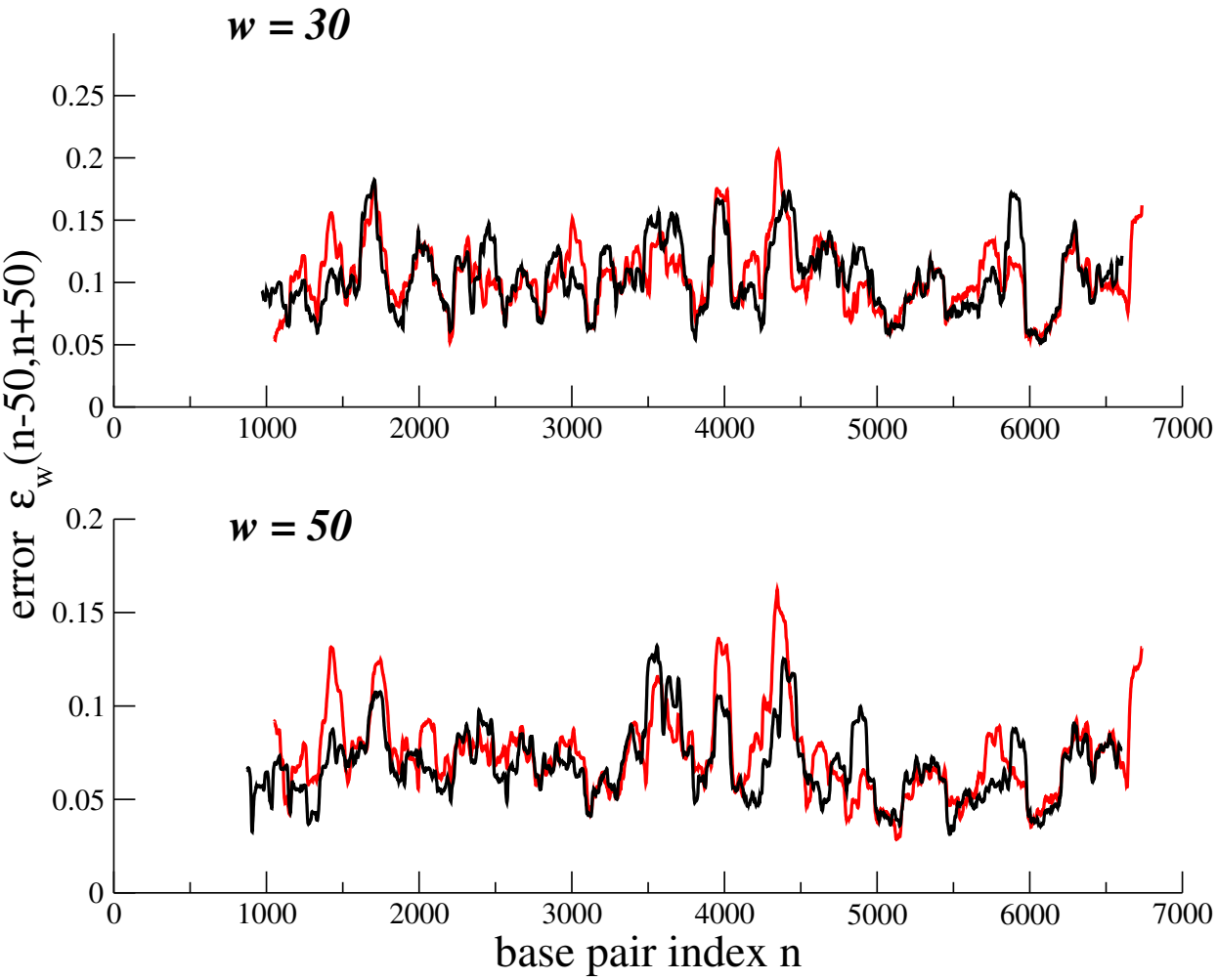

FIG. S15: Position-dependent errors in the inference for Molecule 2. Error $\epsilon_w(n-50, n+50)$ on the inferred bp free energy with respect to the true value, averaged on a window $w = 30$ (top) and $w = 50$ (bottom), vs. bp index n.

VI. REALIGNMENT OF THE UNZIPPING FORCE CURVES AND COMPARISON WITH MFOLD PAIRING ENERGY AT 1M

The best pairing parameters fitted in [1] correspond to a global shift of the free-energy landscape with respect to the MFold predictions, as shown in Fig. S19. This shift can be compensated by a global offset δf (which takes different values for Molecules 1 and 2) over the unzipping force, possibly due to the experimental uncertainty on the force. To estimate this offset we have calculated the average value of the inferred free energies over the sequence for the two molecules, and calculated the difference, denoted by δg, with the average free energy along the true sequence according to MFold. We have then translated the force curve by a global shift in the force, $\delta f = \delta g/(2\,\ell_{ss})$. We

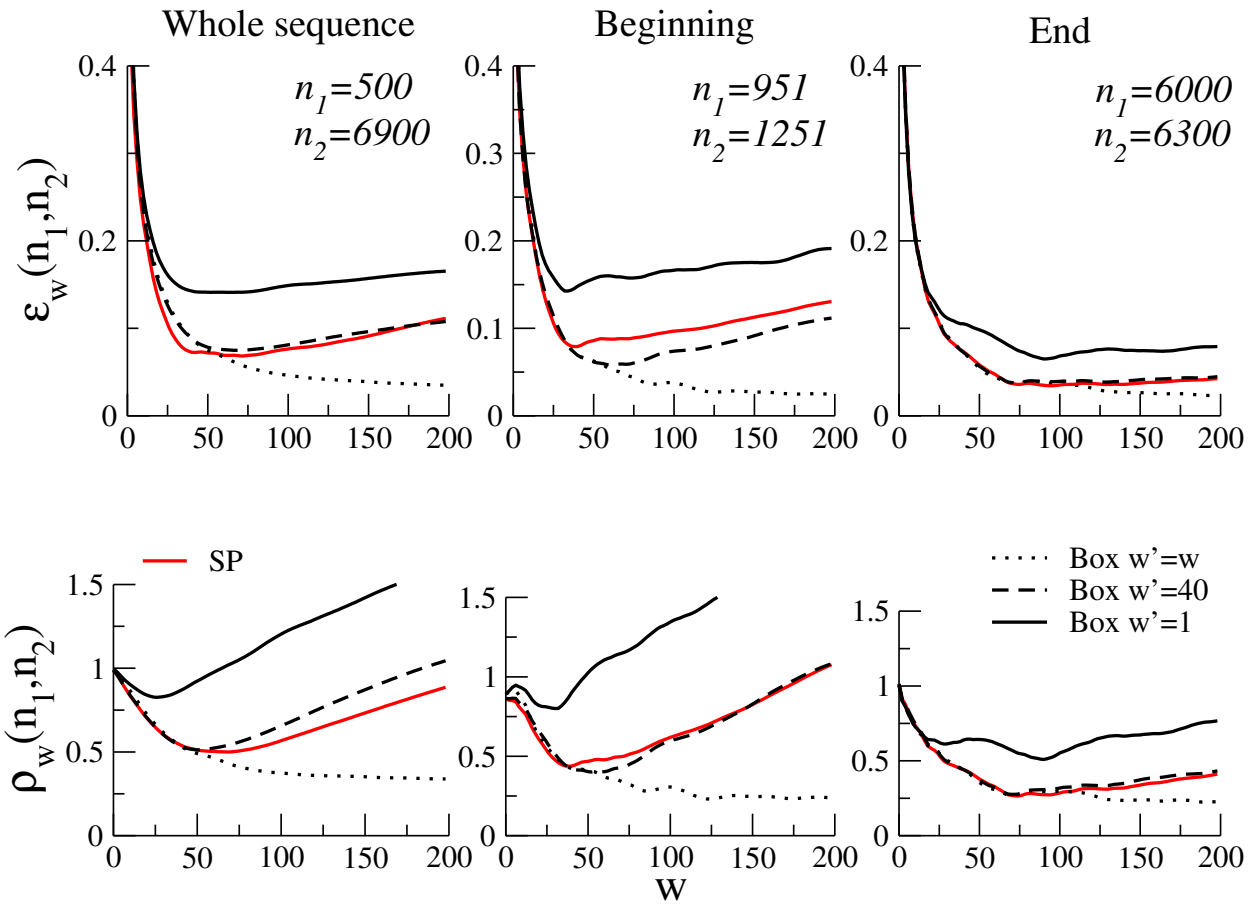

FIG. S16: Average error of the inferred free energy for molecule 2. Top: average error ϵ_w; Bottom: averaged relative error ρ_w, as functions of the window size w of the running average, and for different size w' of the running average on the Box inferred free energy. The errors are computed over the whole sequence (left), and for 300 bases at the beginning (middle) and at the end (right) of the sequence; SP approximation: red line, Box approximation: black line.

obtain $\delta f = 1.2$ pN for Molecule 1, and $\delta f = 0.7$ pN for Molecule 2.

For the alignment of the force signals along the L-axis, we have followed two procedures:

- a very simple and minimal shift done by 'hand', consisting in a displacement shift for Molecule 1 by $\delta L = 30$ nm if $n < 1500$, and $\delta L = 50$ nm if $n > 1500$, and a displacement shift of Molecule 2 by $\delta L = 20$ nm;
- a more sophisticated realignment with the Needleman-Wunsch algorithm [6] described in main text (Methods Section).

The force signals obtained with both alignment procedures are shown in Fig. S20.

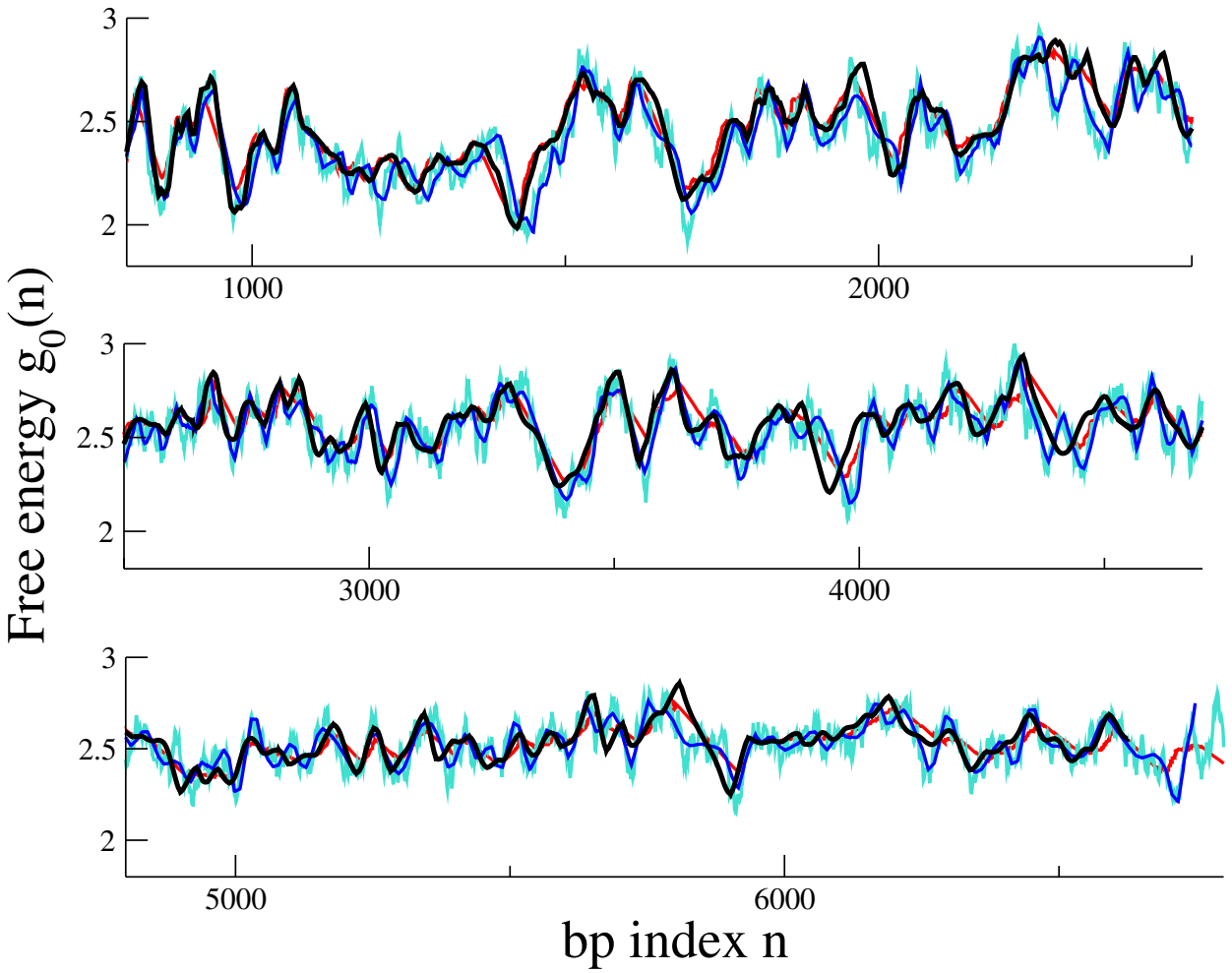

FIG. S17: Inference of the base pair energy from the unzipping force for Molecule 1. Comparison of the SP (red line) and the Box (black line) landscapes with the true free energy (turquoise) (sliding average over 40 bp) and the Box approximation for synthetic data (blue).

After these realignments we have inferred the free-energy landscapes shown in Fig. S21 and S22. Even if local errors in the alignments are still present the agreement with the free-energy landscape obtained with MFold is remarkable (and obtained with no fitting of parameters). Residual alignment errors can be observed, e.g. around position $n = 1500$ with the Needleman-Wunsch procedure (which could be cured by lowering the force increment $\simeq 0.2$ pN used to discretize the force signal prior to alignment) and around position $n = 1700$ with the 'hand-made' alignment.

Figure S21 and Fig. S23 show that the errors for the SP inferred free energy landscapes, after manual realignment, are similar to the one obtained with the 'best' fitted free energies. The experimental drift is by far the major problem in the analysis of unzipping forces. As we dispose here of two unzipping curves only, drift problems cannot be completely solved by aligning these

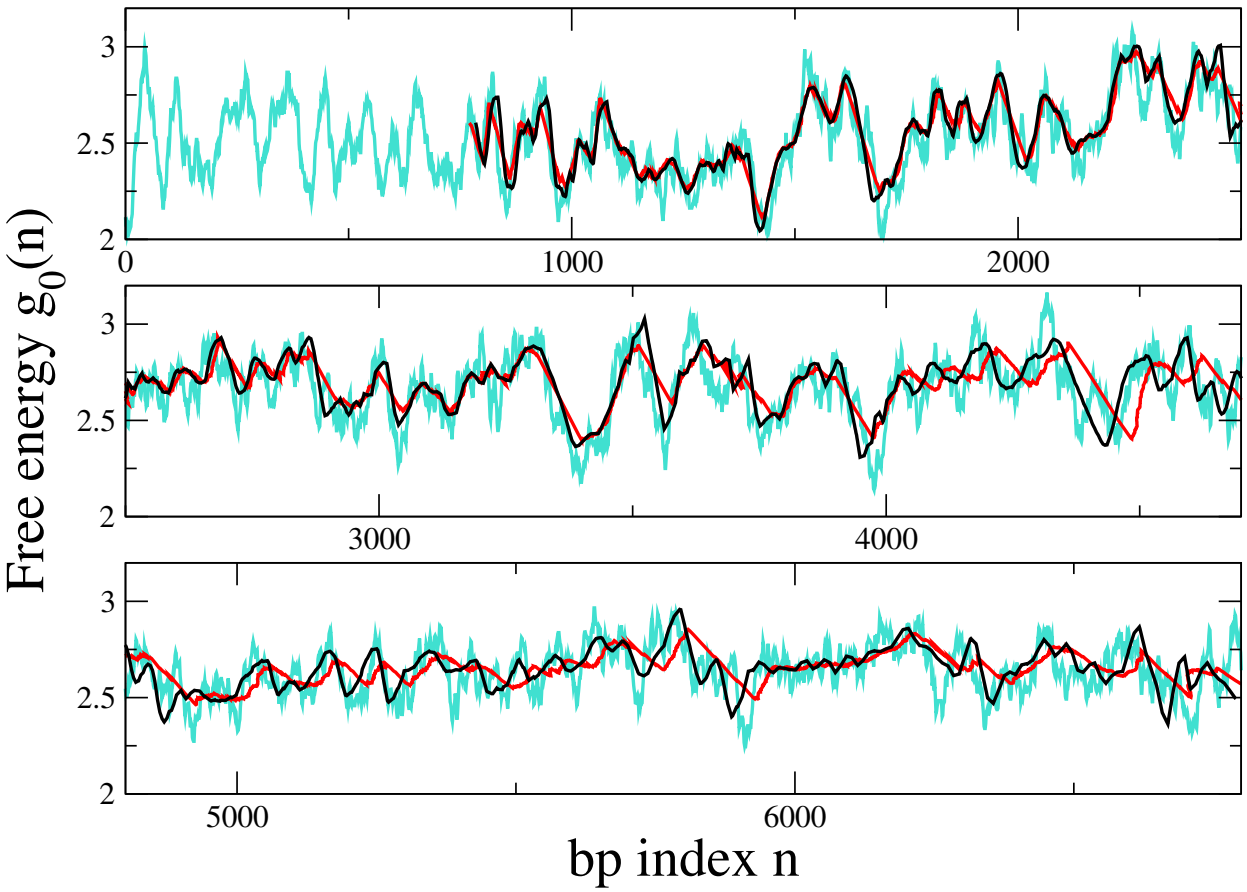

FIG. S18: Inference of the base pair energy from the unzipping force for Molecule 2. Comparison of the SP (red line) and the Box (black line) landscapes with the true free energy (sliding average over 40 bp.)

two curves. Alignment of multiple unzipping curves would be very useful to further decrease the effects of drift.

As explained in the main text the Needleman-Wunsch algorithm is used to align the force signals after discretization of the force values in N_f values. In Fig. S24 the parameter used in the main text, $N_f = 22$, $\sigma^2 = 5$ and gap penalty $S_{gap} = -20$ (middle panel) are compared to $N_f = 22$, $\sigma^2 = 0.25$ and gap penalty $S_{gap} = -100$ (top panel), and $N_f = 4$, $\sigma^2 = 0.1$ and gap penalty $S_{gap} = -20$. The two choices of parameters with $N_f = 22$ give almost identical aligned forces. The alignment with $N_f = 4$ with a smaller resolution on the force is quite similar, even if slightly worse, especially at the end of the unzipping curve.

In Fig. S25 we show the total difference Δg in the inferred free energies between the B-F bacterium and all the other sequences in the database compared to the number of mismatches, with a force alignment based on forces discretization on $N_f = 4$ intervals. Results are very similar

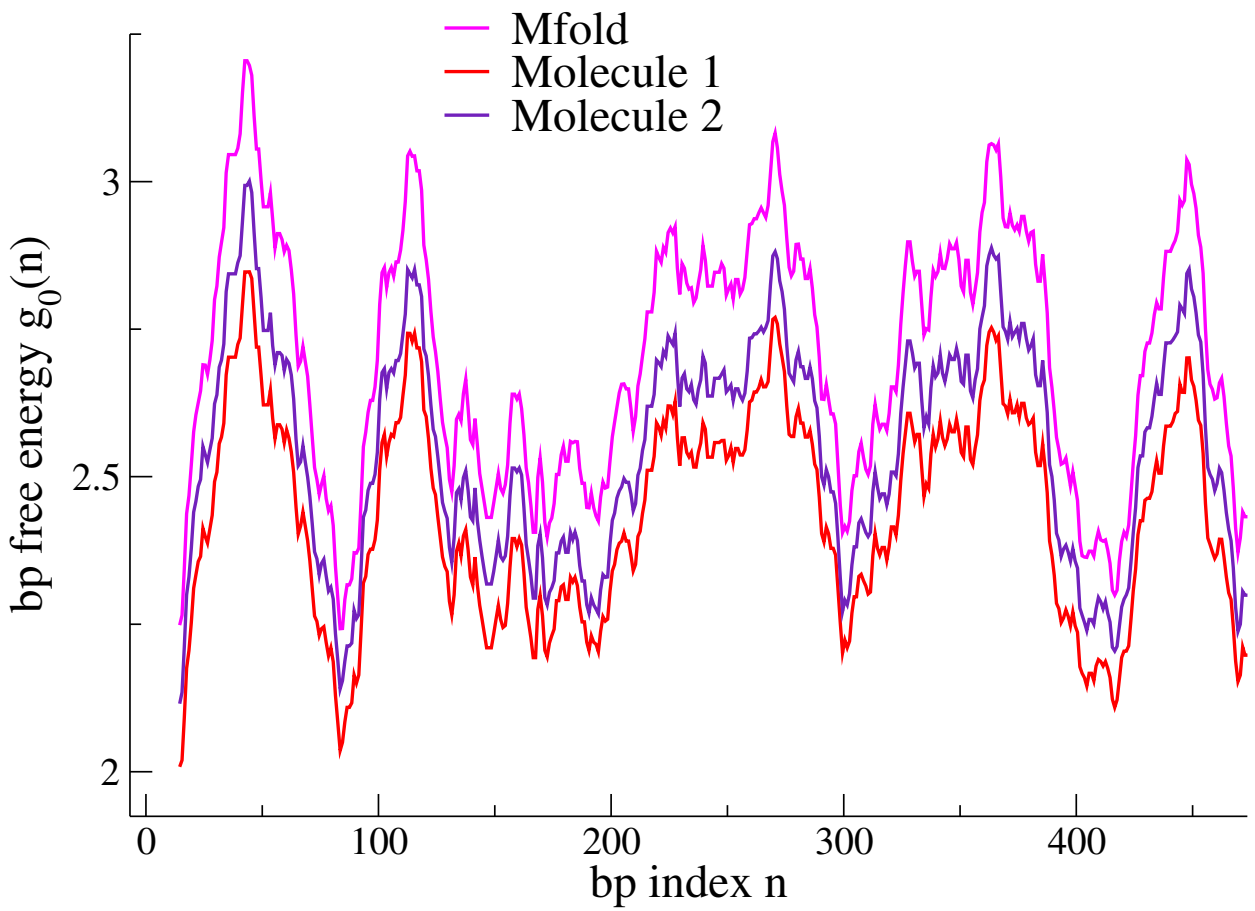

FIG. S19: Base pair free energies along the initial portion of the sequence for the optimal pairing parameters for Molecule 1 and Molecule 2 (extracted from [1]) and for the MFold energetic parameters. The sliding average is computed over $w = 30$ bp.

to what is obtained with $N_f = 22$ force increments, see Fig. 7 of the main text.

VII. SYNTHETIC UNZIPPING FORCE SIGNALS FOR BACTERIA N-A, B-F, B-H AND B-S, AND WHOLE-DATABASE SCREENING

A. Inference of B-F free-energy landscape: comparison between $K_{trap} = 0.08$ and $K_{trap} = 0.3$ pN/nm

Fig. S26 shows the free-energy landscape, for the first 200 bp, of bacterium B-F inferred from synthetic data obtained with a trap stiffness $K_{trap} = 0.08$ pN/nm used in [1] (left) and with trap stiffness $K_{trap} = 0.3$ pN/nm (right). Landscapes are inferred with the SP (top) and the Box (bottom) procedures. We see that on the first 200 base pairs the SP approximation reproduces, for

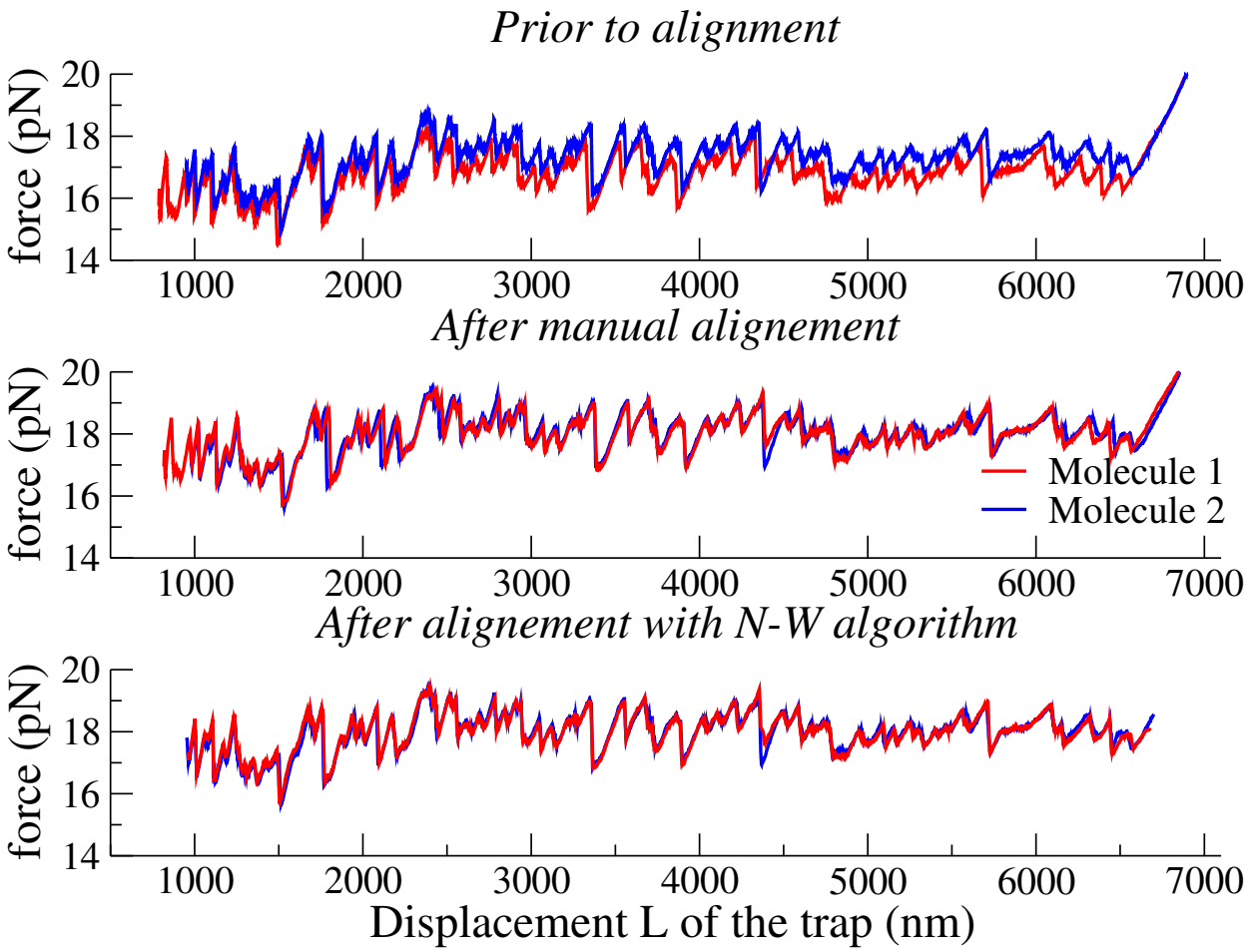

FIG. S20: Top: Experimental force vs. trap displacement for the two unzipping experiments Molecule 1 and Molecule 2, with the corrections to remove the drift done in [1]. Middle: Same data after the manual alignment described in Section 6. Bottom: Same data after alignment with the Needleman-Wunsch (N-W) algorithm, see Section 6.

the first 200 base pairs, the free-energy landscape on a scale of 30 bp, and the Box approximation on a scale of 10 bp for $K_{trap} = 0.08$ pN/nm. With a stiffer trap, $K_{trap} = 0.3$ pN/nm, the SP approximation reproduces the free energy landscape on a scale of 10 bp, and the Box approach on a scale of $\simeq 2$ bp.

B. Force alignments to compare 16S genes in test (B-F) and reference (N-A, B-H, B-S) sequences

In Fig. S27 (left) we show the theoretical force signals for the four 16S genes of the bacteria N-A, B-F, B-H, and B-S, before and after the alignment with the Needleman-Wunsch algorithm

described in the main text and in Section VI. This alignment allows to infer well aligned free-energy landscapes even if the sequences are slightly different due to insertions and deletions of some nucleotides in the course of evolution. It is interesting to note that if the free-energy landscapes are first inferred from non-aligned force signals and are then compared, *e.g.* based on standard pairwise sequence alignments, the two resulting SP landscapes are neither adequately aligned with one another, nor with the true free-energy landscape.

C. Differences between B-F and N-A sequences through unzipping experiments

As shown in Fig. S28 and in Fig. 5 of the main text, the free energy landscape of the N-A bacterium is very different from the one of B-F. The number of mismatches (black crosses in Fig. S28, bottom panel) between the two sequences is, indeed, of 339 bases. N-A and B-F free-energy landscapes can be clearly distinguished on a 30 base-pair scale. The SP free energies are also very different along the sequence, see Fig. S28 (bottom). Their difference (orange line) clearly reflects the difference between the 'true' free-energy landscapes (turquoise line). As expected the total difference between the SP free energies of the two genes ($\simeq 161$ k_BT for $\simeq 1540$ base pairs) is smaller then the true total difference ($\simeq 470$ k_BT), as SP underestimates differences in the landscapes associated to barriers.

D. Comparison of B-F and B-H free-energy landscapes for trap stiffnesses $K_{trap} = 0.08$ and 0.3 pN/nm

Fig. S29 and Fig. S30 compare the free-energy landscapes of B-F and B-H bacteria when using the SP and Box approximations. The true free-energy landscapes are plotted in top panels. They are obtained from the aligned B-F and B-H sequences and the pairing parameters of Table S4, and are averaged over a sliding window w=30 bp for the comparison with the SP approximation and w=10 bp for the comparison with the Box approximation. In the middle panels the free energy landscape are inferred from the aligned force signals. In the bottom panels the free energy landscape differences are plotted, as in the Fig. 6 of the main paper.

The differences between the free energies for bacteria B-F and B-H, inferred with the SP and Box methods, and with the two trap stiffnesses $K_{trap} = 0.08$ and 0.3 pN/nm are shown for the first 200 base pairs in Fig. S31 and Fig. S32. In these plots the comparison is made with the true free-energy landscape without any sliding average.

E. Comparison of B-F free-energy landscape with the ones of B-S

Figure S33 and Fig. S34 show the true free-energy landscapes computed from the sequences B-F and B-S and MFold, compared to the outcomes of the SP and Box inferences based on synthetic unzipping data. The bottom panels shows the free energy differences as in the Fig. 6 of the main paper. In Fig. S34(bottom) we show the differences in real free energy landscapes without any sliding average (turquoise line) and the one obtained with the box approximations.

F. NCBI database and whole-database screening of N-A 16S gene

While the NCBI database [7] contains about 2500 sequences of 16S genes, we exclude sequences containing an N symbol (corresponding to an unknown nucleotide in that position), one sequence much smaller than the others (112 bases), and 6 sequences with more than 2000 nucleotides. We are therefore left with 2076 sequences in the data base.

As shown in Fig. S35 the comparison of the N-A gene landscape to all the other sequence landscapes in the bacterial database shows similar features to what is obtained for the test gene B-F, see Fig. 7 of the main text. In the N-A case, however, the gap with the most similar sequence is larger then the estimated experimental resolution in the experiment of Huguet and collaborators [1] (red line in Fig. S35).

---

[1] Huguet, J.M., Bizarro, C.V., Forns, N., Smith, S.B., Bustamante, C. and F. Ritort. 2010. Single-molecule derivation of salt dependent base-pair free energies in DNA. Proc Natl Acad Sci U S A. 107:15431-6.

[2] M. Zuker, 2000. Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol 10:303.

[3] Smith, S.B., Cui, Y. and C. Bustamante. 1996. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. Science 271:795-798.

[4] Barbieri, C., Cocco, S., Monasson, R. and F. Zamponi. 2009. Dynamical modelling of molecular constructions and setups for DNA unzipping. Phys. Biol. 6:025003.

[5] Bockelmann, U., Essevaz-Roulet,B., and F. Heslot. 1997. Molecular Stick-Slip Motion Revealed by Opening DNA with Piconewton Forces. Phys. Rev. Lett. 79:4489-4492.

[6] Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48 (3): 44353.

[7] See RefSeq Targeted Loci Project web page, and 16S Bacterial Ribosomal RNA project: http://www.ncbi.nlm.nih.gov/genomes/static/refseqtarget.html

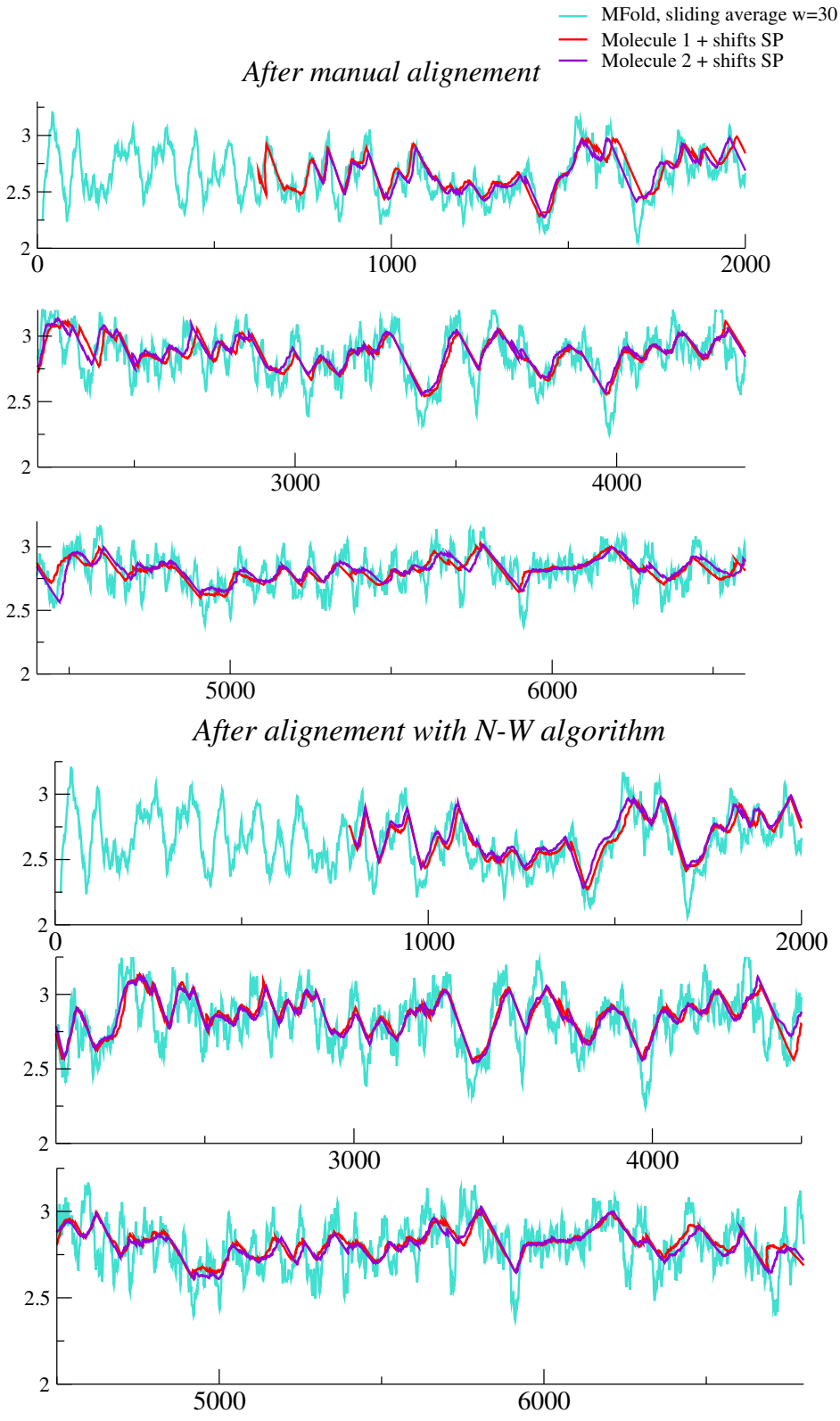

FIG. S21: Saddle point inference for Molecules 1 (red line) and 2 (blue line) using the MFold energetic parameters and after further shifts on the pairing energies and on the trap position. The true sequence landscape (sliding average over $w = 30$ bp) is shown with the turquoise line.

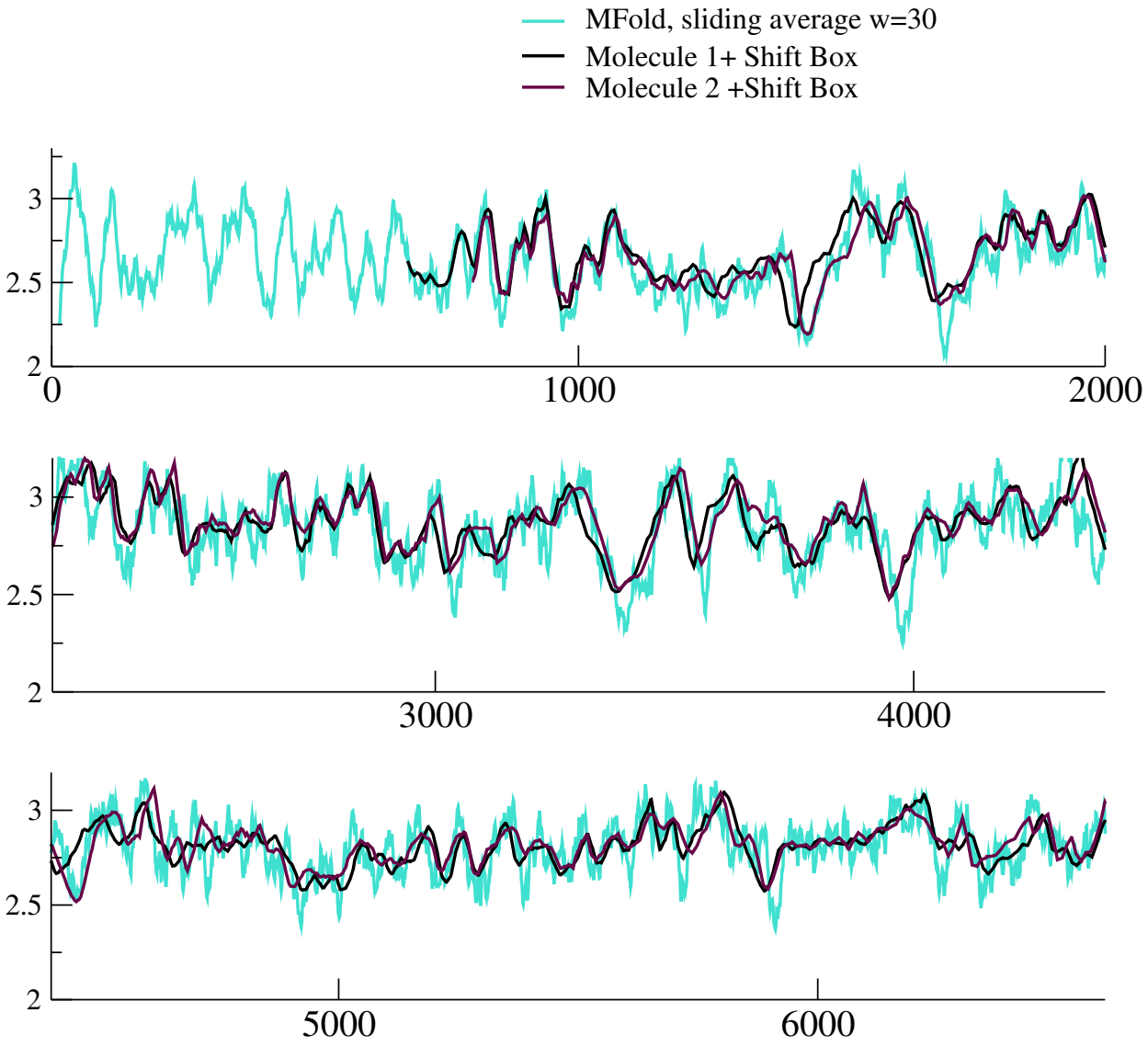

FIG. S22: Box inference for Molecules 1 (red line) and 2 (blue line) from MFold energetic parameters and after global shifts on the force curves and manual alignment on the trap positions. The true sequence landscape (sliding average over $w = 30$ bp) is shown with the turquoise line.

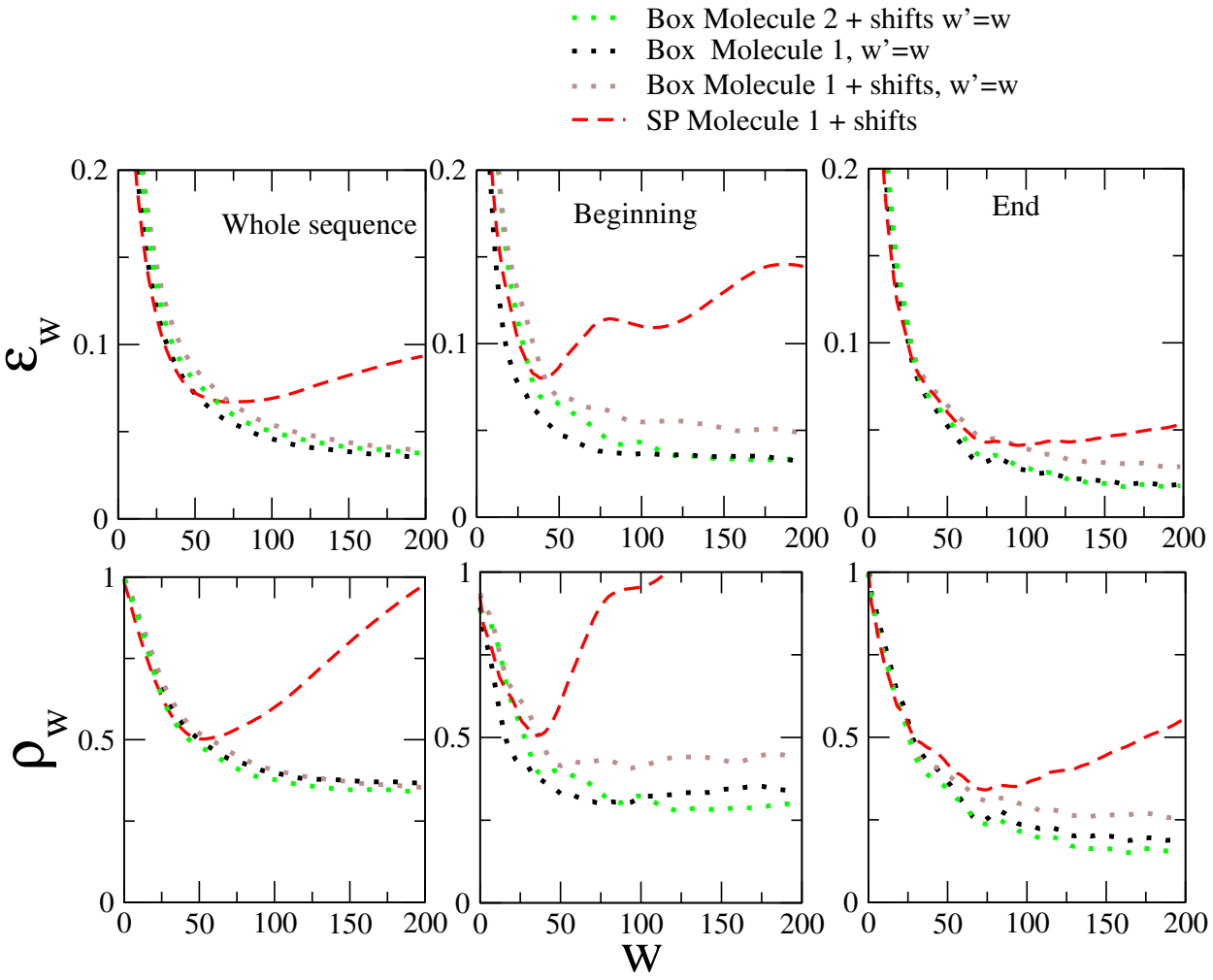

FIG. S23: Absolute (top) and relative (bottom) errors on the inferred free energies as a function of the window size w of the running average, for the whole sequence and for 300 bases at the beginning and the end of the sequence. The data correspond to Molecule 1 (with the free energies found by Huguet et al. [1], black line) and to Molecule 1 with manual shifts (MFold energetic parameters g_0).

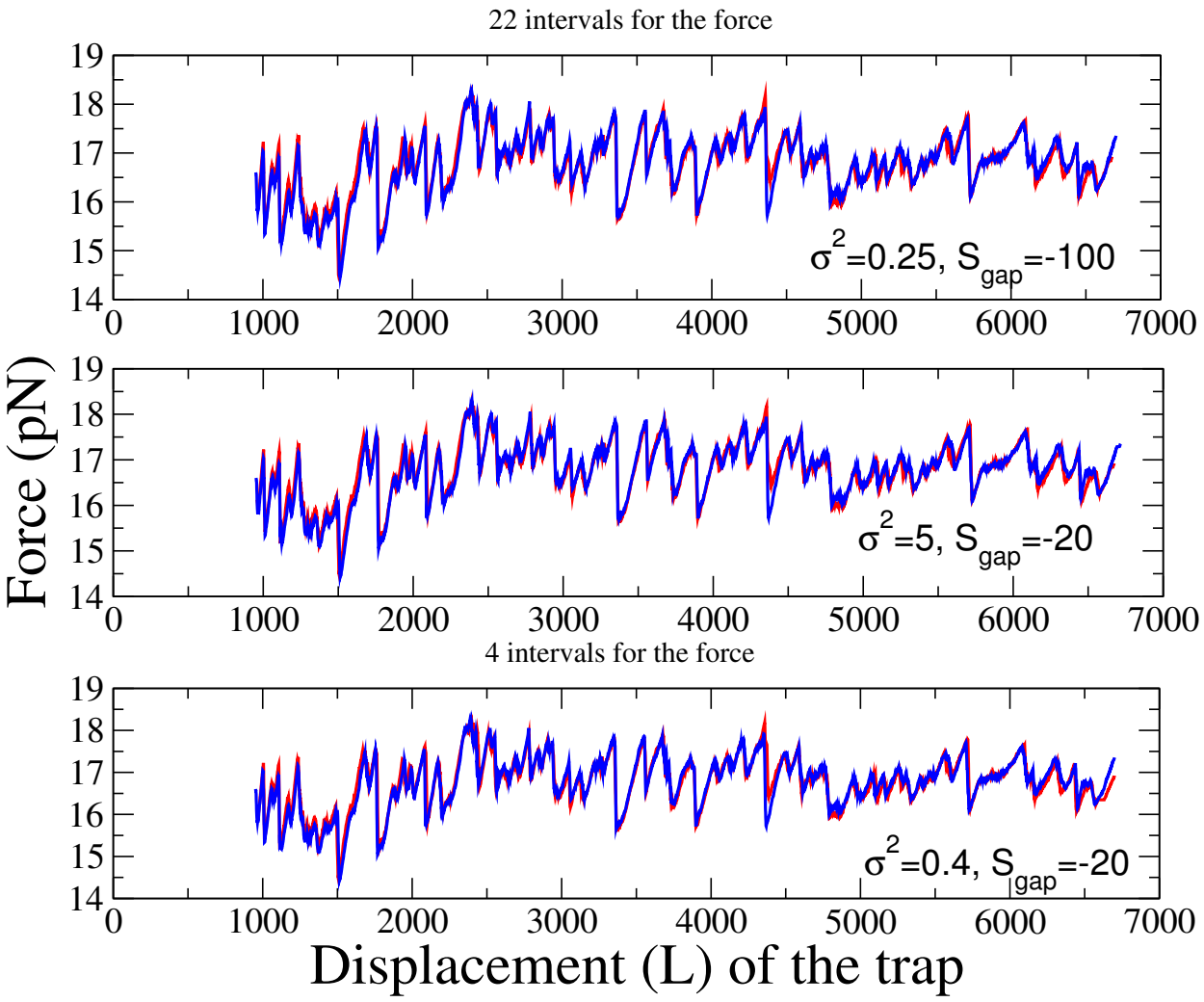

FIG. S24: Re-alignement of the experimental forces with 22 (top & middle panels) and with 4 (bottom panel) force intervales. The values of σ^2 and of the gap penalty S_{gap} are shown in the panels.

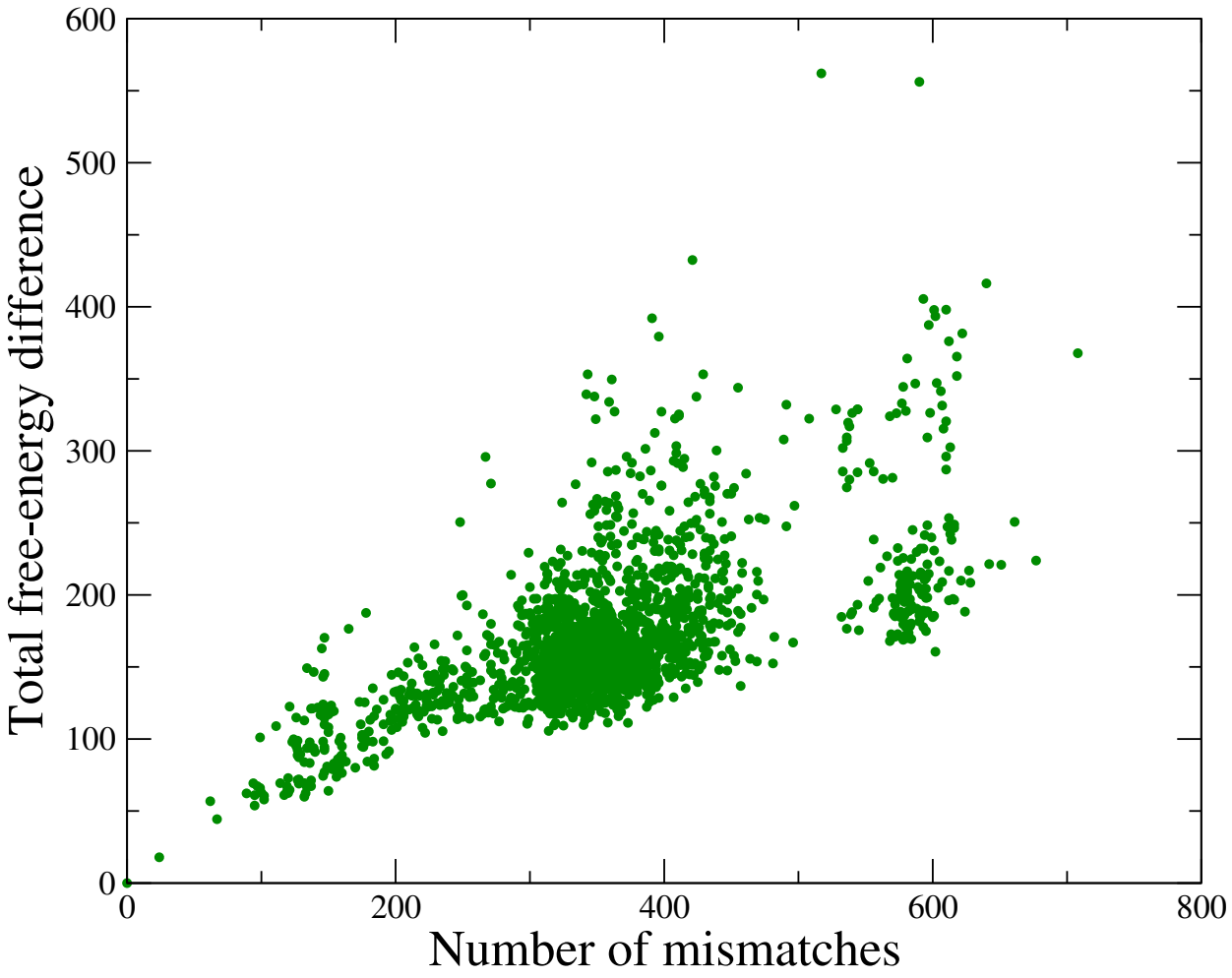

FIG. S25: Total difference Δg in free energy along the aligned sequence vs. number of mismatches, when discretizing the forces with $N_f = 4$ values only. The test sequence is bacterium B-F.

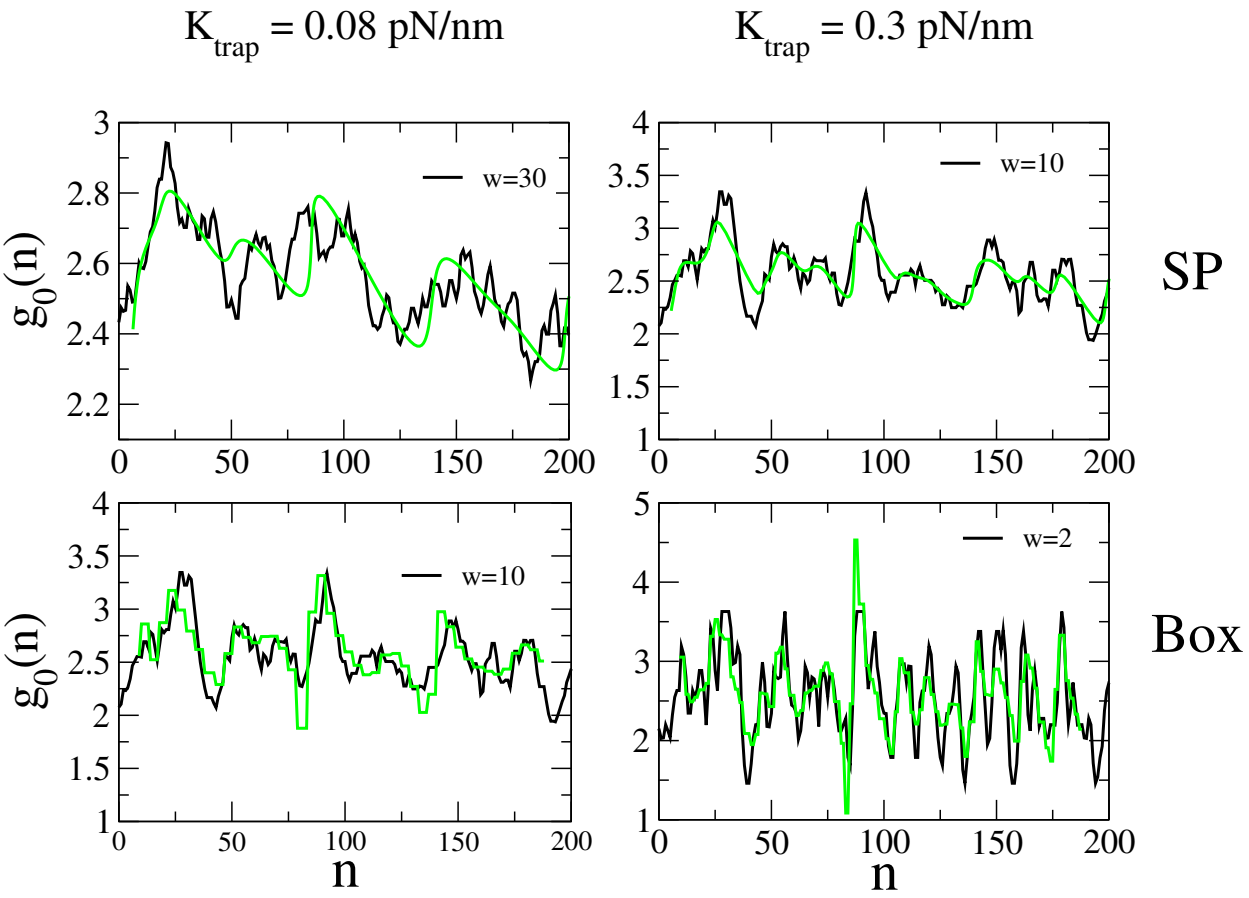

FIG. S26: Free-energy landscape (green curves) on the first 200 base pairs from synthetic unzipping data generated from the sequence of the B-F bacterium, with trap stiffness $K_{trap} = 0.08$ pN/nm (left) and $K_{trap} = 0.3$ pN/nm (right), inferred with the SP (top) and the Box (bottom) approximations. Black curves show the sliding averages of the 'true' free energies over w base pairs (values of w are shown in the panels).

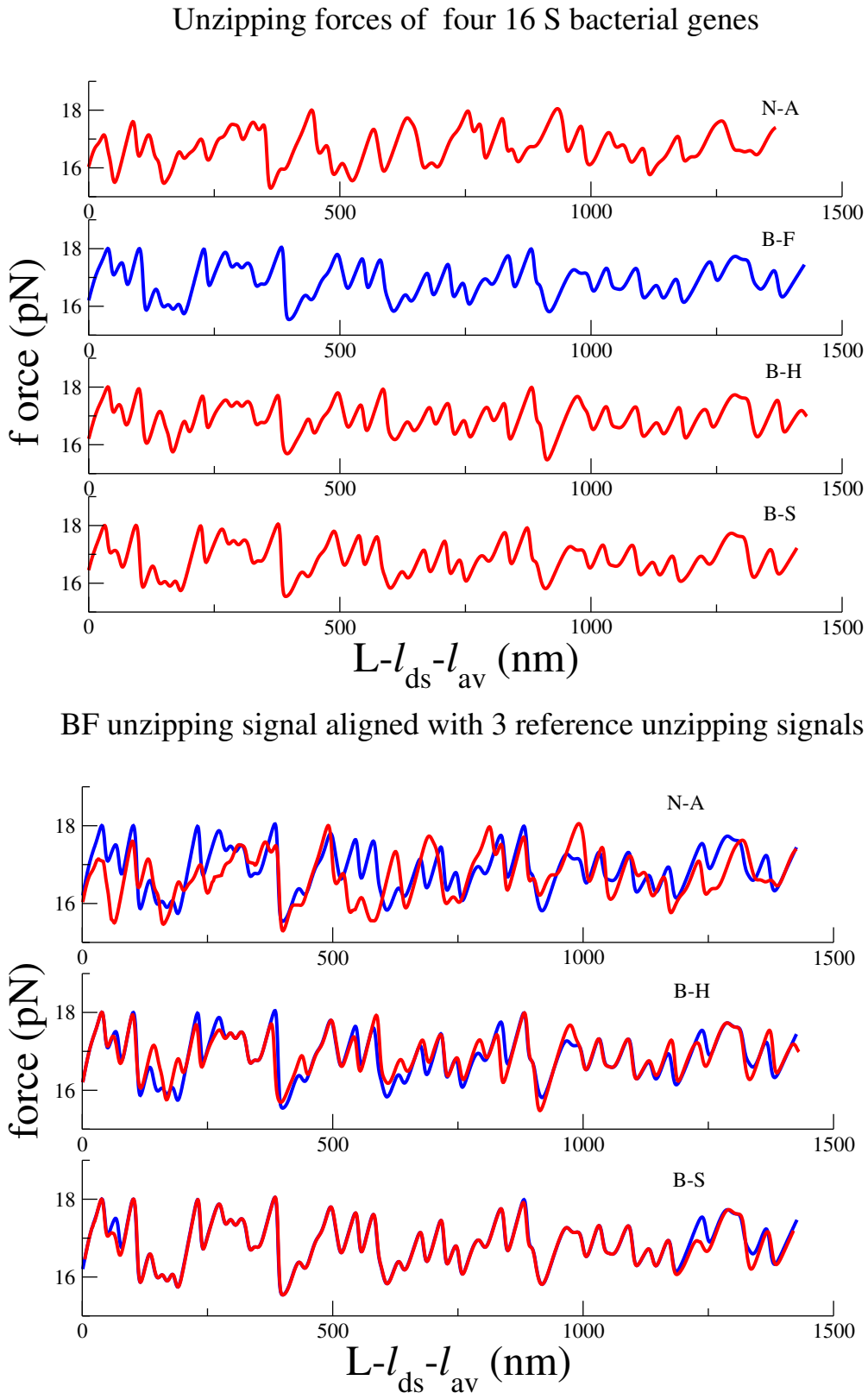

FIG. S27: Top: Unzipping force signals corresponding to N-A, B-F, B-H and B-S bacteria. Bottom: alignment of the B-F unzipping force curve (blue) with the N-A (top), B-H (middle) and B-S (bottom) force curves.

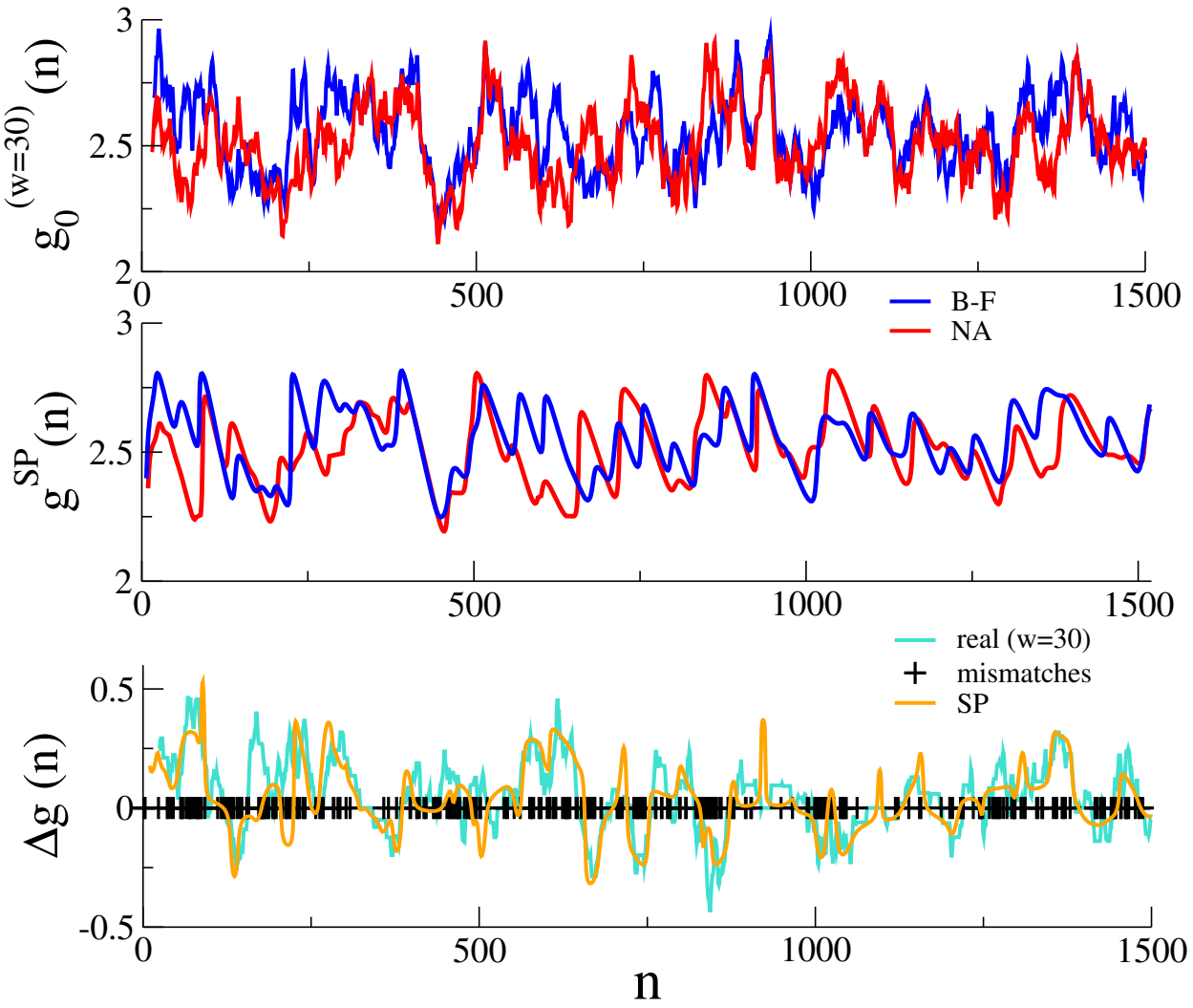

FIG. S28: Comparison of free-energy landscapes for bacteria B-F and N-A. Top: free energy with a sliding average over $w = 30$ base pairs, obtained from the sequences and the pairing parameters of MFold at 150 mM NaCl, after having aligned the two sequences. Middle: inferred SP free-energy landscape obtained from the synthetic force signals computed for the two sequences and then aligned (with parameter $\sigma^2 = 5$ and $S_{gap} = -20$). Bottom: difference (turquoise line) between the aligned free-energy landscapes of B-F and N-A of the top panel with a sliding average $w = 30$, compared to the difference between the inferred SP free-energy landscape (orange line). Mismatches between the two sequences are shown with black crosses.

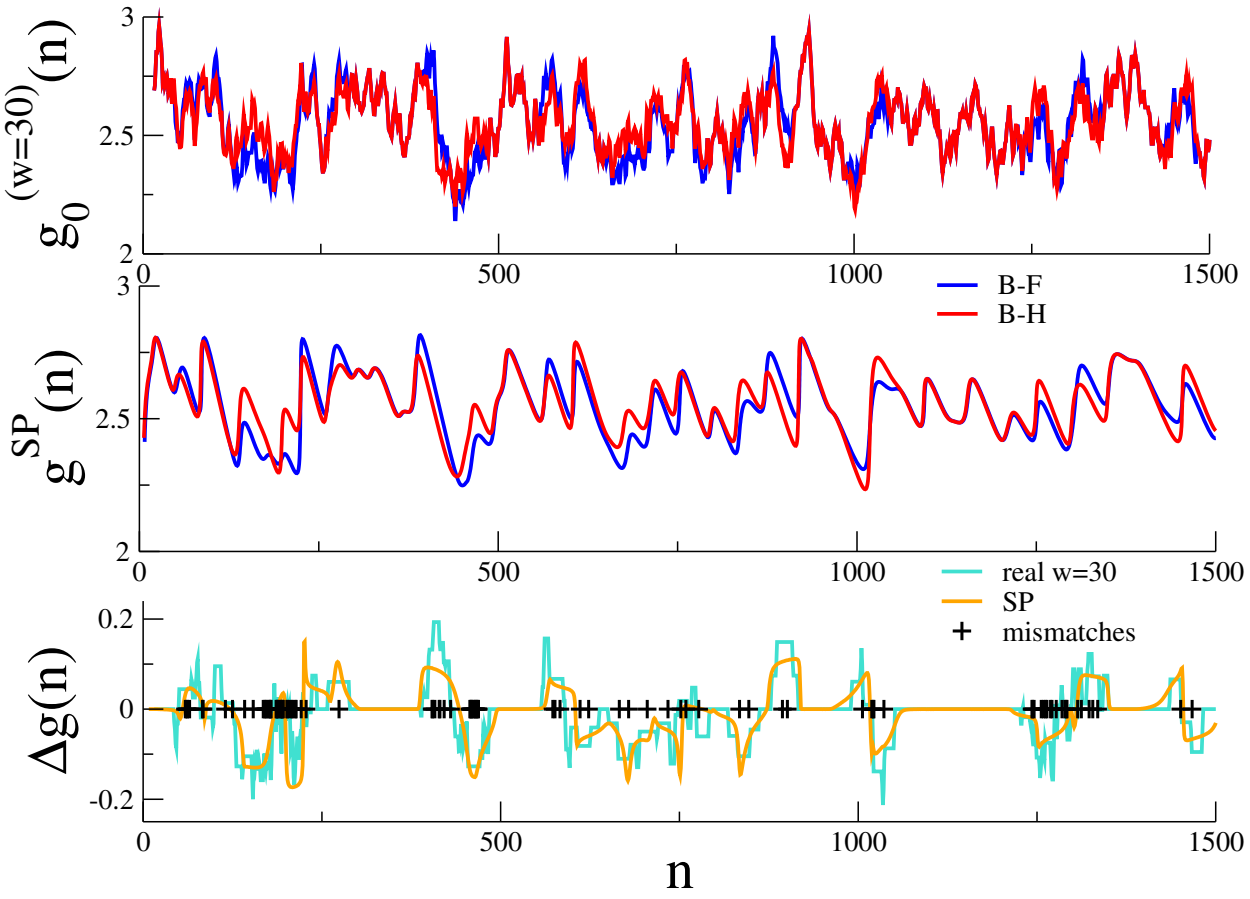

FIG. S29: Comparison of 16S gene of bacteria B-F and B-H from the SP inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 30$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred SP free energy landscape from the synthetic force signals after their alignment. Bottom panel: difference (turquoise line) between the aligned free energy landscapes of B-F and B-H of the top panel compared with the difference (orange line) between the inferred SP free energy landscapes of the middle panel (as the top left panel of Fig.6 in the main paper). Black crosses: mismatches between the two sequences.

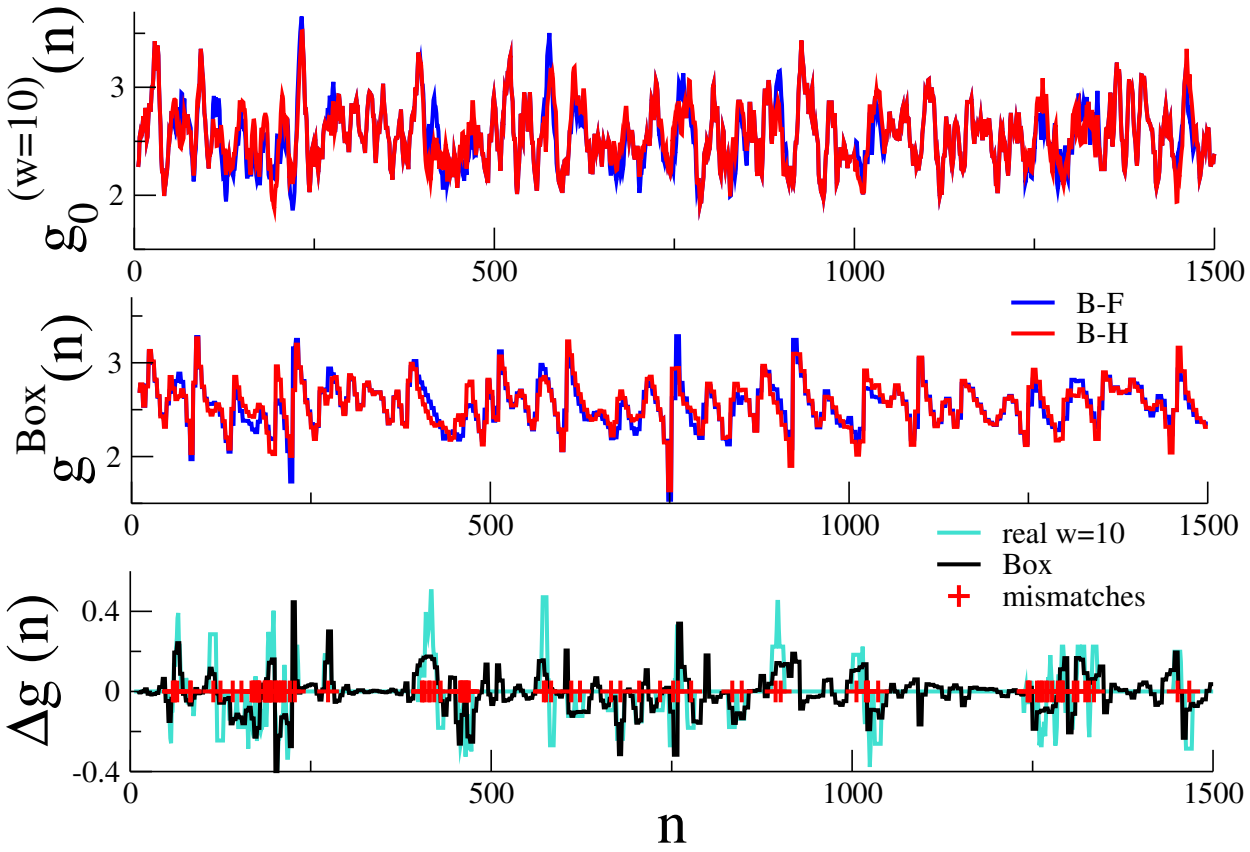

FIG. S30: Comparison of 16S gene of bacteria B-F and B-H from the Box inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 10$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred Box free energy landscape from the synthetic force signals after their alignment. Bottom panel: difference (turquoise line) between the aligned free energy landscapes of B-F and B-H of the top panel compared with the difference (black line) between the inferred Box free energy landscapes of the middle panel. Red crosses: mismatches between the two sequences.

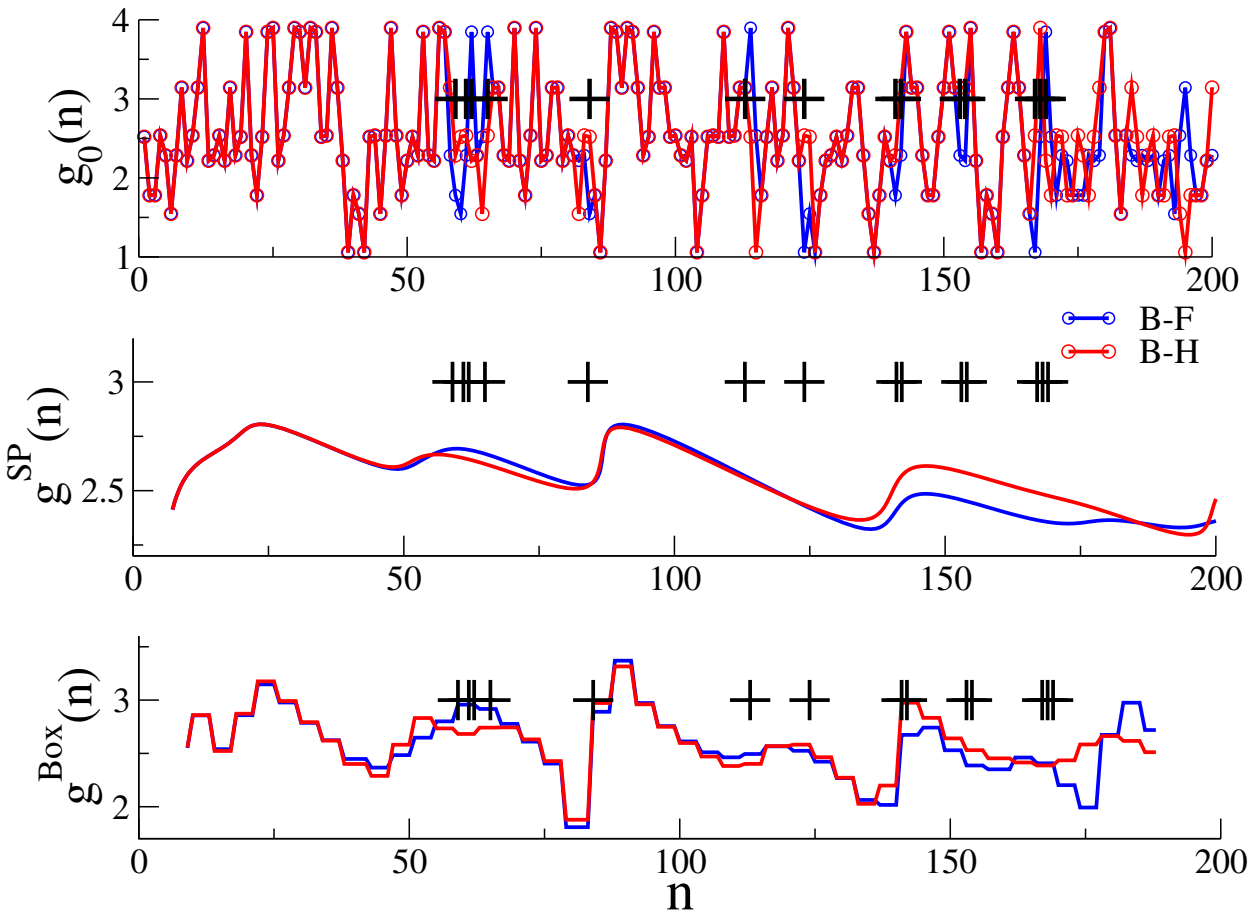

FIG. S31: Trap stifness $K_{trap} = 0.08$ pN/nm. Magnification over the first 200 bases of the sequence: comparison between the real free energy differences (without any sliding average), the SP inference and the Box inference for B-F and B-H bacteria. Dark crosses locate mismatches between the two sequences.

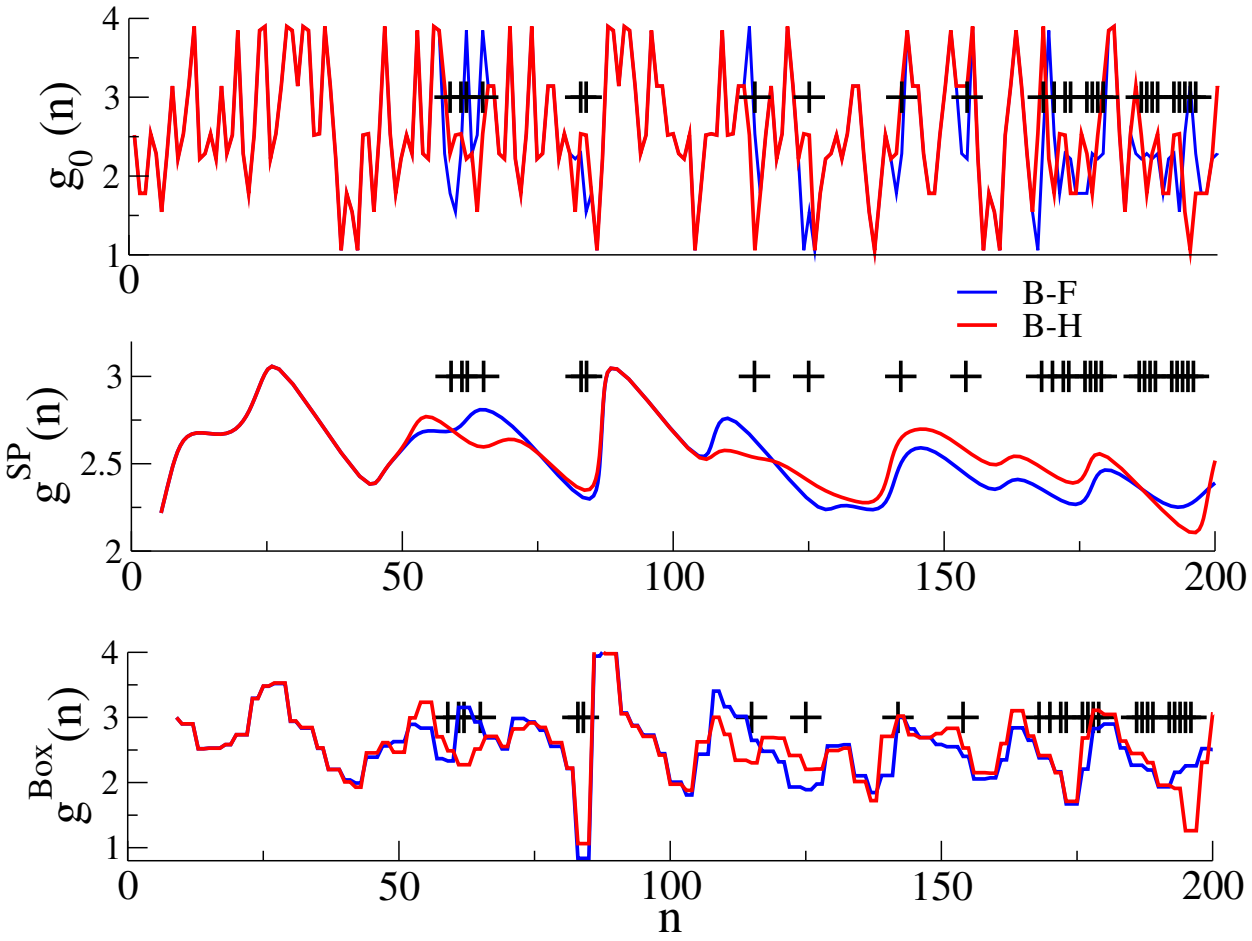

FIG. S32: Trap stifness $K_{trap} = 0.3$ pN/nm. Focus on the first 200 bases of the sequence: comparison between the real free energy differences (without any sliding average),the SP inference and the Box inference for B-F and B-H bacteria. Dark crosses locate mismatches between the two sequences.

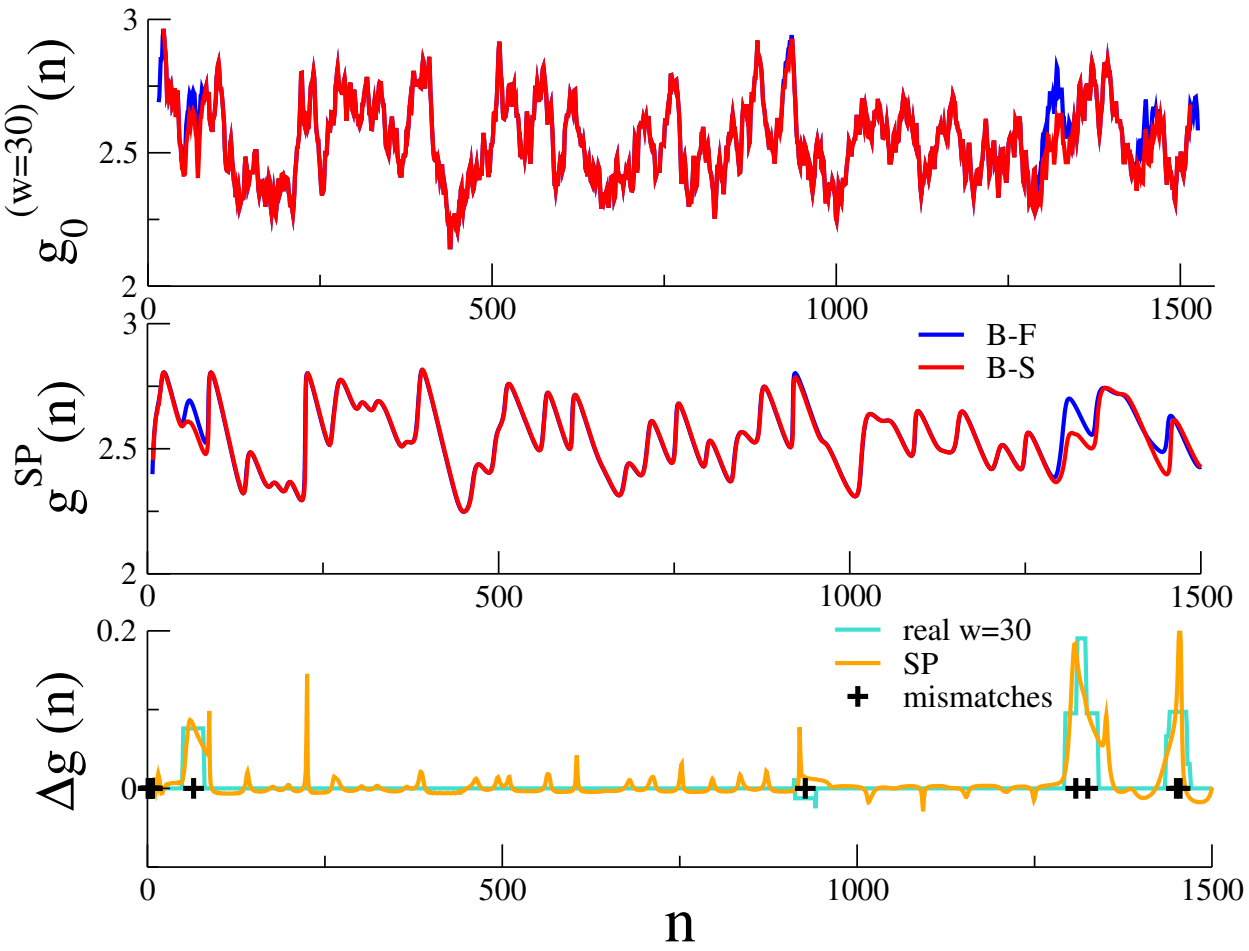

FIG. S33: Comparison of 16S gene of bacteria B-F and B-S from the SP inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 30$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred SP free energy landscape from the synthetic force signals after their alignment. Bottom panel: difference (turquoise line) between the aligned free energy landscapes of B-F and B-S of the top panel compared with the difference (orange line) between the inferred SP free energy landscapes of the middle panel (as the top left panel of Fig.6 in the main paper). Black crosses: mismatches between the two sequences.

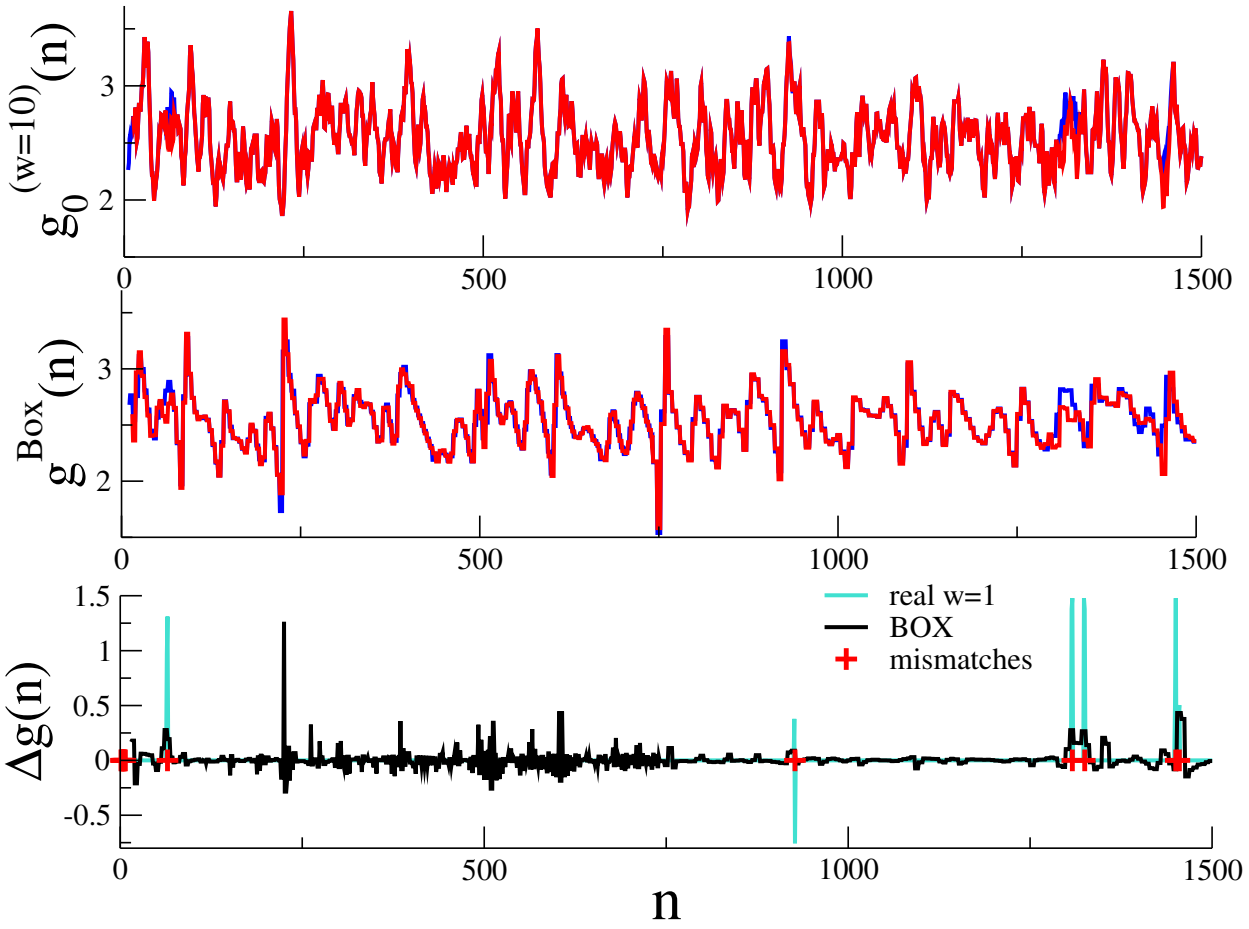

FIG. S34: Comparison of 16S gene of bacteria B-F and B-S from the Box inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 10$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred Box free energy landscape from the synthetic force signals after their alignment. Bottom panel: difference (turquoise line) between the aligned free energy landscapes of B-F and B-S without any sliding average (w=1) and difference (black line) between the inferred Box free energy landscapes . Red crosses: mismatches between the two sequences.

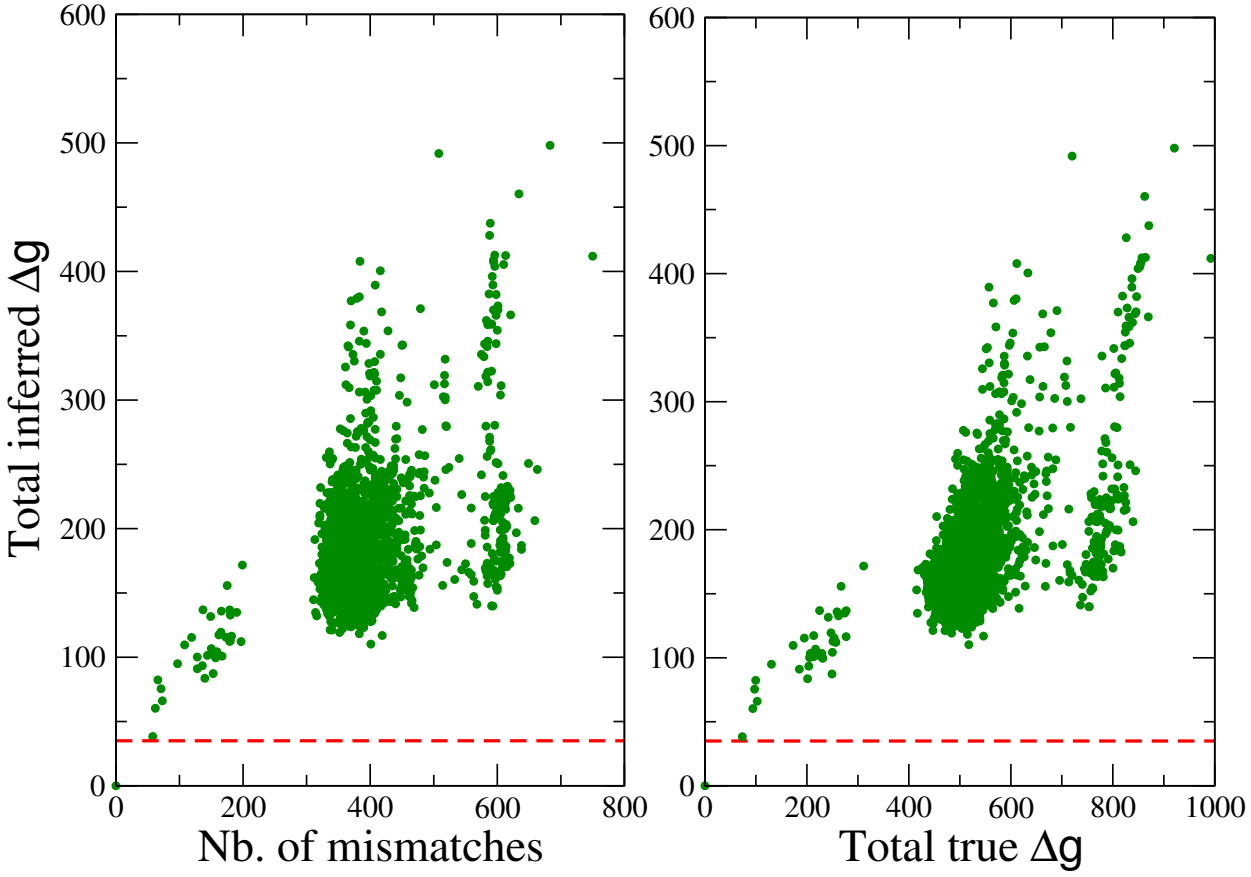

FIG. S35: Comparison of the SP inferred free-energy landscape for bacterium N-A with the other bacteria in the database. Total differences in free energies vs. number of mismatches (left) and vs. the true diferences in free energies along the sequences (right), computed after pairwise alignments.