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}.

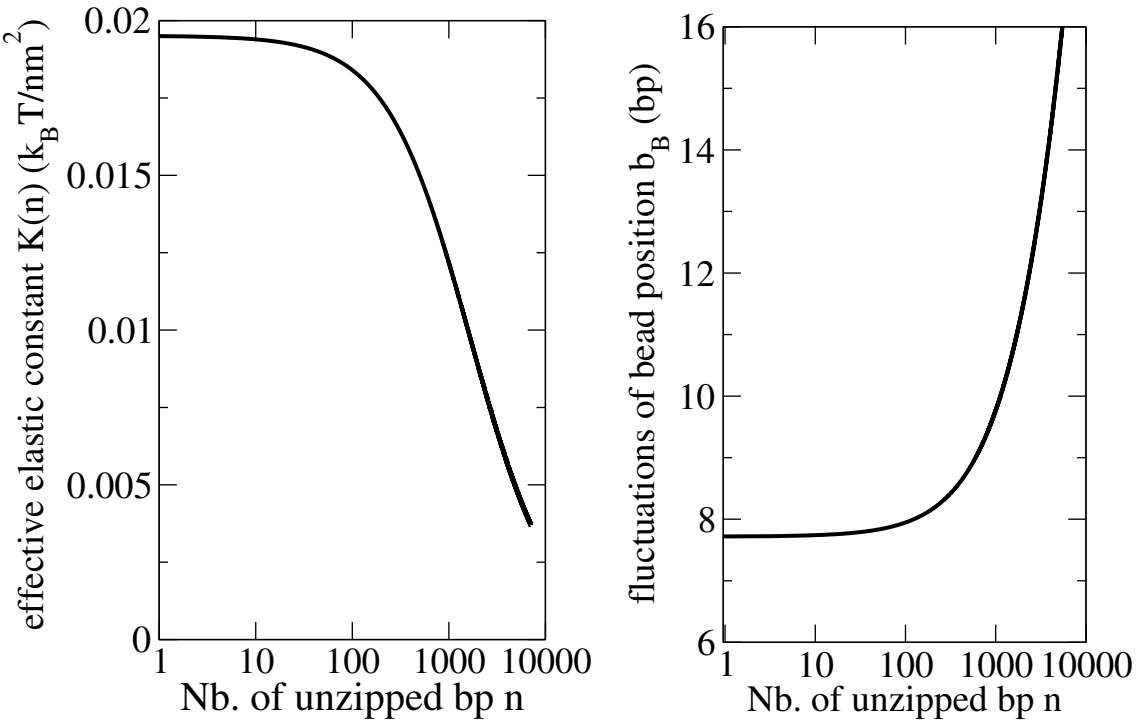

FIG. S1: Left: Stiffness of the experimental setup, $K(n)$, in units of k_BT/nm^2, as a function of the number of unzipped base pairs. Right: fluctuation $b_B(n)$ of bead in units of the extension ℓ_{ss} of an open bp as a function of the number of unzipped base pairs.

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),

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

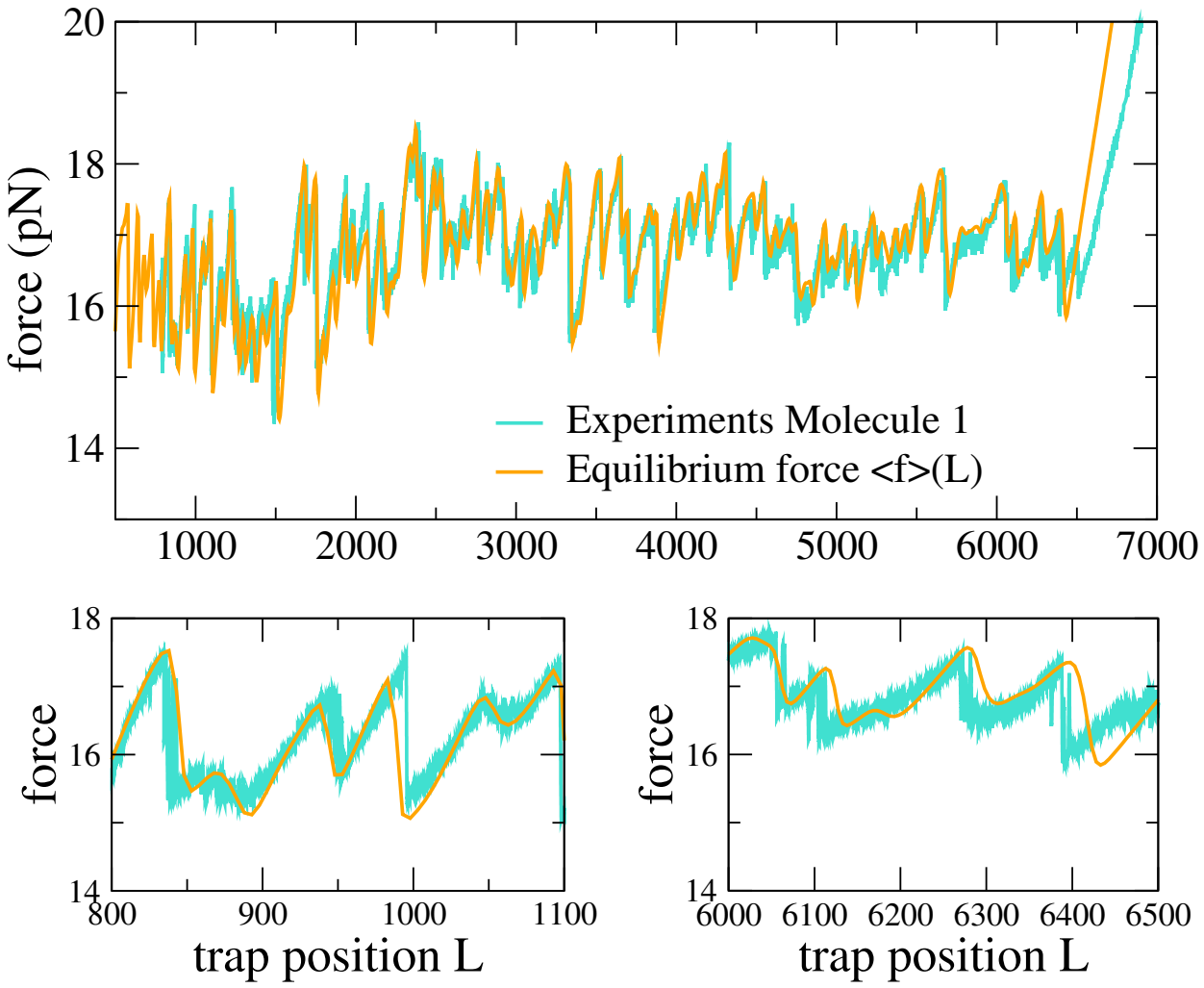

FIG. S2: Equilibrium force in the harmonic approximation compared to the experimental force. Top: Unzipping force, as a function of the trap position L (in nm). Turquoise line: experimental results for Molecule 1, Orange line: average force at equilibrium $\langle f \rangle(L)$. Bottom: magnification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

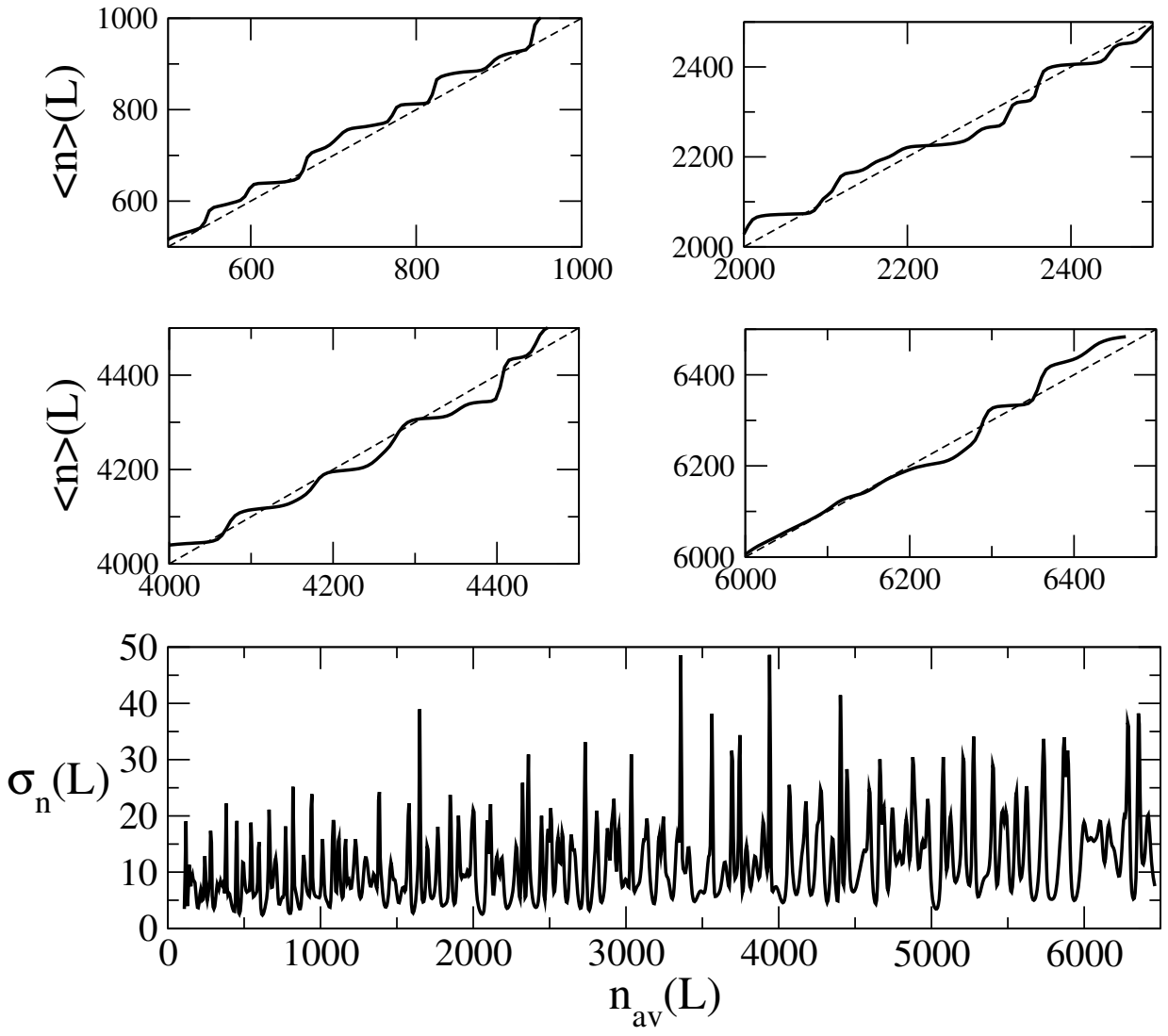

FIG. S3: Top and middle panels: Average number of open base pairs $\langle n \rangle(L)$ in the harmonic model (black) as a function of the its average sequence counterpart, $n_{av}(L)$, see Eq. (18). Dashed lines show the $n_{av} = \langle n \rangle$ curves. Bottom: Standard deviation of the number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

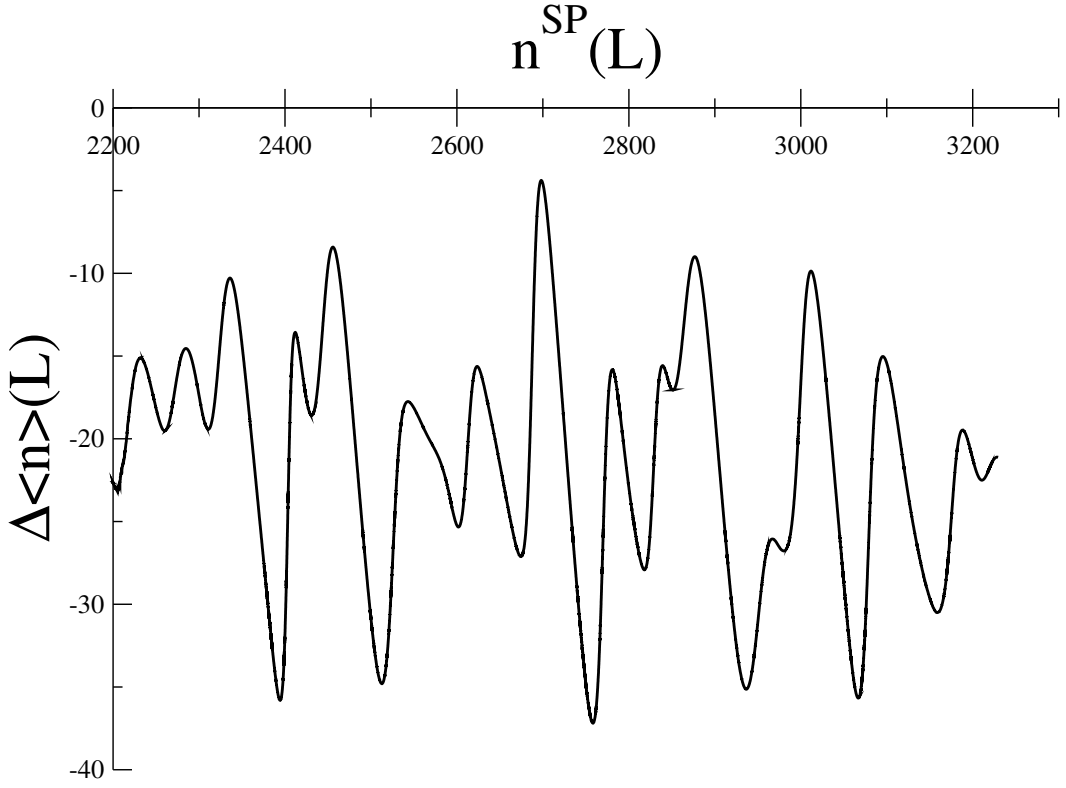

FIG. S4: Parametric representation of the deviation $\Delta\langle n\rangle(L)$ between the average number of open bp and the SP value vs. $n^{SP}(L)$ for a portion of the sequence landscape inferred with the SP approximation.

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:

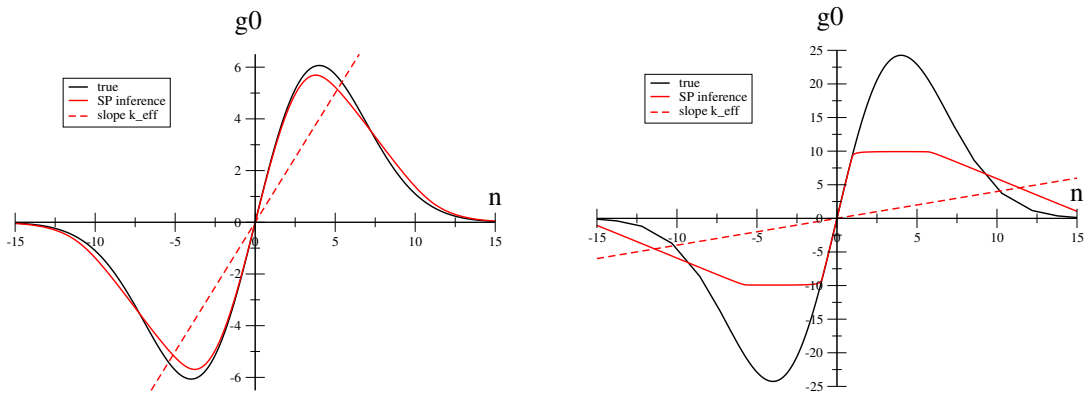

FIG. S5: Examples of W-S barriers. Left: $\Delta G = -10$, Right: $\Delta G = -40$. Other parameter are $\Delta n = 1$, $K_{eff} = K\,\ell^2 = 1$.

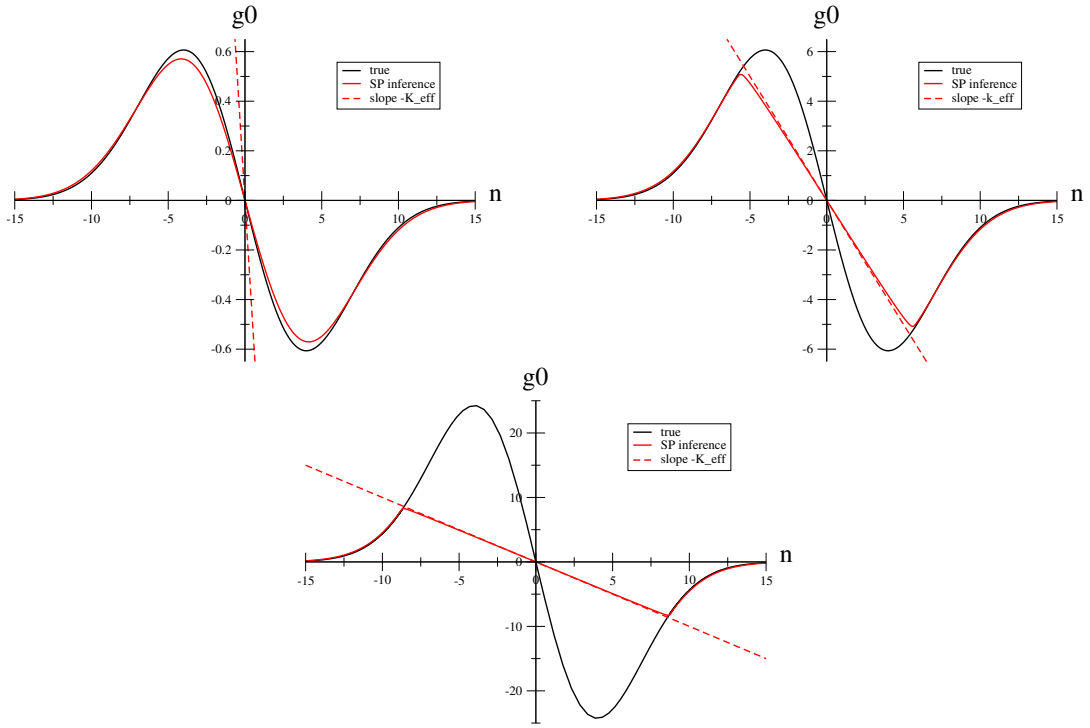

FIG. S6: Examples of S-W barriers. Top Left: $\Delta G = 1$, Top Right: $\Delta G = 10$, Bottom: $\Delta G = 40$. Other parameter are $\Delta n = 1$, $K_{eff} = K\ell^2 = 1$.

- 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,

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

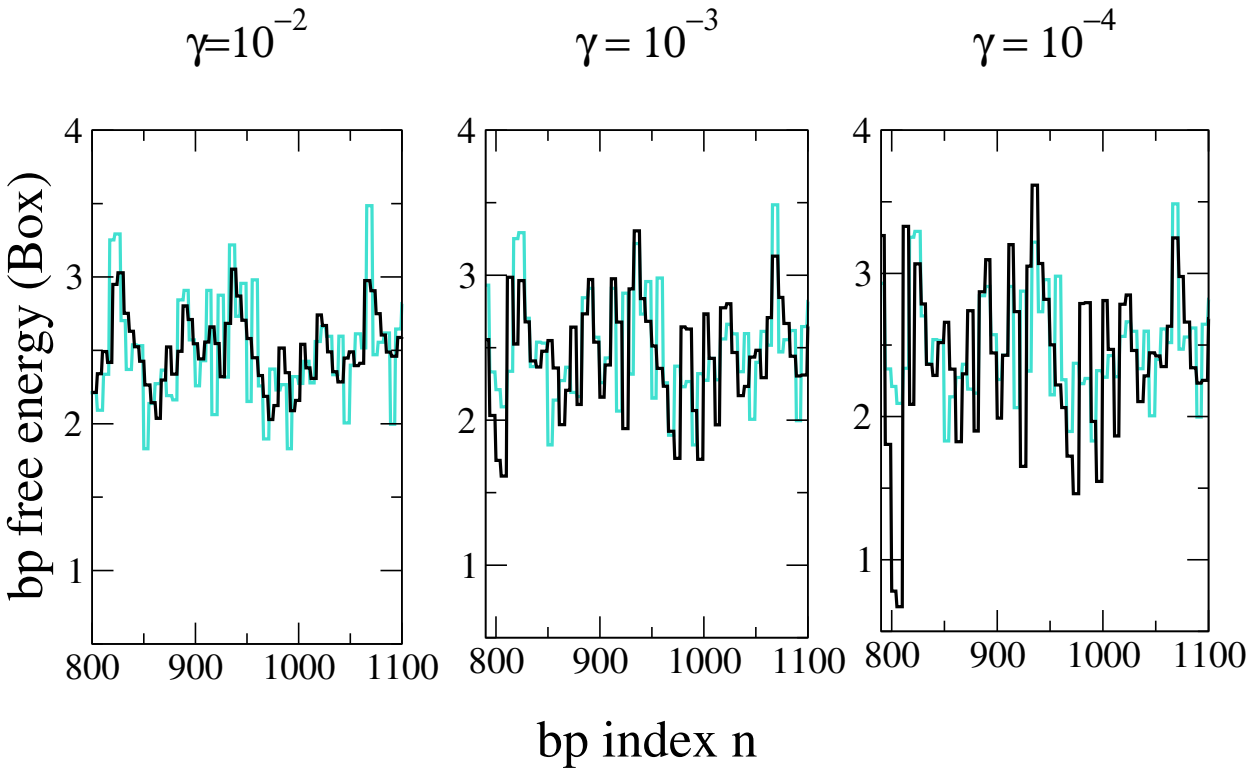

FIG. S7: Base pair free energies at the beginning of Molecule 2 (turquoise: true values, black: outcome of the Box inference procedure with $b = 5$). Left: penalty parameter $\gamma = 10^{-2}$. Middle: penalty parameter $\gamma = 10^{-3}$. Right: penalty parameter $\gamma = 10^{-4}$

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

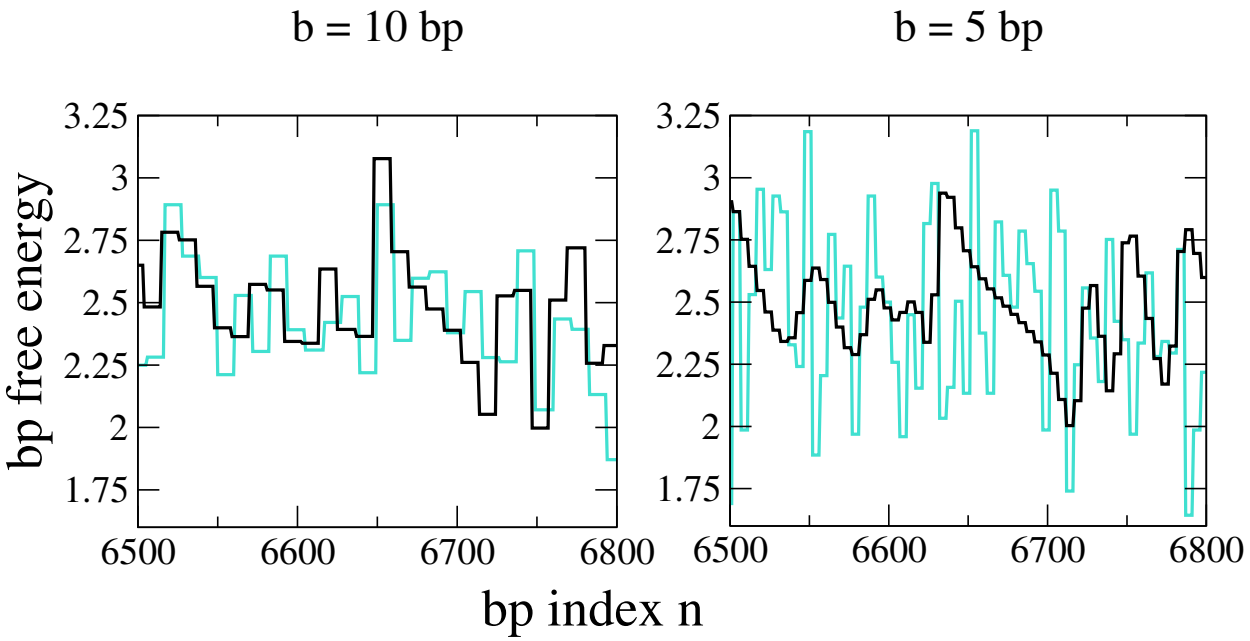

FIG. S8: Inference of the base pair energy at the end of Molecule 1. Left: trial function constant over $b = 10$ bp (black), comparison with the box average of the true free energy over 10 bp (turquoise). Right: trial function constant over $b = 5$ bp (black), comparison with the box average of the true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

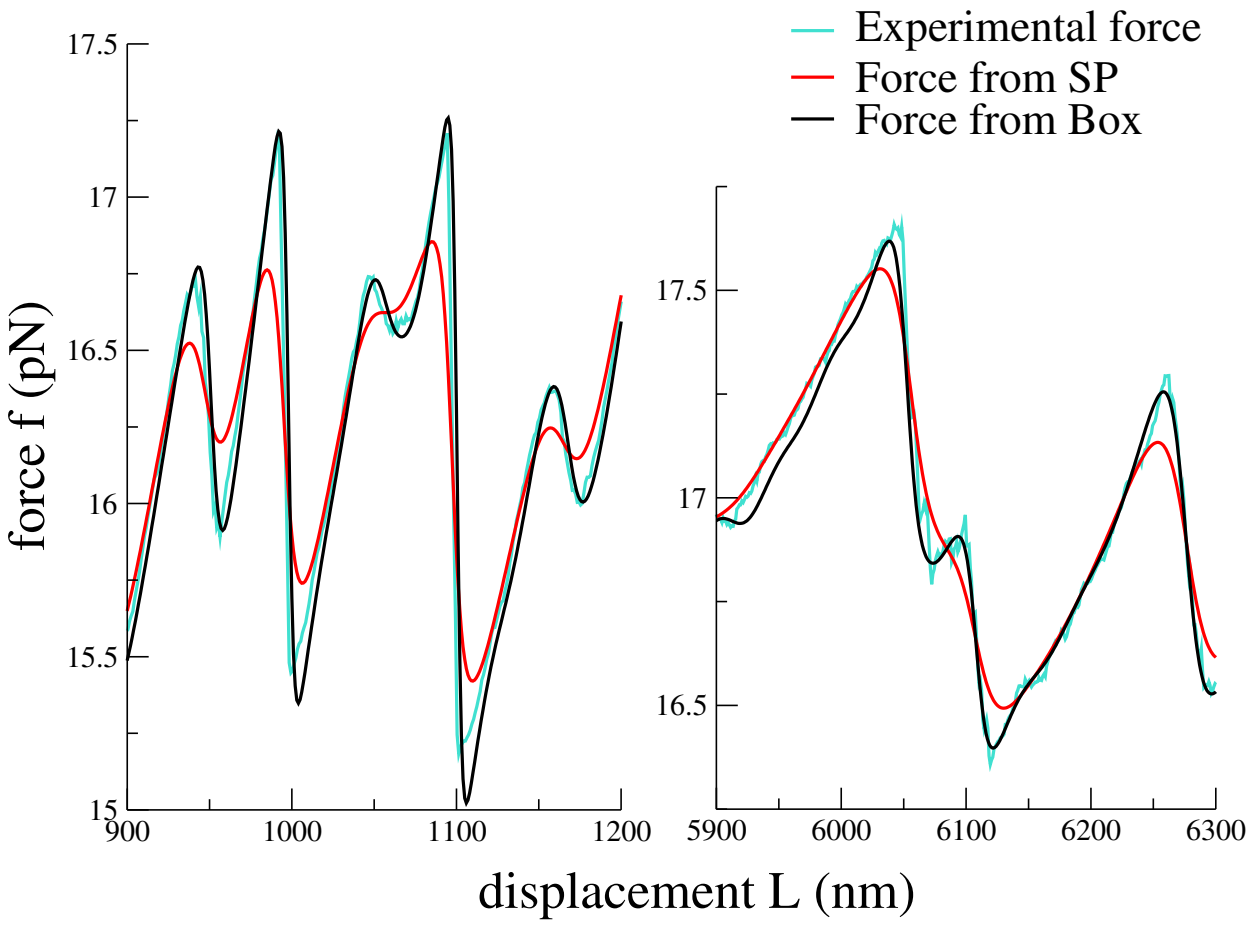

FIG. S9: Average force vs. trap position: Experimental values $f_{exp}(L)$ for Molecule 1 are shown in turquoise, while the equilibrium force $\langle f \rangle(L)$ computed the bp free energies inferred with the SP and Box approximations are shown in, respectively, red and black.

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.

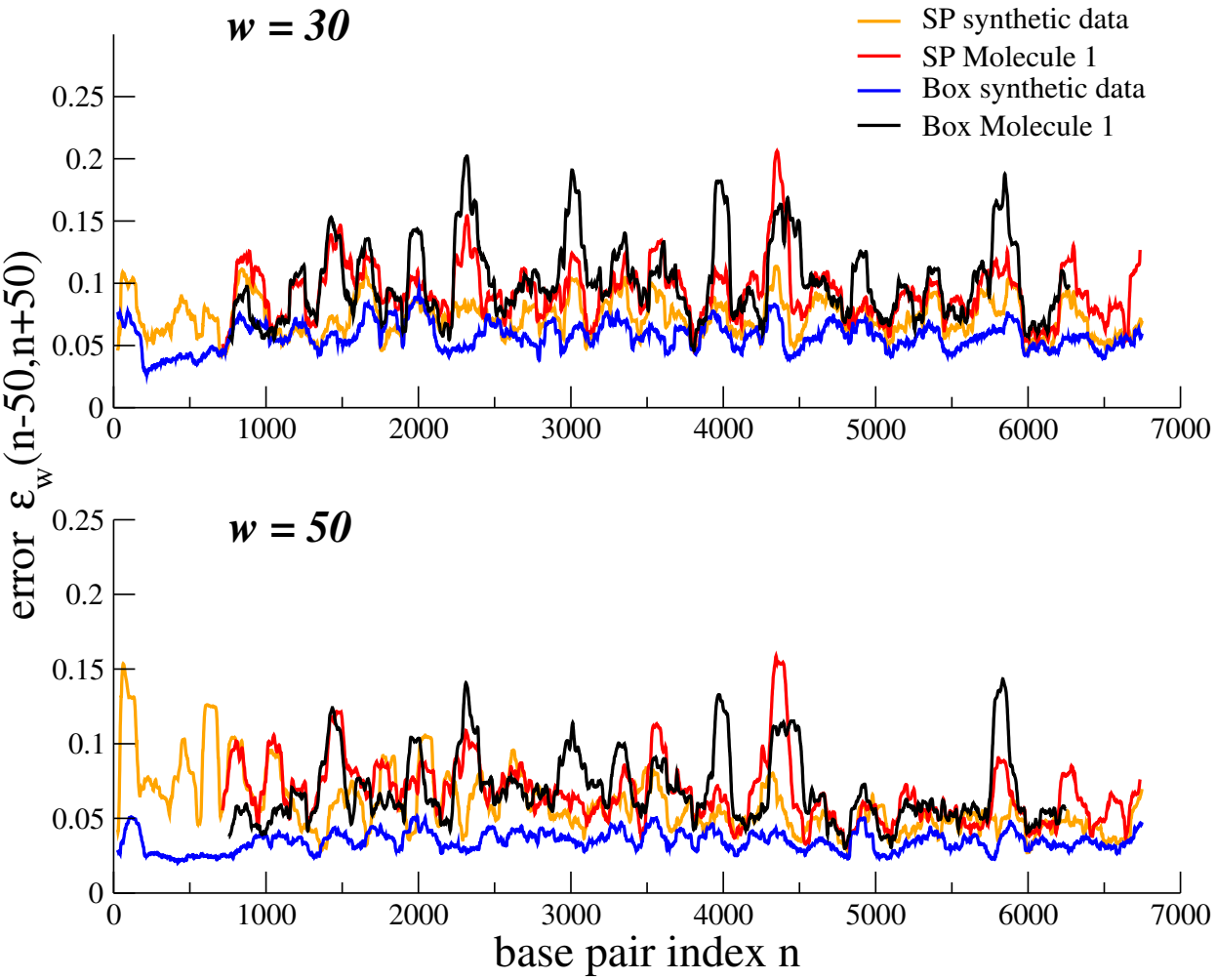

FIG. S10: Position-dependent errors in the inference. Error $\epsilon_w(n-50, n+50)$ on the inferred bp free energy with respect to the true value, averaged on a window $w = 30$ (top) and $w = 50$ (bottom), vs. bp index n.

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

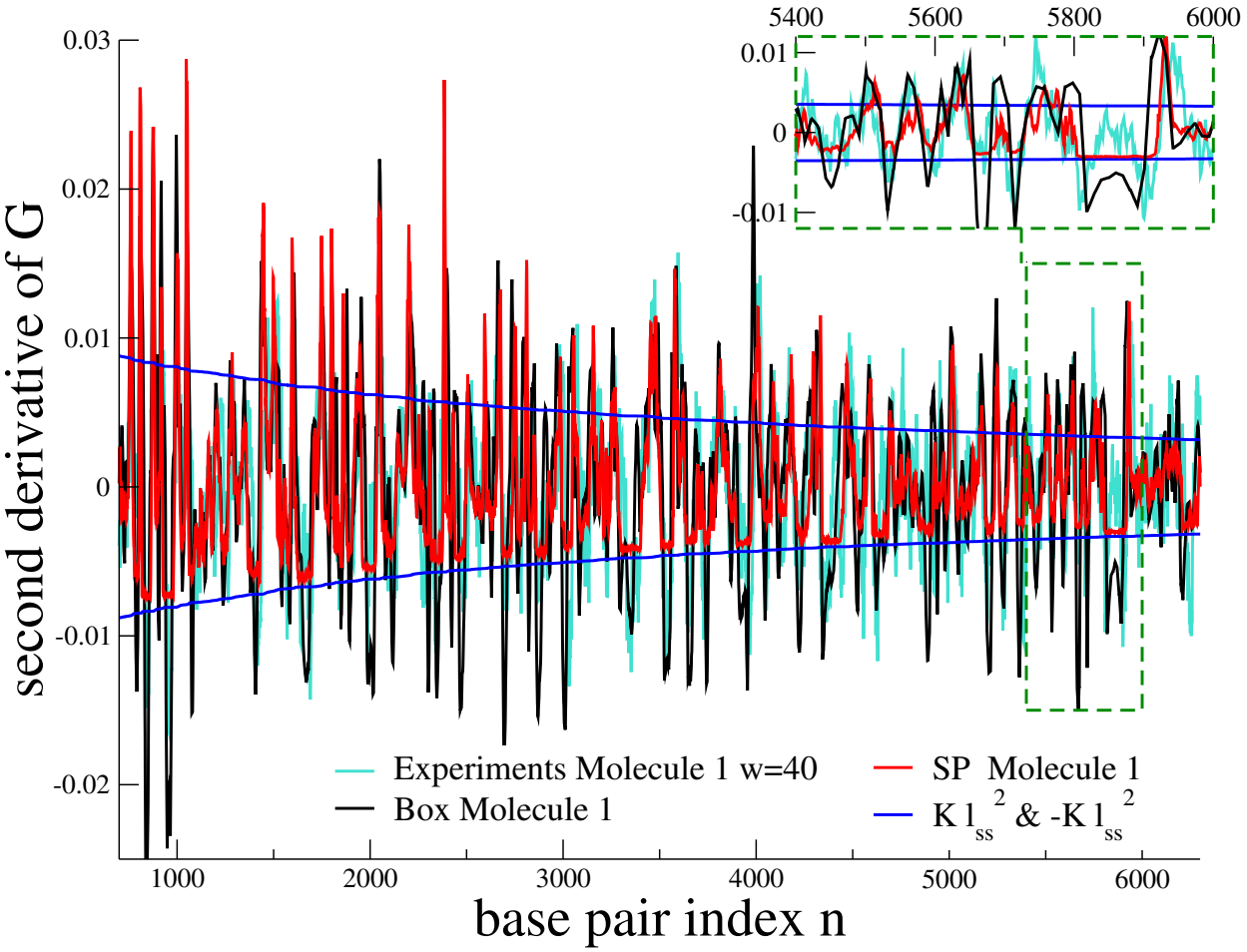

FIG. S11: Second derivative of the inferred cumulative free-energy landscape $G(n)$ (convoluted over 10 bp) plotted as a function of the bp index n for the SP (red curve) and Box (black curve) approximations, and for the experimental data (turquoise line, $w = 40$). Blue lines show the effective stiffness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

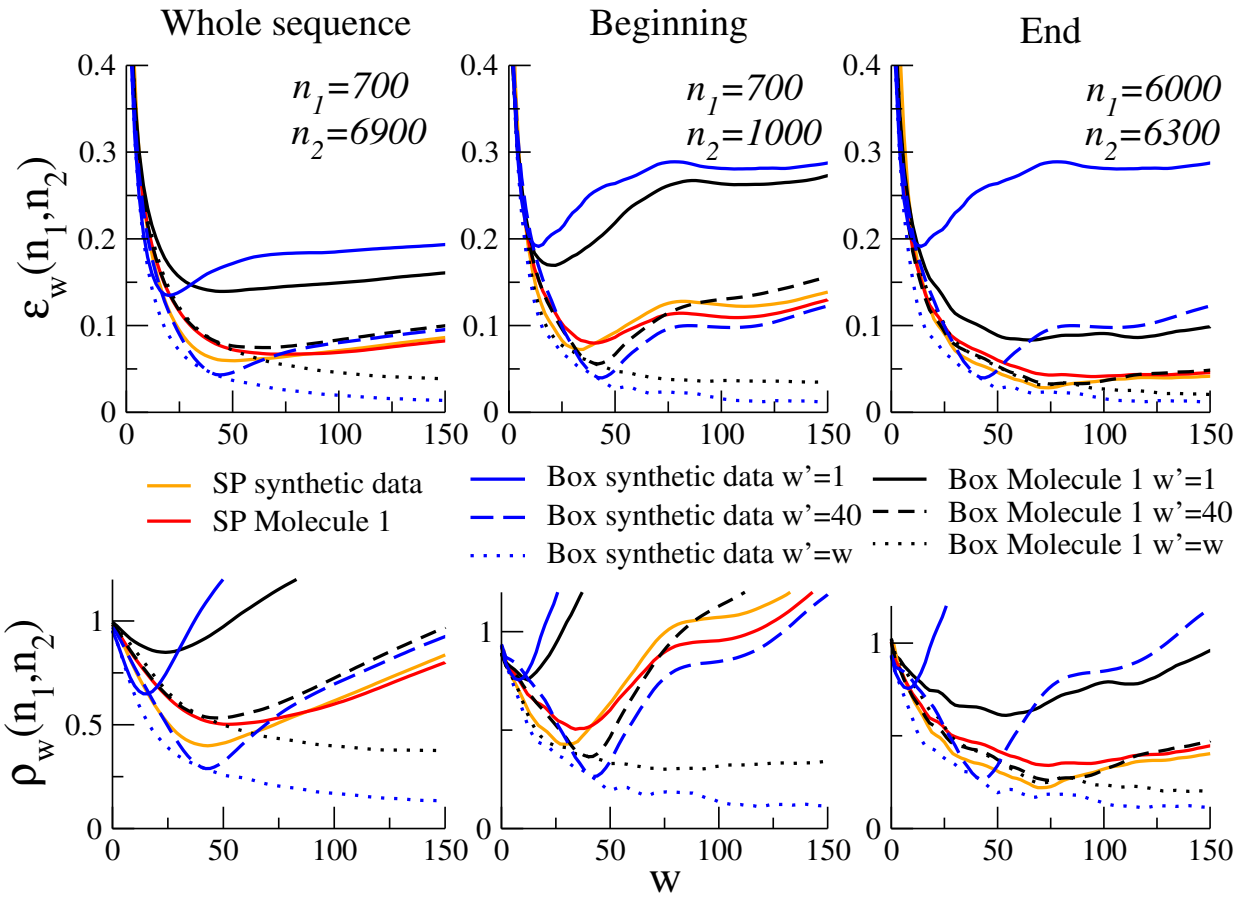

FIG. S12: Average errors in the inference: Average error over the inferred free energy ϵ_w (Top) and averaged relative error ρ_w (Bottom) as a function of the window size w of the running average on the true free energy landscape, and for different window size w' of the running average on the inferred 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.

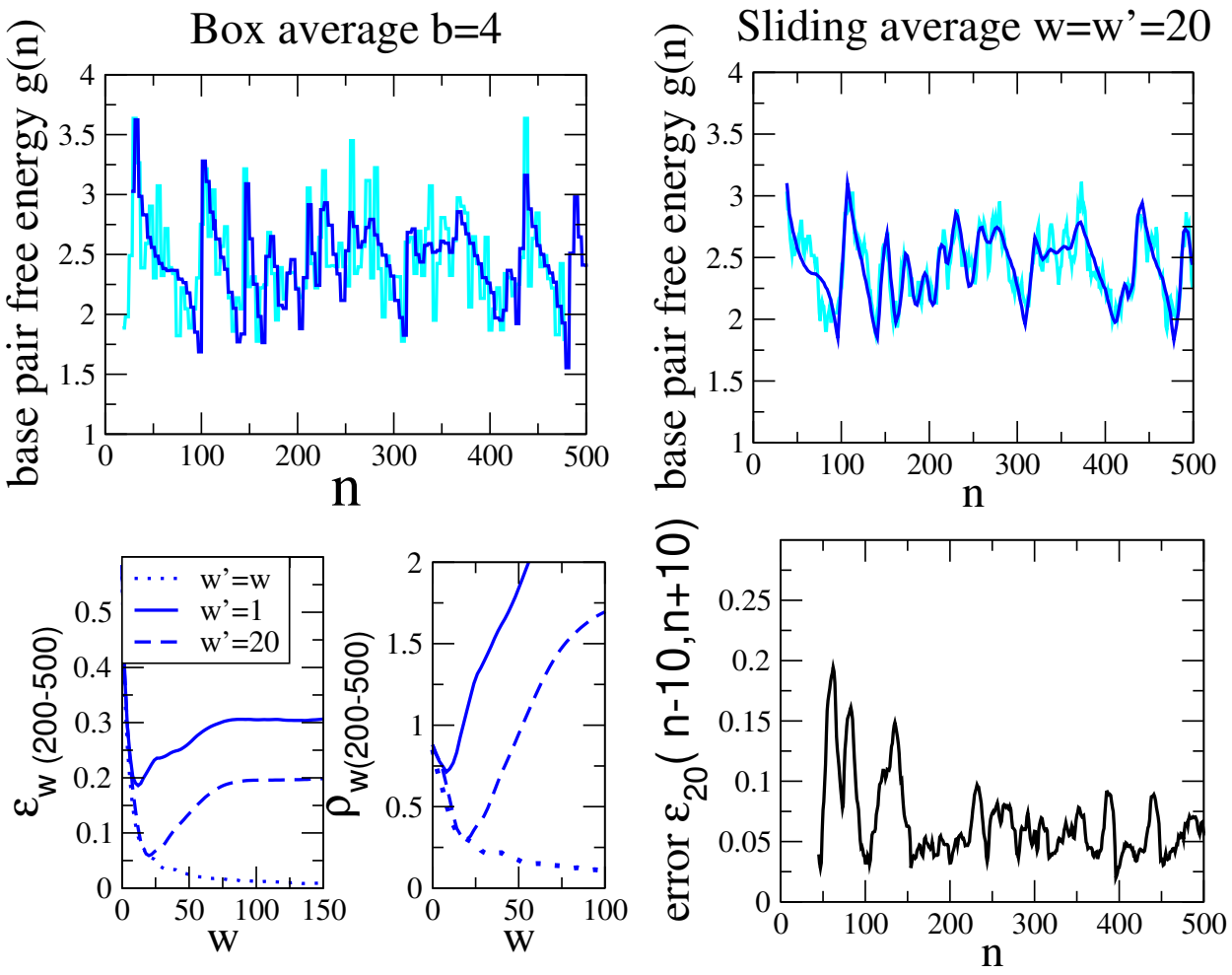

FIG. S13: Inference from the unzipping force for synthetic data. Top: Left: Free energy inferred for the first 500 base pairs (blue) compared with a box average of the true sequence with the same b (turquoise). Right: sliding average of the free energy over $w' = 20$ bp (turquoise) compared to the sliding 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.

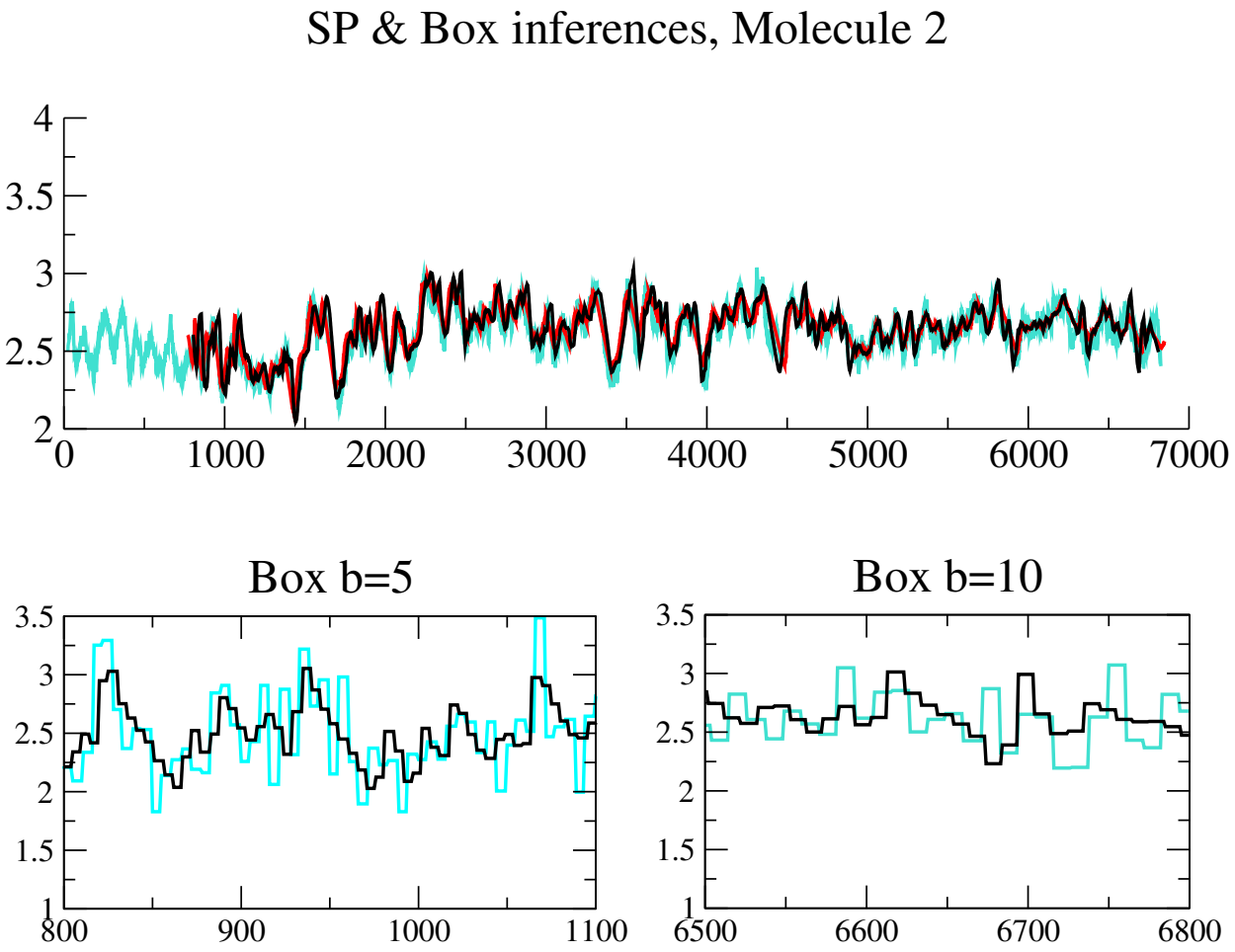

FIG. S14: Inference of the base pair energy from the unzipping force for molecule 2. Top: comparison of the SP (red line) and the Box (black line) landscapes with the true free energy (sliding average on 50 bp). Bottom: Box inference for the first (Left, $b = 5$) and the last (Right, $b = 10$) base pairs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

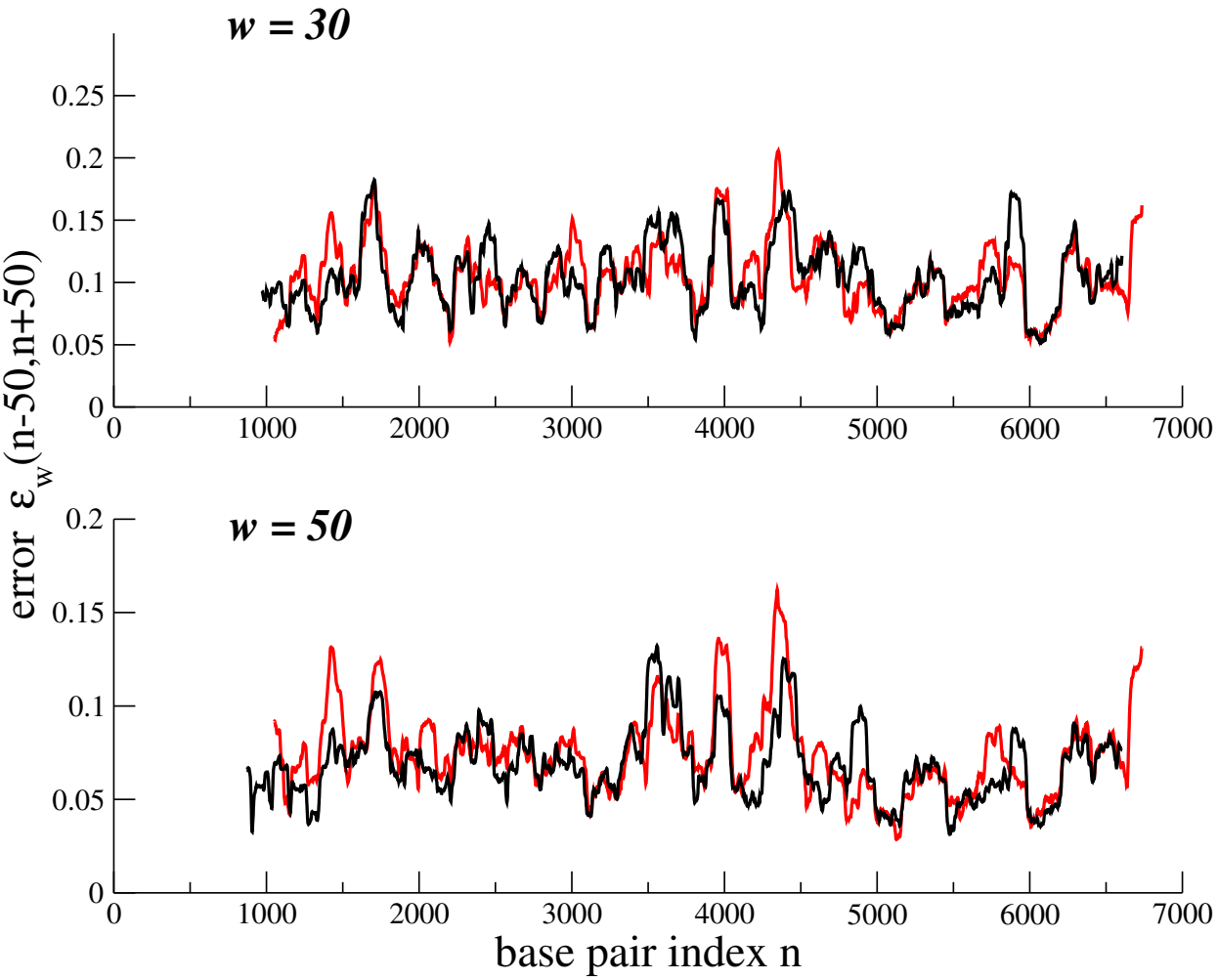

FIG. S15: Position-dependent errors in the inference for Molecule 2. Error $\epsilon_w(n-50, n+50)$ on the inferred bp free energy with respect to the true value, averaged on a window $w = 30$ (top) and $w = 50$ (bottom), vs. bp index n.

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

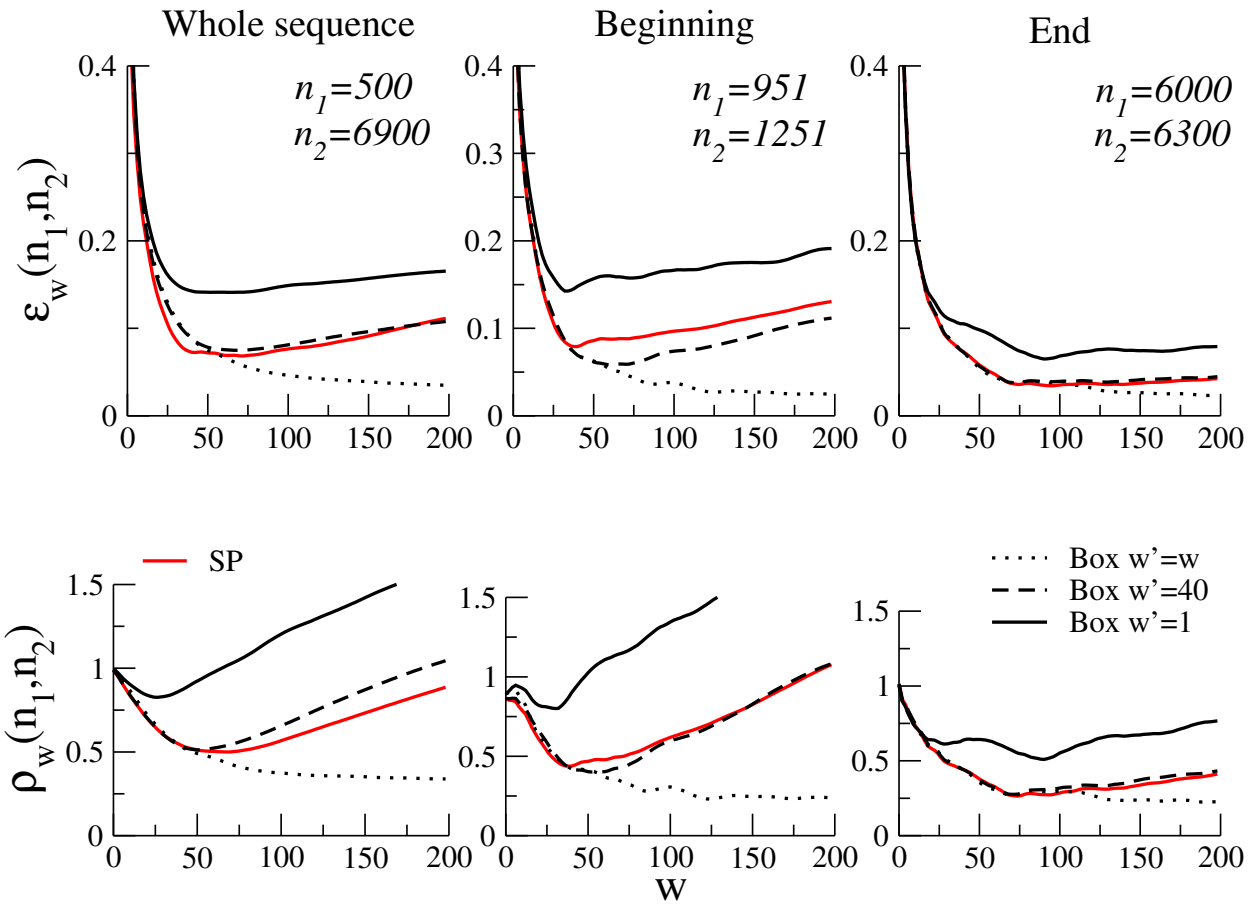

FIG. S16: Average error of the inferred free energy for molecule 2. Top: average error ϵ_w; Bottom: averaged relative error ρ_w, as functions of the window size w of the running average, and for different size w' of the running average on the Box inferred free energy. The errors are computed over the 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.

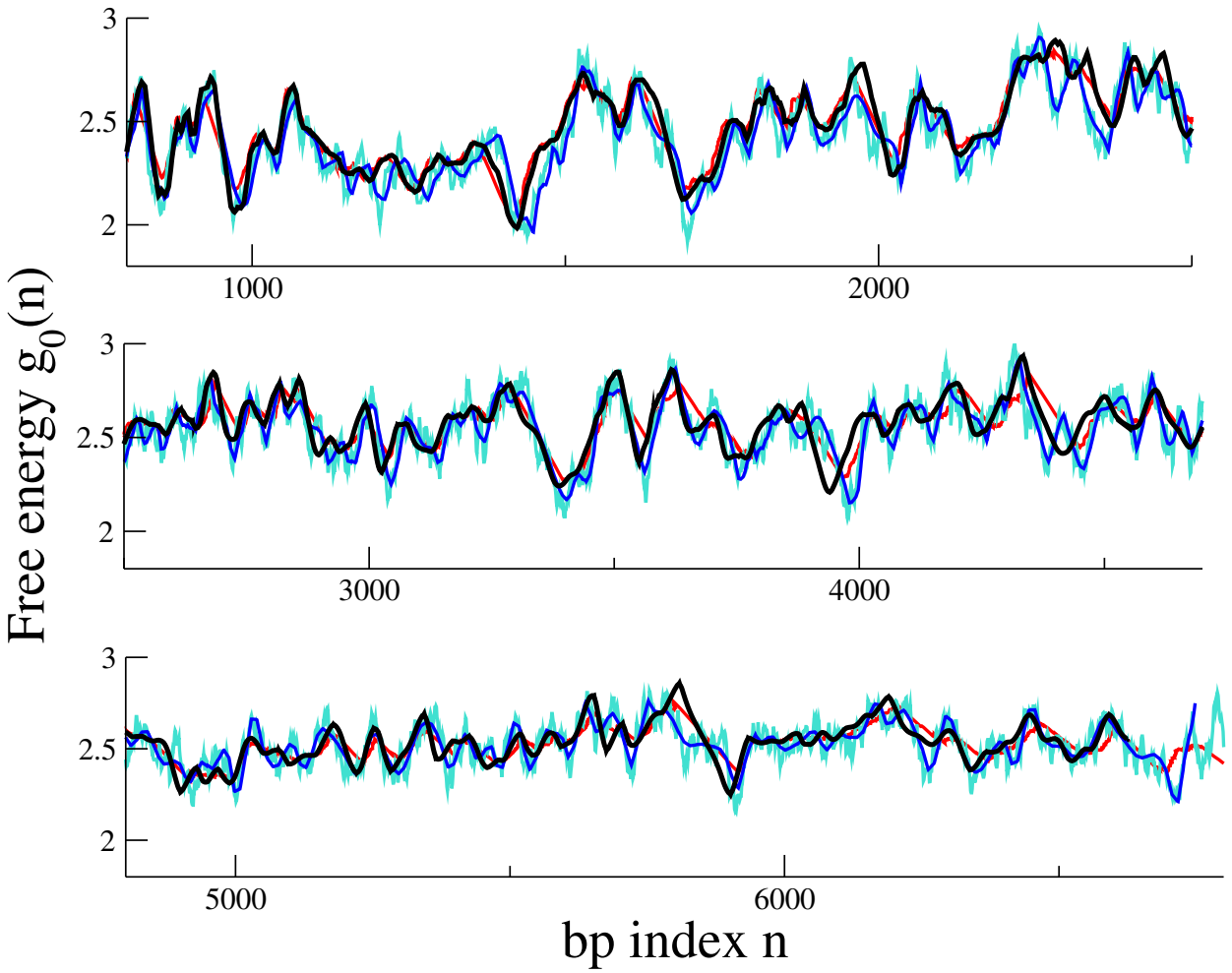

FIG. S17: Inference of the base pair energy from the unzipping force for Molecule 1. Comparison of the SP (red line) and the Box (black line) landscapes with the true free energy (turquoise) (sliding average over 40 bp) and the Box approximation for synthetic data (blue).

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

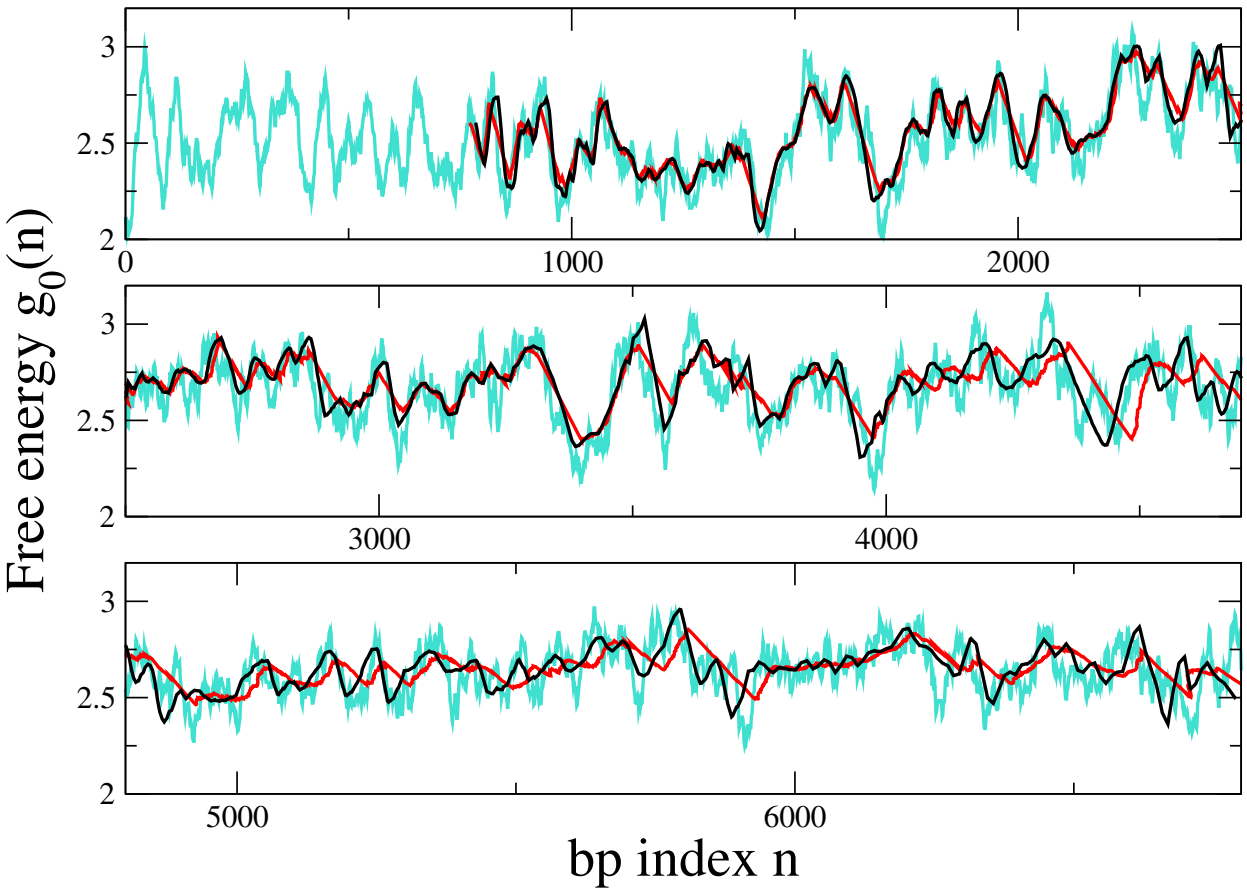

FIG. S18: Inference of the base pair energy from the unzipping force for Molecule 2. Comparison of the SP (red line) and the Box (black line) landscapes with the true free energy (sliding average over 40 bp.)

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

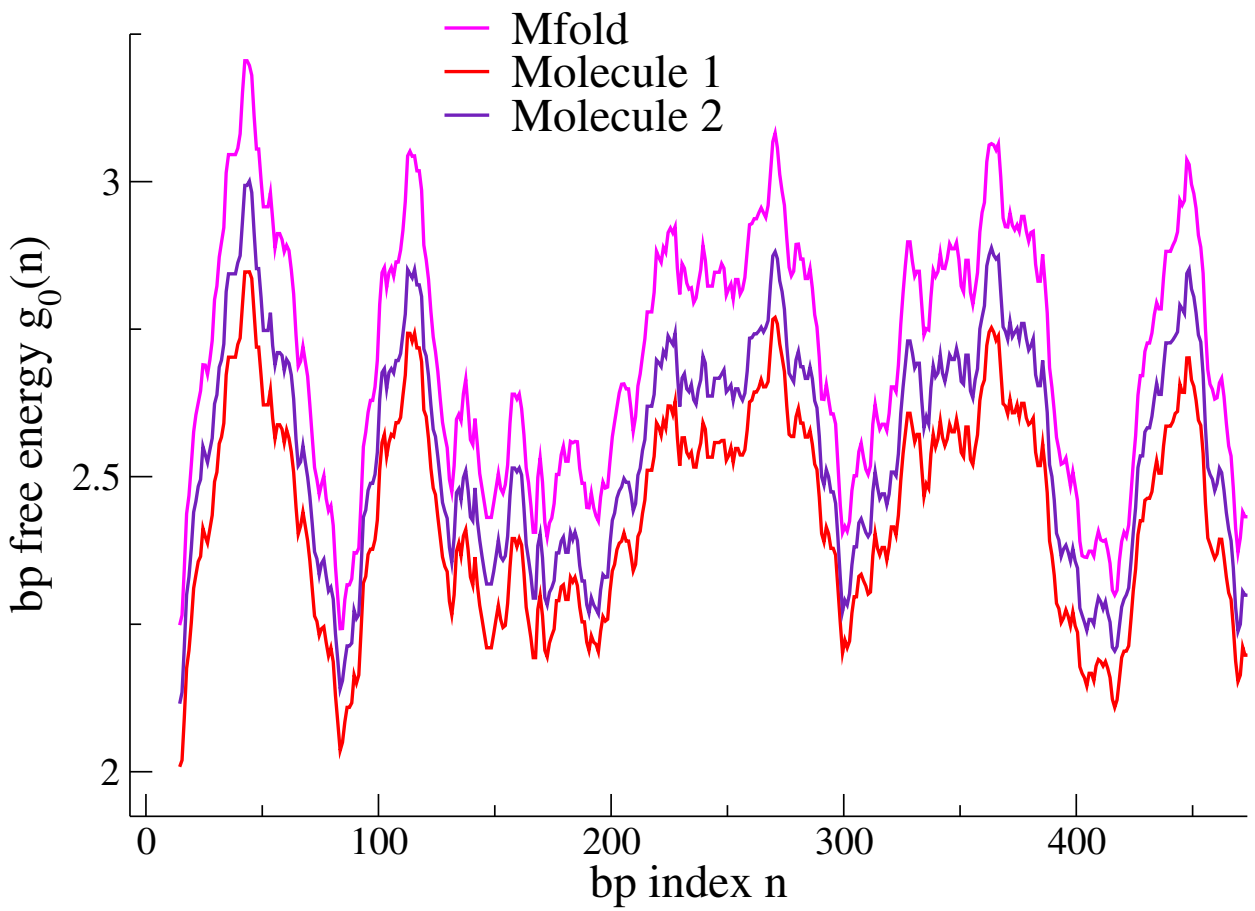

FIG. S19: Base pair free energies along the initial portion of the sequence for the optimal pairing parameters for Molecule 1 and Molecule 2 (extracted from [1]) and for the MFold energetic parameters. The sliding average is computed over $w = 30$ bp.

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

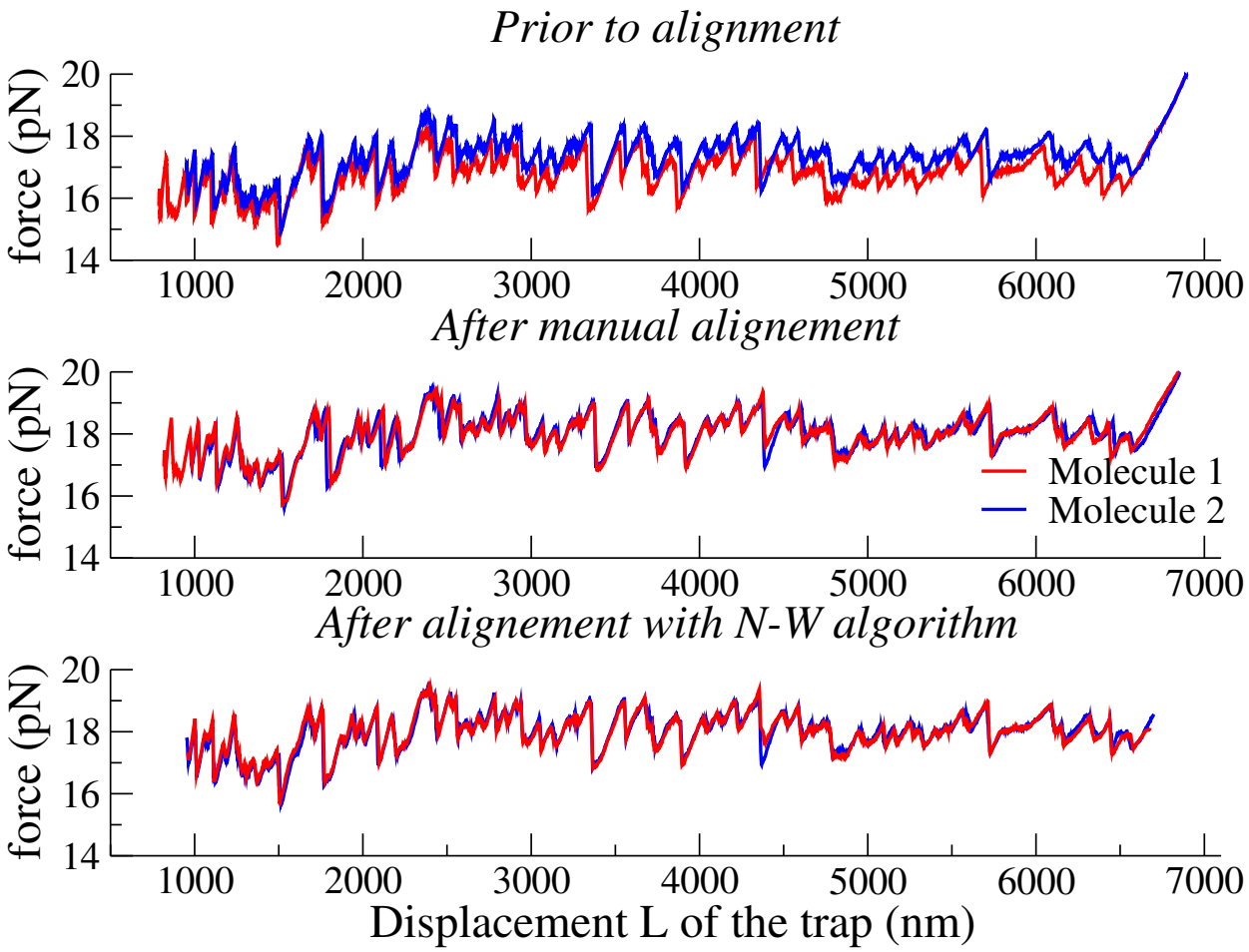

FIG. S20: Top: Experimental force vs. trap displacement for the two unzipping experiments Molecule 1 and Molecule 2, with the corrections to remove the drift done in [1]. Middle: Same data after the manual alignment described in Section 6. Bottom: Same data after alignment with the Needleman-Wunsch (N-W)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

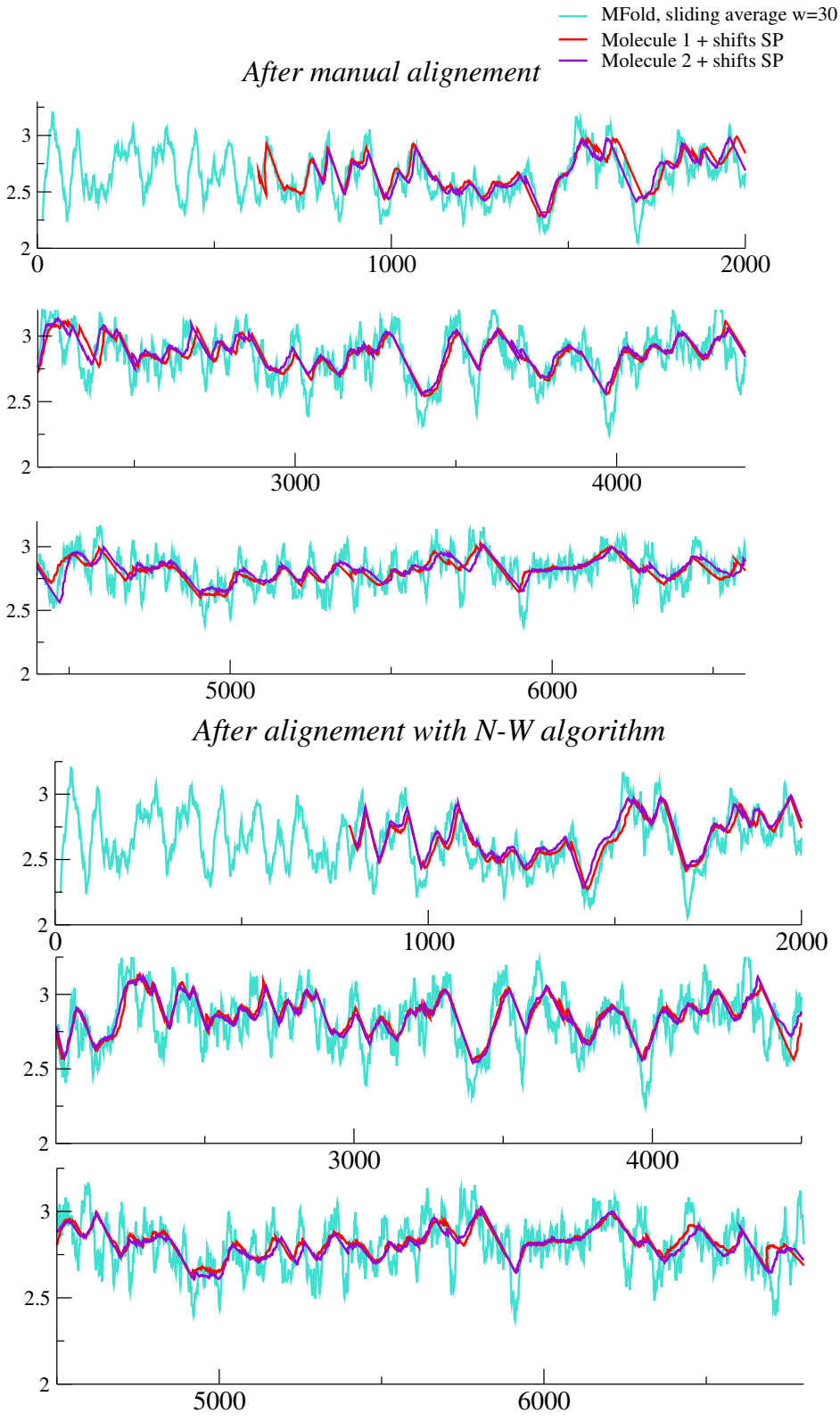

FIG. S21: Saddle point inference for Molecules 1 (red line) and 2 (blue line) using the MFold energetic parameters and after further shifts on the pairing energies and on the trap position. The true sequence landscape (sliding average over $w = 30$ bp) is shown with the turquoise line.

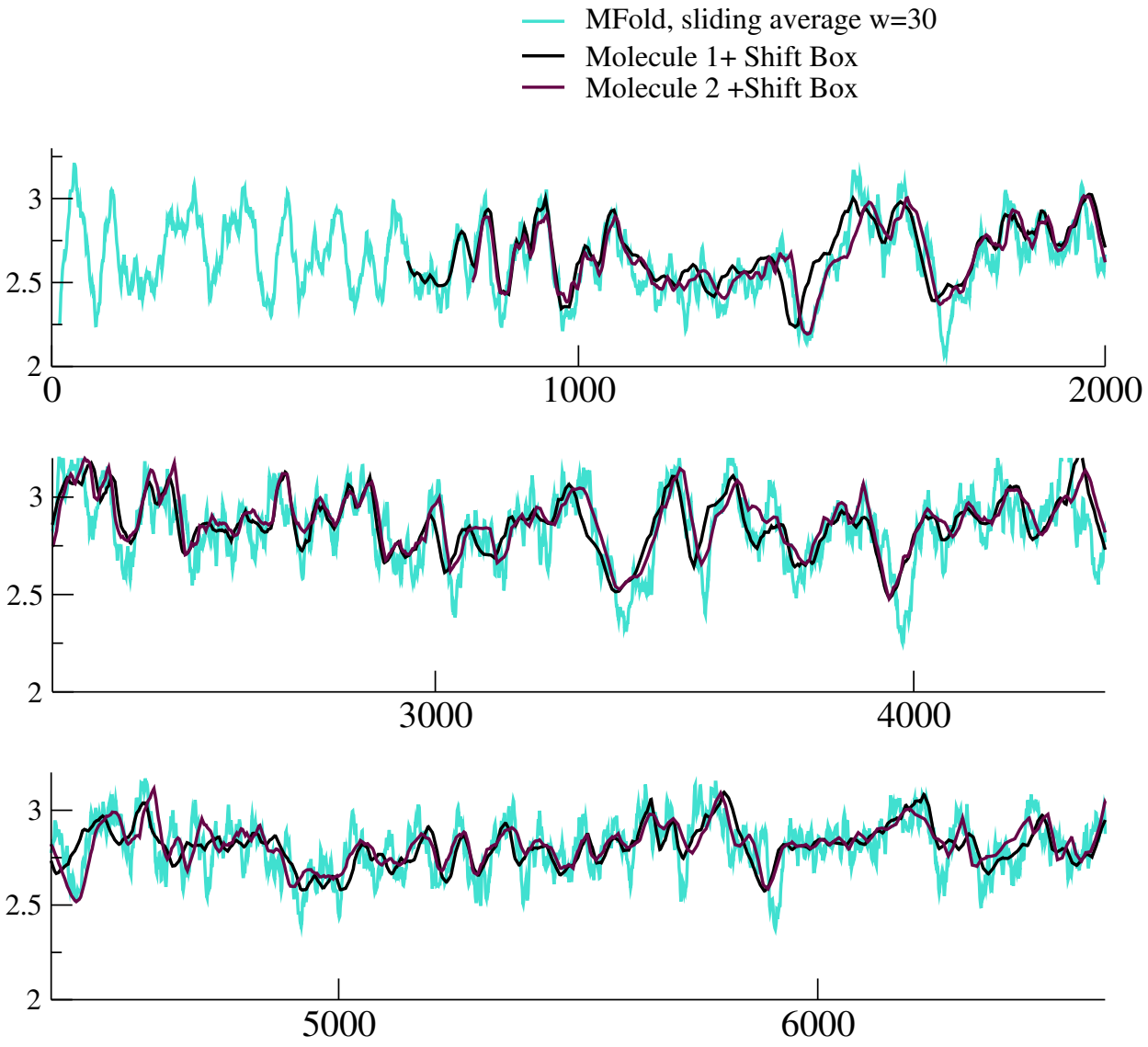

FIG. S22: Box inference for Molecules 1 (red line) and 2 (blue line) from MFold energetic parameters and after global shifts on the force curves and manual alignment on the trap positions. The true sequence landscape (sliding average over $w = 30$ bp) is shown with the turquoise line.

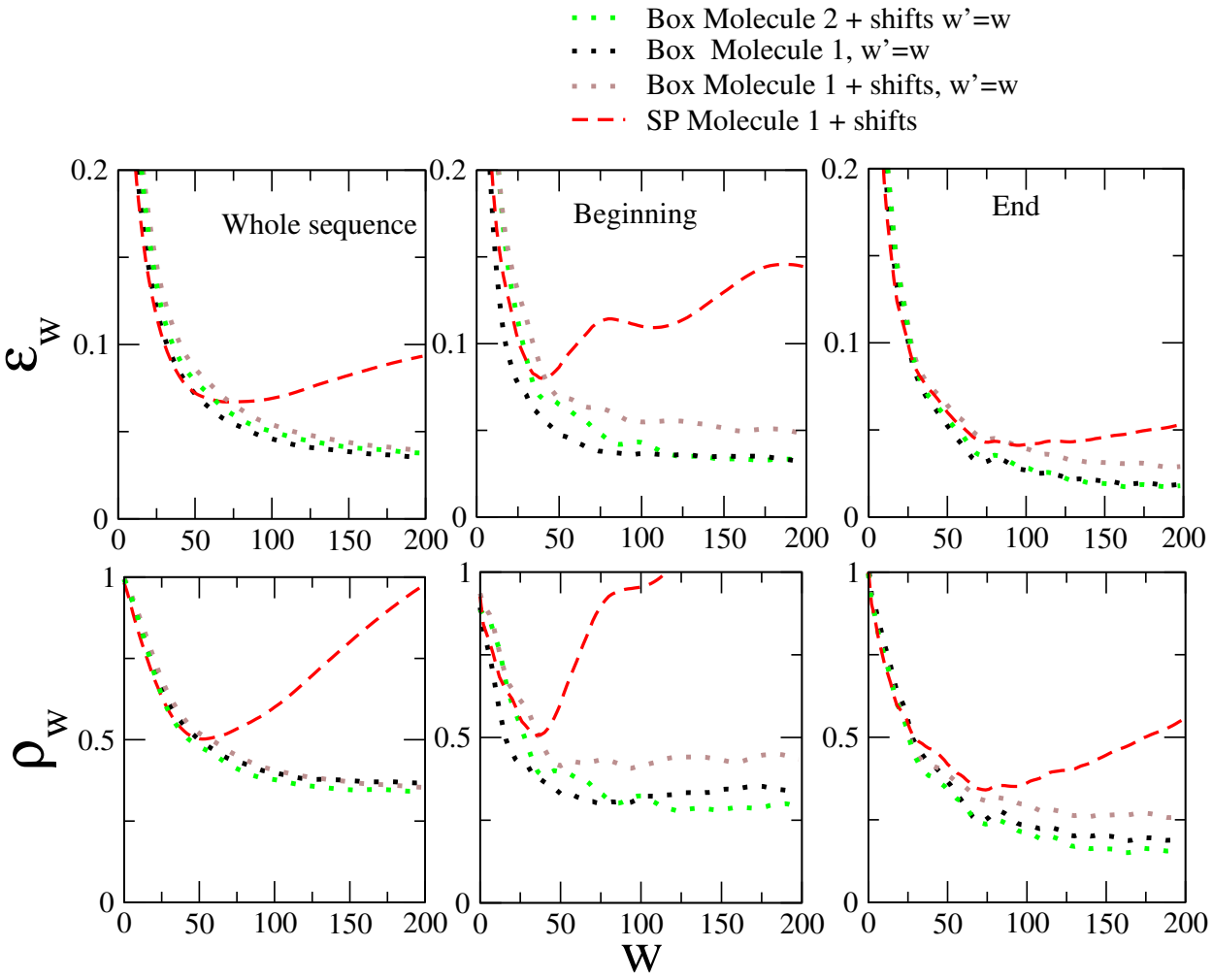

FIG. S23: Absolute (top) and relative (bottom) errors on the inferred free energies as a function of the window size w of the running average, for the whole sequence and for 300 bases at the beginning and the end of the sequence. The data correspond to Molecule 1 (with the free energies found by Huguet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

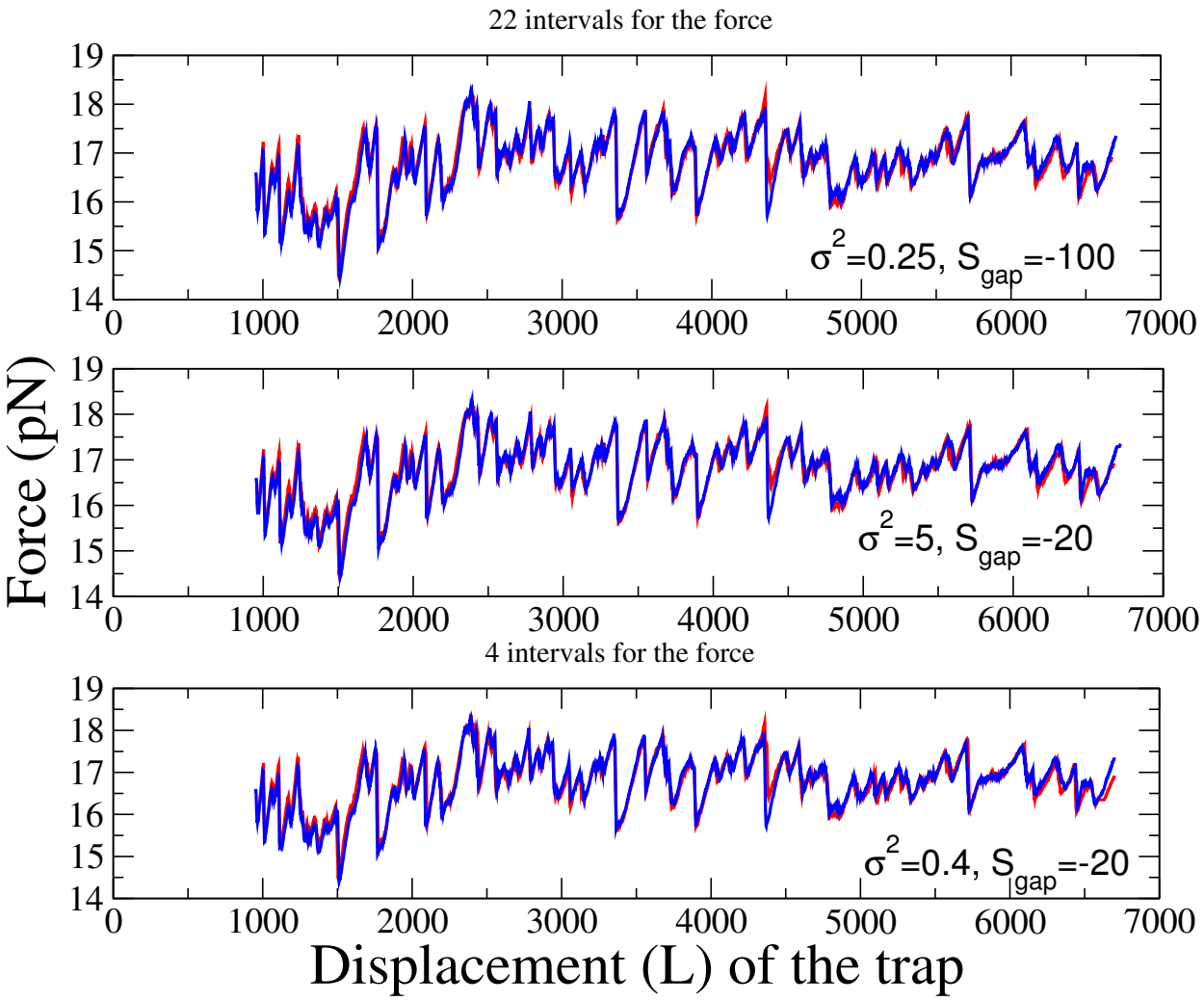

FIG. S24: Re-alignement of the experimental forces with 22 (top & middle panels) and with 4 (bottom panel) force intervales. The values of σ^2 and of the gap penalty S_{gap} are shown in the panels.

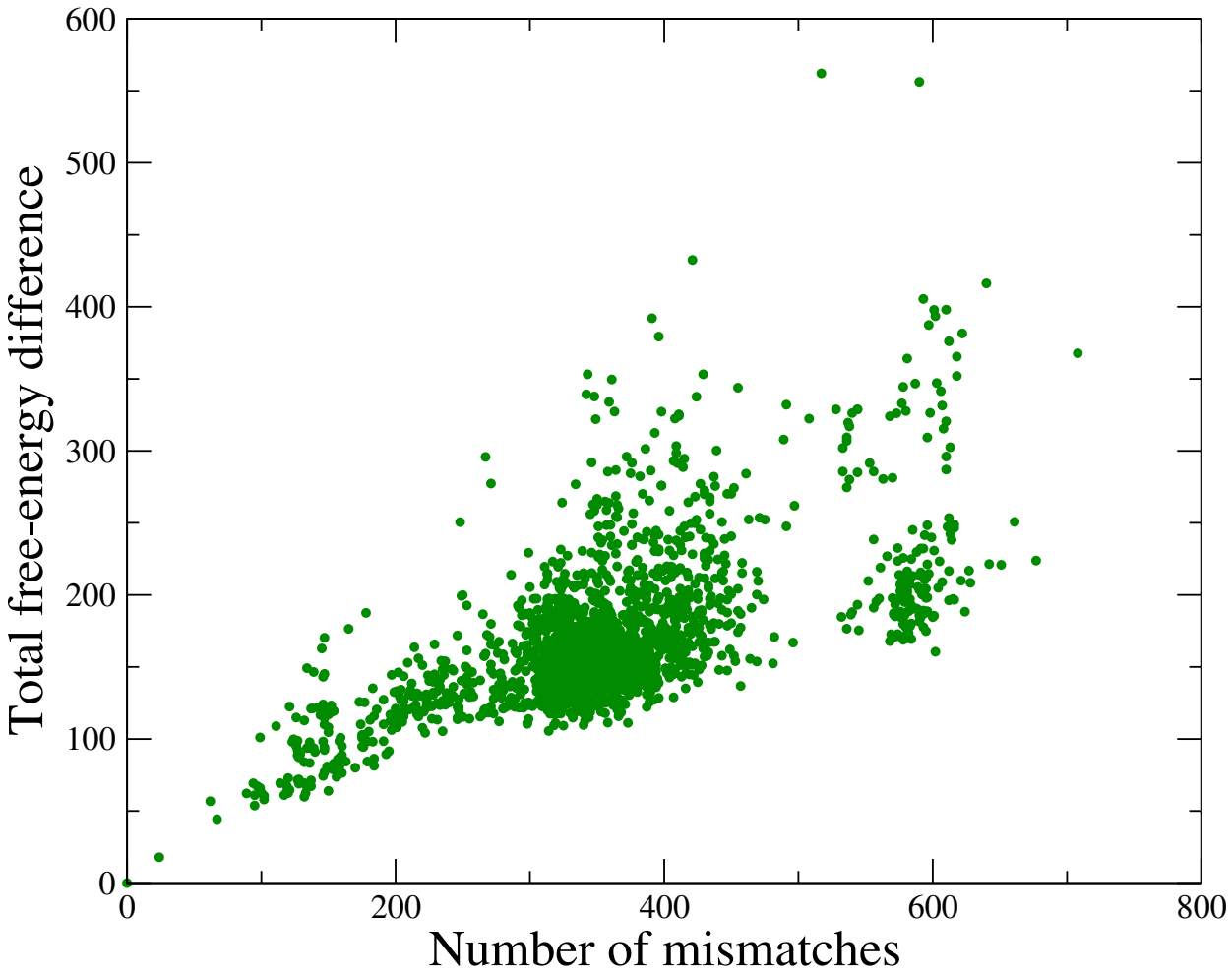

FIG. S25: Total difference Δg in free energy along the aligned sequence vs. number of mismatches, when discretizing the forces with $N_f = 4$ values only. The test sequence is bacterium B-F.

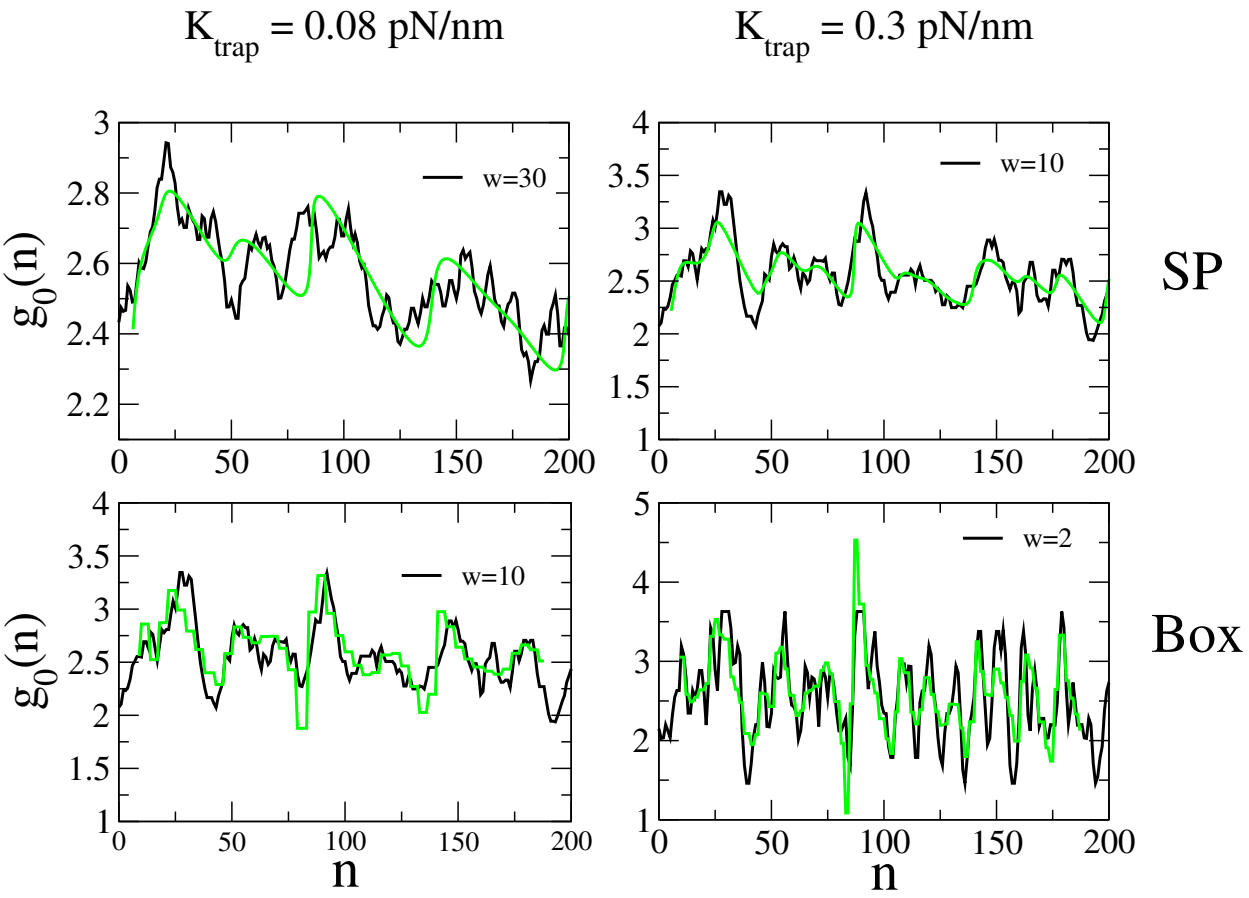

FIG. S26: Free-energy landscape (green curves) on the first 200 base pairs from synthetic unzipping data generated from the sequence of the B-F bacterium, with trap stiffness $K_{trap} = 0.08$ pN/nm (left) and $K_{trap} = 0.3$ pN/nm (right), inferred with the SP (top) and the Box (bottom) approximations.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

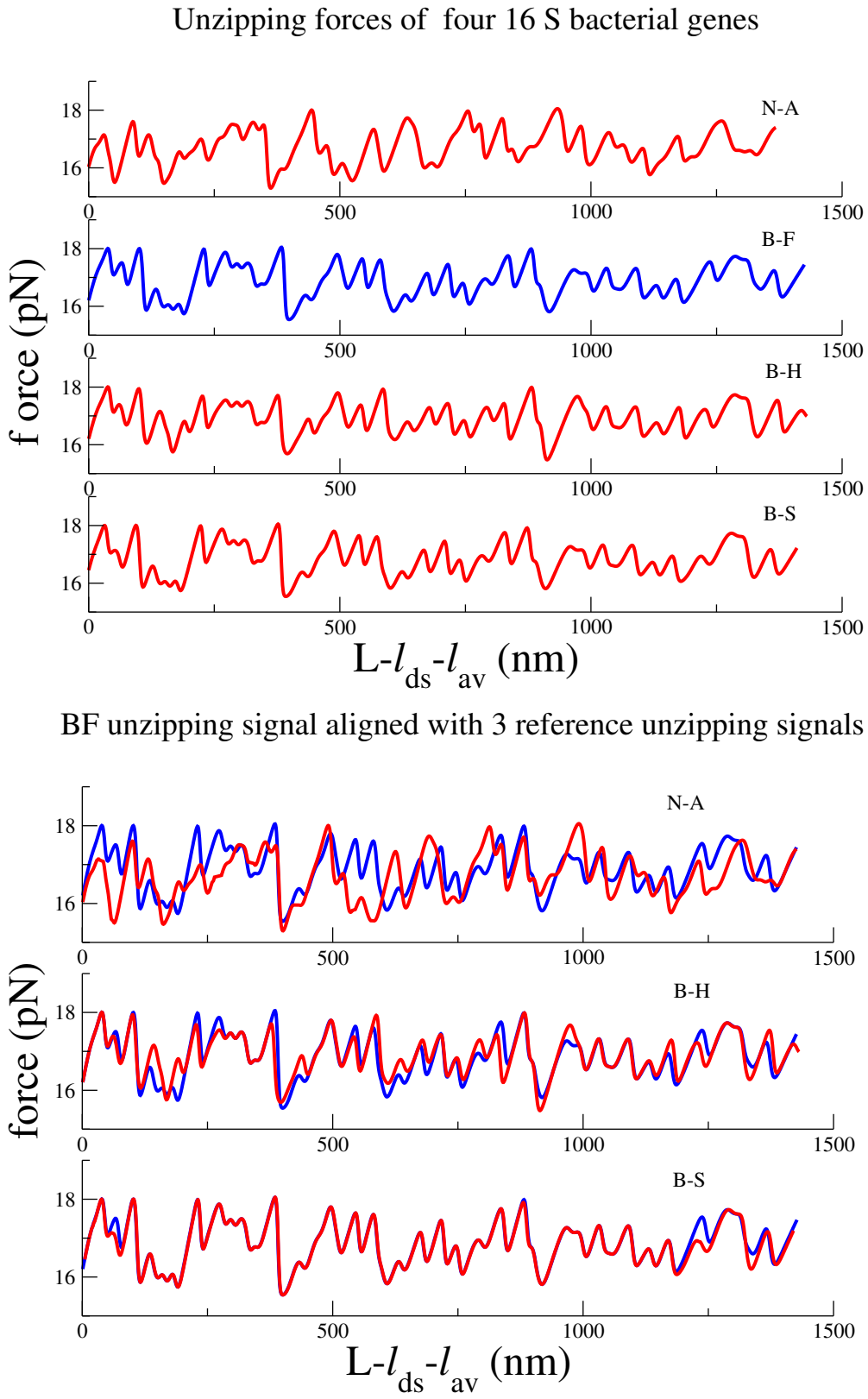

FIG. S27: Top: Unzipping force signals corresponding to N-A, B-F, B-H and B-S bacteria. Bottom: alignment of the B-F unzipping force curve (blue) with the N-A (top), B-H (middle) and B-S (bottom) force curves.

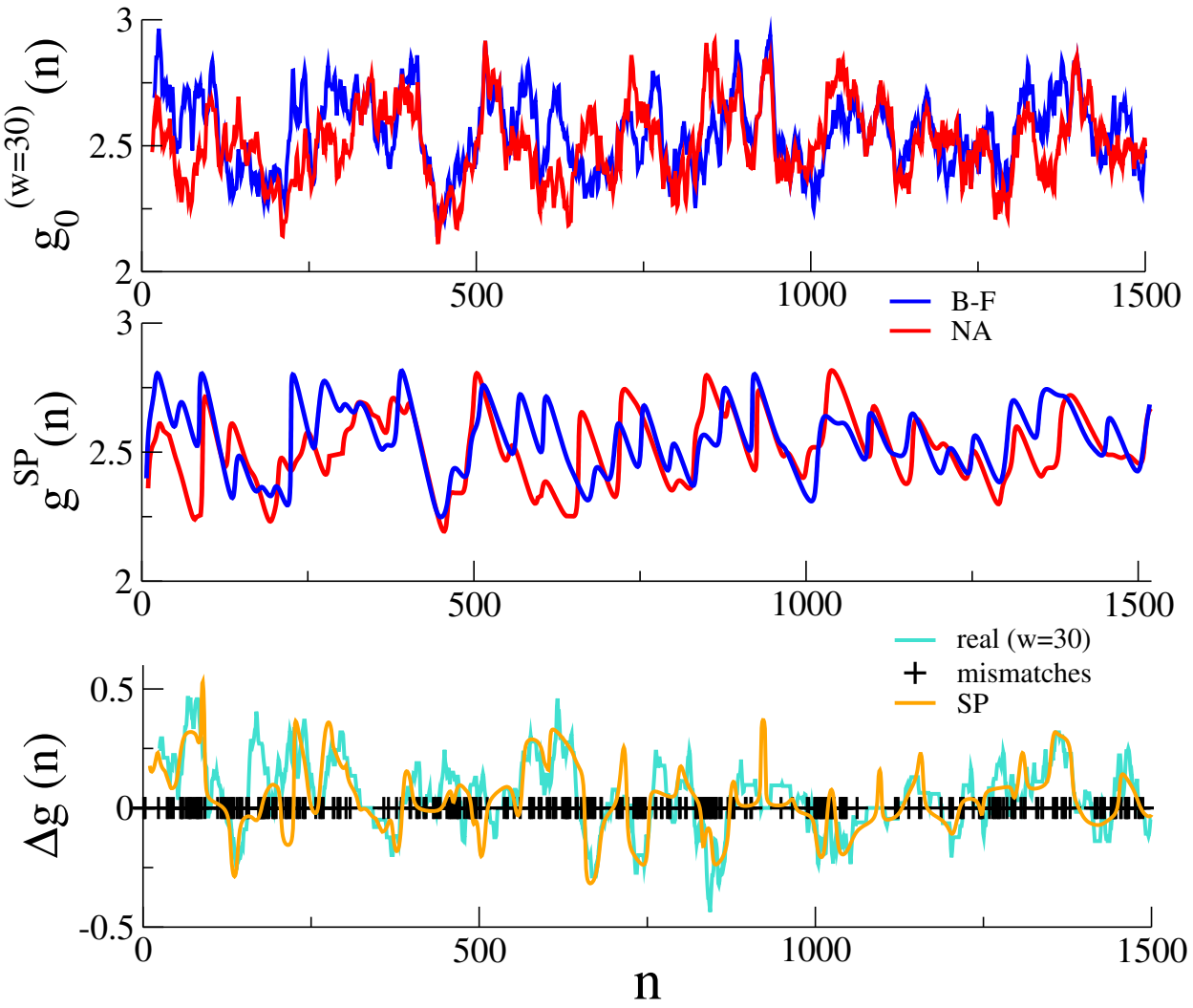

FIG. S28: Comparison of free-energy landscapes for bacteria B-F and N-A. Top: free energy with a sliding average over $w = 30$ base pairs, obtained from the sequences and the pairing parameters of MFold at 150 mM NaCl, after having aligned the two sequences. Middle: inferred SP free-energy landscape 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.

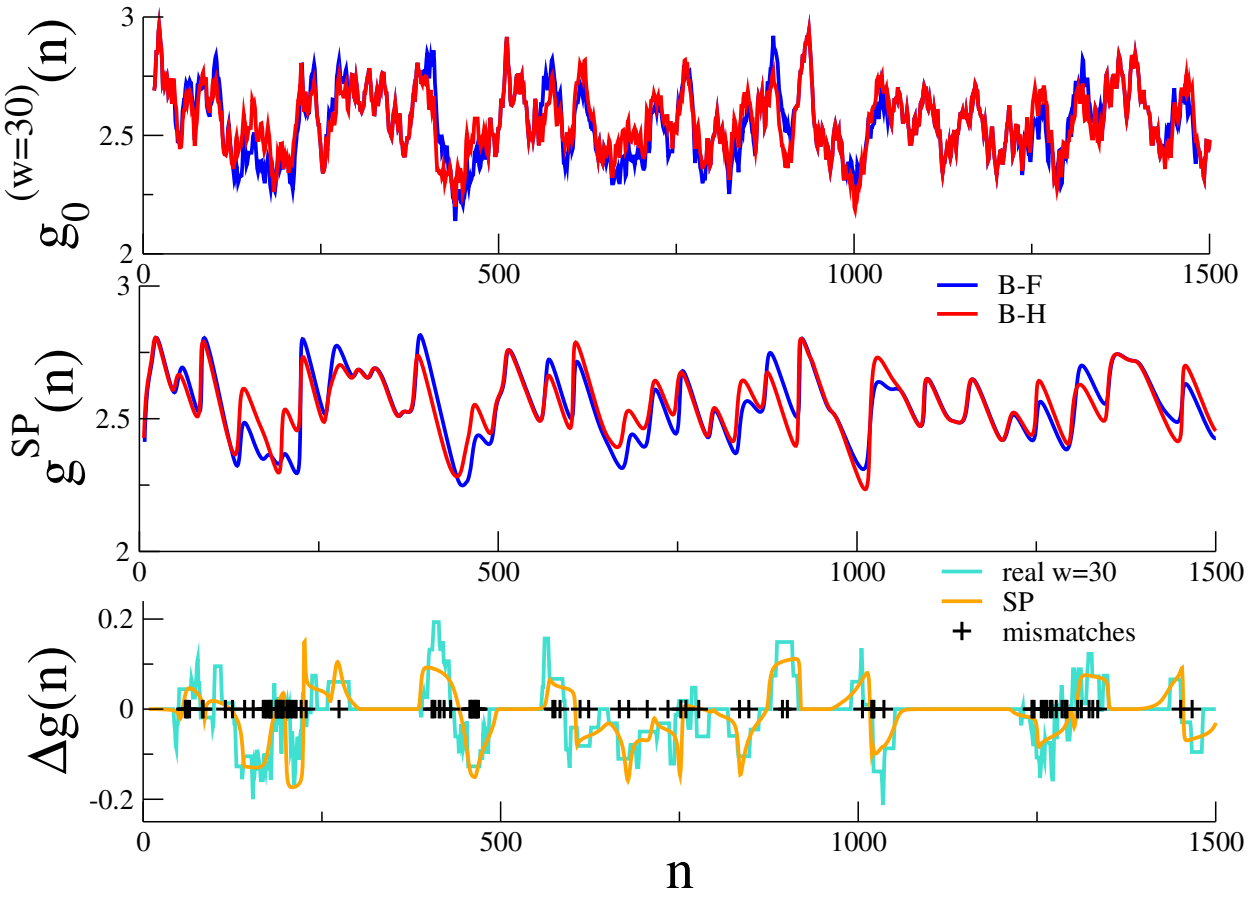

FIG. S29: Comparison of 16S gene of bacteria B-F and B-H from the SP inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 30$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred 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.

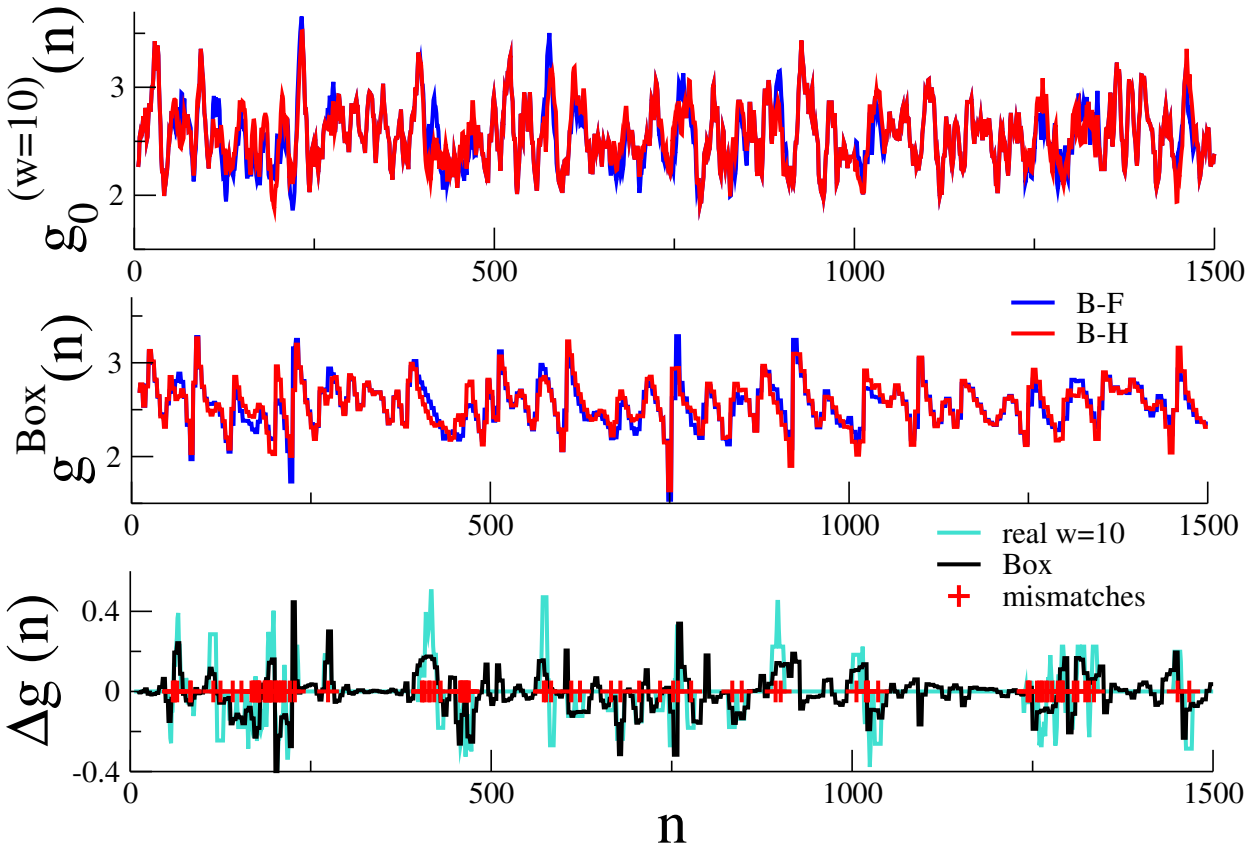

FIG. S30: Comparison of 16S gene of bacteria B-F and B-H from the Box inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 10$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred 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.

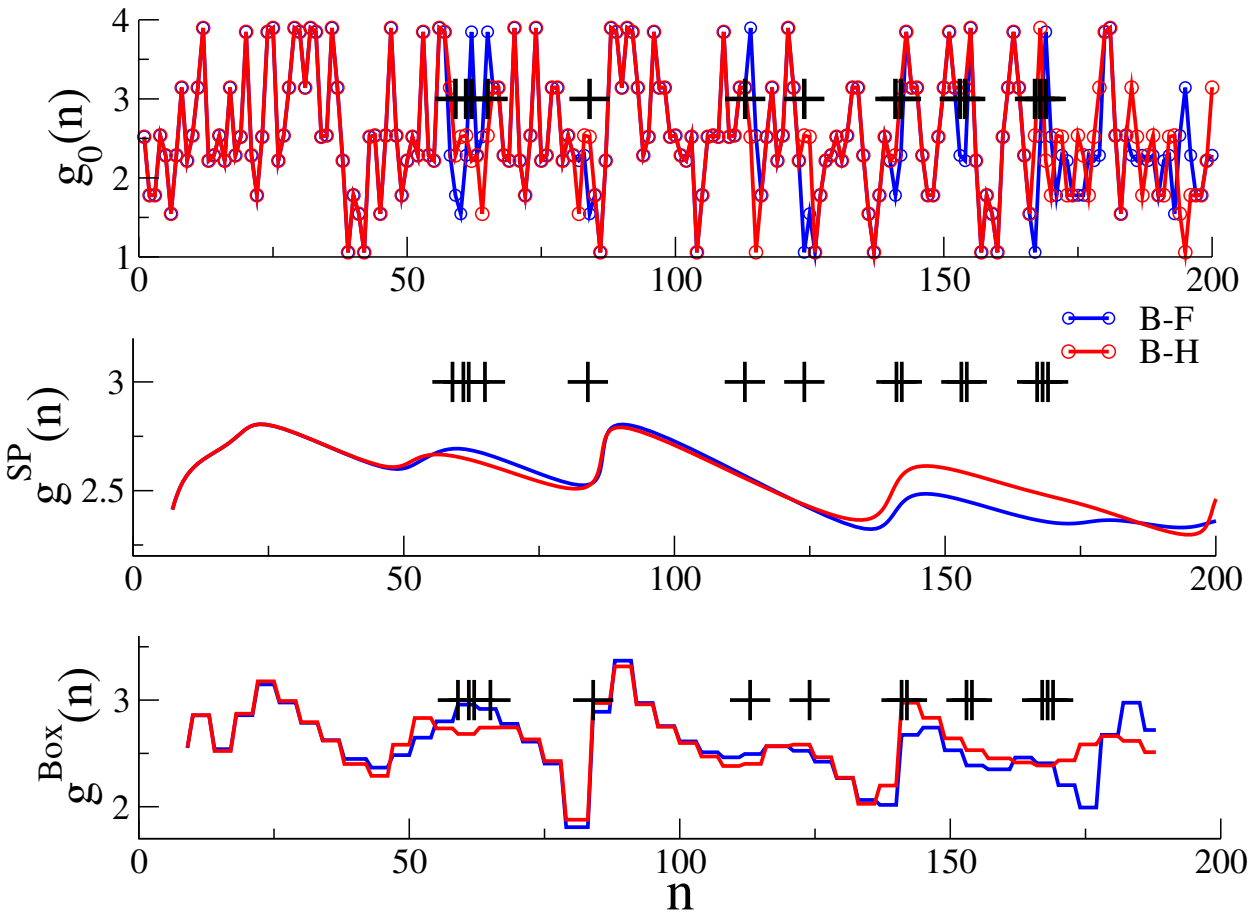

FIG. S31: Trap stifness $K_{trap} = 0.08$ pN/nm. Magnification over the first 200 bases of the sequence: comparison between the real free energy differences (without any sliding average), the SP inference and the Box inference for B-F and B-H bacteria. Dark crosses locate mismatches between the two sequences.

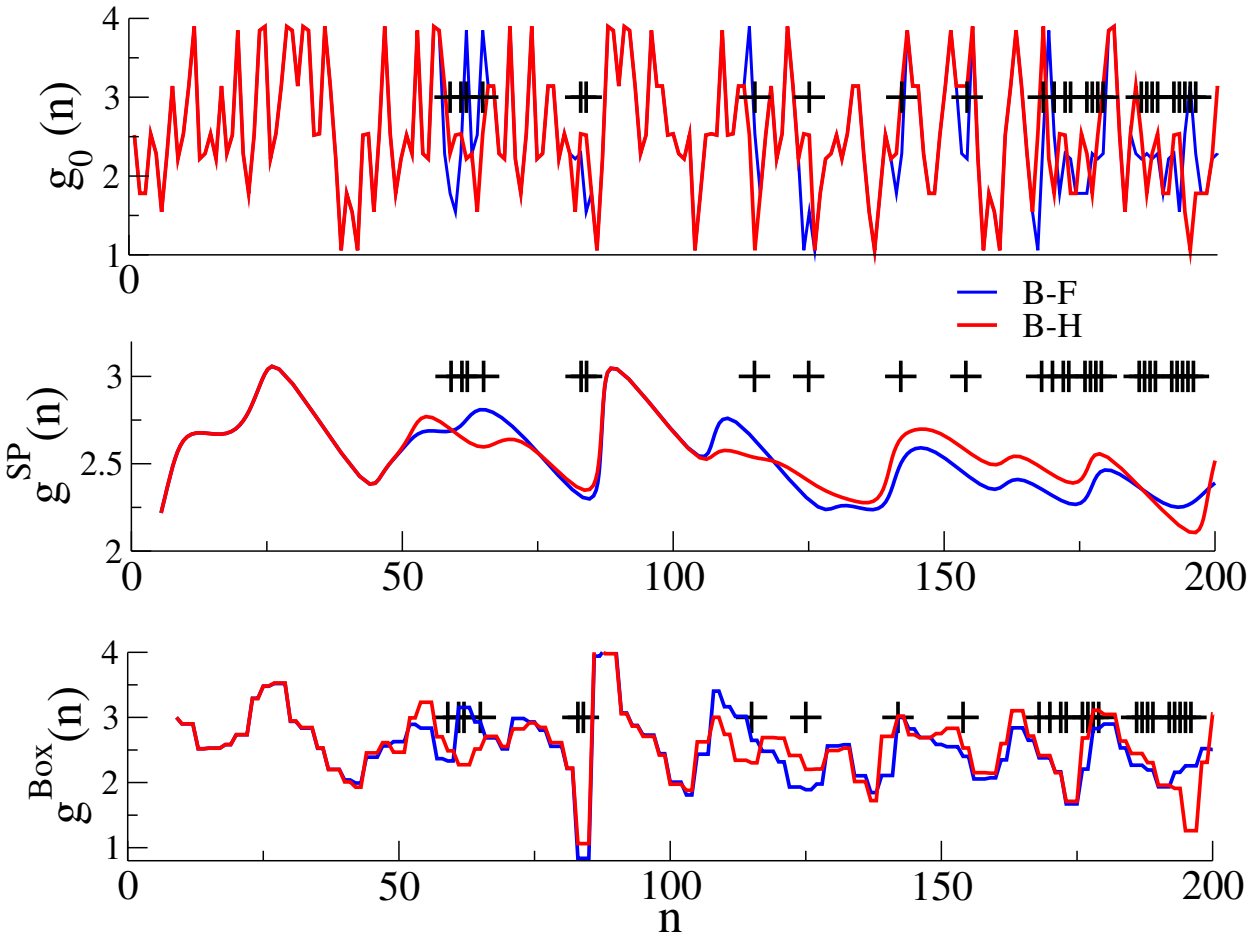

FIG. S32: Trap stifness $K_{trap} = 0.3$ pN/nm. Focus on the first 200 bases of the sequence: comparison between the real free energy differences (without any sliding average),the SP inference and the Box inference for B-F and B-H bacteria. Dark crosses locate mismatches between the two sequences.

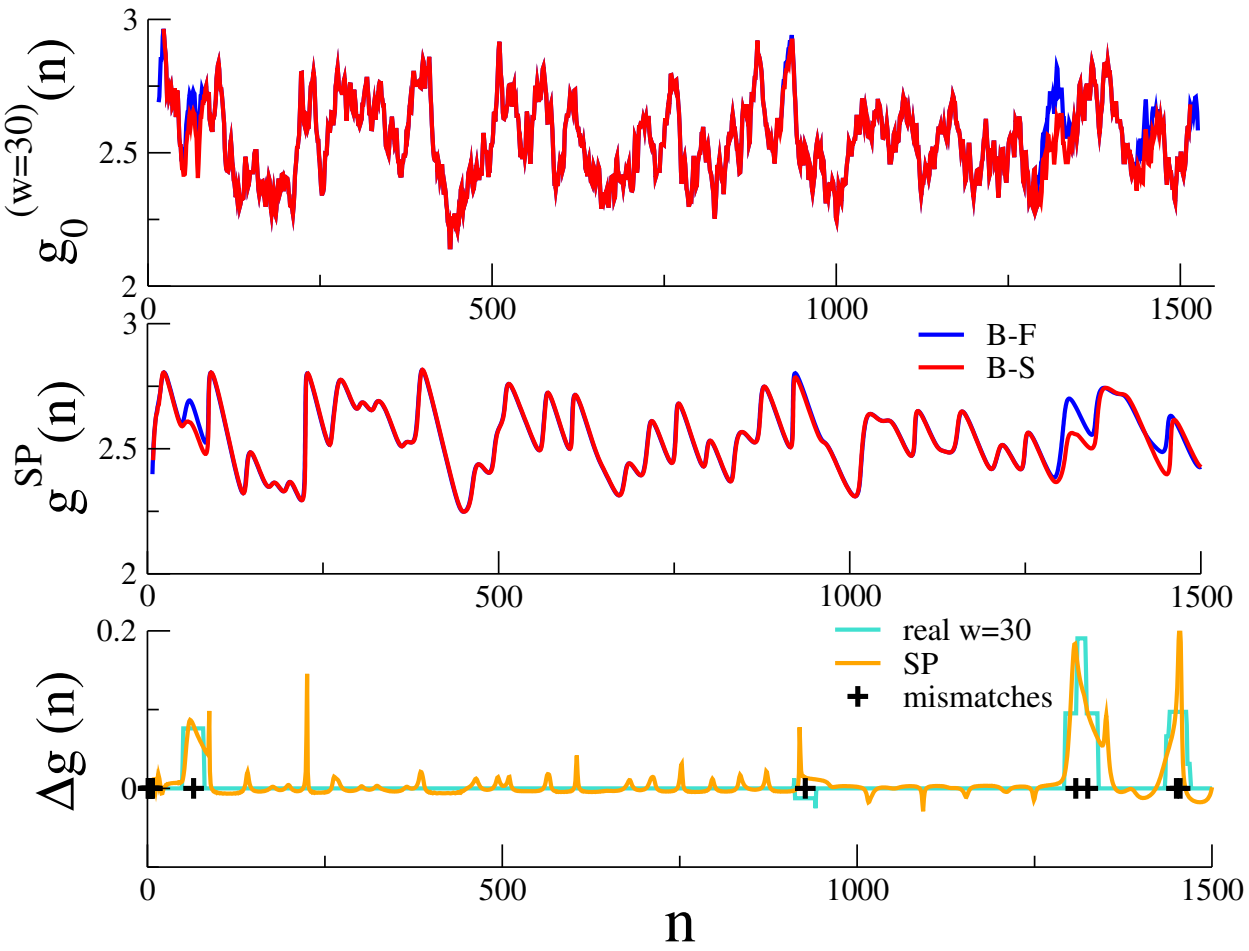

FIG. S33: Comparison of 16S gene of bacteria B-F and B-S from the SP inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 30$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred 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.

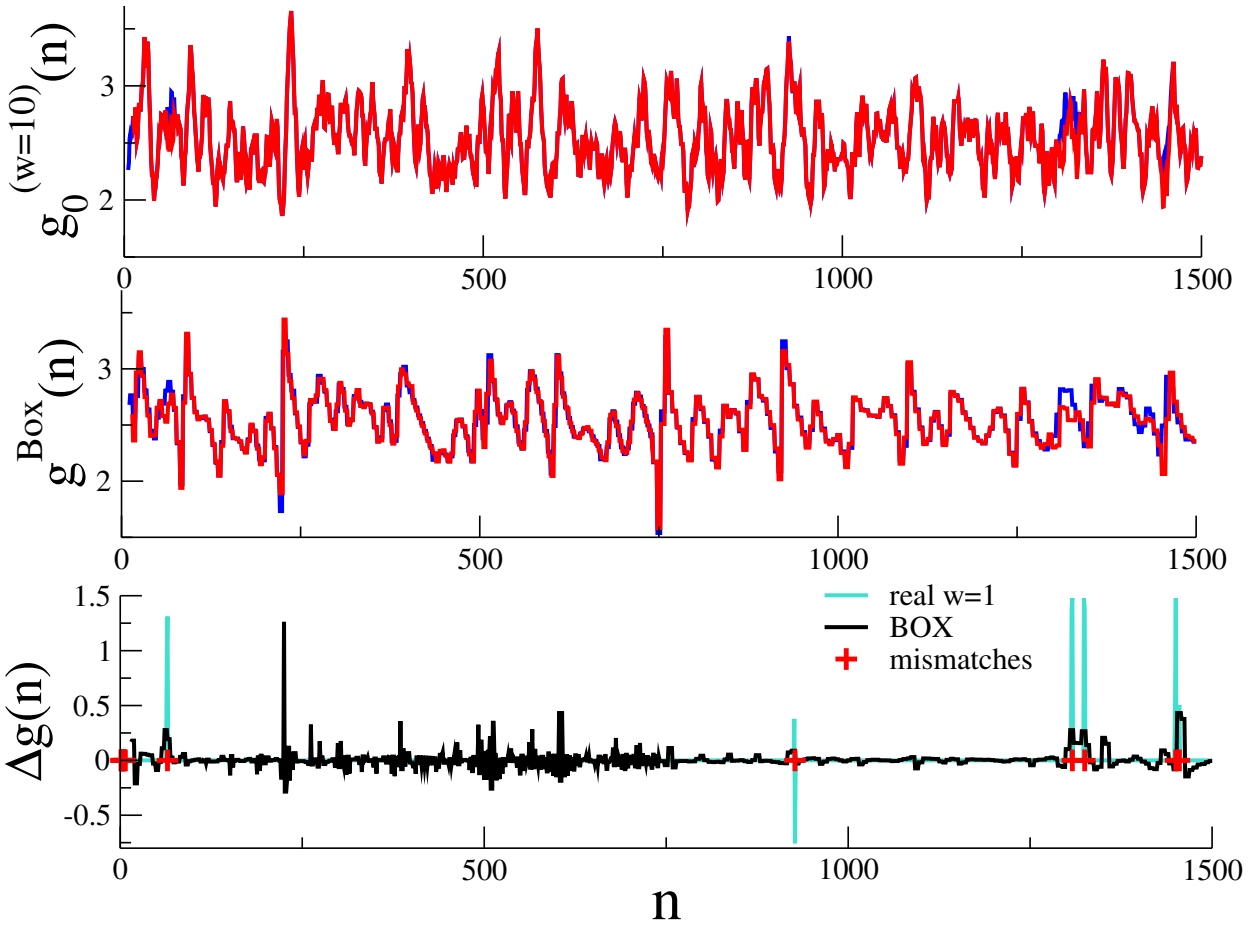

FIG. S34: Comparison of 16S gene of bacteria B-F and B-S from the Box inference on aligned unzipping force signals. Top panel: Pairing free energy with a sliding average over $w = 10$ base pairs, obtained from the aligned sequences and the pairing parameters of Mfold at 150 mM NaCl. Middle panel: inferred 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.

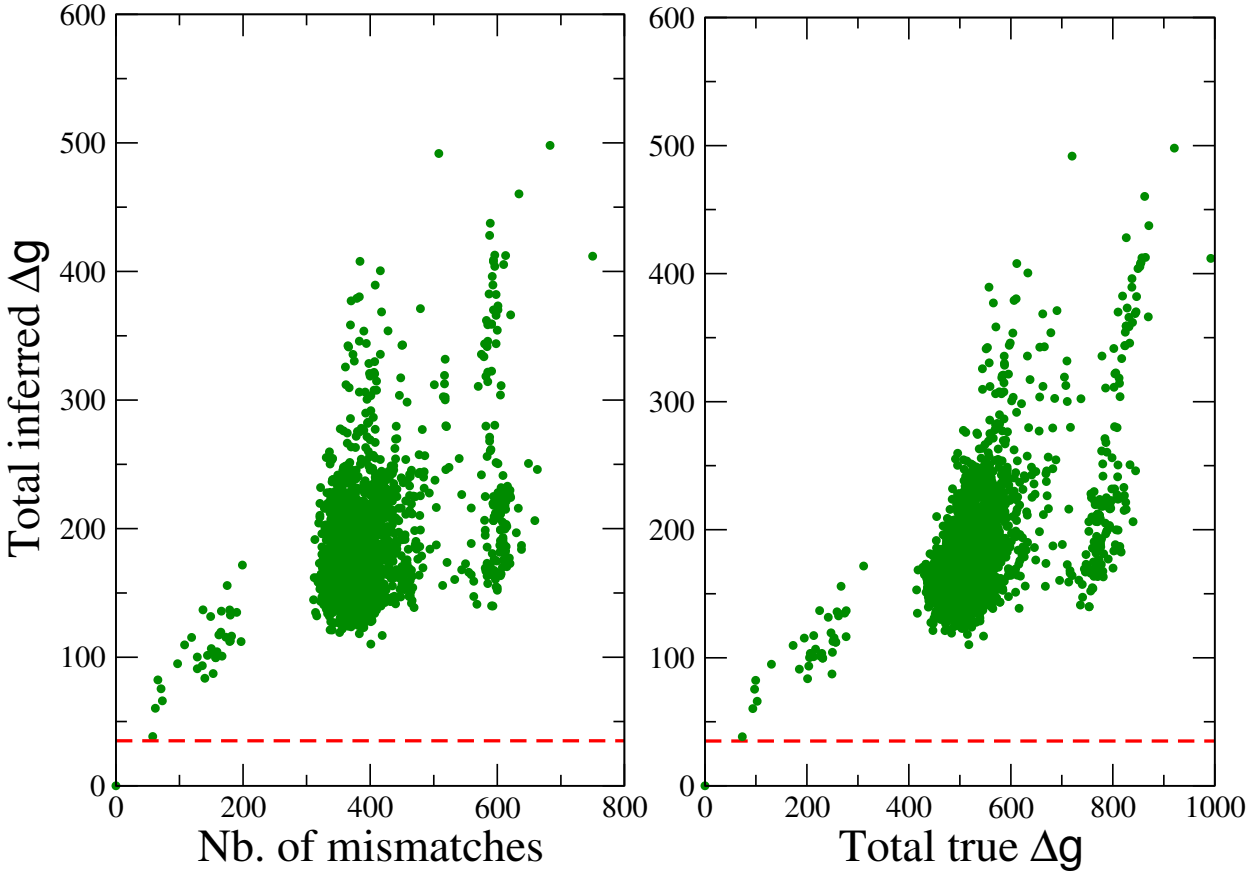

FIG. S35: Comparison of the SP inferred free-energy landscape for bacterium N-A with the other bacteria in the database. Total differences in free energies vs. number of mismatches (left) and vs. the true diferences in free energies along the sequences (right), computed after pairwise alignments.