PNAS
www.pnas.org

# Supplementary Information for

## Stem-loop formation drives RNA folding in mechanical unzipping experiments

**Paolo Rissone, Cristiano V. Bizarro, Felix Ritort**

**F. Ritort**
**E-mail: ritort@ub.edu**

**This PDF file includes:**

## Supporting Information Text

**Hairpin synthesis.** We synthesized an RNA hairpin made of a stem of 2027 equally represented canonical Watson-Crick base pairs (Table S2), ending in a GAAA tetraloop and inserted between short hybrid DNA/RNA handles (29bp). The synthesis of long RNA hairpins is a challenging task (see main text) and required the development of a suitable protocol which can be split into 7 main steps and is schematically depicted in Fig.S1A,B.

***PCR amplification of target sequence.*** The target sequence was selected inside a $\lambda$-DNA (J02459) region (30286 – 38650) that was previously shown to be efficiently transcribed in both strands by T7 RNA polymerase, the same enzyme used in our experiments (1). A PCR reaction was performed to obtain an amplicon of 2027bp in length using 1$\mu$M of Univ_hairpin_F and 1$\mu$M of EcoRI_2.0kb_R primers (Table S1), 25ng of $\lambda$-DNA (Dam-) as DNA template, 1.5mM $MgCl_2$, 1X Opti and 1X HiSpec buffers, dNTPs 0.2mM each, and 4U of Eco Taq Plus DNA Polymerase (Ecogen). Primer sequences were selected using Primer3Plus software (2). EcoRI sites were added at 5′ termini of both primers (Table S1, sequence in bold). Cycling parameters were as follow: initial denaturation step (94°C) for 1 min 30 sec, enzyme addition (hotstart), 30 cycles of denaturation at 94°C for 45 sec, annealing at 60°C for 1 min and extension at 72°C for 6 min, with a final extension step at 72°C for 7 min.

***Synthesis of pRNA1 and pRNA2 constructs.*** The 2kbp PCR amplicon was digested with EcoRI (NEB, New England Biolabs) and cloned into vector pBR322/EcoRI (3, 4). Plasmid DNA was purified using Illustra PlasmidPrep Mini Spin Kit (GE Healthcare). The insert orientation was evaluated by digesting plasmids with EcoRV and according to it constructs were defined as pRNA1 or pRNA2 (Fig.S1). These constructs were used as templates for PCR reactions using primers RNA1_T7Forw and RNA1_Rev (pRNA1) or RNA2_T7Forw and RNA2_Rev (pRNA2) (Table S1).

***PCR amplification of templates for in vitro transcription.*** The pRNA1 and pRNA2 constructs were used as templates for PCR reactions using primers RNA1_T7Forw and RNA1_Rev (pRNA1) or RNA2_T7Forw and RNA2_Rev (pRNA2) (see Table S1). RNA_T7Forw primer contains a cytidilate nucleotide (in bold) upstream the minimal T7 RNA Polymerase Promoter (5′-ctaatacgactcactataggga-3′) to improve transcription efficiency (5), followed by a pBR322-annealing sequence (5′-ataaaaataggcgtatcacgag-3′). This sequence codes for part of RNA1 handle. Primer RNA1_Rev anneals at its 3′ termini (5′-gaaaaacgcctcgagtgaag-3′) with the Univ_hairpin_F binding site located at the end of pRNA1 insert opposite to RNA1_T7Forw binding site (see Fig.S1). The 5′ termini of RNA1_Rev (5′-ctcatctgtttccagatgag-3′) codes for the last 8bp of the RNA hairpin near the loop and the GAAA tetraloop itself (in bold, reverse complement). The sequence 5′-ggga-3′ was introduced between 5′ and 3′ portions of RNA1_Rev in order to introduce the sequence 5′-uccc-3′ into RNA1. This tetranucleotide RNA sequence is located between the hairpin stem portion formed by sequences from the 2kbp insert and the last 8bp-stem and loop regions coded by RNA1_Rev sequence, and serve to base pair the first four nucleotides (5′-ggga-3′) at the 5′ end of RNA2 molecule. RNA2_T7Forw anneals with same sequence that pairs with RNA1_Rev primer (5′-gaaaaacgcctcgagtgaag-3′), but it is used in PCR reactions with pRNA2 construct. As in the case of RNA1_T7Forw, its 5′ termini contain an optimized T7 promoter containing an upstream "c" nucleotide (5′-ctaatacgactcactataggga-3′). RNA2_Rev primer contains an pBR322-annealing region and codes for the RNA2 handle. PCR reactions were performed using the same conditions previously described. Amplification products were purified from PCR mixtures using the GFX PCR DNA and Gel Band Purification Kit (GE Healthcare).

***In vitro transcription of RNA1 and RNA2 molecules.*** *in vitro* transcription reactions were performed using the T7 MEGAscript-High Yield transcription Kit (ThermoFisher Scientific/Ambion) according to manufacturer's recommendations. Samples were incubated with 3$\mu$L of TURBO DNase (2U/$\mu$L) for 15 min at 37°C, and synthesized RNA was precipitated by addition of 90$\mu$L of LiCl Precipitation solution (7.5M lithium chloride, 50mM EDTA). Reactions were incubated overnight at -20°C, centrifuged for 15 min at 13,000 rpm, washed twice with 70% ethanol and resuspended in 15$\mu$L of nuclease-free water.

***Treatment of RNA2 molecules with Calf Intestinal Phosphatase (CIAP) and Polynucleotide Kinase (PNK).*** *in vitro* transcribed RNA 2 molecules were treated with 1U of CIAP (Roche) for 1h at 50°C to remove their 5′ triphosphate ends. Dephosphorylated RNA2 molecules were treated with Polynucleotide Kinase (PNK) to produce RNA molecules containing 5′ monophosphate termini according to manufacturer's recommendations. The reactions were heat-inactivated by incubating for 20 min at 65°C and precipitated with LiCl as described in the following section.

***Digoxigenin 3′ tailing of S Handle A oligonucleotide.*** S Handle A (Table S1) tailing with digoxigenins were performed by using the DIG oligonucleotide Tailing Kit 2nd Generation (Roche), according to manufacturer's recommendations. DIG-labeled S Handle A was purified by using the Qiaquick Nucleotide Removal kit (Qiagen).

***Assembling RNA1 and RNA2 molecules to form the 2kbp RNA hairpin.*** The assembly of RNA hairpin was performed in one annealing step, where RNA1, CIAP and PNK-treated RNA2, S Handle A and biotin-labeled S Handle B2 oligonucleotides (Table S1) were incubated together. A total of 20$\mu$g of RNA1 and 20$\mu$g of CIAP/PNK-treated RNA2 were incubated with 5$\mu$L of DIG-tailed S Handle A (2$\mu$M), 5$\mu$L of 5′-Bio-S Handle B2 (2$\mu$M), 2$\mu$L of Tris 1M, pH 7.0, 2$\mu$L of NaCl 5M and water to a final volume of 80$\mu$L. Reactions were incubated for 1 h at 65°C and cooled to 10°C at a rate of 0.5°C/min using a thermocycler. After a final cooling step at 10°C for 1 h and 30 min, the samples were subjected to microdialysis. The annealing reaction was pipetted over a 0.05$\mu$m Millipore membrane which was put in a plate containing 50mL of 20mM Tris.HCl, 5mM NaCl, pH 7.5, and allowed to stay for 1 h. Microdialyzed, annealed molecules were then incubated with 1$\mu$L of T4 RNA ligase 2 (RNL2 1U/$\mu$L) (NEB) and 1X RNL2 Reaction Buffer for 2 h at 37°C to covalently join RNA1 and RNA2 molecules. Ligated RNA hairpin

molecules were microdialyzed against Tris.HCL 100mM, EDTA 1mM as described above and stored at -20°C or directly used in single-molecule experiments.

### Recovery of the equilibrium FDC

To recover the equilibrium FDC from the irreversible experimental data, we developed an approach based on the extended fluctuation relations (EFR). Here we report in detail the implementation of this method.

Let us first recall the notation and the relations introduced in the main text. Given an irreversible region of the FDC limited by starting (left, $L$) and ending (right, $R$) equilibrated states, let $\mathcal{S}$ be the set of all the states in that region $\mathcal{S} = (I_0 = L, I_1, I_2 \ldots, I_P, I_{P+1} = R)$ being $(I_1, \ldots, I_P)$ the partially equilibrated intermediates. During the experimental forward process (F) the trap position $\lambda$ is moved at a constant speed starting in $I_0$ at $\lambda_0$ and ending in $I_p$ at $\lambda$. Similarly, in the time-reversed protocol (R) the trap position is moved back at the same speed starting in $I_p$ at $\lambda$ and ending in $I_0$ at $\lambda_0$. Thus, the EFR reads

$$\frac{\phi_F^{I_0\to I_p}}{\phi_R^{I_p\to I_0}} \frac{P_F^{I_0\to I_p}(W)}{P_R^{I_p\to I_0}(-W)} = \exp\left[\frac{W - \Delta G_{I_0 I_p}(\lambda)}{k_B T}\right], \qquad [1]$$

where $P_F^{I_0\to I_p}(W)$ ($P_R^{I_p\to I_0}(-W)$) is the partial distribution of the work $W$ measured along the F (R) protocol, $\Delta G_{I_0 I_p}(\lambda) = G_{I_p}(\lambda) - G_{I_0}(\lambda_0)$ is the free-energy difference between states $I_p$ at $\lambda$ and $I_0$ at $\lambda_0$ and $\phi_F^{I_0\to I_p}$ ($\phi_R^{I_p\to I_0}$) is the fraction of paths starting in $I_0$ ($I_p$) at $\lambda_0$ ($\lambda$) and ending in $I_p$ ($I_0$) at $\lambda$ ($\lambda_0$). $k_B$ is the Boltzmann constant and $T$ is the environment temperature.

Let us now introduce the extended Bennett acceptance ratio (EBAR) method (6, 7). By multiplying Eq. (1) by the function $f(W) = \left(1 + \frac{\phi_R}{\phi_F}\frac{n_F}{n_R}\exp\frac{W-\Delta G_{I_0 I_p}}{k_B T}\right)^{-1}$ and integrating over the work one gets

$$\frac{u}{k_B T} = -\log\left(\frac{\phi_F^{I_0\to I_p}}{\phi_R^{I_p\to I_0}}\right) + z_R(u) - z_F(u)\,, \qquad [2]$$

where the variance of the free-energy estimator is minimized by the equations

$$z_R(u) = \log\left\langle \frac{\exp -\frac{W_i^R}{k_B T}}{1 + \frac{\phi_R}{\phi_F}\frac{n_F}{n_R}\exp -\frac{W_i^R + u}{k_B T}} \right\rangle_R = \log\frac{1}{n_R}\sum_{i=1}^{n_R}\left(\frac{\exp -\frac{W_i^R}{k_B T}}{1 + \frac{\phi_R}{\phi_F}\frac{n_F}{n_R}\exp -\frac{W_i^R + u}{k_B T}}\right) \qquad [3a]$$

$$z_F(u) = \log\left\langle \frac{1}{1 + \frac{\phi_R}{\phi_F}\frac{n_F}{n_R}\exp\frac{W_i^F - u}{k_B T}} \right\rangle_F = \log\frac{1}{n_F}\sum_{i=1}^{n_F}\left(\frac{1}{1 + \frac{\phi_R}{\phi_F}\frac{n_F}{n_R}\exp\frac{W_i^F - u}{k_B T}}\right) \qquad [3b]$$

for the choice $u = \Delta G_{I_0 I_p}$. Here $\langle\cdot\rangle_{F(R)}$ denotes the thermodynamic average over the forward (reverse) process and $n_{F(R)}$ is the number of forward (reverse) trajectories that are in state $p$ for a given $\lambda$. Note that $n_{F(R)}$ corresponds to the total number of trajectories in a region (roughly a hundred per region, as reported in the main text) only at equilibrium ($I_0$), whereas, for a given $\lambda$, it ranges from 1 to 20-25 per state. Although these numbers are small, it has been shown that they lead to reasonable free energy estimates when applying the extended fluctuation relations to analogous cases (8, 9). Thus, Eq. (2) is a self-consistent relation which returns the free energy of the transition $I_0 \to I_p$ at position $\lambda$ by using information from both the folding and refolding trajectories (bidirectional estimator). Importantly, if, for given $\lambda$, a state $I_p$ only occurs in the forward (reverse) process, $n_R = 0$ ($n_F = 0$) in Eq. (3) causing Eq. (2) to fail. Because hysteresis differently affects the forward and the reverse processes, the number of intermediates is, in general, different between unzipping and rezipping, making EBAR unsuitable to our purpose. This is clearly shown in Fig.S2A, where the unzipping FDCs always exhibit three states whereas the majority of the rezipping trajectories directly go from $R$ to $L$.

To solve this problem we introduced the extended form of the Jarzynski free-energy estimator (10)

$$\left\langle \exp -\frac{W_{F(R)}}{k_B T} \right\rangle_{F(R)} = \phi_{F(R)}^{S_{0(p)}\to S_{p(0)}} \exp -\frac{\Delta G_{I_0 I_p}(\lambda)}{k_B T}\,, \qquad [4]$$

which allows to compute $\Delta G_{I_0 I_p}(\lambda)$ by only taking into account the trajectories of the forward (reverse) protocol that visit state $I_p$ (unidirectional estimator). However, unidirectional free-energy estimators have a slow convergence and need a big number of trajectories to obtain reliable free-energy measures, i.e. are affected by non-negligible bias (11). On the contrary, bidirectional free-energy estimators, such as the EBAR, have a faster convergence and lead to much smaller errors (12). Therefore, we correct the bias of the free energies obtained with the Jarzynski estimator by using the results of the EBAR as reference. This has been obtained with following method. Let us consider a state $I_p$ stretching over the positions set $\{\lambda\}$. Firstly, we computed the energies $\{\Delta G(\lambda)_{I_0 I_p}\}^{\mathrm{EBAR}}$ by solving Eq. (2) for all those $\lambda$ having $n_F(\lambda), n_R(\lambda) \neq 0$. Then, the energies of the forward, $\{\Delta G_{I_0 I_p}(\lambda)\}_S^{\mathrm{Jar}}$, and reverse, $\{\Delta G_{I_p I_0}(\lambda)\}_R^{\mathrm{Jar}}$, processes have been separately computed for all the $\lambda$ by using Eq. (4). By using the energy values computed with the EBAR $\{\Delta G_{I_p I_0}(\lambda)\}^{\mathrm{EBAR}}$ as reference, the closest intersection point with the Jarzynski energy sets $\{\Delta G_{I_p I_0}(\lambda)\}_F^{\mathrm{Jar}}$ ($\{\Delta G_{I_p I_0}(\lambda)\}_R^{\mathrm{Jar}}$) is computed. Eventually, the bias is corrected by applying a rigid shift

to $\{\Delta G_{I_pI_0}(\lambda)\}_F^{\rm Jar}$ ($\{\Delta G_{I_pI_0}(\lambda)\}_R^{\rm Jar}$) in order to match the Bennett set. This method is schematically shown in Fig.S3 for each state (Native, Intermediate, Unfolded) exhibited by the 3-states region in Fig.S2.

Given the set $\{\Delta G_{I_0I_p}(\lambda)\}$ for all the $I_p$, the equilibrium free energy is recovered as

$$\Delta G_{\rm eq}(\lambda) = -k_BT \log\left(\sum_{p=0}^{P+1} \exp -\frac{\Delta G_{I_0I_p}(\lambda)}{k_BT}\right) \quad [5]$$

for any $\lambda$ (black line in Fig.S2C). Eventually, by solving the equation

$$f_{\rm eq}(\lambda) = \frac{\partial}{\partial\lambda}\Delta G_{\rm eq}(\lambda)\,, \quad [6]$$

we computed the equilibrium FDC (black line in Fig.S2D).

This method allowed us to compute the equilibrium FDC in all the 8 irreversible regions in sodium and the 4 irreversible regions in magnesium (Fig.2,3 of the main text). Eventually, the full equilibrium FDC has been recovered by the piecewise merging of the equilibrated FDC segments.

## Conversion from number of bases to trap-pipette distance

The conversion between the number of unzipped basepairs $n$ and the trap-pipette distance $\lambda$ is standard in experimental setups such as optical tweezers and AFM, where the position of the force device (optical trap and cantilever) is controlled (13). In a nutshell, the relative distance between the optical trap and the bead in the pipette at a given force $f$ is given by,

$$\lambda(f) = x_b(f) + x_h(f) + x_{\rm ssRNA}(n, f)\,, \quad [7]$$

where $x_b(f)$ is the bead displacement, $x_h(f)$ is the handle's extension and $x_{\rm ssRNA}(n, f)$ is the ssRNA extension given by Eq.(6) (main text) with $L + i = 2n$ (equal to the total number of unzipped bases). Here only the relative value of $\lambda$ matters, however in our study we choose $\lambda = 0$ for the case when the two beads of the experimental setup (bead in trap, bead in pipette) are in contact. The elastic contributions $x_b(f)$ and $x_h(f)$ are given by, $x_b(f) = f/k_b$ for a trap of stiffness equal to $k_b$ and $x_h(f)$ is given by the worm-like chain (WLC) model for the hybrid DNA/RNA handles of persistence length ($P_{\rm DNA/RNA} = 10$nm) and contour length per base-pair ($L_{\rm DNA/RNA} = 7.8$nm) that we used in the thermodynamic analysis part of the main text (Section *Derivation of the NNBP energies for RNA*). The above relation between $\lambda$ and $n$ is univoque and has been used to convert $n$ into $\lambda$.

## Control tests for the correlation analysis

To corroborate the validity of the correlation analysis we performed tests on different controls. First, starting from the original RNA sequence, for each $L$-segment (in the range $L = [8, 28]$) we generated a new hairpin by randomly shuffling segments of a given length $L$ along the sequence. In this way, a different shuffled sequence is obtained for each value of $L$. Given the loop-BEL of each shuffled sequence, we computed the maximum average rolling correlation $\langle r_w\rangle_{\rm UR}^{\rm Max}$ (shift $s = 0$) with respect to the hysteresis between unfolding and refolding (Eq.(7) with $\alpha\beta = {\rm UR}$) for a window of size $w = 100$. Fig.S7A shows the results of this analysis (orange circles) along with the analogous results obtained for the original sequence (blue squares). As discussed in the main text (see Fig.5D), the correlation of the original sequence increases with the stem-loops size $L$ and has a maximum at $\langle r_w\rangle_{\rm UR}^{\rm Max} \simeq 0.25$ for $L \approx 18 - 22$ bases. On the contrary, the correlation for the shuffled sequences varies in the range [-0.06:0.07] with no apparent trend. As a second control, we generated three different random RNA hairpins with the same GC content of the original sequence. Then, for the case $L = 20$ (which roughly corresponds to the maximum correlation observed) we computed the loop-BEL and the average rolling correlation with the hysteresis. The results are shown in Fig.S7B. Analogously to the previous case, the random sequences (red triangles) do not appear to be correlated with the hysteresis. In fact, the correlation falls in the range [-0.04:-0.005] so that the average of $\langle r_w\rangle_{\rm UR}^{\rm Max}$ over the random sequences is roughly equal to $-0.02$ which magnitude (in absolute number) is 10 times lower than the value of the original sequence ($\simeq 0.25$).

Let us point out that the measured values of the average rolling correlation (or analogous direct quantities) are, in general, much smaller than the fraction of highly correlated points (Fig.S5). The normalized correlation (as defined by the Pearson coefficient) between the hysteresis and the loop-BEL is sensitive to several factors. First, the hysteresis landscape computed in Eq. (7) (see main text) is much more accurate in the regions where the irreversibility is large, i.e. where the difference between the unzipping and rezipping FDCs is large. In fact, when the hysteresis is small, i.e. along the reversible regions, Brownian fluctuations and instrumental effects contribute to reducing the signal-to-noise ratio of the measured correlation of the hysteresis landscape with the loop-BEL. In the reversible regions, thermal (Brownian) fluctuations in the unzipping and the rezipping FDCs mask the correlation between the (low) hysteresis and the loop-BEL. The loop-BEL is noiseless by construction, whereas the computed hysteresis is not, so the correlation along the reversible regions is dominated by the noise. Moreover, instrumental effects are also detrimental in estimating such a correlation in the reversible regions. In fact, slight misalignments between the experimental trajectories give contributions to the measured hysteresis that are comparable with those due to residual irreversibility effects. Finally, correlation measurements require matching the experimental measure of the hysteresis profile and the loop-BEL, further reducing correlation estimates. These sources of error render the Pearson coefficient in the reversible regions inaccurate. Therefore, it is not surprising that the Pearson coefficient measured over the entire profile and the rolling one averaged over a given window are both small, whereas the one averaged over the entire profile but restricted to the irreversible regions is markedly larger as shown in Fig.S8.

## Sequence dependency of the hysteresis

Here we show results for all segments of length $N \geq 6$ bases containing consecutive purine (for stacking) and consecutive Watson-Crick complementary bases (for hybridization) along the two unpaired strands of the RNA hairpin. The aim is to identify differences in purine content (for stacking) and Watson-Crick base pairs (for hybridization) between the two kinds of regions, to demonstrate that irreversible effects are sequence dependent. For the stacking motifs we do not discriminate between purines G and A being both counted in the same set. For example, irreversible region 1 contains 220 base pairs and a total of 3 segments of 6 consecutive purines (a single GGGGGG and GGAAAG on one strand and AGGGGA on the other strand) and 2 segments of 7 consecutive purines (AGGAGAA, AGAGAAA on one strand). An analogous count is made on segments capable of forming Watson-Crick complementary bases on the two strands by counting the number of segments containing consecutive A,U or G,C on each of the two strands, again without discriminating their specific order. For example, the same irreversible region 1 contains 1 segment of 9 consecutive G,C (CGCGGGGGG) and 1 segment of 10 consecutive G,C (CGCCGCCGCG). In contrast, the reversible region 3/4 (meaning that it separates irreversible regions 3 and 4) contains 97 base pairs and has only 1 segment of 6 consecutive A,U (UUAAAA). Table S3 summarises all results and Fig.S9 illustrates them with frequency histograms. Note that the fraction of the different segments of purine-stacks and Watson-Crick bases in Fig.S9 are normalized by the total number of bases $\Delta n$ contained in each region (irreversible or reversible). Thus, for a given region, the fraction of bases of a given type (stacking or base-pairing) is defined as $f_N = (M_N N)/\Delta n$, where $M_N$ is the total number of segments of length $N$ of that type (reported in Table S3). Given the values of $f_N$, for each $N$ we computed the average fraction of bases (for stacking or base-pairing) over all the irreversible (reversible) regions

$$\overline{f_N} = \frac{1}{n_{\text{regions}}} \sum_{\text{regions}} f_N \,, \qquad [8]$$

where $n_{\text{regions}}$ is the number of irreversible (reversible) regions ($n_{\text{regions}} = 8$ and $n_{\text{regions}}$, respectively). Finally, we defined the average segment length $\langle N \rangle$ of a given type (stacking or base-pairing) for the irreversible (reversible) regions as the weighted average over $\overline{f_N}$, which is

$$\langle N \rangle = \frac{\sum_N N \overline{f_N}}{\sum_N \overline{f_N}} \qquad [9]$$

Analogously, the variance of the segment length for the irreversible (reversible) regions and for the two types of analysis (stacking or base-pairing) has been computed as

$$\langle N^2 \rangle - \langle N \rangle^2 = \frac{\sum_N N^2 \overline{f_N}}{\sum_N \overline{f_N}} - \left( \frac{\sum_N N \overline{f_N}}{\sum_N \overline{f_N}} \right)^2 \qquad [10]$$

The results are shown in Fig.S9 and point out that stacking and base-pairing effects are larger in the irreversible regions than in the reversible ones. Overall, we found that stacking and base-pairing contribute to the observed hysteresis facilitating the formation of stem-loop structures along each single-strand.

## Experiments in short RNA hairpins

In a previous work (14) a sequence known as CD4 hairpin was studied over three orders of magnitude of monovalent and divalent salt conditions in the diluted regime finding that the average unzipping force in magnesium is larger than the average unzipping force in sodium at the equivalent salt concentrations as derived from thermodynamics. Results are shown in Fig.S10 where we plot the results obtained in that reference under the light of the newly derived $77 \pm 49 : 1$ salt rule. As we can see the values of the folding free energies nicely match each other according to the salt rule (left panel). Also rupture forces during unzipping and rezipping fulfill the rule albeit with a small systematic difference. This is probably due to the fact that the rupture force is a nonequilibrium quantity. The salt rule, although generally satisfied for the kinetics, is not as clean as for the free energy, probably because many other factors affect kinetics as compared to thermodynamics (Fig.S10, central and right panels). In particular, at equivalent monovalent salt conditions, the average unzipping force in magnesium (filled blue squares) is larger than the average unzipping force in sodium (filled orange circles). Furthermore, the average rezipping force in magnesium (empty blue squares) is larger than the average unzipping force in sodium (empty orange circles). This is observed for two different pulling rates (central panel, 1.8pN/s; right panel, 12.5 pN/s) underlining that the amount of hysteresis (related to the difference in the average unzipping and rezipping forces) is always larger for magnesium at the equivalent salt concentrations as derived from thermodynamics.

Here we compare the kinetic of the CD4 hairpin with the results obtained by pulling a short RNA hairpin sequence specifically designed to contain base stacking and hybridisation motifs, mimicking the kinetics of a single irreversible region in the RNA hairpin. The short RNA hairpin (52 bases) contains a stem of 20 complementary base pairs that ends in a dodecaloop GAAAAAAAAAAA that creates stacking between the 11 adenines. We will denote it as hairpin A (Fig.S11A). This hairpin is very different from the CD4 haripin, which has fully complementary stem ending in a tetraloop (GAAA) but does not contain relevant contiguous stacking and hybridising base pair motifs along the unpaired ssRNA. In fact, unzipping experiments of the CD4 shown that only the native structure is formed. In contrast, upon pulling hairpin A under similar experimental conditions (100nm/s, 10mM $MgCl_2$ and 298K) the hairpin forms an alternative misfolded structure (red unzipping curves

and green misfolded band in Fig.S11B, right panel). While misfolding is commonly observed in magnesium (roughly 30% frequency), in sodium it is rarely observed (red unzipping curves in Fig.S11B, left panel). We stress the importance of base pairing interactions within the individual RNA strands. Interestingly, the same experiments carried out on the above mentioned CD4 hairpin but with the stem ending in a dodecaloop (that we will denote as hairpin B) does not show the misfolded state neither in sodium nor in magnesium (Fig.S12). Interestingly, hairpin B has equally enhanced stacking effects as much as hairpin A does (both contain the A-rich dodecaloop), however Watson-Crick base pairing on the 7 A,U in each of the single strands at the beginning of the stem in hairpin A is not present in hairpin B. The repeated AU motif in hairpin A, plus the stacking stabilisation of the A-rich dodecaloop, both induce the formation of a competing structure. The large dispersion of unzipping forces of the misfolded state (7-17pN) might be interpreted as arising from a mechanically rigid structure with a short distance to the transition state*. Alternatively, the lower value of the average unzipping force of the misfolded state (as compared to the 22pN of the native) and its large dispersion (in the range 7-17pN) might be interpreted as due to the fact that the misfolded state is not unique. In this case, the misfolded state contains multiple competing structures stabilized by the weaker A-U bonds along the unpaired strands.

*This result can be inferred in the Bell-Evans model in the Gaussian approximation, which shows that the standard deviation of the rupture force distribution equals $\frac{k_B T}{x^\ddagger} \log(\frac{3+\sqrt{5}}{2})$ with $x^\ddagger$ the distance to the transition state.

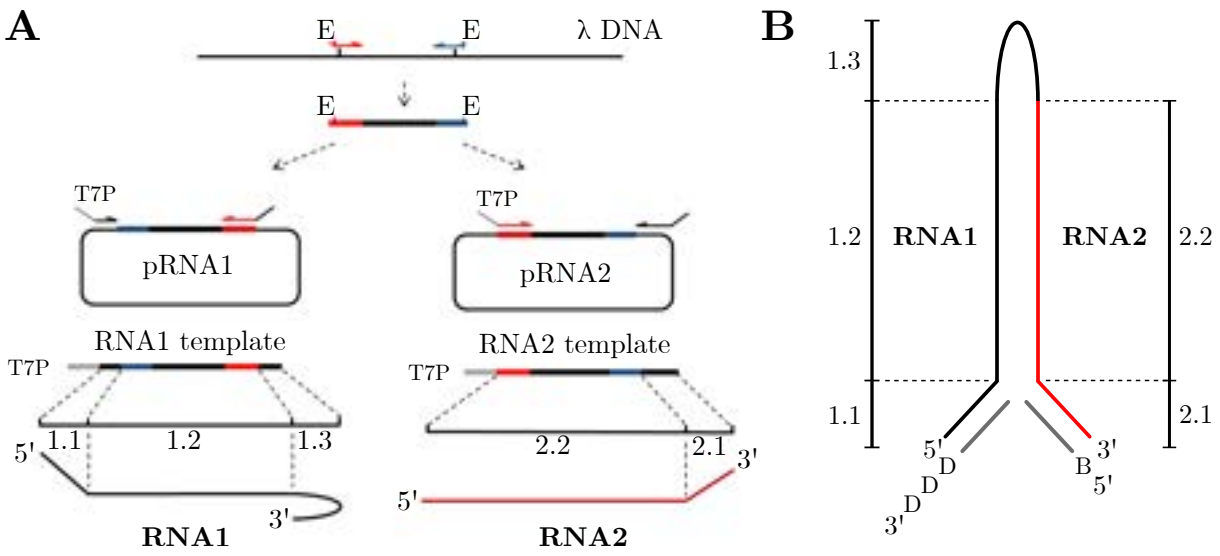


**Fig. S1. (A)** RNA hairpin synthesis. A PCR amplicon of 2027bp in length obtained from $\lambda$-DNA and containing EcoRI sites (E) at its termini was cloned into pBR322 plasmid in both orientations, generating the pRNA1 and pRNA2 constructs. The pRNA1 and pRNA2 constructs were then used as templates for PCR reactions. Both PCR products contained a minimal T7 RNA Polymerase Promoter (T7P) carried by the forward primers. The PCR products were used as templates for *in vitro* transcription reactions that were performed to synthesize RNA1 and RNA2 molecules. Regions 1.1 and 2.1 are derived from pBR322 sequence, regions 1.2 and 2.2 from $\lambda$-DNA sequence, and region 1.3 from the RNA1 reverse primer. **(B)** RNA hairpin structure and assembly. The hairpin is formed by annealing molecules RNA1 and RNA2, a digoxigenin (DIG)-labeled and a biotin (BIO)-labeled oligonucleotide. RNA1 molecule contains three regions: region 1.1 anneals with DIG-labeled oligonucleotide to form RNA1 handle. Region 1.2 anneals with RNA2 and together with region 2.2 from RNA2 forms most of the hairpin stem. Finally, region 1.3 forms the hairpin loop and the upper part of the stem. Apart from region 2.2, RNA2 molecule also contains a $3'$ portion (region 2.1) that anneals with BIO-labeled oligonucleotide to form the RNA2 handle.

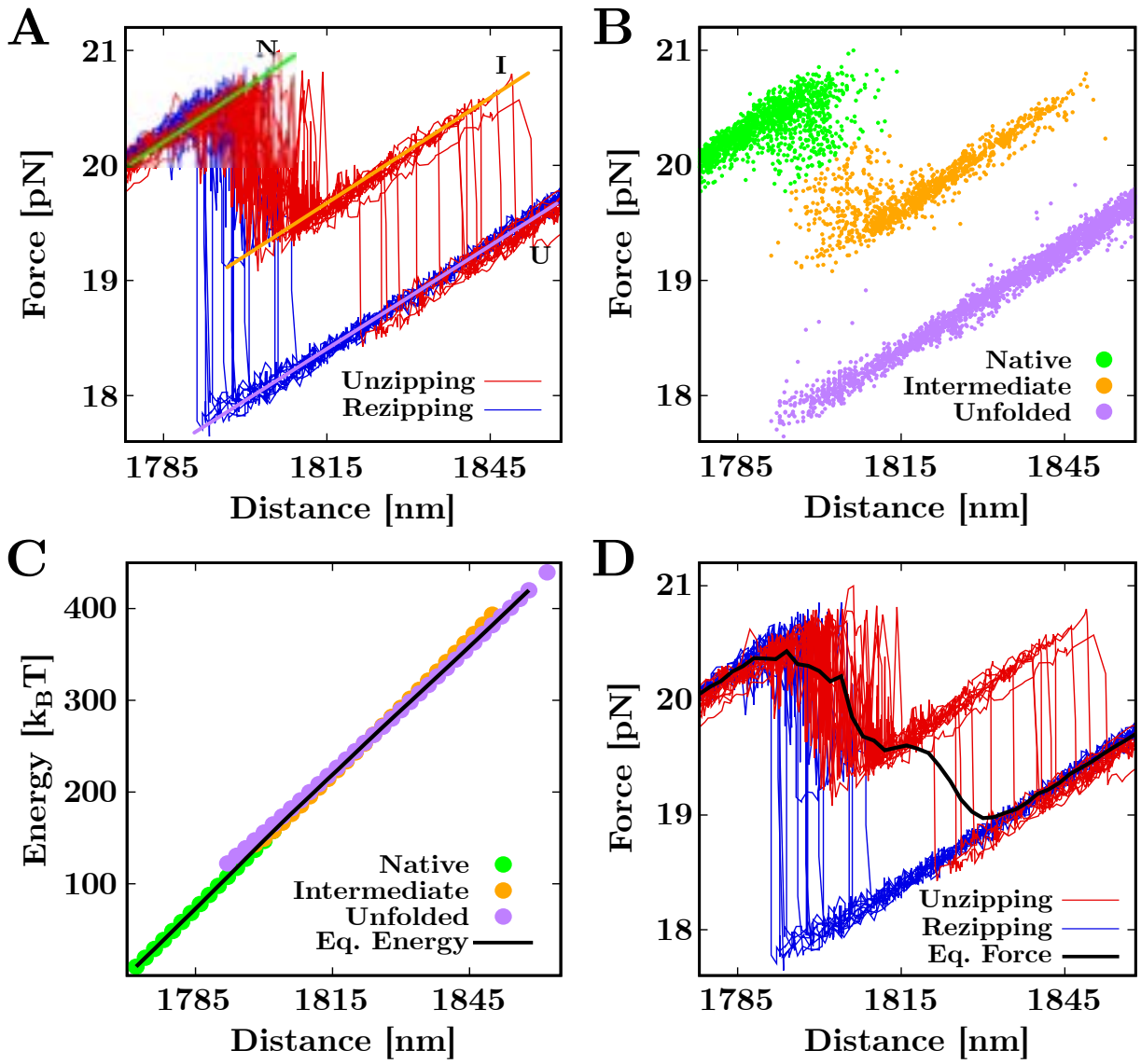


**Fig. S2.** Reconstruction of the equilibrium FDC in the 3-states region measured in 500mM NaCl (see Fig.3 in the main text). The protocol involves four main steps. **(A)** Given the unzipping/rezipping experimental trajectories, each state (Native, Intermediate, Unfolded) is identified and labeled (green, orange and purple lines, respectively). **(B)** The experimental points are assigned to the closest state through a least-square fit. **(C)** The free energy of each state (U,I,N) is computed as the result of the combination of the Bennett and Jarzynski equations (see text). The equilibrium free-energy between all the states (black line) is computed through Eq. (5). **(D)** Eventually, the equilibrium FDC (black line) is recovered by computing Eq. (6).

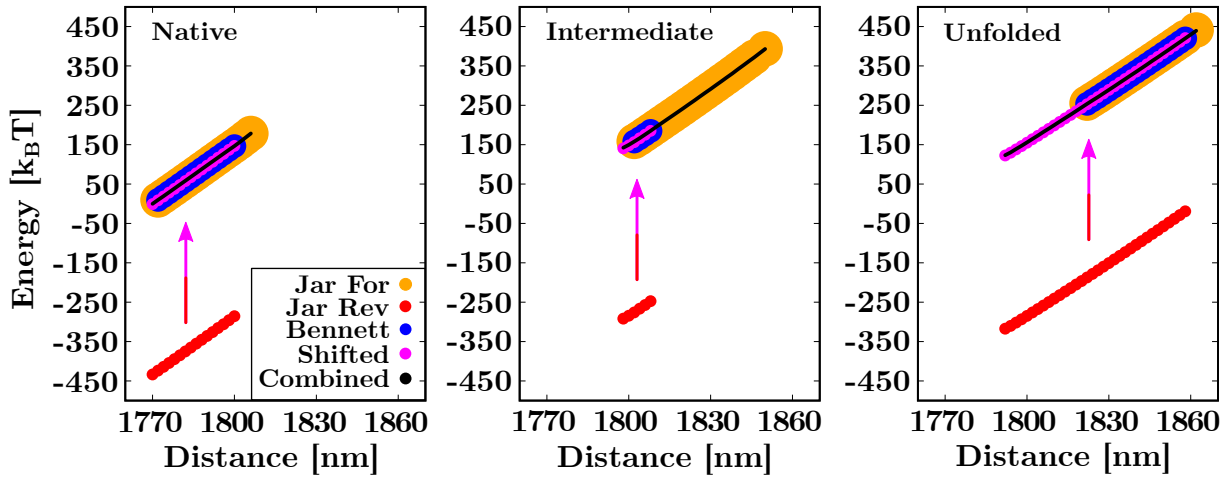


**Fig. S3.** free energy computation of the 3-states (Native, Intermediate, Unfolded) region showed in Fig.S2. The free energy of each state is computed by combining $\mathrm{Eq.}\ (2)$ and $\mathrm{Eq.}\ (4)$. EBAR method (blue dots) only holds if $n_F, n_R \neq 0$ for each position $\lambda$, often leading to free energy estimations limited to a restricted data fraction (see the Intermediate state panel). The forward and reverse Jarzynski estimators are used to compute the energies of the forward (yellow dots) and reverse (red dots) trajectories for each $\lambda$. The (biased - see text - ) results are eventually corrected (violet dots) according to the computed EBAR values, used as reference. This procedure gives the complete free-energy set of each state (black dots).

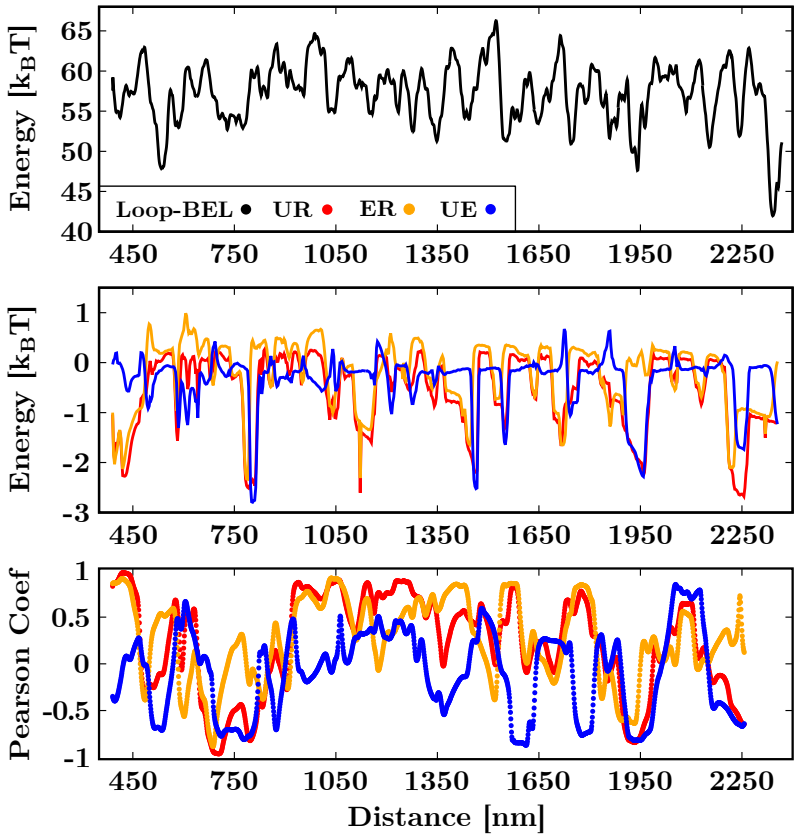


**Fig. S4.** Correlation between the loop-BEL and the hysteresis profiles. **(Top)** Loop-BEL $\Delta G_L(\lambda)$ computed for the RNA 2kbp sequence according to Eq. (4) for $L = 20$ (see main text). **(Middle)** Hysteresis profiles $\Delta G^{\rm Hyst}_{\alpha\beta}(\lambda)$ with $\alpha\beta =$ UR (red), ER (orange), UE (blue) computed for the 500mM NaCl experimental trajectories by solving Eq. (7) (see main text). **(Bottom)** Pearson correlation coefficients $r_w(\lambda)$ resulting from the comparison between loop-BEL and the $\Delta G^{\rm Hyst}_{\alpha\beta}(\lambda)$ over windows of length $w \approx 100$, as described in the main text. Maximal correlation appears for $\alpha\beta = {\rm UR}, {\rm ER}$ revealing the unzipping process as the main source of hysteresis.

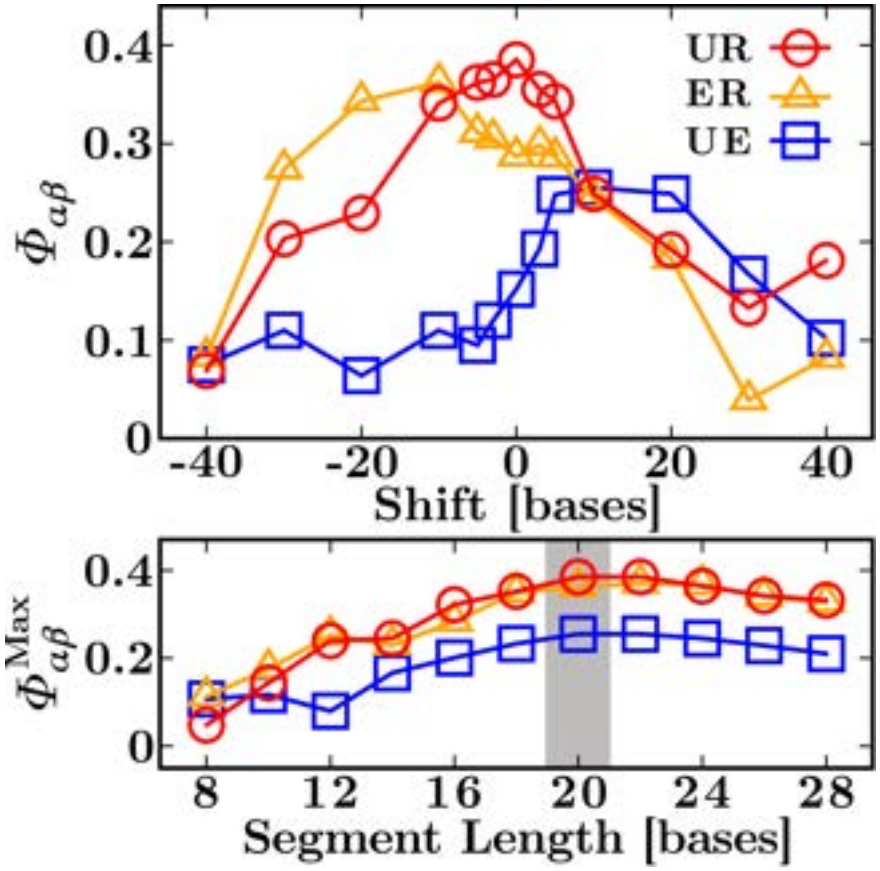


**Fig. S5.** Correlation between loop-BEL and hysteresis profile evidenced by the probability $\phi_{\alpha\beta}$ that $r_w(\lambda) \geq 0.5$ at a given $\lambda$. **Top.** $\phi_{\alpha\beta}$ as a function of the shift $s$ (in bases) of the loop-BEL relative to the hysteresis profiles for the case $L = 20$ (see main text). **Bottom.** Dependence of $\phi^{\mathrm{Max}}_{\alpha\beta}$ with the length $L$ of the segments forming the stem-loops. All curves exhibit a single broad maximum for $L \approx 20$, showing that this is the characteristic stem-loop size that slows down RNA folding to the native stem.

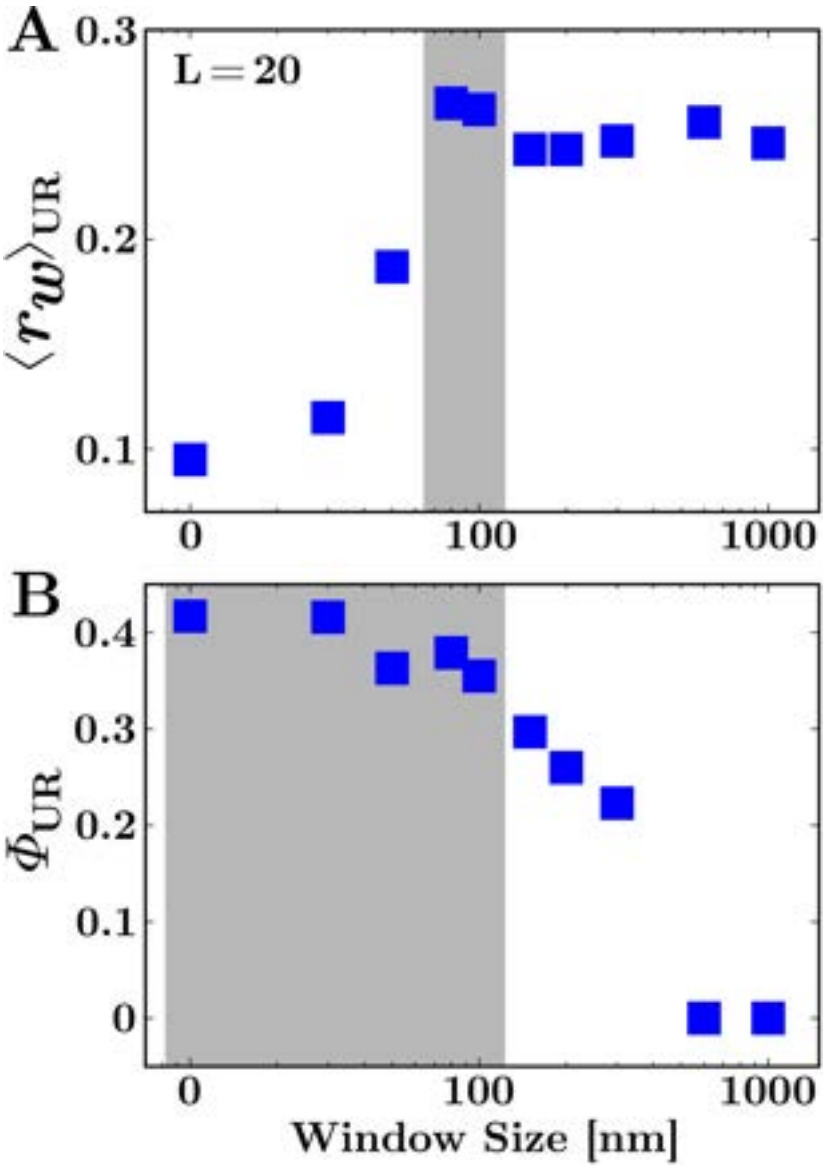


**Fig. S6.** Analysis of the optimal window size $w$ for the computation of the correlation profile $r_w(\lambda)$ in Fig.S4 **(A)** Average rolling correlation $\langle r_w \rangle_{\alpha\beta}$ (computed between the loop-BEL and the hysteresis profile for $\alpha\beta = \mathrm{UR}$ and stem-loops of size $L = 20$) as a function of the window size $w$. The correlation rapidly increases with $w$ and exhibits a maximum in the range [100,150]nm (grey band). **(B)** $\phi_{UR}$ as a function of the window size $w$ for the case $L = 20$. The correlation is maximum in the range [10,150]nm (grey band) and is damped for larger values. Despite $\phi_{UR}$ is stable for a broader interval of $w$ than $\langle r_w \rangle_{\mathrm{UR}}$, both quantities exhibit maximal correlation at $w \sim 100$nm, which roughly correspond to the average size of the released (annealed) base-pairs during the unzipping (rezipping) process.

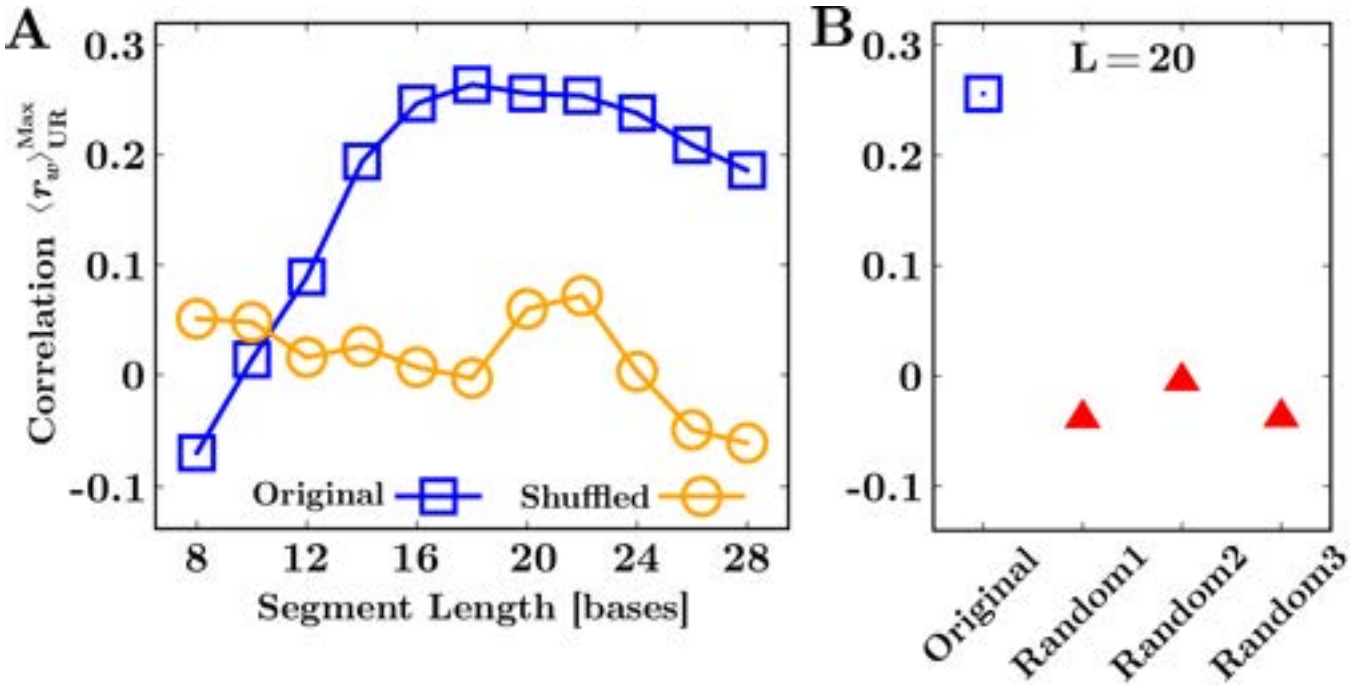


**Fig. S7.** Maximum average rolling correlation $\langle r_w \rangle_{\mathrm{UR}}^{\mathrm{Max}}$ (shift $s = 0$) between hysteresis and loop-BEL for different control sequences. **(A)** Average rolling correlation computed for the shuffled sequences (orange circles) and the original sequence (blue squares) for each value of $L$=[8,28]. **(B)** Average rolling correlation of the hysteresis with the Loop-BEL for the random sequences (see text) at the maximally correlated case $L = 20$. The analogous value for the original sequence (blue square) is also reported for a direct comparison.

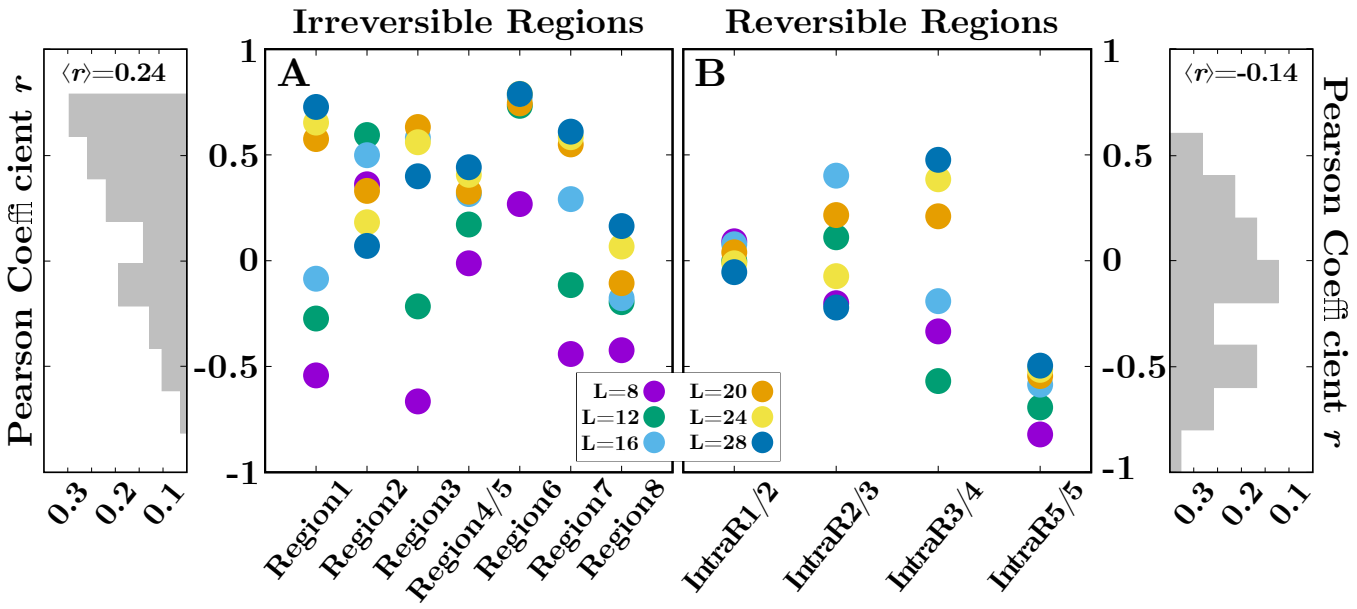


**Fig. S8.** Pearson correlation of the loop-BEL with the hysteresis between unfolding and refolding trajectories ($\alpha\beta = \mathrm{UR}$ in Eq.(7) of the main text) for all the irreversible **(A)** and reversible **(B)** regions for stem-loops sizes in the range [8:28]. The loop-BEL and hysteresis exhibit a higher correlation in the irreversible regions than in the reversible ones as shown by the coefficient distributions (grey bars in the leftmost end rightmost panels) that are peaked at $r_{\mathrm{irr}}^{\mathrm{Max}} \approx 0.7$ and , $r_{\mathrm{rev}}^{\mathrm{Max}} \approx -0.1$ respectively. As shown in the paper, the correlation grows with the size of the $L$-segments and is maximal for $L > 16$. The $r$ value averaged over all regions and stem-loops sizes for the irreversible and reversible regions ($\langle r \rangle = 0.24$ and $\langle r \rangle = -0.14$, respectively) is also shown.

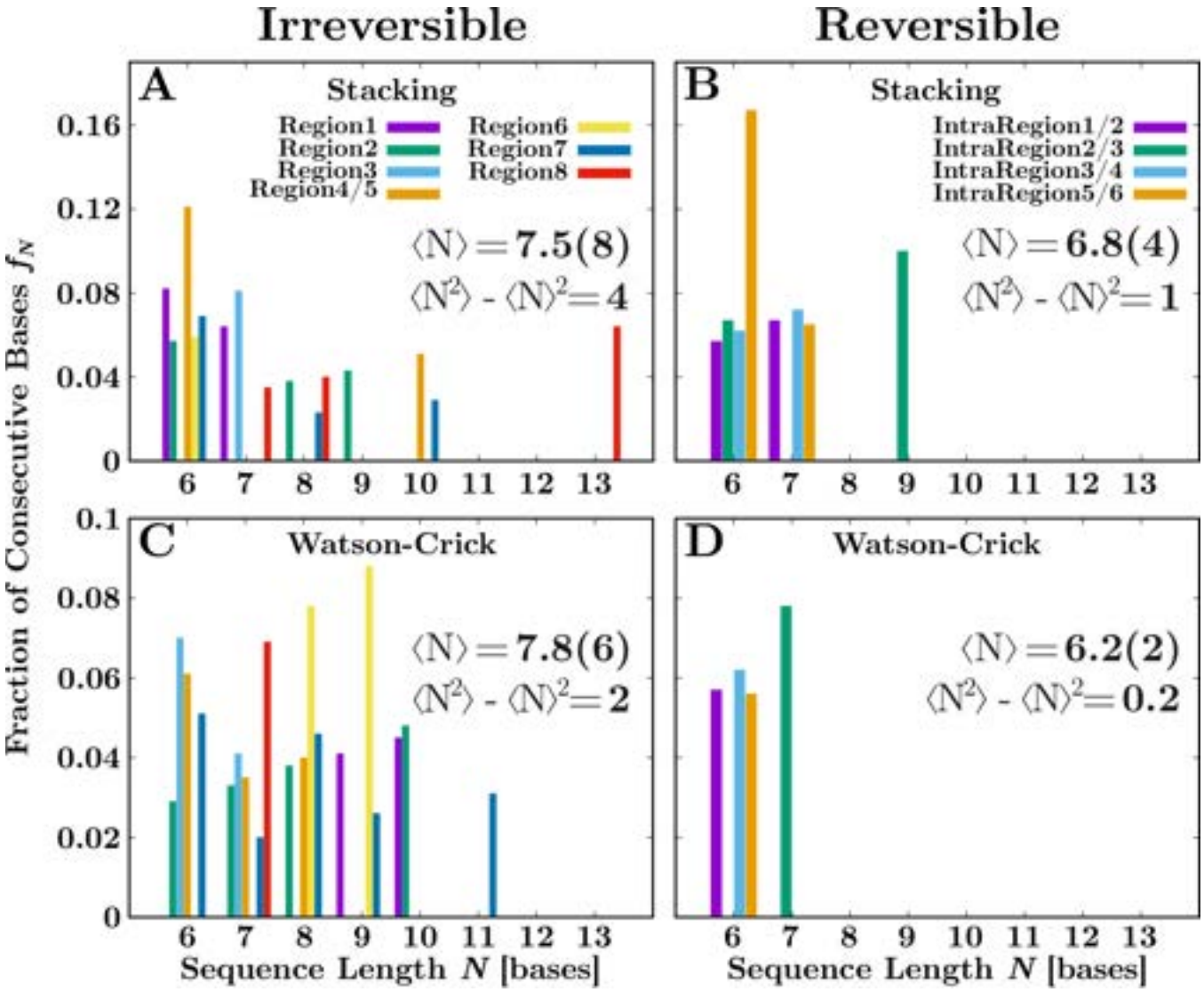


**Fig. S9.** Histogram of the results from the 2kbp RNA sequence analysis of the irreversible and reversible regions. Sequence motifs of $N \geq 6$, containing stacking (G,A) and hydrogen bonding (A,U or C,G) sequences along the unpaired strands, were analysed. **(A,B)** Stacking analysis. Fraction of consecutive stacked purines (A,G) as a function of the segment length per each irreversible (left) and reversible (right) region. **(C,D)** Watson-Crick base pairing analysis. Fraction of consecutive (A,U or C,G) as a function of the segment length per each irreversible (left) and reversible (right) region. The analysis shows both a larger average segment length $\langle N \rangle$ and variance of the segments length $\langle N^2 \rangle - \langle N \rangle^2$ in the irreversible regions. The error (in brackets) is the statistical uncertainty in the last digit.

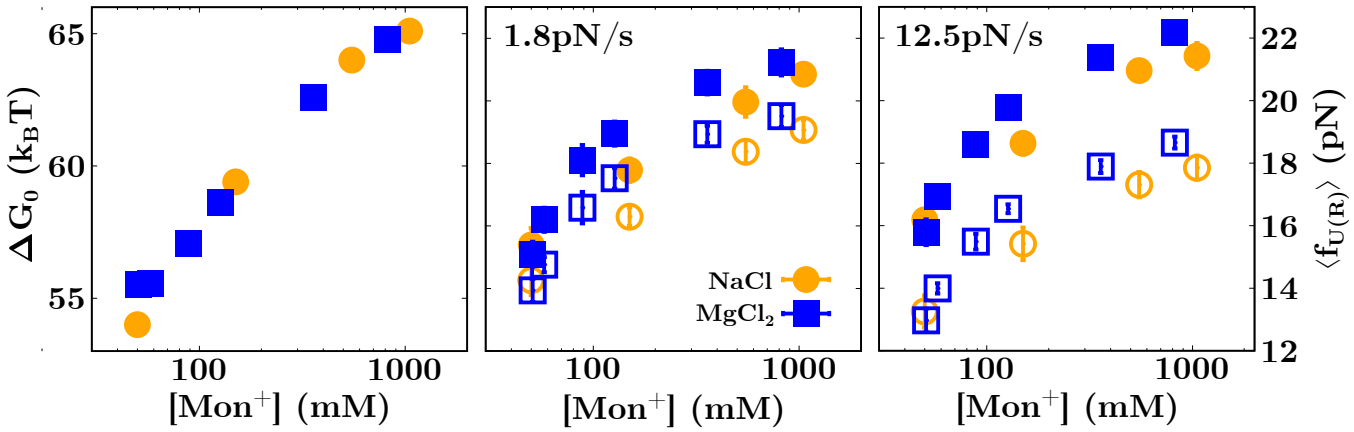


**Fig. S10.** Results from unzipping experiments of hairpin CD4 (14) in sodium (blue squares) at 50mM, 150mM, 550mM, 1050mM and magnesium (orange circles) at 0.01mM, 0.10mM, 0.50mM, 1mM, 4mM, 10mM. **(Left)** Free energies of hybridization of the hairpin for sodium and magnesium in monovalent salt equivalents (according to the measured $77 \pm 49$ equivalence rule). **(Middle)** Unzipping (full symbols) and rezipping (empty symbol) average rupture forces for sodium and magnesium in monovalent salt equivalents at 1.8pN/s pulling speed. **(Right)** Unzipping/rezipping average rupture forces for sodium and magnesium in monovalent salt equivalents at 12.5pN/s pulling speed.

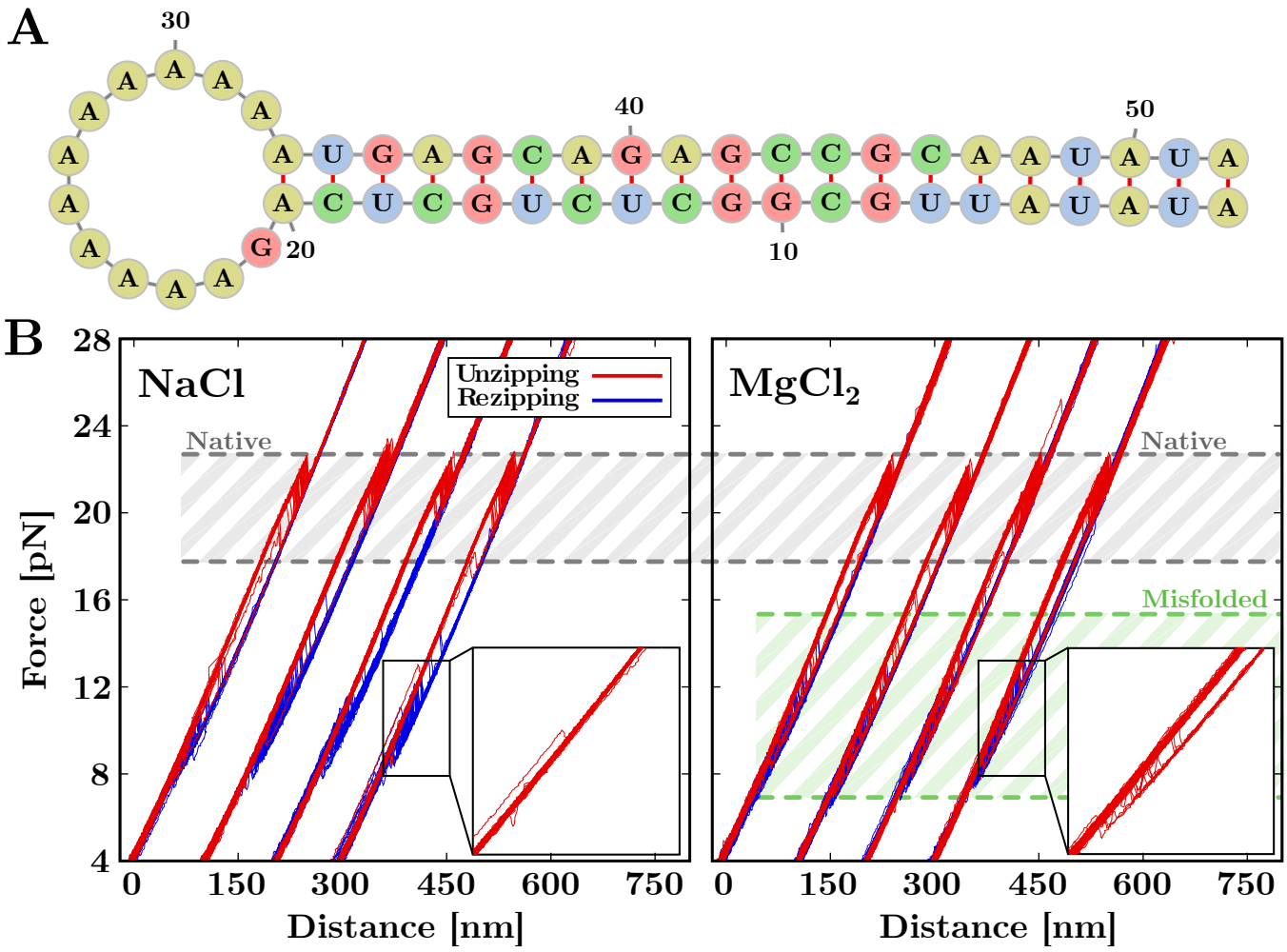


**Fig. S11.** RNA hairpin A. **(A)** Representation of the native conformation of the hairpin A stem ending with the dodecaloop GAAAAAAAAAAA. Hairpin A contains a large fraction of stacking (in the loop) and contiguous and repeated Watson-Crick base pairs (of the AU type) in the unpaired strands. **(B)** Unzipping (red) and rezipping (blue) trajectories measured for the two salt conditions 1M NaCl (left panel) and 10mM $MgCl_2$ (right panel). While in sodium only the native conformation appears (gray shaded band) with a unzipping force rip at 20-22pN, experiments with magnesium reveal a misfolded state (green shaded band) occurring in the force range 7-17pN. The zooms show the unzipping FDC in the low range of forces where the misfolded state is observed in magnesium. The hairpin cartoon is based on the representation given by the Vienna RNA Web Services (15).

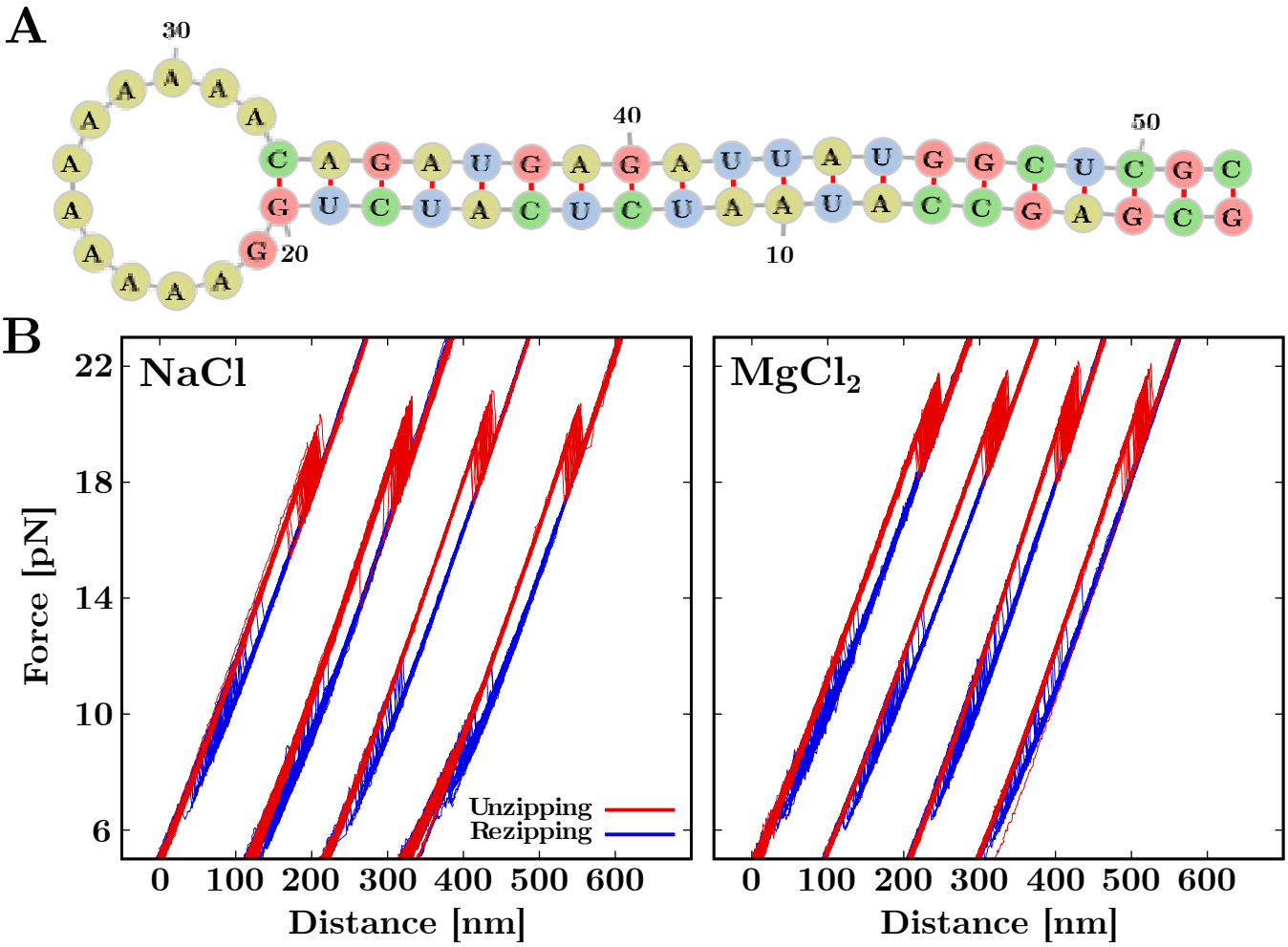


**Fig. S12.** RNA hairpin B. **(A)** Representation of the native conformation of the hairpin B, composed by the CD4 stem ending with the dodecaloop GAAAAAAAAAAA. **(B)** Unzipping experiments a the equivalent salt conditions of NaCl (300mM) and magnesium (4mM) according to the measured $77 \pm 49$ salt equivalence rule. The small difference in the unzipping forces in the two cases (higher for magnesium) must be attributed to the non equilibrium effect shown in Fig.S10. The hairpin cartoon is based on the representation given by the Vienna RNA Web Services (15).

**Table S1. List of the oligonucleotides used in the synthesis of the 2027bp RNA hairpin.**

| Oligonucleotides | Sequence |
|---|---|
| Univ_hairpin_F | $5'$-ac**gaattc**gaaaaacgcctcgagtgaag-$3'$ |
| EcoRI_2.0kb_R | $5'$-ac**gaattc**ttggggtgtgtgatacgaaa-$3'$ |
| RNA1_T7Forw | $5'$-ctaatacgactcactatagggaataaaaataggcgtatcacgag-$3'$ |
| RNA1_Rev | $5'$-ctcatctgtttccagatgaggggagaaaaacgcctcgagtgaag-$3'$ |
| RNA2_T7Forw | $5'$-ctaatacgactcactatagggagaaaaacgcctcgagtgaag-$3'$ |
| RNA2_Rev | $5'$-gaacatacgaaacggatgataagctgtcaaaca-$3'$ |
| S Handle A | $5'$-acgaaagggcctcgtgatacgcctattttt-$3'$ |
| S Handle B2 | $5'$-Bio-gaacatacgaaacggatgataagctgtcaa-$3'$ |

**Table S2. Occurrence of NN motifs in the RNA sequence**

| NNBP | Frequency (%) |
|---|---|
| AA/UU | 17.1 |
| CA/GU | 14.7 |
| GA/CU | 12.7 |
| AU/UA | 9.6 |
| GU/CA | 10.0 |
| CC/GG | 8.7 |
| CG/GC | 4.1 |
| AG/UC | 11.0 |
| GC/CG | 5.6 |
| UA/AU | 6.3 |

The abundance is similar for all the different NN motifs. Notice that for those motifs where it appears to be higher (AA/UU, CA/GU, GA/CU, GU/CA, AG/UC, CC/GG) there is the double degeneracy due to Watson-Crick symmetry (e.g. the fraction of AA/UU includes AA/UU and UU/AA).

**Table S3. Analysis of the 2kbp RNA hairpin sequence**

| | Hysteresis | | | | | | | No Hysteresis | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **R1** | **R2** | **R3** | **R4** | **R5/6** | **R7** | **R8** | **IntraR1/2** | **IntraR2/3** | **IntraR3/4** | **IntraR5/6** |
| **$\Delta$n (bases)** | 220 | 210 | 172 | 198 | 102 | 350 | 202 | 105 | 90 | 97 | 108 |
| **N (bases)** | **Number of Stacked Segments ($M_N$)** | | | | | | | | | | |
| **6** | 3 | 2 | 0 | 4 | 1 | 4 | 0 | 1 | 1 | 1 | 3 |
| **7** | 2 | 0 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| **8** | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| **9** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **10** | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **11** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **13** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **N (bases)** | **Number of Watson-Crick Segments ($M_N$)** | | | | | | | | | | |
| **6** | 0 | 1 | 2 | 2 | 0 | 3 | 0 | 1 | 0 | 1 | 1 |
| **7** | 0 | 1 | 1 | 1 | 0 | 1 | 2 | 0 | 1 | 0 | 0 |
| **8** | 0 | 1 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| **9** | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **10** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **11** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **13** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Analysis of the irreversible and reversible regions along the 2kbp RNA sequence. The irreversible regions are labeled from 1 to 8 whereas the reversible ones are labeled as 'IntraR' meaning, for example, that reversible region IntraR1/R2 lies in between irreversible regions 1 and 2. The bases length $\Delta n$ of each region is also reported. **(Top)** Number of segments of $N$ consecutive stacked purines (A,G) along both strands of the stem. **(Bottom)** Number of segments of $N$ consecutive Watson-Cricks bases (A,U and C,G) along both strands of the stem.

## References


1. J Abels, F Moreno-Herrero, T Van der Heijden, C Dekker, NH Dekker, Single-molecule measurements of the persistence length of double-stranded RNA. Biophys. J. **88**, 2737–2744 (2005).
2. A Untergasser, et al., Primer3plus, an enhanced web interface to Primer3. Nucleic Acids Res. **35**, W71–W74 (2007).
3. F Bolivar, RL Rodriguez, MC Betlach, HW Boyer, Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. Gene **2**, 75–93 (1977).
4. F Bolivar, et al., Construction and characterization of new cloning vehicle. II. A multipurpose cloning system. Gene **2**, 95–113 (1977).
5. MM Baklanov, LN Golikova, EG Malygin, Effect on DNA transcription of nucleotide sequences upstream to T7 promoter. Nucleic Acids Res. **24**, 3659–3660 (1996).
6. CH Bennett, Efficient estimation of free energy differences from monte carlo data. J. Comput. Phys. **22**, 245–268 (1976).
7. MR Shirts, E Bair, G Hooker, VS Pande, Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. Phys. Rev. Lett. **91**, 140601 (2003).
8. A Alemany, A Mossa, I Junier, F Ritort, Experimental free-energy measurements of kinetic molecular states using fluctuation theorems. Nat. Phys. **8**, 688 (2012).
9. J Camunas-Soler, A Alemany, F Ritort, Experimental measurement of binding energy, selectivity, and allostery using fluctuation theorems. Science **355**, 412–415 (2017).
10. C Jarzynski, Nonequilibrium equality for free energy differences. Phys. Rev. Lett. **78**, 2690 (1997).
11. M Palassini, F Ritort, Improving free-energy estimates from unidirectional work measurements: theory and experiment. Phys. Rev. Lett. **107**, 060601 (2011).
12. FM Ytreberg, RH Swendsen, DM Zuckerman, Comparison of free energy methods for molecular systems. J. Chem. Phys. **125**, 184114 (2006).
13. A Severino, AM Monge, P Rissone, F Ritort, Efficient methods for determining folding free energies in single-molecule pulling experiments. J. Stat. Mech.: Theory Exp **2019**, 124001 (2019).
14. CV Bizarro, A Alemany, F Ritort, Non-specific binding of Na+ and Mg2+ to RNA determined by force spectroscopy methods. Nucleic Acids Res. **40**, 6922–6935 (2012).
15. P Kerpedjiev, S Hammer, IL Hofacker, Forna (force-directed rna): Simple and effective online rna secondary structure diagrams. Bioinformatics **31**, 3377–3379 (2015).