A thermodynamic model of bacterial transcription

Jin Qian,* David Dunlap,† and Laura Finzi‡
Physics Department, Emory University

Transcriptional pausing is highly regulated by the template DNA and nascent transcript sequences. Here, we propose a thermodynamic model of transcriptional pausing, based on the thermal energy of transcription bubbles and nascent RNA structures, to describe the kinetics of the reaction pathways between active translocation, intermediate, backtracked, and hairpin-stabilized pauses. The model readily predicts experimentally detected pauses in high-resolution optical tweezers measurements of transcription. Unlike other models, it also predicts the effect of tension and the GreA transcription factor on pausing.

INTRODUCTION

During bacterial transcription there are frequent pauses in the forward translocation of RNA polymerase. Pauses have been observed widely *in vivo* and *in vitro* and vary in durations from milliseconds to minutes [1, 2]. Long pauses, which may last tens of seconds, are classified as Class I 'hairpin-stabilized' and Class II 'backtracked' signals. Short pauses, which typically last less than one second, are proposed to be intermediate precursors of long pauses [3]. Class I and Class II pauses have been structurally characterized and mechanistically explored [4, 5]. They are thought to be regulated by the sequence of the DNA template, the structure of the nascent transcript, and the availability of transcription factors [6–8].

Previous models of the kinetics of back-tracked pauses reproduces some types of experimentally detected pauses [9–12] but fail to predict other types of pausing and pause duration, and do not treat external tension and transcription factors. Here, we propose a model based on our current biochemical understanding of transcription pausing mechanisms and train the model with high-resolution transcription traces. The model successfully predicts the experimentally observed pause sites and duration, and provide a mechanistic explanation to the effect of external tension and transcription factors. The model shows broad predictive power in RNAP elongation and is readily extendable to incorporate the initiation and termination stages.

MODEL DESCRIPTION

The energy of the ternary elongation complex (TEC) is estimated as the sum of four contributions: the energy of the (i) transcription bubble, (ii) DNA-nascent RNA hybrid, (iii) nascent RNA, and (iv) RNAP binding:

$$\begin{aligned} G_{\rm TEC} &= G_{\rm bubble} + G_{\rm hybrid} \\ &\quad + G_{\rm RNA} + G_{\rm RNAP\_binding}. \end{aligned} \tag{1}$$

This estimate of the energy of a TEC is mostly sequence-dependent, as the first two terms are clearly sequence-dependent, and the secondary structure of nascent RNA (the third term) is also directed by sequence. The fourth term corresponds to interactions be- tween the nucleic acids and RNAP subunits and is considered sequence-independent (taken to be zero here, as in Yager&von Hippel, Tadigotla, and Bai) [9–11]. To determine the configuration of a transcription bubble and the details of the energy profile of a TEC, we used an approach based on statistical mechanics, the basis of which was described by Tadigotla [10] (SI).

The model considers two translocation states (n): the active (0) and the backtracked state (-), plus one conformational state: the hairpin-stabilized pause (hsp). The hairpin-stabilized (Class I), the backtracked (Class II) and the pre-translocated pauses are sequence-induced. The interconnection among these states is shown in Supplementary Figure 1b. The first two have been studied extensively and their existence is supported by crystallographic evidence, while the pre-translocated pause is a theoretical prediction.

The active translocation of RNAP is modeled by the Michaelis-Menten (M-M) equation

$$\begin{aligned} k_{\rm forward} &= \frac{k_{\rm max}\ [\rm NTP]}{K_{\rm d}^{\rm eff} + [\rm NTP]}, \\ K_{\rm d}^{\rm eff} &= K_{\rm d}\,(1 + K_{\rm i})\,, \end{aligned} \tag{2}$$

where $k_{\rm max}$ is the rate of NTP catalysis, $K_{\rm d}$ is the NTP dissociation constant, and $K_{\rm i}$ is the equilibrium constant between two adjacent translocation states. This is a good approach, whether the translocational register of a TEC is considered to be determined by the presence of incoming NTP, as in the Brownian-ratchet model [13], or by the release of pyrophosphate [14], as long as $k_{\rm max}$ and $K_{\rm d}$ are given a different physical meaning in each case (SI).

Backtracking was modeled using the ArrheniusEquation (3) with activation barrier up to $40 - 50\ k_{\rm B}T$ for each step of backward translocation [9]. This value may be unreasonably high given that the free energy of base pairing in a transcription bubble is typically less than $-20\ k_{\rm B}T$ [10].

$$k_{\rm bt} = k_1\ \exp\left(\triangle G_{\rm bt}/k_{\rm B}T\right) \tag{3}$$

We take the same Arrhenius approach but treat the first step of backtracking differently from the subsequent ones (Supplementary Figure 1c), based on the assumption that initially the 3' end of the nascent transcript disrupts the active site and invades the secondary channel of RNAP [15] , while additional backtracking stabilizes the interaction of RNA within the secondary channel. The details are given in the SI.

Hypertranslocation, which refers to the forward translocation of RNAP without concurrent RNA elongation at the active site, is translocationally similar to backtracking. However, we do not consider hypertranslocation for two reasons. First, hypertranslocation may not be a general phenomenon during transcription [16] , and it cannot be distinguished from backtracking in force spectroscopy assays. Second, hypertranslocation is never energetically favored, because there is less base-pairing than in the active state.

We take an allosteric view when modeling the hairpin-stabilized pause [4, 17]. The pathway is modeled as a fast equilibrium between two configurational states, the active state and an inactive state with an RNA hairpin allosterically disrupting the active site through a hairpin-flap interaction. The equilibrium is followed by a rate-limiting catalytic step (Supplementary Figure 1d). The equilibrium is considered rapid compared to the formation of chemical bonds that stabilize the inactive state.

A stable RNA hairpin structure can increase the entry rate into a hairpin-stabilized, inactive state and significantly bias the equilibrium towards that paused state. To correctly model the formation of the hairpin-stabilized pause, we must evaluate the co-transcriptional folding of nascent RNA. To this end, we rely on KINEFOLD algorithms [18], which provide a simulation of co-transcriptional RNA folding at a constant RNA elongation rate. We also test the model with RNA folding generated by the lowest-energy method.

Since the experimental data we used to validate the model were acquired under tension of magnitude up to 25pN, the effect of external tension on the thermodynamics of TEC needs to be considered. For the forward translocation and backtracking pathways, we employed the idea that the energy barrier is modulated by the work produced by tension [19]. The hairpin-stabilized pause was assumed to be independent of any applied tension, since it does not involve RNAP translocation.

It is important to notice that transcription is a process that involves only very small numbers of reacting molecules, thus the law of mass chemical reaction is not suitable. Rather, we apply two stochastic kinetics methods: (i) continuous-time Markov chain and (ii) stochastic simulation. The continuous-time Markov chain allows us to analytically solve for the expected time spent in each state at a certain position (SI). The stochastic simulation sheds light on how individual pausing events unfold.

The model is encapsulated in a MATLAB class object, which can generate a predicted residence time histogram with the input of a template sequence and a guess of unknown parameters. Thus, the model can be trained with the real-time single molecule experimental results of bacterial transcription. We used the traces obtained in high-resolution optical tweezers transcription experiments by Gabizon et al. [20] without or with factors that are known to interact with RNA and affect RNAP pausing, such as GreB or RNase A. The transcription experiments were performed on a DNA template (8XHis) containing the T7A1 promoter followed by eight tandem repeats of a 239 bp sequence containing the *his*-leader pause site and four other known sequence-dependent pause sites [1] . The temporal resolution is high enough to detect pausing events longer than ~100ms. This allows sampling of the residence time at the one base-pair resolution. Alignment of the traces under different forces and transcription factor conditions generates the residence time histogram (Figure 1A) as described previously.

COMPARISON OF THE MODEL WITH EXPERIMENTAL DATA

To test the validity of the model, we optimized the values of the model parameters to yield a residence time histogram that resembled the experimental one Figure 1. Since the model includes many parameters, we avoided overfitting when tuning the parameters(SI). Fortunately, the experimental data under different factor conditions could manifest the mechanisms of the pauses. Also, the analysis of the backtracking dynamics helps differentiate backtracked pauses from others. For example, pauses at position "a" are likely due to pre-translocation in Figure 1 , since their duration is barely affected by the addition of GreB. Pauses at position "b" are likely due to backtracking, as their duration responds to the presence of GreB, and they are preceded by backward RNAP translocation, as previous analysis suggests [20]. Pauses "d" and *his* are hairpin-stabilized, as they almost disappear in the presence of RNase. The nature of pauses at "c" and at other less significant sites are less clear because they respond insignificantly to changes in experimental conditions.

The effect of tension is correctly modeled by introducing two different effective lengths EL_{f} and EL_{bt} for forward and backtracking translocation, respectively. Figure 2 shows the comparison of the duration of pauses at site "a" (pre-translocated pause), "b" (backtracked pause), "d" and *his* under different values of tension. Our model is clearly able to reproduce, therefore predict, the lifetime of pauses observed experimentally. Notice that the effective length for the forward translocation pathway is shorter than 1 base pair, while the external force acts on an effective length shorter than 0.1 bp during backtracking (Table S1). The fitted values of effective length agree with previous work [13, 21] . These results indi-

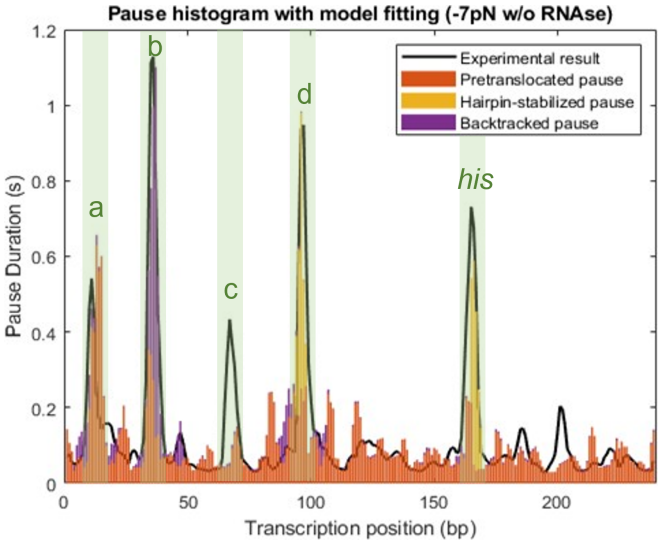

FIG. 1. Stacked histogram produced by the model for the condition of opposing 7 pN. The residence time due to different pausing mechanisms is represented by different colors. The experimental result is shown by the black line.

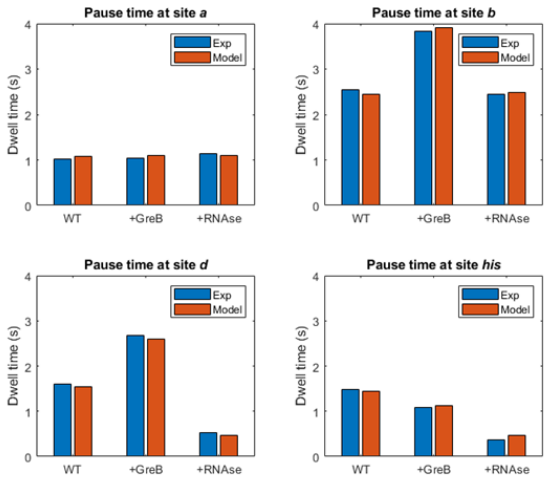

FIG. 3. Comparison between the pause times obtained experimentally and those predicted by the model under different transcription factor conditions.

cate that opposing tension extends the duration of backtracked pauses mostly due to a decrease in transcription rate. It also supports the idea that the entry into long-lived pauses, such as backtracked pauses, follows entry in short-lived pauses.

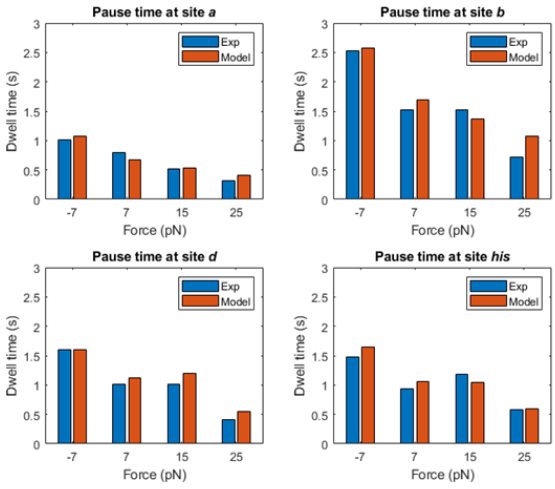

FIG. 2. Comparison between experimental result and model prediction of the duration of pauses at different sites under different force conditions.

Figure 3 shows the comparison of pause duration in the presence or absence of transcriptional factors GreB and RNAse. In general, the presence of GreB extends the dwell time at these pause sites [20]. The model reproduces this effect by decreasing the forward translocation rate (i.e., decrease $k_{\max}$and/or increase K_{d}). The presence of RNase significantly decreases the dwell time at sites "d" and *his* but has little effect on the duration of pauses at other sites. The model can achieve this effect by setting the energy of RNA secondary structure to zero.

To further test the validity of the model, we used Monte Carlo simulations to generate a large number of transcription traces, and we compared the dynamics of backtracking in experimental and simulated traces. The pauses at site "b" in simulated traces were analyzed for backtrack depth and backtrack duration(Figure 4). The clear agreement between simulated and experimental results lends further support to the model.

The power of the model is demonstrated by the fact it accurately predicts major pauses in the transcription of an unfamiliar 210 bp sequence. This sequence just prior to the repeat region of the 8XHis template was not included in the data used to optimize the model parameters. In Figure 5, significant pauses near 15, 40, and 130 bp shown in the dwell histogram acquired after aligning experimental transcription records (red, see methods) were successfully predicted (blue).

STRENGTHS AND LIMITATIONS OF THE MODEL

The model identifies pausing sites and correctly characterizes the mechanism of transcription pausing. Note that, the model identifies sites of slow forward translocation rates as pre-translocation pauses. In other reports, these are often referred to as ubiquitous pauses. The model also distinguishes the backtracked pause and the hairpin-stabilized pause, which are typically viewed as long-lived pauses.

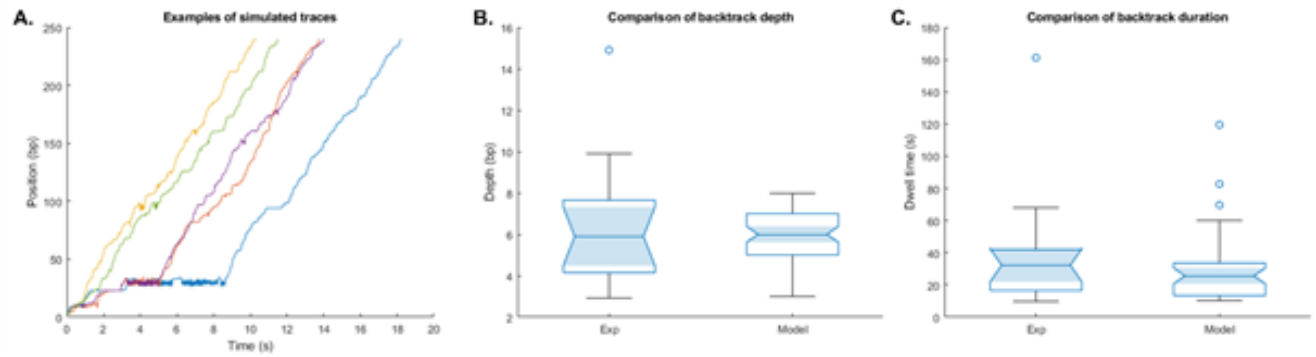

FIG. 4. Comparison between the backtracking dynamics in experimental and simulated traces . (A) Examples of simulated traces generated by Monte Carlo simulation. (B) Distribution of backtrack depth observed experimentally and predicted by the model. (C) Distribution of backtrack duration observed in the experiments and predicted by the model.

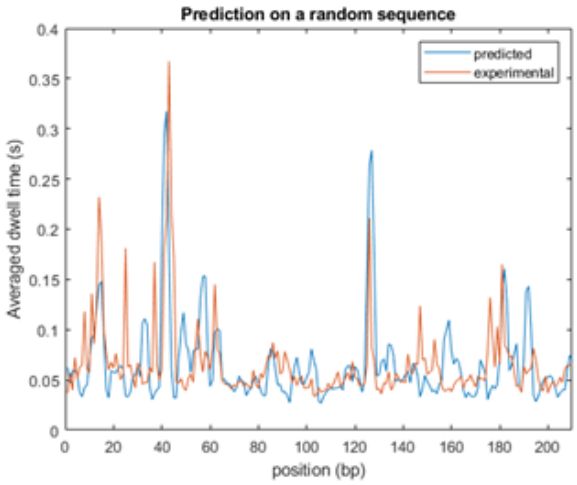

FIG. 5. Comparison of the dwelltime histogram of experimental data and model predictions.

Our results support a previous theoretical analysis of transcriptional pauses which suggests that a long-lived pause is stabilized from a short-lived ubiquitous pause [3]. For example, at pause site "b", backtracking is favored over forward translocation because of the low forward translocation rate. Indeed, the energetic parameters of the model would predict comparable backtracking rates at the 35 bp site (pause "b") and at the 160 bp site, but the fast forward translocation rate at the 160 bp site opposes backtracking. Using the canonical Michaelis-Menten expression, we determined that the forward translocation rate along the template varies from less than 3 nt/s to 50 nt/s. This implies that a slowly transcribing complex may enter into a long-lived pause at one site, even if the back-tracking energy barrier at this position is higher than the barrier height at a position where transcription is faster.

The model predicts that the effective length of force is about half bp for forward translocation pathway, but less than 0.1 bp for the backtracking pathway. This result suggests that external forces insignificantly affect the back-tracking rate. During backtracking RNAP must ratchet backwards on the DNA and disrupt the RNA-DNA hybrid near the active site. The rate is determined in large measure by the denaturation of the hybrid complex. Thus, external forces cannot alter this process as much as biasing the equilibrium constant in the forward translocation pathway. This result further supports that backtracked pausing is favored by slow forward translocation.

The hairpin-stabilized pause requires the interaction between a transcript hairpin and the RNAP flap domain. Previous models simulated the folding of nascent transcripts using the lowest-energy method [10, 13, 22]. However, that method cannot locate the correct positions of hair- pins. In our sequence, the lowest energy model predicts strong hairpins with optimal stem lengths at positions 27 and 101. However, the experimental data do not suggest the existence of hairpin-stabilized pauses at these positions. Alternatively, we used the co-transcriptionally folded RNA structure in which a new nucleotide is added every 40 ms. In this case, hairpins at position 27 and 101 are unlikely to interact with RNAP, because they are not likely to form before RNAP has past these positions. Instead, hairpins form quickly at positions 86 and 136, which correspond to pauses at sites "d" and *his*, respectively. Clearly, considering co-transcriptional folding of nascent RNA leads to more accurate prediction of hairpin-stabilized pauses.

The model shows promise in predicting the transcriptional kinetics at pause sites "a", "b", "d" and *his*. Nonetheless, the current model cannot characterize the pauses observed at site "c" and at other less significant sites. The duration of the pause at site "c" is largely unaffected by the addition of either GreB or RNase, suggesting a mechanism besides backtracking or hairpin-stabilized pausing. that is not captured in the current model. Alternatively, such a pause might result from

misalignments when processing the experimental traces.

COMPARISON WITH OTHER MODELS

In recent decades, many attempts have been made to model the kinetics of transcription. Bai et al., Tadigotla et al. and others independently proposed models in which the kinetic of transcription is treated as a competition between the active transcription pathway and a branched pathway [9, 10, 12, 13, 22]. Although their models yield results in statistical agreement with experimental results, the predicted pauses differ from those observed in single-molecule measurements. In addition, previous models mostly focus on short ubiquitous and Class II backtracked pauses, without considering Class I, hairpin-stabilized pauses and the effects of tension and transcription factors.

The model described in this report drives the kinetics of transcription from a competition between one active and two branched pathways, and the details of the pathways differ from previous efforts. For example, backtracking is treated as a two-step mechanism in which the first step backwards must overcome a higher energy barrier than the successive steps. Unlike earlier models, the kinetics of Class I pauses are also included. By fitting specific kinetic parameters under specific experiment conditions, the resulting model achieves not only statistical agreement with experimental results, but reveals quantitative, detail regarding the effects of DNA sequences, applied tension, and transcription factors.

CONCLUSION AND OUTLOOK

The success of the model indicates that a thermodynamic consideration of the transcription complex can faithfully describe transcription kinetics. By incorporating both Class I and Class II pauses, the model refines our current understanding of active pathway and branched pathways in transcription and can be used to predict the occurrence of Class I and II pauses that regulate transcription.

Further improvements in our biochemical understanding of transcriptional pauses, in the quality of experimental data, and in the model itself, will likely improve its predictive power. For example, the model might predict the pause at site "c" if the mechanism of this pause is determined and incorporated. Longer spans of high resolution observations of transcription would also improve optimization of the model and the accuracy of predictions.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (NIH) grants R01 GM084070 to LF. We are grateful to Carlos Bustamante and Alex Tong for generously providing high resolution transcription traces

* jin.qian@emory.edu
† ddunpal@emory.edu
‡ lfinzi@emory.edu

[1] K. M. Herbert, A. Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick, and S. M. Block, Sequence-Resolved Detection of Pausing by Single RNA Polymerase Molecules, Cell **125**, 1083 (2006).
[2] E. A. Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick, and S. M. Block, Direct observation of base-pair stepping by RNA polymerase, Nature **438**, 460 (2005).
[3] I. Artsimovitch and R. Landick, Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals, Proceedings of the National Academy of Sciences **97**, 7090 (2000).
[4] I. Toulokhonov, I. Artsimovitch, and R. Landick, Allosteric Control of RNA Polymerase by a Site That Contacts Nascent RNA Hairpins, Science **292**, 730 (2001).
[5] J. Y. Kang, T. V. Mishanina, M. J. Bellecourt, R. A. Mooney, S. A. Darst, and R. Landick, RNA Polymerase Accommodates a Pause RNA Hairpin by Global Conformational Rearrangements that Prolong Pausing, Molecular Cell **69**, 802 (2018).
[6] M. H. Larson, R. A. Mooney, J. M. Peters, T. Windgassen, D. Nayak, C. A. Gross, S. M. Block, W. J. Greenleaf, R. Landick, and J. S. Weissman, A pause sequence enriched at translation start sites drives transcription dynamics in vivo, Science **344**, 1042 (2014).
[7] K. M. Herbert, J. Zhou, R. A. Mooney, A. L. Porta, R. Landick, and S. M. Block, E. coli NusG Inhibits Backtracking and Accelerates Pause-Free Transcription by Promoting Forward Translocation of RNA Polymerase, Journal of Molecular Biology **399**, 17 (2010).
[8] R. V. Dalal, M. H. Larson, K. C. Neuman, J. Gelles, R. Landick, and S. M. Block, Pulling on the Nascent RNA during Transcription Does Not Alter Kinetics of Elongation or Ubiquitous Pausing, Molecular Cell **23**, 231 (2006).
[9] L. Bai, A. Shundrovsky, and M. D. Wang, Sequence-dependent Kinetic Model for Transcription Elongation by RNA Polymerase, Journal of Molecular Biology **344**, 335 (2004).
[10] V. R. Tadigotla, D. O. Maoiléidigh, A. M. Sengupta, V. Epshtein, R. H. Ebright, E. Nudler, and A. E. Ruckenstein, Thermodynamic and kinetic modeling of transcriptional pausing, Proceedings of the National Academy of Sciences **103**, 4439 (2006).
[11] T. D. Yager, V. Hippel, and P. H, A thermodynamic analysis of RNA transcript elongation and termination in Escherichia coli, Biochemistry **30**, 1097 (1991).
[12] L. Bai, R. M. Fulbright, and M. D. Wang, Mechanochemical Kinetics of Transcription Elongation, Phys. Rev. Lett

98, 68103 (2007).

[13] D. O. Maoiléidigh, V. R. Tadigotla, E. Nudler, and A. E. Ruckenstein, A Unified Model of Transcription Elongation: What Have We Learned from Single-Molecule Experiments?, Biophysical Journal **100**, 1157 (2011).

[14] Y. W. Yin and T. A. Steitz, The Structural Mechanism of Translocation and Helicase Activity in T7 RNA Polymerase, Cell **116**, 393 (2004).

[15] B. E. Nickels and A. Hochschild, Regulation of RNA Polymerase through the Secondary Channel, Cell **118**, 281 (2004).

[16] M. H. Larson, W. J. Greenleaf, R. Landick, and S. M. Block, Applied Force Reveals Mechanistic and Energetic Details of Transcription Termination, Cell **132**, 971 (2008).

[17] A. Chauvier, J. F. Nadon, J. P. Grondin, A. M. Lamontagne, and D. A. Lafontaine, Role of a hairpin-stabilized pause in the Escherichia coli thiC riboswitch function, RNA Biology **16**, 1066 (2019).

[18] A. Xayaphoummine, T. Bucher, and H. Isambert, Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots, Nucleic Acids Res **33** (2005).

[19] I. Tinoco and C. Bustamante, The effect of force on thermodynamics and kinetics of single molecule reactions, Biophysical Chemistry , 513 (2002).

[20] R. Gabizon, A. Lee, H. Vahedian-Movahed, R. H. Ebright, and C. J. Bustamante, Pause sequences facilitate entry into long-lived paused states by reducing RNA polymerase transcription rates, Nature Communications **9**, 2930 (2018).

[21] K. C. Neuman, E. A. Abbondanzieri, R. Landick, J. Gelles, and S. M. Block, Ubiquitous Transcriptional Pausing Is Independent of RNA Polymerase Backtracking, Cell **115**, 437 (2003).

[22] A. Bochkareva, Y. Yuzenkova, V. R. Tadigotla, and N. Zenkin, Factor-independent transcription pausing caused by recognition of the RNA-DNA hybrid sequence, The EMBO Journal **31**, 630 (2012).