PHYSICAL REVIEW E **81**, 030904(R) (2010)

Influence of secondary structure on recovery from pauses during early stages of RNA transcription

A. V. Klopper, J. S. Bois, and S. W. Grill
Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38, D-01187 Dresden, Germany
and Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, D-01307 Dresden, Germany
(Received 31 March 2009; revised manuscript received 29 October 2009; published 18 March 2010)

The initial stages of transcription by RNA polymerase are frequently marked by pausing and stalling events. These events have been linked to an inactive backtracked state in which the polymerase diffuses along the template DNA. We investigate theoretically the influence of RNA secondary structure in confining this diffusion. The effective confinement length peaks at transcript lengths commensurate with early stalling. This finite-size effect accounts for slow progress at the beginning of transcription, which we illustrate via stochastic hopping models for backtracking polymerases.

DOI: 10.1103/PhysRevE.81.030904 PACS number(s): 87.14.gn, 87.14.ej, 87.15.bd, 87.15.Cc

The transcription of information encoded in the genome into RNA form plays a pivotal role in sustaining biological cellular function [1,2]. The process is executed with remarkable efficiency by the macromolecular machine RNA polymerase, which steps along a DNA template, assembling and extruding a complementary RNA transcript (Fig. 1). Following the initiation of transcription, a high density of polymerases on the template [3] indicates that progress is slow when the RNA transcript is very short. Elongation reaches normal speeds some time after the RNA has polymerized enough to emerge from the enzyme. This suggests that transcription may be self-regulated by its RNA product. Here, we investigate the notion that this regulation is controlled by an intrinsic physical mechanism pertaining to the ability of the RNA molecule to fold.

The region in the vicinity of the initiation site, termed the promoter, is characterized by *promoter-proximal pausing*, responsible for the initially slow transcription. While pausing is prevalent throughout the entire process, promoter-proximally paused polymerases experience exceptional difficulty in recovering from pauses in order to resume synthesis. By contrast, the polymerase productively elongates the RNA transcript once it has formed a length of around 50 nucleotides. It has been shown that the inability to recover can be induced via the truncation of the nascent RNA during normal elongation, which naturally supports the suggestion that elongation is controlled by the length of the emergent RNA transcript [4].

A large and important class of pauses involves the polymerase *backtracking* along the DNA template, maintaining the length of the RNA-DNA hybrid by feeding the newly formed RNA through a channel at the front of the enzyme [Fig. 1(b)] [5–7]. In the backtracked state, the polymerase is thought to move diffusively along the DNA [8–12] until it returns to the catalytic pathway and resumes synthesis. This recovery can be assisted by auxiliary proteins which realign the RNA by cleaving the portion of transcript in front of the RNA-DNA hybrid [5,13]. The short pieces of RNA are thought to have regulatory functions, and recent studies have linked their production to the positioning of nucleosomal roadblocks [14,15]. These experiments indicate that the short transcripts may be generated when nucleosome-induced backtracking results in transcript cleavage because the polymerase cannot recover without assistance.

Since diffusive return times in confined geometries scale with system size [16], one expects that the likelihood of *unassisted* recovery will be relatively small when the polymerase has ample space for backtracking. The sequence-specific pairings between nucleotides on an RNA molecule may influence backtracking by effectively shortening the length scale on which the polymerase can diffuse [17–19]. Furthermore, the absence of such a mechanism would be particularly prevalent at the first nucleosome, where the transcript is too short to form stable secondary structures. This provides a physical basis for the difficulty experienced by promoter-proximally paused polymerases in resuming synthesis. Here, we study the statistical mechanics of RNA secondary structures and reveal a finite-size effect in the confinement length of backtracked polymerases. This effect accounts for slow progress at the beginning of transcription, as we show through an analysis of stochastic hopping models of backtracking.

With this in mind, we now examine the length-dependent conformational characteristics of RNA. As depicted in Fig. 1, we define the number of consecutive unpaired bases of RNA adjacent to the polymerase, $\lambda[\mathcal{S}]$, for a given secondary structure $\mathcal{S}$. Secondary structure describes the microstate of the transcript and may be defined as a set of base pairs (i,j), with $i<j$ by convention, in which each base appears at most once. The mean number of unpaired bases adjacent to the polymerase is drawn from an equilibrium ensemble,

$$\langle \lambda(N) \rangle = Z(N)^{-1} \sum_{\mathcal{S} \in \Omega(N)} \lambda[\mathcal{S}] e^{-E[\mathcal{S}]}, \tag{1}$$

where $\Omega(N)$ defines the set of all allowed secondary structures for a transcript of N bases. The energy $E[\mathcal{S}]$ of a given structure is defined with respect to the energy associated with an unpaired base (in units of $k_B T$), and the corresponding partition function $Z(N)$ is given by

$$Z(N) = \sum_{\mathcal{S} \in \Omega(N)} e^{-E[\mathcal{S}]}. \tag{2}$$

The causal aspect of cotranscriptional folding (the simultaneous stepwise production and folding of the RNA transcript) suggests that a native ground state for the RNA strand may not necessarily be the most likely conformation since

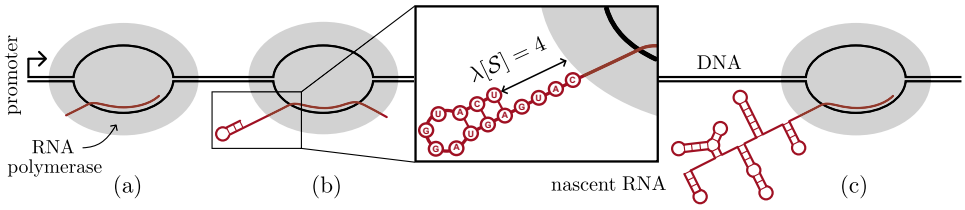

FIG. 1. (Color online) Stages of transcriptional elongation: (a) promoter clearance, (b) promoter-proximal pausing, and (c) productive elongation. The polymerase typically protects up to $N_{\mathrm{prot}} \sim 23$ bases of the RNA in elongation mode [20]. Inset: the number of consecutive unpaired bases adjacent to the polymerase in a particular secondary structure $\mathcal{S}$ is denoted by $\lambda[\mathcal{S}]$. In this example, $\lambda[\mathcal{S}]=4$.

the folding occurs in a time-dependent fashion. Perhaps more likely is a scenario in which sectional folds constituting nonnative metastable conformations are adopted by the transcript [21]. However, the dearth of microstates accessible to the short transcript in the vicinity of the promoter suggests that an equilibrium description is a good approximation for our purposes.

We can divide the sum in Eq. (1) into structures for which the base closest to the polymerase is paired (and thus $\lambda[\mathcal{S}]=0$) and those for which it is unpaired. The terms in the former set vanish, and we can write the remaining sum in terms of secondary structures for transcripts comprising $N-1$ bases. We obtain

$$\langle \lambda(N) \rangle = \frac{Z(N-1)}{Z(N)}(1 + \langle \lambda(N-1) \rangle). \tag{3}$$

This is consistent with the fact that the mean will be linear in N while the transcript is too small to form stable base pairings. During this early phase, the system has access to a single microstate corresponding to $Z(N)=1$. The linearity breaks down upon introduction of additional microstates due to self-interaction. This occurs when the number of transcribed bases exceeds a minimum, which is related to the RNA persistence length. We define this minimum as μ, such that any paired bases i and j must satisfy $|j-i| \geq \mu$. We can iterate the relation in Eq. (3) and note that $\langle \lambda(0) \rangle = 0$ to arrive at

$$\langle \lambda(N) \rangle = \frac{\sum_{k=0}^{N-1} Z(k)}{Z(N)}. \tag{4}$$

For large transcripts, both the numerator and denominator in this expression are rapidly increasing functions of N. An extensive free energy in the thermodynamic limit corresponds to a power-law partition function of the form

$$Z(N) \sim \alpha^N, \quad \alpha > 1. \tag{5}$$

Such an ansatz leads to a finite saturation of $\langle \lambda(N) \rangle$ in the limit of long transcripts,

$$\langle \lambda(N) \rangle \sim \frac{1-\alpha^{-N}}{\alpha - 1} \underset{N \to \infty}{\longrightarrow} \frac{1}{\alpha - 1}. \tag{6}$$

This has interesting implications for the limit in which the polymerase elongates productively. Here, the addition of a nucleotide incites the growth of the self-interacting part of the transcript and has no effect on the length of the unpaired extremal portion of RNA near the polymerase [22].

The crossover from linear behavior to this saturation falls precisely in the region we wish to probe. While it is clear that both the numerator and denominator in Eq. (4) increase monotonically for $N > \mu$, the behavior of their quotient is sensitively dependent on both μ and the effective energy associated with base-pairing interactions. The crossover region is characterized by competition between the restricted microstate space for small transcripts and the relative statistical weight of self-interacting structures. The mean exhibits a maximum where $\partial_N \langle \lambda(N) \rangle = 0$ such that

$$\partial_N \left(\ln \sum_{k=0}^{N-1} Z(k) \right) \approx \partial_N \ln Z(N), \tag{7}$$

where ∂_N denotes a continuum approximation of the derivative with respect to N.

Depending on the intrinsic structural characteristics of the transcript, this may occur prior to saturation, forming a maximum for finite N. Introducing an appropriate model for the energy $E[\mathcal{S}]$ of each secondary structure, Eq. (4) can be evaluated explicitly. Ideally, this model should account for all contributions to secondary structure free energy, including favorable base-stacking interactions and unfavorable bending and looping of the RNA.

A first approximation dispenses with sequence information and identifies the energy of each possible structure as a sum of all base-pairing energies E_0, such that the energy associated with a secondary structure depends only on its magnitude, $E[\mathcal{S}] = E_0 |\mathcal{S}|$. In this case, the base-pairing energy describes an *effective* energy, which encodes both energy penalties and energy payoffs in a single value, common to all pairs along the transcript. The partition function corresponding to this model forms a sum of polynomials in e^{-E_0} with combinatorial coefficients. The resulting mean number of consecutive unpaired bases adjacent to the polymerase is plotted as a dotted line in Fig. 2.

Within this simplified framework, we can include important physical characteristics influencing state space by excluding certain secondary structures from the ensemble. This includes intertwining loop structures (or *pseudoknots*), which are infrequently seen in naturally occurring RNA and typically omitted from analysis [23]. We form the reduced ensemble by considering a restricted set of base pairs for which any two pairs (i,j) and (i',j'), with $i<i'$, can either be independent where $i<j<i'<j'$, or nested where $i<i'<j'<j$. The partition function for a strand containing

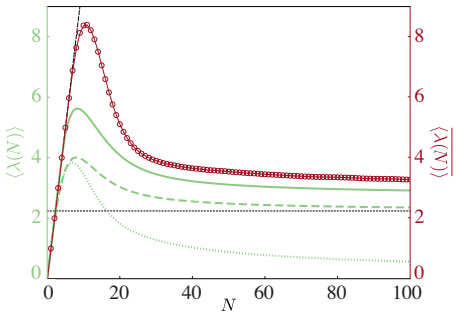

FIG. 2. (Color online) Mean number of consecutive unpaired bases adjacent to the polymerase as a function of the number of unprotected bases, N. Curves illustrate $\langle\lambda(N)\rangle$ for homogeneous sequences using a simple base-pairing energy model with $E_0=3$ and $\mu=1$ [green (light) dashed line], $\mu=4$ [green (light) solid line], and $\mu=1$ with all pairings included [green (light) dotted line]; $\langle\lambda(N)\rangle=N$ (black dashed line) and the predicted asymptotic value for $\mu=1$ (black dotted line). Red (dark) circles depict $\langle\lambda(N)\rangle$ for random sequences computed via an exact enumeration algorithm.

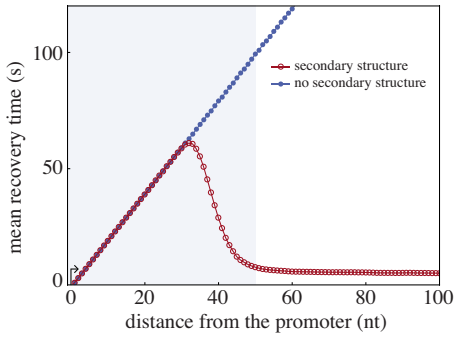

FIG. 3. (Color online) Mean recovery time for a polymerase backtracking on an ensemble of random sequences with (red empty circles) and without (blue filled circles) confining secondary structure as a function of distance from the promoter in units of nucleotides. Shading indicates the region of slow elongation.

N bases can then be written recursively in the form

$$Z(N) = Z(N-1) + \sum_{k=1}^{N-\mu} Z(k-1)Z^{\mathrm{cl}}(N-k+1), \qquad (8)$$

where $Z(0)=Z(1)=1$ and $Z^{\mathrm{cl}}(N)$ denotes the closed partition function associated with a substrand whose ends form a base pair. Here, we have built in an effective persistence length by restricting the upper limit of the sum to $N-\mu$.

This model has been studied extensively [24–27] and the case for which an effective energy is associated with each base pair yields $Z^{\mathrm{cl}}(N)=e^{-E_0}Z(N-2)$. This has been approximated in the thermodynamic limit for $\mu=1$, revealing an asymptotic length dependence of $N^{-3/2}(1+2e^{-E_0/2})^N$ [25–27]. In the large-N limit, we should thus expect to see the behavior described in Eq. (6) with $\alpha=(1+2e^{-E_0/2})$. For small N, the energy required to bend the transcript tends to dominate any base-pairing payoff [28]. With this in mind, we take $E_0>0$ (a base-pairing penalty) and plot an exact enumeration of $\langle\lambda(N)\rangle$ along with the predicted asymptote in Fig. 2. The mean peaks at a value determined by the effective energy attributed to base pairing. We compare the case for which $\mu=1$ to that in which it is a more realistic $\mu=4$. The deviation from linear behavior coincides with the introduction of multiple microstates. Both cases exhibit a saturation to finite $\langle\lambda(N)\rangle$ in the limit of large N. In the former case, this value is approximated well by the predicted asymptotic value. The saturation contrasts strongly with the case for which all possible base pairs are included (dotted line in Fig. 2). There, the decay is much faster and reflects the fact that we overlook any topological difficulty posed by base pairing.

While these simple models exhibit the expected linear regime and saturation at finite N, they lack the scope to differentiate between base pairs which attract an energy payoff and those which cost energy. An increase in the payoff associated with stacked base pairs is expected to augment the statistical weight associated with polymerase-proximal conformations, resulting in a narrower peak saturating at a lower value of $\langle\lambda(N)\rangle$. This suggests that a sequence prone to stacking (i.e., one containing long complementary subsequences) may curb the tendency to backtrack by reducing the unpaired strand length adjacent to the polymerase early in the process of transcription. The expression in Eq. (8) can be computed more accurately using a sequence-dependent empirically parametrized [29] exact enumeration algorithm for secondary structure prediction [30]. Utilizing the NUPACK software [31], we are able to encode full sequence and structural information, and we do so by constructing an ensemble of 4×10^5 randomly sequenced transcripts and averaging $\langle\lambda(N)\rangle$ over this set. We obtain a disorder-averaged quantity $\langle\lambda(N)\rangle$, which is plotted in Fig. 2. The resulting curve is remarkably similar to those given by the reduced models, suggesting that the small-N behavior is an inherent property of linear base-pairing polymers.

We can similarly calculate an expected time of recovery from a backtrack on a given sequence, assuming that the secondary structure imposes a confining length scale on the diffusive motion of the polymerase. We undertake a simple hopping simulation corresponding to an unbiased random walk in one dimension subject to a reflective boundary condition. The recovery time is given by the mean first passage time for a polymerase starting at site $N-1$ (in the reference frame of the RNA, upon entering a backtrack) to reach N and resume elongation. Stepping times are exponentially distributed with a mean of 1 s [12]. The reflective boundary lies at the promoter for configurations without base pairings and $\langle\lambda(N)\rangle$ nucleotides upstream from the active site for all other configurations. In terms of $\langle\lambda(N)\rangle$, for a given sequence the effective confinement of the walk can be written as

$$L_{\mathrm{eff}} = \langle\lambda(N)\rangle + N_{\mathrm{prot}}/Z(N), \qquad (9)$$

where $N_{\mathrm{prot}}\sim 23$ is the number of protected bases of RNA [20]. We take an ensemble of random sequences and plot the disorder-averaged mean recovery time in Fig. 3. The shaded area indicates the region within which promoter-proximal pausing is experimentally observed [1]. Clearly, there is a marked increase in the expected recovery time for a backtracking polymerase in this region. The typical time required to exit the backtracked state via cleavage has not been accu-

rately measured but is expected to be on the order of 10 s [11]. It would therefore appear that cleavage is far more likely in the promoter-proximal region, owing to the absence of secondary structure. Polymerases undergoing unrestricted walks are expected to have monotonically increasing recovery times as they move further away from the promoter, while those confined by secondary structure have a reduced and length-independent expected recovery time beyond the promoter-proximal region.

We have presented a simple statistical physics argument implicating the self-interactive nature of RNA in the regulation of early transcription. In all physically realistic models used to describe the secondary structure of the nascent RNA strand, we find that the mean number of consecutive unpaired bases adjacent to the polymerase displays a peak where N is small and decays to a finite value in the thermodynamic limit. This behavior is induced by the restrictions imposed by the conformational space of a base-pairing linear polymer. The effect manifests itself in a markedly increased expected recovery time for a backtracking polymerase within the region characterized by promoter-proximal pausing. Our simulation results are directly comparable with single-molecule experiments. A close analysis of the recovery characteristics of transcripts too short or otherwise unable to fold due to sequence incompatibilities would provide the relevant statistics needed to verify our findings experimentally and would confirm that the absence of pairwise interactions of RNA prevents a backtracking polymerase from resuming elongation in the promoter-proximal region.

The authors wish to thank A. Carlsson, M. Depken, E. Galburt, F. Jülicher, B. Lindner, K. Neugebauer, D. Ó Maoiléidigh, and J. Parrondo for insightful discussion. J.S.B. acknowledges the Human Frontier Science Program for funding.

[1] T. Margaritis and F. C. P. Holstege, Cell **133**, 581 (2008).
[2] S. J. Greive and P. H. von Hippel, Nat. Rev. Mol. Cell Biol. **6**, 221 (2005).
[3] L. J. Core and J. T. Lis, Science **319**, 1791 (2008).
[4] A. Újvári, M. Pal, and D. S. Luse, J. Biol. Chem. **277**, 32527 (2002).
[5] N. Komissarova and M. Kashlev, Proc. Natl. Acad. Sci. U.S.A. **94**, 1755 (1997).
[6] N. Komissarova and M. Kashlev, J. Biol. Chem. **272**, 15329 (1997).
[7] E. Nudler, A. Mustaev, E. Lukhtanov, and A. Goldfarb, Cell **89**, 33 (1997).
[8] K. M. Herbert, W. J. Greenleaf, and S. M. Block, Annu. Rev. Biochem. **77**, 149 (2008).
[9] L. Bai, T. J. Santangelo, and M. D. Wang, Annu. Rev. Biophys. Biomol. Struct. **35**, 343 (2006).
[10] J. W. Shaevitz, E. A. Abbondanzieri, R. Landick, and S. M. Block, Nature (London) **426**, 684 (2003).
[11] E. A. Galburt, S. W. Grill, A. Wiedmann, L. Lubkowska, J. Choy, E. Nogales, M. Kashlev, and C. Bustamante, Nature (London) **446**, 820 (2007).
[12] M. Depken, E. A. Galburt, and S. W. Grill, Biophys. J. **96**, 2189 (2009).
[13] R. J. Sims III, R. Belotserkovskaya, and D. Reinberg, Genes Dev. **18**, 2437 (2004).
[14] R. J. Taft *et al.*, Nat. Genet. **41**, 572 (2009).
[15] R. J. Taft, C. D. Kaplan, C. Simons, and J. S. Mattick, Cell Cycle **8**, 2332 (2009).
[16] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, Berlin, 1983).
[17] V. R. Tadigotla, D. Ó Maoiléidigh, A. M. Sengupta, V. Epshtein, R. H. Ebright, E. Nudler, and A. E. Ruckenstein, Proc. Natl. Acad. Sci. U.S.A. **103**, 4439 (2006).
[18] D. Ó Maoiléidigh, Ph.D. thesis, Rutgers University, 2006.
[19] T. C. Reeder and D. K. Hawley, Cell **87**, 767 (1996).
[20] J. Andrecka, R. Lewis, F. Brückner, E. Lehmann, P. Cramer, and J. Michaelis, Proc. Natl. Acad. Sci. U.S.A. **105**, 135 (2008).
[21] D. Thirumalai and C. Hyeon, Biochemistry **44**, 4957 (2005).
[22] Independently recognized by A. Ben-Shaul (unpublished).
[23] I. Tinoco, Jr. and C. Bustamante, J. Mol. Biol. **293**, 271 (1999).
[24] P. G. Higgs, Phys. Rev. Lett. **76**, 704 (1996).
[25] P. G. de Gennes, Biopolymers **6**, 715 (1968).
[26] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).
[27] R. Bundschuh and T. Hwa, Phys. Rev. E **65**, 031903 (2002).
[28] V. A. Bloomfield, D. M. Crothers, and I. Tinoco, Jr., *Nucleic Acids: Structures, Properties, and Functions* (University Science Press, Mill Valley, CA, 2001).
[29] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, J. Mol. Biol. **288**, 911 (1999).
[30] J. S. McCaskill, Biopolymers **29**, 1105 (1990).
[31] R. M. Dirks and N. A. Pierce, J. Comput. Chem. **24**, 1664 (2003).