Thermodynamic and kinetic modeling of transcriptional pausing

Vasisht R. Tadigotla*†, Dáibhid Ó Maoiléidigh*†‡, Anirvan M. Sengupta*‡, Vitaly Epshtein§, Richard H. Ebright*¶‖, Evgeny Nudler§, and Andrei E. Ruckenstein*‡**

*BioMaPS Institute for Quantitative Biology, ‡Department of Physics and Astronomy, ¶Department of Chemistry and Chemical Biology, and ‖Howard Hughes Medical Institute and Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; and §Department of Biochemistry, New York University Medical Center, New York, NY 10016

Communicated by Peter H. von Hippel, University of Oregon, Eugene, OR, January 20, 2006 (received for review September 10, 2005)

We present a statistical mechanics approach for the prediction of backtracked pauses in bacterial transcription elongation derived from structural models of the transcription elongation complex (EC). Our algorithm is based on the thermodynamic stability of the EC along the DNA template calculated from the sequence-dependent free energy of DNA–DNA, DNA–RNA, and RNA–RNA base pairing associated with (*i*) the translocational and size fluctuations of the transcription bubble; (*ii*) changes in the associated DNA–RNA hybrid; and (*iii*) changes in the cotranscriptional RNA secondary structure upstream of the RNA exit channel. The calculations involve no adjustable parameters except for a cutoff used to discriminate paused from nonpaused complexes. When applied to 100 experimental pauses in transcription elongation by *Escherichia coli* RNA polymerase on 10 DNA templates, the approach produces statistically significant results. We also present a kinetic model for the rate of recovery of backtracked paused complexes. A crucial ingredient of our model is the incorporation of kinetic barriers to backtracking resulting from steric clashes of EC with the cotranscriptionally generated RNA secondary structure, an aspect not included explicitly in previous attempts at modeling the transcription elongation process.

cotranscriptional folding | statistical mechanics

Transcription is the first step in protein synthesis and the step at which most regulation of gene expression takes place. The transcription process is carried out by RNA polymerase (RNAP) (1–3), a multisubunit molecular motor, the basic structure of which is conserved from bacteria to eukaryotes (4). Over the past decade a great deal has been learned about the structure of RNAP, particularly in the context of yeast (5, 6), thermophilic bacteria (7, 8), and *Escherichia coli* (9–12). Given the high degree of structural conservation of RNAP, and the synthesis of structural, biochemical, and kinetic information from bacteria and yeast, we are able to begin building mechanistic models of transcription valid across species.

This article focuses on the general question of whether our current knowledge is sufficient to produce predictive quantitative models of transcription. In particular, we consider the possibility of predicting pause positions where RNAP halts either reversibly (pauses) or irreversibly (arrests) (13). Pauses, which are characterized by highly variable durations and efficiencies (3), are an ubiquitous aspect of transcription elongation and are known to play regulatory roles particularly in synchronizing transcription with other biological processes, such as translation in bacteria (14), factor-dependent and factor-independent termination (15, 16), and interactions with regulatory proteins (17). Even though pauses are not associated with a consensus sequence, pause positions along the template are strongly sequence-dependent. This sequence dependence is encoded indirectly, through the effect of base-pairing nucleic acid interactions on the motion of RNAP along DNA.

The thermodynamic properties of transcription elongation have been studied in the pioneering articles of von Hippel and collaborators (18, 19) (see also ref. 3 and references therein). von Hippel and collaborators have also shed light on the kinetics of RNA synthesis (20) and have discussed the kinetics of pausing in the context of factor-mediated antitermination (21). Although this body of work has had significant impact on our understanding of transcription in general and of transcription elongation in particular, no attempt at a quantitative prediction of transcriptional pausing has appeared in the literature until the recent article of Bai *et al.* (22). This work sought to address the problem of pause prediction within a simplified kinetic model ignoring all effects of RNA folding. Although a nontrivial step forward, the model presented in ref. 22 appears to have limited predictive power (see below).

The limited success in predicting transcriptional pauses underscores the complexity of the problem. Both transcription elongation and pausing are intrinsically nonequilibrium, kinetic phenomena that are affected by both the transcribed DNA sequence and the folding of the upstream RNA. Although the free energy associated with RNA folding is included in discussions of the thermodynamics of RNA synthesis (20), the role of slow folding–unfolding kinetics in the regulation of transcription elongation and pausing have been ignored up to this point. The two principal aims of this article are (*i*) to critically evaluate the capability of the thermodynamic model to predict transcriptional pausing both in the presence and in the absence of RNA folding effects and (*ii*) to compare the thermodynamic predictions with those of a highly simplified kinetic model that incorporates the qualitative effect of kinetic barriers induced by RNA folding.

Model of Transcription Elongation

The basis for our quantitative analysis is the structural and mechanistic model of the elongation complex (EC) (3, 5, 10, 23), sketched in Fig. 1. The EC consists of a melted DNA duplex region of 12–14 nt (transcription bubble) enclosed within RNAP and stabilized by interactions with the enzyme and with the last 8 or 9 nt of the synthesized RNA transcript (the DNA–RNA hybrid). The RNA transcript upstream of the hybrid exits RNAP via the "RNA exit channel"; whereas the duplex DNA downstream of the bubble threads through a "sliding clamp" in the enzyme, which holds on tightly to the DNA while allowing for smooth sliding during transcription elongation. The "secondary channel" provides the access of the incoming nucleoside triphosphate (NTP) to the active center where the catalysis of phosphodiester bond formation takes place, resulting in the elongation of the RNA transcript by one nucleotide. Immediately after the transcript elongation step the EC is in the so-called "pretranslocated" state (translocational state 0)

Conflict of interest statement: No conflicts declared.

Abbreviations: EC, elongation complex; MBF, multiple bubbles with RNA folding; MBNF, multiple bubbles without RNA folding; NTP, nucleoside triphosphate; PPV, positive predictive value; RNAP, RNA polymerase; SBF, single bubble with RNA folding; SBNF, single bubble without RNA folding; TP, true positives.

†V.R.T. and D.Ó M. contributed equally to this work.

**To whom the correspondence should be addressed. E-mail: andreir@physics.rutgers.edu.

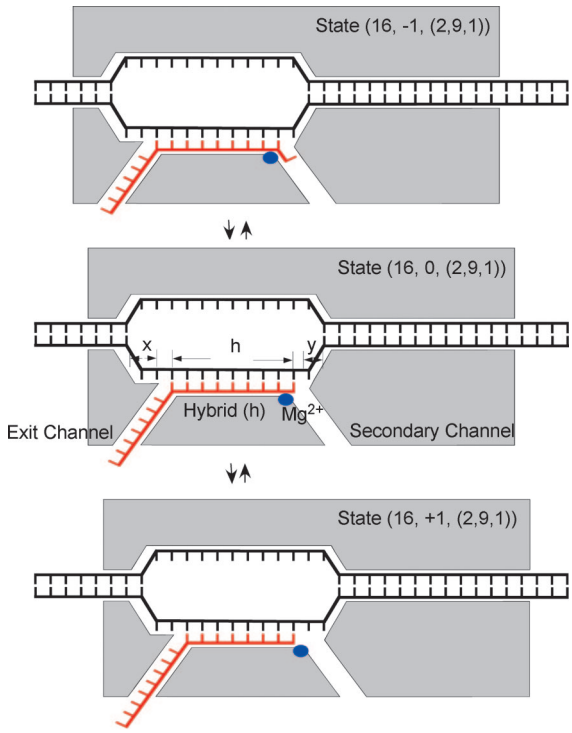

Fig. 1. Model of the EC. Shown is a schematic representation of the EC, in a state labeled as $(m, n, b) \equiv (m, n, (x, h, y))$. The figure shows three translocational states, $n = -1$, $n = 0$, and $n = +1$, for a fixed transcript length of 16; at position 0 the 3′ end of RNA occupies the active center of the enzyme (the blue dot); x and y indicate, respectively, the number of unpaired DNA bases upstream (to the left) and downstream (to the right) of the hybrid of length h, and thus the size of the bubble is given by $s_b = x + h + y$. In averaging over the bubble configurations, x, h, and y were varied between 2 and 5, 8 and 9, and 1 and 2, respectively, resulting in a variation of s_b from 11 to 16.

in which the 3′ end of the transcript overlaps the catalytic site. The next incorporation step requires that RNAP translocate forward by one base pair, into the "posttranslocated" state (translocational state +1), making the catalytic center available for the binding of the next NTP.

There are two basic proposals for how translocation occurs: the "powerstroke" mechanism makes use of the energy released during phosphodiester bond formation to drive translocation between states 0 and +1 (24, 25); whereas in the "Brownian ratchet" mechanism bidirectional translocational steps occur stochastically as a result of reversible thermal (Brownian) motion of the polymerase along DNA (22, 26–28). In the latter case the phosphodiester bond formation rectifies the motion and resets the system into a state with a longer transcript length, poised to incorporate the next NTP. In this article we adopt the Brownian ratchet picture, recently supported by single-base-pair-resolution single-molecule experiments (29).

According to this latter picture, between incorporation steps, RNAP can undergo further reverse translocation ("backtracking") accessing translocational states −1, −2, −3, . . . (see Fig. 1) and can potentially undergo further forward translocation ("hypertranslocation") accessing translocational states +2, +3, +4, . . . During both backtracking and hypertranslocation the catalytic center of the enzyme looses contact with the 3′ end of RNA, and the EC leaves the productive elongation pathway. Both motions are believed to be involved in transcription regulation: it has been proposed that hypertranslocation provides a mechanism for intrinsic termination of transcription (30), whereas backtracking plays a role in controlling transcription fidelity by affecting the speed of elongation and facilitating the editing action of cleavage factors (31).

The motion of RNAP along the nucleic acid scaffold can be visualized in terms of thermally induced transitions between free energy minima associated with translocational states, n, along the template, for a fixed length of the RNA transcript, m. The free energy difference between two neighboring translocational states, n and $n + 1$, is associated with (*i*) the breaking the DNA–DNA base pair in front and annealing the base pair behind the moving bubble, (*ii*) the change in the RNA–DNA hybrid base-pairing, (*iii*) the folding of the RNA transcript protruding out of the exit channel (1, 3, 18), and (*iv*) the possible change in RNAP–DNA and RNAP–RNA interactions.

The above picture defines a free energy landscape that, for a strictly uniform sequence [e.g., poly(A)], would be flat for positions upstream of position 0: in this case the energy cost in breaking the base pair in the direction of motion is exactly compensated by the energy gain from annealing the base pair at the other end of the bubble while the hybrid energy remains unchanged [note that for a poly(A) sequence there is no folding of the RNA transcript]. On the other hand, the cost of shortening of the hybrid associated with each forward translocation step results in an uphill free energy profile downstream of position 0. Clearly, for real biological DNA templates the sequence dependence profoundly affects the free energy landscape. In particular, the free energy contributions from (*i*)–(*iii*) above involve the nucleic acid duplex stabilities, which are highly sequence-dependent (32, 33); whereas the contribution from (*iv*) above includes the contacts of nucleic acid backbones with RNAP and is not expected to be strongly sequence-dependent.

Equilibrium Modeling of Transcriptional Pausing

In this article we will focus on the identification of backtracked pauses [also referred to as class II pauses (34)], i.e., pauses resulting from the backtracking of RNAP at positions along the sequence corresponding to a particularly weak hybrid. The model presented here, although sufficient to identify potential sites for hairpin-induced (class I) pauses (34), is too simplistic to determine their stability because these pauses are expected to involve sequence-nonspecific interactions between the RNA fold and subunits of the enzyme (35).

Strictly speaking, achieving equilibrium at each position along a template (for a fixed transcript length) requires a sufficiently long time so that RNAP has a chance to equilibrate over all translocational states while, at the same time, the nascent RNA transcript is able to reach its lowest free energy folded conformation for each and every translocation position. In such an idealized equilibrium situation, RNAP would pause at all positions corresponding to deep minima of the free energy landscape upstream of translocational state 0. In practice, given the slow unfolding rates of RNA secondary structure with respect to typical translocation rates (see *Kinetic Modeling of Transcriptional Pausing*), reaching equilibrium at each translocation position is unlikely to occur on experimentally relevant time scales. In this case, RNAP is only expected to backtrack until it encounters a fold of the RNA secondary structure, which provides a kinetic barrier to further backtracking. Moreover, elongation proceeds without pausing through the shallower minima of the free energy landscape, implying that in a real experiment one should observe fewer pauses than predicted by a strict equilibrium analysis.

To identify backtracked pauses from an equilibrium calculation, we have formulated a minimal set of heuristic rules that mimic the fact that in real experiments RNAP only equilibrates over a finite region in the vicinity of translocational position 0. Our working hypotheses are as follows: (*i*) candidate backtracked pause positions correspond to backtracked minima of the free energy profile; (*ii*) consecutive pauses resulting from the repeated backward shift of a particular pause by 1 nt with each transcript elongation step are clustered together and counted as a single pause [we note that criterion (*i*) implies that a cluster grows until a new minimum appears closer to position 0]; and (*iii*) a free energy minimum

satisfying both (*i*) and (*ii*) is classified as a pause if the time of recovery of an EC backtracked to that position is longer than a cutoff, to be determined by optimizing the statistical significance of the results.

Statistical Mechanics Approach. We are now in position to discuss the calculation of the free energy profile as a function of position along the template. The state of the EC is labeled by m, the transcript length; n, the translocation position; and b, the bubble configuration. The equilibrium assumption implies that the probability of a state (m, n, b) of the EC is given by the Boltzmann distribution,

$$P_{n,b}^{(m)} = Z_m^{-1} e^{-\frac{G_{m,n,b}}{k_B T}}; \quad Z_m = \sum_{n,b} e^{-\frac{G_{m,n,b}}{k_B T}}, \qquad [1]$$

where k_B is Boltzmann's constant, and T is the temperature. The bubble configuration $b \equiv (x, h, y)$ is described in terms of $x(y)$, the number of unpaired DNA bases upstream (downstream) of the hybrid of length h, resulting in a bubble of size $s_b = x + h + y$ (see Fig. 1). $G_{m,n,b}$, the free energy associated with the state (m, n, b), can be decomposed as

$$G_{m,n,b} = G_{m,n,b}^{\mathrm{DNA-DNA}} + G_{m,n,b}^{\mathrm{RNA-DNA}} + G_{m,n,b}^{\mathrm{RNA-RNA}} + G^{NS}. \qquad [2]$$

Here, DNA–DNA, RNA–DNA, and RNA–RNA labels, respectively, the sequence-dependent free energy cost in forming the transcription bubble of size s_b, the free energy of the RNA–DNA hybrid and the free energy associated with the secondary structure of the RNA upstream of the exit channel. G^{NS} incorporates all sequence nonspecific interactions of nucleic acids with the enzyme and is taken to be a constant in the current calculation.

Multiple Bubble Configurations. Previous models of transcription have used a fixed-length transcription bubble (12–17 bp) and a fixed length RNA–DNA hybrid (8–12 bp) (18, 19, 22). These assumptions are unjustified from a physical point of view: in principle, both the size of the bubble and the size of the hybrid can change spontaneously as a result of thermal fluctuations with probabilities that decrease exponentially with the associated free energy cost. The notion of multiple bubbles is consistent with changes in the size of the transcription bubble recently detected experimentally during transcription initiation (R. H. Ebright, personal communication). To account for the variation in bubble/hybrid size, we have averaged over 16 different bubble configurations (see Fig. 1) corresponding to bubble lengths 11–16 nt and a hybrid of 8–9 nt. The results are insensitive to the choice of the upper bound on the bubble size (because larger bubbles appear with exponentially small probability), and all other bounds are consistent with the constraints imposed by the structural model. It then follows from Eq. **1** that the equilibrium probability for an EC at position n, for a fixed transcript length but including multiple bubble configurations, is given by

$$P_n^{(m)} = \textstyle\sum_b P_{n,b}^{(m)} \equiv Z_m^{-1} e^{-\frac{G_{m,n}^{eff}}{k_B T}}. \qquad [3]$$

Thus, in the presence of bubble and hybrid size fluctuations the pause positions will be identified from the minima of the effective free energy, $G_{m,n}^{\mathrm{eff}} = -k_B T[\ln P_n^{(m)} + \ln Z_m]$. The DNA–DNA (bubble) and RNA–DNA (hybrid) contributions to $G_{m,n}^{\mathrm{eff}}$ are calculated by using a nearest-neighbor thermodynamic model for DNA–DNA and DNA–RNA duplex stabilities (at 37°C) using parameters from ref. 32, whereas the RNA–RNA (folding) contribution is obtained from the VIENNARNA package (33). The number of hypertranslocated configurations is bounded by the length of the hybrid, whereas the backtracked ones are fixed at 9, to account for paused complexes known to backtrack by >6 nt (36). Increasing backtracking beyond 9 nt does not affect our results in the presence of RNA folding (see below). In addition, we allow the entire RNA transcript beyond the exit channel to fold.

Criteria for Pausing. As already implied above, our pausing criterion is based on the intuitive notion that, for transcript lengths associated with transcriptional pauses, the incorporation rate of the next nucleotide drops substantially below the maximum elongation rate along the template. The steady state elongation rate of a transcript of length m to the next transcript length, $m + 1$, for a particular sequence and NTP concentration, [N], is given by $k_E^m = [N]/([N] + K_d^{\mathrm{eff}}(m))$ (taking the PP_i release rate, $\gamma_P = 1$), with an effective dissociation constant,

$$K_d^{\mathrm{eff}}(m) = K_d \Bigg[\Bigg(1 + e^{\frac{(G_{m,+1}^{\mathrm{eff}} - G_{m,0}^{\mathrm{eff}})}{k_B T}} + e^{\frac{(G_{m,+1}^{\mathrm{eff}} - G_{m,-1}^{\mathrm{eff}})}{k_B T}} + e^{\frac{(G_{m,+1}^{\mathrm{eff}} - G_{m,-2}^{eff})}{k_B T}} + e^{\frac{(G_{m,+1}^{\mathrm{eff}} - G_{m,-3}^{\mathrm{eff}})}{k_B T}} + \ldots \Bigg) + \Bigg(e^{-\frac{(G_{m,+2}^{\mathrm{eff}} - G_{m,+1}^{\mathrm{eff}})}{k_B T}} + e^{-\frac{(G_{m,+3}^{\mathrm{eff}} - G_{m,+1}^{\mathrm{eff}})}{k_B T}} + e^{-\frac{(G_{m,+4}^{\mathrm{eff}} - G_{m,+1}^{\mathrm{eff}})}{k_B T}} + \ldots + e^{-\frac{(G_{m,+8}^{\mathrm{eff}} - G_{m,+1}^{\mathrm{eff}})}{k_B T}} \Bigg) \Bigg]. \qquad [4]$$

To account for the RNA-folding kinetic barriers resulting from the finite time scales accessible experimentally, we restrict the backtracked states for each transcript length in Eq. **4** to the sequence downstream of the position where RNAP first clashes into a RNA secondary structure. At these positions, further backtracking of RNAP incurs a penalty for breaking one or more base pairs of the fold. This penalty leads to a decrease in the backtracking rate by at least a factor of 10 corresponding to breaking of one or more RNA–RNA base pairs [with binding energy E, $2\ k_B T < E < 5.75\ k_B T$ (33) per base pair].

Identifying whether RNAP pauses at a particular transcript length, m, on a given template involves calculating the maximum elongation rate over all m, $\max_m(k_E^m)$, and counting the particular m as a *pause site* if the elongation rate at that transcript length is sufficiently slow, $k_E^m < \xi \max_m(k_E^m)$. [Note that, as in the gel experiments used to identify transcriptional pauses (36–38), a *pause site* refers to the transcript length for which pausing occurs, without reference to the precise backtracked position along the template.] The cutoff parameter, ξ ($0 < \xi < 1$), is determined by optimizing the statistical significance of our predictions over all experimental templates.

Statistical Significance. Evaluating the statistical significance requires optimizing some appropriate combination of conventional statistical measures: the numbers of *true positives*, TP (correctly predicted experimental pauses); *false positives*, FP (predicted pauses not seen experimentally); and *false negatives*, FN (missed experimental pauses). Typical choices are the *positive predictive value*, PPV = TP/(TP + FP) (the fraction of predicted pauses that are correct), and the *sensitivity*, σ = TP/(TP + FN) (the fraction of experimental pauses correctly identified). Although the performance of the algorithm could be optimized, for example, by maximizing PPV and σ simultaneously, we find it more transparent to optimize and plot a single statistical measure of performance. In particular, we choose to minimize the proportion of incorrect to correct predictions, η_1 = (FP + FN)/TP. We have also examined a number of other measures of performance, such as η_2 = PPV + σ (see the

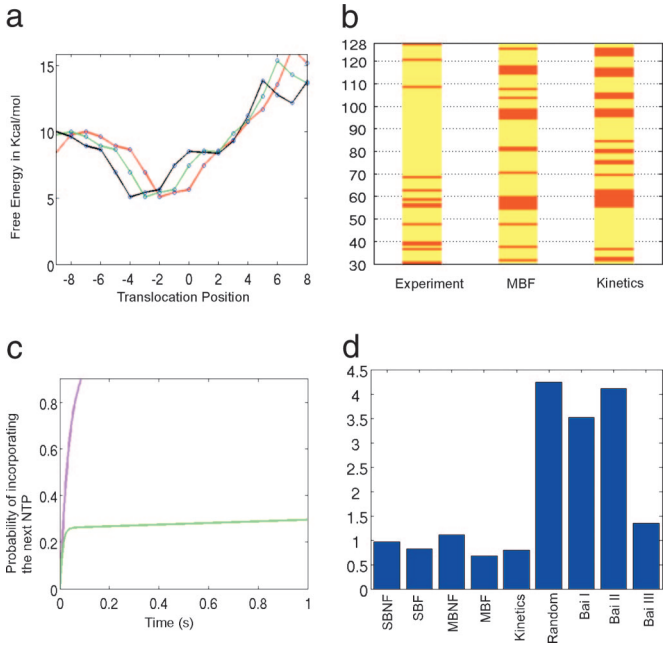

Fig. 2. Equilibrium and kinetic results. (*a*) Example of a pause cluster, +32 (red), +33 (green), +34 (black) on template seq11: these are consecutive pauses backtracked by 2, 3, and 4 nt, respectively, resulting from the backward shift of a local free energy minimum. (*b*) Illustration of pause clusters on D111 comparing experimental data with the multiple-bubble-folding model (MBF) and kinetics (see text). Red (yellow) indicates pause (nonpause) sites. (*c*) Incorporation probability curves as a function of time for transcript length 85 on D167. In the absence of folding barriers, the recovery rate is slow, consistent with a backtracked pause at position 85 (green curve); in the presence of folding barriers, backtracking is suppressed, and the EC recovers quickly, eliminating the pause (magenta curve). (*d*) Overall pause statistics on all 10 templates using the proportion of incorrect to correct pause predictions, η_1. For comparison, *d* includes results from kinetics (see *Kinetic Modeling of Transcriptional Pausing*), results from the converged average of many random assignments of pause sites (Random), and results obtained on our 10 templates by using the model of Bai *et al.* (22) (Bai I, Bai II, and Bai III).

supporting information, which is published on the PNAS web site), all of which yield qualitatively similar results. To underscore the statistical significance, we compare our predictions for four models (see below) with the corresponding results obtained by randomly assigning the same number of pause sites as found experimentally, from a discrete uniform distribution along each template. Finally, we demonstrate that pause patterns are encoded in a sequence-specific manner, by comparing our results with those obtained when our algorithm is applied to randomized sequences with the same length and AT/GC content as our templates.

Equilibrium Results. The model was used to analyze 10 templates with a total of 100 known pauses (see the supporting information): seq10 (36), seq11, seq12, seq13 (V.E. and E.N., unpublished work), D123, D167, D111, D112, D104, and D387 (37). We apply our pausing criteria to four thermodynamic models (implemented in MATLAB): (*i*) single bubble (2, 9, 1) without RNA folding (SBNF), (*ii*) single bubble with RNA folding (SBF), (*iii*) multiple bubbles without RNA folding (MBNF), and (*iv*) multiple bubbles with RNA folding (MBF). Fig. 2*a* illustrates the clustering of pauses as a result of the backward shift of a particular free energy minimum with increasing transcript length: the pause sites 32, 33, and 34 from seq11 originate from the shift of the backtracked pause at site 32 by 1 and 2 nt, respectively. For each of the sites 35, 36, and 37, the first free energy minimum is located at translocation position 0, and thus these sites are not classified as pauses. The next paused cluster arises for site 38, in which case the first free energy minimum is shifted backward by 2 nt.

We note that both the ±3-nt error in assigning experimental pause sites (37) and the clustering of pauses implicit in our algorithm result in a larger fraction of the sequence covered by predicted pauses than observed experimentally. The error of ±3 nt implies that meaningful results cannot involve >30% sequence coverage by predicted pauses, a constraint that is imposed on the optimization of statistical significance (for comparison, experimental pauses cover 12% of the sequence space). In determining the cutoff parameters, we used $K_d = 20\ \mu M$ (see the supporting information) and NTP concentrations appropriate for each experimental condition (10 μM for Chamberlin's data, 40 μM for sequence seq10, and 30 μM for sequences seq11–seq13). Our results correspond to cutoff fractions $\xi = 0.0012$, 0.015, 0.010, and 0.075 for SBNF, SBF, MBNF, and MBF, respectively, with the corresponding percentages of sequence covered, 29%, 28%, 30%, and 23%.

For illustration, Fig. 2*b* compares the experimental pauses (lane 1) with our best (MBF) thermodynamic result (lane 2) for template D111. In addition, the pattern of pause clusters obtained by applying our algorithm to a randomized sequence, with the same AT/GC content as D111, is significantly different from those generated from the actual template. This difference can be quantified by measuring the dot-product-overlap between the two patterns associated with sequence S (see the supporting information), $d[S, S'(S)] = \vec{S} \cdot \vec{S}' / \vec{S}' \cdot \vec{S}' = 0.30$ [here $S'(S)$ is the randomized sequence obtained from S].

Fig. 2*d* shows the total proportion of incorrect to correct predictions, η_1, over the 10 templates. The best results are obtained for the case of MBF, in which case the algorithm identifies 84% of the experimental pauses (i.e., sensitivity, $\sigma = 84\%$), with a PPV, quoted as a percentage, of 68%. The PPV reflects the fact that, as alluded to above, the algorithm predicts more pauses than are seen experimentally. These results are highly statistically significant as can be seen from a comparison with the average of many random assignments of pause sites, which leads to a PPV of 32% and a σ value of 32%. The other three thermodynamic models are also statistically significant when compared with the random assignment of pauses, with (PPV, σ) for SBNF, SBF, and MBNF of (60%, 78%), (64%, 80%), and (60%, 70%), respectively. We note that the cutoffs for the different models span a large range of values. In particular, progressively lower cutoff values were required to reduce the sequence coverage to or below 30% for MBF, SBF, MBNF, and SBNF, respectively. Because (*i*) the lower bound on the duration of experimentally defined pauses is given by the time resolution, of the order of 1 s (38), and (*ii*) at saturating NTP the average elongation time is one order of magnitude lower, it follows that the elongation rate at typical pause sites should be of the order of 10% of the maximum rate. This is in agreement with the cutoff found for MBF, greater (by a factor of 5) than the cutoffs for SBF and MBNF, and substantially greater than the cutoff for SBNF, pointing to MBF as the most physically reasonable of the four models. [It is worth mentioning that, although for SBF and MBF the amount of backtracking is completely determined by the sequence-dependent positions of RNA-folding barriers, allowing backtracking beyond 9 nt (to 15 nt) in the case of SBNF and MBNF does not affect the results provided that the corresponding cutoffs are further decreased below their already unphysically small values.] In addition, the optimization of η_2 (see the supporting information), and other statistical measures we have examined, reinforces MBF as the model with the highest predictive power. It is also worth mentioning that the global dot-product-overlap between MBF pause patterns on actual and randomized templates is $d_{global} = \Sigma_S\, d[S, S'(S)]/10 = 0.26$, lending support to the idea that pause sites are encoded by sequence-dependent signals on DNA.

Finally, MBF reproduces the correct NTP concentration depen-

dence of pausing by yeast *Pol III* on the *SUP4* template (38). It is encouraging that (*i*) we predict the correct pause sites and trend in their number (the number of experimental pauses decreases from 10 at 100 μM to 4 at 1 mM, whereas the corresponding predicted numbers decrease from 7 to 4) and (*ii*) the statistical significance increases with decreasing NTP concentration. This latter dependence is as expected, because a decreasing NTP concentration is accompanied by a decreasing elongation rate, in which case the experimental situation should be better approximated by an equilibrium model (D.ÓM., V.R.T., and A.E.R., unpublished work).

Kinetic Modeling of Transcriptional Pausing

As already stressed above, transcriptional pausing is an intrinsically kinetic phenomenon strongly affected by the complex dynamics of folding–unfolding of the RNA available beyond the exit channel. Indeed, this view is supported by *in vitro* experiments demonstrating the "antiarrest" effect of RNA folding in the process of transcription by mammalian RNAP downstream of the mouse β-globin promoter (39). Also, it is found that the transition to competent elongation of early RNAP II transcription complexes requires the synthesis of RNA transcripts longer that 50 nt (40); for shorter transcripts, a significant fraction of ECs on the same DNA sequence backtrack and sometimes even become arrested. It is appealing to interpret the inhibition of backtracking in the case of longer transcripts as being due to kinetic barriers induced by the slow kinetics of cotranscriptional folding–unfolding of the upstream RNA.

The kinetics of transcription elongation in the presence of cotranscriptional RNA folding is tractable in two special limiting cases: (*i*) in the case in which the relaxation of the RNA to its native fold would occur faster than any of the translocation steps, the motion of RNAP can be described as translocation between the minima of the free energy functional in Eq. **4**; and (*ii*), in the opposite extreme limit, in which translocation rates are orders-of-magnitude faster than folding rates, folding does not affect RNAP motion. In practice there is no clear separation of time scales between translocation and folding kinetics, and the interplay between them must be taken into account for a proper understanding of the kinetics of transcription elongation under biologically relevant conditions.

Kinetic Model. Rather than following the full kinetics of the elongation process, we focus on the simpler problem of the kinetics of the recovery of paused complexes: we imagine that the EC was "walked" to a particular position along DNA corresponding to a fixed value of the transcript length, at which point the system is starved of NTP. Once RNAP has had a chance to equilibrate fully, so that the probability of occupying a particular position relaxes to the form given in Eq. **3**, we add the next complementary NTP and monitor the incorporation rate. The rate of recovery of a complex equilibrated at a particular transcript length gives a quantitative measure of pausing at that particular position.

Our strategy is to propagate the probability for a given transcript length according to the Master equation (41) starting from an initial condition given by the equilibrium probability distribution in Eq. **3**. In principle, each of the components of the initial distribution, corresponding to a different translocation position, is associated with a different equilibrium RNA folding configuration of the available transcript. The main idea is that, as long as we propagate the system for a time short compared with that required for a fold rearrangement, the forward kinetic rates will not be affected by RNA folding, whereas backtracking against a fold involves a free energy penalty resulting in a substantial kinetic barrier. The implication of this oversimplified picture is that unfolding of typical folds is slow compared with a typical translocation step. This assumption can be motivated by noting that translocation is not the rate-limiting step in transcription elongation (42). In *E. coli*, elongation proceeds at $\approx$10–35 nt/s at saturating (millimolar) NTP concentrations (43) while the prefactor in the Arrhenius rate of unfolding of short RNA duplexes is $\approx 30\ s^{-1}$. The latter follows from the estimated half-life of a GC-rich 10-bp RNA duplex of 100 years (44)!

The independent evolutions of the individual components of the initial equilibrium distribution (see the supporting information) are combined to produce the time-dependent probability from which we can extract the elongation rate for a given transcript length. For simplicity, we only follow the behavior of the most probable bubble configuration, (9, 2, 1). We assume translocation rates are of the Arrhenius form with a prefactor of $10^7\ s^{-1}$ and barrier heights obtained by adding $6.5\ k_BT$ to the mean of initial and final state free energies involved in the transition (this is the average barrier height that must be overcome for each translocation). The latter assumption is consistent with the order of magnitude of the maximum energy required to translocate RNAP by one base pair along the template: this involves the breaking of the first DNA–DNA base pair at the edge of the bubble in the direction of motion and breaking of one base pair of the RNA–DNA hybrid on the opposite side of the bubble. Our choices for kinetic rates are consistent with published experimental data on incorporation kinetics (see the supporting information).

Kinetic Results. Here, we only summarize the principal findings of our simplified kinetic model and defer the details to the supporting information and future work (D.ÓM., V.R.T., and A.E.R., unpublished data). The predicted pause sites for template D111 are shown in lane 3 of Fig. 2*b* and correlate well with the experimental pause sites. The kinetic scheme yields a sensitivity of 80% with a PPV of 65%, which match the results (of 80% and 64%, respectively) obtained with SBF. Fig. 2*d* also shows agreement between SBF and kinetics, further validating the conceptual framework of the kinetic model. The other important feature of our results concerns the effect of the kinetic folding barriers on the incorporation rate and thus on the probability of pausing. Fig. 2*c* compares the probabilities of incorporating the next NTP as a function of time in the cases with and without kinetic folding barriers for transcript length 85 on template D167. Note that in the absence of folding barriers, the EC at this position results in a paused complex whereas in the presence of barriers backtracking is strongly inhibited and the pause sites is eliminated. This emphasizes the role of RNA folding in restricting the excursions of RNAP away from the elongation pathway. Also, this behavior is consistent with the reduction in the number of predicted pause sites when RNA folding is included in the equilibrium calculations.

Comparison with Previous Attempts at Modeling Transcriptional Pausing

Recently, Bai *et al.* (22) modeled transcriptional elongation kinetics in the absence of RNA folding, in an attempt to identify pause sites of *E. coli* RNAP on four different templates. Two features of the work in ref. 22 are difficult to justify from a physical point of view: (*i*) the barrier height between the pretranslocated (0) and post-translocated (+1) states is assumed to be very small, implying that the corresponding rate is much faster than all other rates in the problem; and (*ii*) all other barrier heights derived from fitting experimental data are unreasonably high, of the order of 40–50 k_BT, comparable with the base-pairing free energy cost for the formation of a 14-bp bubble!

We implemented the kinetic model of ref. 22 with a Monte Carlo (MC) Gillespie simulation to check the performance of the algorithm on our 10 templates using the same parameters as in ref. 22, at appropriate NTP concentrations (see the supporting information). Pause sites were defined by identifying those transcript lengths for which the pause duration and pause probability fall *above* thresholds, τ and P_{thresh}, respectively. Even when using very conservative thresholds, τ = 15 s (the shortest pause duration

quoted in ref. 22) and $P_{\text{thresh}} = 0.95$ (Bai I), the resulting PPV is 49%, with a sensitivity, σ, of only 29%, to be compared with the results (32% and 32%, respectively) obtained from the average over random assignments of pause positions. Choosing conservative pause duration thresholds of 0.3–0.5 s (only five times the maximum single nucleotide incorporation time) and a reasonable $P_{\text{thresh}} = 0.5$ (Bai II) only yielded a PPV of 50% and a σ of 24%. We found that substantially improving the performance of the algorithm above the random assignment of pauses required choosing the same pause duration cutoffs ($\tau = 0.3$–0.5 s) and an unreasonably high P_{thresh} of 0.9 (Bai III), resulting in a PPV of 52% and a σ of 70%. The corresponding values of η_1 for the three cutoff choices Bai I, Bai II, and Bai III are shown in Fig. 2*d*. We believe that the poor performance of the model presented by Bai *et al.* is due to the low probability of backtracking and small resulting pausing probabilities, features which can be traced back to the unreasonably large values of translocational barriers.

Discussion

We discussed two algorithms for identifying pause sites during transcription elongation by *E. coli* RNAP, both of which led to statistically significant results: the first is equilibrium-based and associates pauses with deep local minima in the free energy profile describing the thermodynamic stability of the EC as a function of position along the template. The second algorithm is based on a simplified kinetic model describing the NTP-driven recovery rate of ECs stalled and allowed to equilibrate in the absence of NTP at each position along the template. The essential ingredient of both equilibrium and kinetic models is the presence of sequence-specific kinetic barriers due to RNA cotranscriptional folding, which strongly inhibit backtracking of RNAP on all templates. This article relies on a strict separation of times scales between the slow unfolding kinetics of RNA secondary structure and the rapid translocational equilibration of RNAP. In practice, this assumption cannot be satisfied at each position along the template; and, moreover, kinetic folding barriers typically involve excited rather than lowest free energy conformations of the RNA transcript, as assumed in this work. Improving on these approximations requires a highly nontrivial, detailed kinetic treatment. It is important to stress that, in the absence of folding barriers or of another mechanism inhibiting backtracking (such as interference with the ribosomal translational machinery in bacteria), RNAP would backtrack considerably. As a result the equilibrium approach (with physically reasonable cutoffs) would predict a large number of false positives.

The other novel aspect of our equilibrium algorithm is that it accounts for thermal fluctuations in the size of the transcription bubble. Even though this affects the precise backtracked position of a few specific pauses, it only has a small, qualitative effect on the statistical significance of our results.

We expect that the sequence-specific effects discussed here are essential in determining pause sites *in vivo* in both prokaryotes and eukaryotes, particularly in the context of pausing on stable RNA genes and untranslatable control elements. Although the current equilibrium approach involves a single adjustable parameter (the cutoff on the elongation rate), the quality of predictions may be enhanced by (*i*) treating RNA–RNA, DNA–DNA, and RNA–DNA base-pairing interactions as adjustable parameters, and (*ii*) including simple parameterizations of the sequence nonspecific interactions between RNAP and nucleic acids (ignored in this work), or of interactions between RNAP and regulatory factors present under *in vivo* conditions. In principle, provided that a large amount of additional data are collected, the free parameters could be determined ("learned") from experiment.

A quantitative, detailed understanding of the effects of cotranscriptional folding on transcription elongation kinetics will clearly require increasing the level of sophistication of both single-molecule experiments and computational modeling methods. We are encouraged that both the experimental and theoretical tools (see, for example, refs. 45–47) are becoming available to address these important problems in the near future.

We thank Peter H. von Hippel for many constructive comments on the manuscript. A.E.R. is grateful to Konstantin Severinov for encouragement during a decisive stage of this project. This work was supported by National Institutes of Health Grants GM41376 (to R.H.E.), GM72814 (to E.N.), GM58750 (to E.N.), and GM58750-07S1 (to A.E.R. and E.N.) and a Howard Hughes Medical Investigatorship (to R.H.E.).

1. von Hippel, P. H. (1998) *Science* **281,** 660–665.
2. Landick, R. (2001) *Cell* **105,** 567–570.
3. Greive, S. J. & von Hippel, P. H. (2005) *Nat. Rev. Mol. Cell Biol.* **6,** 221–232.
4. Ebright, R. H. (2000) *J. Mol. Biol.* **304,** 687–698.
5. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. & Kornberg, R. D. (2001) *Science* **292,** 1876–1882.
6. Cramer, P., Bushnell, D. A. & Kornberg, R. D. (2001) *Science* **292,** 1863–1876.
7. Zhang, G., Campbell, E. A., Minakhin, L., Richter, C., Severinov, K. & Darst, S. A. (1999) *Cell* **98,** 811–824.
8. Vassylyev, D. G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M. N., Borukhov, S. & Yokoyama, S. (2002) *Nature* **417,** 712–719.
9. Darst, S. A., Polyakov, A., Richter, C. & Zhang, G. (1998) *J. Struct. Biol.* **124,** 115–122.
10. Korzheva, N., Mustaev, A., Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A. & Darst, S. A. (2000) *Science* **289,** 619–625.
11. Naryshkin, N., Revyakin, A., Kim, Y., Mekler, V. & Ebright, R. H. (2000) *Cell* **101,** 601–611.
12. Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapanidis, A. N., Niu, W., Ebright, Y. W., Levy, R. & Ebright, R. H. (2002) *Cell* **108,** 599–614.
13. Landick, R. (1997) *Cell* **88,** 741–744.
14. Landick, R., Carey, J. & Yanofsky, C. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4663–4667.
15. Wilson, K. S. & von Hippel, P. H. (1994) *J. Mol. Biol.* **244,** 36–51.
16. Nudler, E. & Gottesman, M. E. (2002) *Genes Cells* **7,** 755–768.
17. Nickels, B. E. & Hochschild, A. (2004) *Cell* **118,** 281–284.
18. Yager, T. D. & von Hippel, P. H. (1991) *Biochemistry* **30,** 1097–1118.
19. Wilson, K. S., Conant, C. R. & von Hippel, P. H. (1999) *J. Mol. Biol.* **289,** 1179–1194.
20. Erie, D. A., Yager, T. D. & von Hippel, P. H. (1992) *Annu. Rev. Biophys. Biomol. Struct.* **21,** 379–415.
21. Rees, W. A., Weitzel, S. E., Das, A. & von Hippel, P. H. (1997) *J. Mol. Biol.* **273,** 797–813.
22. Bai, L., Shundrovsky, A. & Wang, M. D. (2004) *J. Mol. Biol.* **344,** 335–349.
23. Nudler, E., Gusarov, I., Avetissova, E., Kozlov, M. & Goldfarb, A. (1998) *Science* **281,** 424–428.
24. Wang, H. & Oster, G. (2002) *Appl. Phys. A* **75,** 315–323.
25. Yin, Y. W. & Steitz, T. A. (2004) *Cell* **116,** 393–404.
26. Guajardo, R. & Sousa, R. (1997) *J. Mol. Biol.* **265,** 8–19.
27. Julicher, F. & Bruinsma, R. (1998) *Biophys. J.* **74,** 1169–1185.
28. Bar-Nahum, G., Epshtein, V., Ruckenstein, A. E., Rafikov, R., Mustaev, A. & Nudler, E. (2005) *Cell* **120,** 183–193.
29. Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. & Block, S. M. (2005) *Nature* **438,** 460–465.
30. Yarnell, W. S. & Roberts, J. W. (1999) *Science* **284,** 611–615.
31. Marr, M. T. & Roberts, J. W. (2000) *Mol. Cell* **6,** 1275–1285.
32. Wu, P., Nakano, S. & Sugimoto, N. (2002) *Eur. J. Biochem.* **269,** 2821–2830.
33. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994) *Monatsh. Chem.* **125,** 167–188.
34. Artsimovitch, I. & Landick, R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7090–7095.
35. Toulokhonov, I. & Landick, R. (2003) *Mol. Cell* **12,** 1125–1136.
36. Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. (1997) *Cell* **89,** 33–41.
37. Levin, J. R. & Chamberlin, M. J. (1987) *J. Mol. Biol.* **196,** 61–84.
38. Matsuzaki, H., Kassavetis, G. A. & Geiduschek, E. P. (1994) *J. Mol. Biol.* **235,** 1173–1192.
39. Reeder, T. C. & Hawley, D. K. (1996) *Cell* **87,** 767–777.
40. Ujvari, A., Pal, M. & Luse, D. S. (2002) *J. Biol. Chem.* **277,** 32527–32537.
41. Gardiner, C. W. (1996) *Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences* (Springer, New York).
42. Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J. & Block, S. M. (1998) *Science* **282,** 902–907.
43. Uptain, S. M., Kane, C. M. & Chamberlin, M. J. (1997) *Annu. Rev. Biochem.* **66,** 117–172.
44. Herschlag, D. (1995) *J. Biol. Chem.* **270,** 20871–20874.
45. Bokinsky, G., Rueda, D., Misra, V. K., Rhodes, M. M., Gordus, A., Babcock, H. P., Walter, N. G. & Zhuang, X. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9302–9307.
46. Lang, M. J., Fordyce, P. M., Engh, A. M., Neuman, K. C. & Block, S. M. (2004) *Nat. Methods* **1,** 133–139.
47. Xayaphoummine, A., Bucher, T., Thalmann, F. & Isambert, H. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 15310–15315.