Supporting Text

Sequences

The transcribed sequences of the various templates used in the paper are given below.

seq10:

AUCGAGAGGGACACGGCGAAUAGCCAUCCCAAUCGACACCGGGGUCCGGG
AUCUGGAUCUGGAUCGCUAAUAACAUUUUUAUUUGGAUCCCCGGGUACCG
AGCUCGAAUUCACUGGCCGUCGUUUUACAACGUCGUGACUGGGAAAACCC
UGGCG

seq11:

AUCGAGAGGGACACGGCGAAUAGCCAUCCCAAUCCGACACCGGGGCAUCGA
GUGGGACACGGCGAAUAGCCAUCCCAAUCGACACCGGGGUCCGGGAUCUG
GAUCUGGAUCGCUAAUAACAGGCCUGCUGGUAAUCGCAGGCCUUUUUAUU
UGGAUCCCCGGGUA

seq12:

AUCGAGAGGGCCACGGCGAACAGCCAACCCAAUCGAACAGGCCUGCUGGUA
AUCGCAGGCCUUUUUAUUUGGAUCCCCGGGUA

seq13:

AUCGAGAGGGCCACGGCGAACAGCCAACCCAAUCCGAACAGCCAUCAUCCU
CAGUAUUCAGGUAGCUGUUGAGCCUGGGGCGGUAGCGUGCUUUUUUCGAA
UUCACUUAAUGGUAAUCUCG

D123:

AUCGAGAGGGACACGGCGAAUAGUGAGAACUUGGCGAGAGAACAACCUCG
AACGCCGCAAGGCACAAGAGAGGGCGGCGUGGCAUAGACGAAAGGAAAAG

GUUAAAGCCAAGAAACUCGCCGCACUUGAACAGGCACUAGCCAACACACUG
AACGCUAUC

D167:
AUCGAGAGGGACACGGCGAAUAGCCAUCCCUAACGUCUACGAUGUACAGC
GCCACGCUGGAUGCUAUACGGUGGUACUUGACGCACUUAAGGAUUGCGAG
CGUUUCAACAAUGAUGCCCAUUAUAAAUACGCUGAGAUUGCAAGCGACAU
CAUUGAUUGC

D111:
AUCGAGAGGGACACGGCGAAUAGCCAUCCCAAUCCACACGUCCAACGGGGC
AACCGUAUGUACACCUGAUGGGUUCGCAAUGAAACAACGAAUCGAACGCC
UUAAGCGUGAACUCCGCAUUAACCGCAAGAUUAACAAGAUAGGUUCCGGC
UAUGACAGA

D112:
AUCGAGAGGGACACGGCGAAUAGCCAUCCCAAUCGACACCGGGGUCAACCG
GAUAAGUAGACAGCCUGAUAAGUCGCACUAGAACAGGCACUAGCCAACAC
ACUGAACGAUAUCUCAUAACGAAGAUAAAGGACACAAUGCAAUGAACAUU
ACCGACAUC

D104:
AUCGAGAGGGACACGGCGAAUAGCCAUCCCAAUCGACACCGGGGUCAACCG
GAUAAGUAGACAGCCUGAUAAGUCGCACGACAGAAAGAAAUUGACCGCGC
UAAGGCCCGUAAAGAACGUCACGAGGGGCGCUUAGAGGCACGCAGAUUCA
AACGUCGCA

D387:
AUCGAGAGGGACACGGCGAAUAGCCAUCCCAAUCGACACCGGGGUCAACCG
GAUAAGUAGACAGCCUGAUAAGUCGCACGAAAAACAGGUAUUGACAAGCG

UCAAGGUAUGCUUAUCGACUUACUGGUCGAGAUGGUCAACAGCGAGACGUGUGAUGGCG

Equilibrium and Kinetics Results on All Templates

Tables 1–10 summarize our results on all 10 templates using the four thermodynamic models (*i*) single bubble (2,9,1), without RNA folding (SBNF); (*ii*) single bubble with RNA folding (SBF); (*iii*) multiple bubbles without RNA folding (MBNF); (*iv*) multiple bubbles with folding (MBF); and (*v*) kinetics.

Dot-Product Overlap

The difference in pause cluster patterns between the actual and randomized templates can be quantified by measuring the dot-product overlap between the two patterns associated with sequence *S*, $d[S,S'(S)] = \vec{S}\square\vec{S}'/\vec{S}'\square\vec{S}'$, here *S'*(S) is the randomized sequence obtained from *S*. Table 11 lists the dot-product overlap between pause clusters calculated with MBF using our pausing criterion on the actual and randomized templates for all ten sequences. A lower overlap indicates a larger dissimilarity between the pause cluster patterns.

Discussion of the Algorithm of Bai *et al.* (1)

We expect that the assumption of much faster equilibration rates between states 0 and +1, is motivated by the discussion of translocation rates in the context of T7 transcription (2). In that case the apparent absence of a sufficiently well defined secondary channel implies that backtracking is strongly suppressed by steric clashes between the 3'end of RNA and RNAP (W. McAllister, personal communication), a situation not met in the case of *E. coli* transcription (3, 4). The second issue, related to the extremely high values of the backtracking and hypertranslocation barriers required to fit experimental gels, maybe related to the fact that the simplified parameterization of kinetics in ref. 1 attempts to mimic the presence of RNA folding barriers not explicitly included in the model.

We implement a Monte Carlo (MC) version of the algorithm in ref. 1, with the fit parameters given in ref. 1 at 24°C, on the same templates considered in the main text (3, 5) with no contribution from RNA folding as in ref. 1.

We use 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). (Application of our approach at 24°C, instead of 37°C, shows little change in the statistics of our predicted pause positions, in agreement with the lack of temperature dependence of backtracked pause patterns found in ref. 5.) We classify as pauses those transcript lengths for which the pause duration is above τ and the pause probability is above P_{thresh}. The pause probability is defined as the fraction of MC trajectories which, on reaching a given transcript length, pause for at least a time τ.

Since the shortest pause identified in ref. 1 has a duration of at least 15 s we first chose the cutoff $\tau = 15$ s. Even when we use a very liberal value of 0.95 for the threshold probability the resulting positive predictive value (PPV) is 49% with a sensitivity (σ) of 21% (Bai I in Fig. 3), implying that, even for this rather liberal way of defining pauses, the algorithm in ref. 1 misses most of the experimentally observed sites.

We further relax the definition of a pause by choosing τ, the cutoff on the pause duration, to be only five times the shortest possible pause duration, namely the inverse of the maximum elongation rate at each NTP concentration. From the Michaelis–Menten form for the latter,

$$(k_{\text{main}}^{\max})^{-1} = \frac{K_d + [NTP]}{k_{\max}[NTP]},$$

we estimate $\tau = 0.3$ s for seq10–13 and $\tau = 0.5$ s for the templates used in ref. 3. Defining pauses by requiring that complexes pause with probability $P_{\text{thresh}} = 0.5$ for at least τ seconds results in a PPV of 50% and a σ of 24% (Bai II in Fig. 3). To increase the predictive power of the model such that it outperforms the average over many random assignments of pauses, we must increase the threshold probability to around 0.9 in which case the PPV is 52% and the σ is 70% (Bai III in Fig. 3). The fact that such a relaxed (and difficult to justify) definition of pausing is required is related to the extremely large barriers for backtracking implied by the algorithm presented in ref. 1.

In our MC implementation of the algorithm of ref. 1 we have treated the states (m, 0), (m, 1) and (m, 1)* (see Fig. 5) as a single composite state in order to have explicit instantaneous equilibration between these three states at each iteration. The probability of translocating forward (hypertranslocation) out of the composite state, (m; (0, 1, 1*)), is almost always larger than the probability of backtracking, this can be seen by calculating the effective rates of leaving the composite state. As found in ref. 1 the effective elongation rate out of the composite state is

$$k_{\text{main}} = \frac{k_{\max}[NTP]}{K_d(1+K_i)+[NTP]}, \text{ where } K_i = e^{(G_{m,1}-G_{m,0})/k_B T}.$$

We can find the effective rates for hypertranslocating to +2 and backtracking to –1 in a similar fashion:

$$k_{\text{forward}} = \frac{k_{m,1\to 2}K_d}{K_d(1+K_i)+[NTP]} \text{ and } k_{\text{back}} = \frac{k_{m,0\to -1}K_d K_i}{K_d(1+K_i)+[NTP]}.$$

Starting in the composite state, the probability of going forward is simply $P_{\text{forward}} = k_{\text{forward}}/(k_{\text{forward}} + k_{\text{back}} + k_{\text{main}})$ and the probability of going back is $P_{\text{back}} = k_{\text{back}}/(k_{\text{forward}} + k_{\text{back}} + k_{\text{main}})$. Since $G_{m,1}$ is usually larger than $G_{m,0}$, when RNA folding is absent and, as the absolute scale of the barrier between translocation states +1 to +2 is 40 k_BT, which is smaller than the absolute scale of the barrier from 0 to –1 of 46.2 k_BT, we find that $k_{1\to 2}$ will usually be much larger than $k_{0\to -1}$. This makes k_{forward} much larger than k_{back} in most cases, which leads to P_{forward} being much larger than P_{back}. For example, for transcript length 32 on seq11 $P_{\text{forward}} = 6.5 \times 10^{-1}$ and $P_{\text{back}} = 7 \times 10^{-4}$. These frequent hypertranslocation attempts are artifacts of the model and would not occur if the forward barriers were made higher than the backtracking barriers.

Kinetic Model Discussion

Here we give a brief description of the details of the kinetic algorithm. A more complete description will be given in a future publication (6). As discussed in the main text, we evolve the components of the equilibrium distribution where we include RNA folding in the free energy (Eq. **2**). We equilibrate around the local minimum in the free energy landscape. The local minimum is defined as a minimum closest to translocation state 0, where moving either forward or backward increases the free energy by more than 2 k_BT. This is only one way to determine the range over which to equilibrate; in practice this is dependent on both the free energy landscape including RNA folding and the RNA-folding barriers. The components of this initial distribution are given by Eq. **1**, where b = (2,9,1) is the only bubble configuration used. An example of a free energy landscape including RNA folding (green curve) and the corresponding initial distribution (red histogram) are shown in Fig. 4.

Each component is evolved separately in a landscape without additional RNA folding from that provided by the initial state (blue curve with an additive constant corresponding to the initial RNA fold, Fig. 4). Adjustments to the backtracking rates are made at translocation positions where additional backtracking of the enzyme would require the breaking of RNA–RNA base pairs of the secondary structure. This is accomplished in our simulations by adding reflecting boundaries at positions one base pair upstream of the where the enzyme first encounters a fold (black vertical lines in Fig. 4). These barriers are only encountered in the process of backtracking and are invisible to components of the original distribution associated with positions of RNAP upstream of the barriers. The enzyme is allowed to break one base pair of the fold before being pushed back by the reflecting barrier. For example, in Fig. 4 the equilibrium probability components at states 0 and +1 see a reflecting barrier at position –1 while the component at –1, on the other hand, sees a reflecting barrier at –3 but not at –1. Note that the barriers shown in Fig. 4 prevents template D167 from being paused at position 85.

An estimate of the maximum energy barrier between two translocation states can be found by noting that at least one hybrid bond and one DNA–DNA bond in the transcription bubble must be broken for each translocation step. We take the extreme view that both of these bonds must be broken before any other bonds are reannealed and thus, ignoring all other free energy contributions to this barrier this estimate corresponds to an upper bound to the sequence dependent translocation barrier. The energy losses corresponding for removing a rGC/dCG from the hybrid and a dCG/dGC from the bubble are 2.4 kcal/mol and 2.8 kcal/mol respectively (7). This gives an estimate of the maximum translocation energy barrier of 8.3 k_BT at 37°C. This provides the estimate for the energy barriers used in the main text.

The kinetic scheme we used is shown in Fig. 5. The pyrophosphate (PPi) release rate is unknown but is thought to be rate-limiting at saturating NTP concentrations (8–11). Single nucleotide incorporation studies of (10, 11) place the rate-limiting step to be at least 700 s^{-1}, which we use as an estimate of the NTP independent PPi release rate. We used an apparent dissociation constant of NTP binding of 20 μM (12), at the lower end of the range of published dissociation constants, to account for the fact that competitive inhibition due to some branched reaction pathways is explicitly accounted for in our model. Two other effects on the effective dissociation constant are worth noting. First, competitive inhibition effects associated with non-complementary NTPs are negligible at the NTP concentrations we consider, as the inhibition constants are on the order of mM (12); and second, the fact that the NTP concentration in the secondary channel may be smaller than in the cellular environment (13) simply results in a rescaling of the dissociation constant and need not be explicitly considered as long as NTP diffusion is not rate-limiting.

The translocation rate prefactor was chosen to be 10^7 s^{-1} along with a dissociation rate of 5×10^4 s^{-1}. These experimentally unknown parameters where chosen so that both translocation and NTP dissociation would be faster than the pyrophosphate release rate. We varied the translocation rates (by varying the constant which determines the barrier heights), the NTP dissociation constant and NTP dissociation rate over several orders of magnitude without any significant change in the statistics of our results (see below).

Kinetic Model Results

As already discussed above, pauses are defined as those sites where complexes do not incorporate the next NTP with a probability greater than or equal to P_{thresh}, and up to a threshold time scale, τ. Pauses were clustered based on adjacency. The thresholds P_{thresh} and τ are determined by maximizing the statistical significance of our results: P_{thresh} is varied between, 0.1 and 1 in steps of 0.1 while τ is varied between 0.1 s and 1 s in steps of 0.1 s, and between 1 s and 10 s in steps of 1 s. We maximize the PPV and σ while minimizing the percentage of sequence space covered constrained by tolerances on σ of 70%, and percentage of sequence space covered, 35%. We allow different thresholds for seq10, seq11–13 and the data of ref. 5 since the corresponding experiments involve different NTP. More precisely, (P_{thresh}, τ) are (0.5, 0.5 s) for seq10, (0.5, 0.6 s) for seq11–13 and (0.4, 3 s) for the data of ref. 5. Using different kinetic parameters leads to different incorporation timescales but we were able to find a pair (P_{incorp}, τ) yielding approximately the same statistics for each set of kinetic parameters. For comparison, the kinetic, thermodynamic and experimental results are shown side by side in Tables 1–10. We also show in Fig. 6, the optimized values of $\eta_2 = \text{PPV} + \sigma$ for the four equilibrium models and the kinetic model.

1. Bai, L., Shundrovsky, A. & Wang, M. D. (2004) *J. Mol. Biol.* **344,** 335–349.
2. Guajardo, R. & Sousa, R. (1997) *J. Mol. Biol.* **265,** 8–19.
3. Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. (1997) *Cell* **89,** 33–41.
4. Greive, S. J. & von Hippel, P. H. (2005) *Nat. Rev. Mol. Cell Biol.* **6,** 221–232.
5. Levin, J. R. & Chamberlin, M. J. (1987) *J. Mol. Biol.* **196,** 61–84.
6. Ó Maoiléidigh, D., Tadigotla, V. R. & Ruckenstein, A. E., in preparation.
7. Wu, P., Nakano, S. & Sugimoto, N. (2002) *Eur. J. Biochem.* **269,** 2821–2830.
8. Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J. & Block, S. M. (1998) *Science* **282,** 902–907.

9. Erie, D. A., Yager, T. D. & von Hippel, P. H. (1992) *Annu. Rev. Biophys. Biomol. Struct.* **21,** 379–415.

10. Holmes, S. F. & Erie, D. A. (2003) *J. Biol. Chem.* **278,** 35597–35608.

11. Foster, J. E., Holmes, S. F. & Erie, D. A. (2001) *Cell* **106,** 243–252.

12. Rhodes, G. & Chamberlin, M. J. (1974) *J. Biol. Chem.* **249,** 6675–6683.

13. Batada, N. N., Westover, K. D., Bushnell, D. A., Levitt, M. & Kornberg, R. D. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17361–17364.