Supporting Information For

Specific Nucleotide Binding and Rebinding to Individual DNA Polymerase Complexes Captured in a Nanopore

Nicholas Hurt¹, Hongyun Wang², Mark Akeson³ & Kate R. Lieberman³*

¹ Department of Chemistry and Biochemistry, ² Department of Applied Mathematics and Statistics, ³ Department of Biomolecular Engineering, Baskin School of Engineering, University of California, Santa Cruz, CA 95064, USA

1. Comparison of abasic and standard DNA residues

![Comparison of abasic and standard DNA residues](image)

**Figure S1.** Comparison of abasic and standard DNA residues. (a) Section of a standard DNA strand, with nucleobases at the 1' position represented as unsubstituted purines. (b) Section of a strand bearing abasic (1', 2'-dideoxy) residues, which are present in the 5ab(12,16) substrate.

*corresponding author: lieberman@biology.ucsc.edu
2. Mathematical model and derivation of mean dwell time

The two-state model is illustrated in the kinetic diagram in Figure S2. We follow the state of the nanopore. In each capture and pulling experiment, we start with an empty pore. If a DNA without KF bound is captured, it will be pulled through the pore quickly. Such events are detected in the data processing and are not used in our modeling analysis since the goal of our mathematical modeling is to study the dGTP binding onto and dissociation from the captured DNA-KF complex. We focus on the case where a DNA-KF complex (with or without a dGTP bound) is captured. In the kinetic diagram, state 1 represents that a binary complex (DNA-KF without a dGTP bound) is on the pore, and state 2 represents that a ternary complex (DNA-KF with a dGTP bound) is on the pore. At the moment when the pore captures a DNA-KF complex from the solution, it has probability $p_1(0)$ of being binary and probability $p_2(0)$ of being ternary. $p_1(0)$ and $p_2(0)$ are determined by the equilibrium between binary and ternary complexes in solution

$$p_1(0) = \frac{K_d^{(B)}}{[dGTP] + K_d^{(B)}}$$

$$p_2(0) = \frac{[dGTP]}{[dGTP] + K_d^{(B)}}$$

where $[dGTP]$ is the concentration of free dGTP in solution and $K_d^{(B)}$ is the affinity of dGTP binding onto DNA-KF complexes in solution. Here, “0” in $p_1(0)$ and $p_2(0)$ refers to time $t = 0$, which we define as the moment when the DNA-KF complex is captured by the pore.

![Figure S2. Kinetic diagram describing the two-state model.](image_url)
For $t > 0$, the DNA-KF complex on the pore fluctuates between state 1 and state 2. In Figure S2, $k_{on}[dGTP]$ is the rate of dGTP (from solution) binding onto and $k_{off}$ is the rate of dGTP dissociating from the DNA-KF complex on the pore. The affinity of dGTP binding onto the DNA-KF complex on the pore is $K_d = \frac{k_{off}}{k_{on}}$. For mathematical convenience, we use $K_d$ and $k_{off}$ (instead of $k_{on}$ and $k_{off}$) as model parameters for the binding and dissociation of dGTP. The key assumption of our model is that KF can dissociate (rate $k_1$ as shown in Figure S1) only in the binary state (state 1). This is a mathematical abstraction for the situation where the rate of dissociation in the ternary state is very small. Thus, the two-state model, as described in Figure S1, has 4 parameters:

$$ \{ k_1, K_d^{(B)}, K_d, k_{off} \} $$

Note that in our model we do not assume $K_d = K_d^{(B)}$. As a result, our model can accommodate the situation $K_d \neq K_d^{(B)}$ (i.e., the affinity of dGTP binding onto the DNA-KF complex on the pore is affected by the voltage pulling).

Let $p_j(t)$ be the probability of state $j$ at time $t$. The governing equation for $p_j(t)$ is

$$ \frac{dp_1}{dt} = -k_1p_1 - \frac{k_{off}}{K_d}[dGTP]p_1 + k_{off}p_2 $$

$$ \frac{dp_2}{dt} = -k_{off}p_2 + \frac{k_{off}}{K_d}[dGTP]p_1 $$

and the initial condition is

$$ p_1(0) = \frac{K_d^{(B)}}{[dGTP] + K_d^{(B)}} $$

$$ p_2(0) = \frac{[dGTP]}{[dGTP] + K_d^{(B)}} $$

In experiments, the dwell time of KF on the pore (the time elapsed from the capture of DNA-KF complex to the dissociation of KF from DNA) can be deduced from the measured time series of electric current through the pore. The dwell time is a random variable. Let $T$ denote the random
dwell time. The probability density of $T$ can be calculated by first solving the governing equation with the initial condition for $p_1(t)$ and $p_2(t)$.

$$\rho_r(t) = -\frac{d}{dt}(p_1(t) + p_2(t))$$

There is no analytic expression for $p_1(t)$ and $p_2(t)$, which hinders our attempt of studying the dependence of dwell time probability density on the model parameters. To better study the dependence of observed dwell time samples on the model parameters, we focus on the mean dwell time, which can be estimated reliably by averaging over many independent samples. The goal of the mathematical derivation below is to express the mean dwell analytically in terms of the 4 model parameters and the dGTP concentration. We start by introducing notations.

Let $\langle T | j \rangle$ denote the conditional mean dwell time given that the system starts in state $j$ at time $t = 0$, and $\langle T \rangle$ denote the overall mean dwell time.

We first derive analytical expressions for $\langle T | 1 \rangle$ and $\langle T | 2 \rangle$. Consider the case where the DNA-KF complex is in state 1 at time $t = 0$ (when it is captured by the pore). For small $\Delta t$, there are several possibilities in time interval $(0, \Delta t)$:

\[
\begin{align*}
\text{KF dissociates,} & \quad \text{prob} = k_i \Delta t + o(\Delta t) \\
\text{dGTP binds (jumps to state 2),} & \quad \text{prob} = k_{on} [dGTP] \Delta t + o(\Delta t) \\
\text{stays in state 1,} & \quad \text{prob} = 1 - (k_i + k_{on} [dGTP]) \Delta t + o(\Delta t) \\
\text{makes two or more transitions,} & \quad \text{prob} = o(\Delta t)
\end{align*}
\]

where $o(\Delta t)$ represents terms of order higher than $\Delta t$. It follows that

\[
\langle T | 1 \rangle = \left[ 1 - \left( k_i + k_{on} [dGTP] \right) \Delta t \right] \cdot \left[ \langle T | 1 \rangle + \Delta t \right] + k_{on} [dGTP] \Delta t \cdot \left[ \langle T | 2 \rangle + \Delta t \right] + o(\Delta t)
\]

\[
= \langle T | 1 \rangle + \Delta t - \left( k_i + k_{on} [dGTP] \right) \Delta t \langle T | 1 \rangle + k_{on} [dGTP] \Delta t \langle T | 2 \rangle + o(\Delta t)
\]

Dividing both sides by $\Delta t$ and taking the limit as $\Delta t$ goes to zero, we get

\[
0 = 1 - \left( k_i + k_{on} [dGTP] \right) \langle T | 1 \rangle + k_{on} [dGTP] \langle T | 2 \rangle
\]

which is a linear equation for $\langle T | 1 \rangle$ and $\langle T | 2 \rangle$. To derive the second equation for $\langle T | 1 \rangle$ and $\langle T | 2 \rangle$, we consider the case where the DNA-KF complex is in state 2 at time $t = 0$. For small $\Delta t$, there are several possibilities in time interval $(0, \Delta t)$:
dGTP dissociates (jumps to state 1), \( \text{prob} = k_{\text{off}} \Delta t + o(\Delta t) \)

stays in state 2, \( \text{prob} = 1 - k_{\text{off}} \Delta t + o(\Delta t) \)

makes two or more transitions, \( \text{prob} = O((\Delta t)^2) \)

It follows that

\[
\langle T \mid 2 \rangle = \left[ 1 - k_{\text{off}} \Delta t \right] \left[ \langle T \mid 2 \rangle + \Delta t \right] + k_{\text{off}} \Delta t \cdot \left[ \langle T \mid 1 \rangle + \Delta t \right] + o(\Delta t)
\]

\[
= \langle T \mid 2 \rangle + \Delta t - k_{\text{off}} \Delta t \langle T \mid 2 \rangle + k_{\text{off}} \Delta t \langle T \mid 1 \rangle + o(\Delta t)
\]

Dividing both sides by \( \Delta t \) and taking the limit as \( \Delta t \) goes to zero, we get

\[
0 = 1 - k_{\text{off}} \langle T \mid 2 \rangle + k_{\text{off}} \langle T \mid 1 \rangle
\]

which becomes

\[
\langle T \mid 2 \rangle = \langle T \mid 1 \rangle + \frac{1}{k_{\text{off}}}
\]

(2)

Substituting (2) into (1), we arrive at a linear equation for \( \langle T \mid 1 \rangle \)

\[
0 = 1 - \left( k_{i} + k_{\text{on}} [dGTP] \right) \langle T \mid 1 \rangle + k_{\text{on}} [dGTP] \left( \langle T \mid 1 \rangle + \frac{1}{k_{\text{off}}} \right)
\]

Solving this linear equation for \( \langle T \mid 1 \rangle \), we obtain

\[
\langle T \mid 1 \rangle = \frac{1}{k_{i}} + \frac{k_{\text{on}} [dGTP]}{k_{\text{off}}} \cdot \frac{1}{k_{i}}
\]

(3)

Using (2) and (3), we write the overall mean dwell time as

\[
\langle T \rangle = p_{1}(0) \langle T \mid 1 \rangle + p_{2}(0) \langle T \mid 2 \rangle
\]

\[
= \frac{K_{d}(B)}{K_{d}(B) + [dGTP]} \cdot \langle T \mid 1 \rangle + \frac{[dGTP]}{K_{d}(B) + [dGTP]} \cdot \left[ \langle T \mid 1 \rangle + \frac{1}{k_{\text{off}}} \right]
\]

which leads to

\[
\langle T \rangle = \frac{1}{k_{i}} + \frac{[dGTP]}{K_{d}} \cdot \frac{1}{k_{i}} + \frac{[dGTP]/K_{d}(B)}{[dGTP]/K_{d}(B) + 1} \cdot \frac{1}{k_{\text{eff}}}
\]

(4)
2. Method for determining the values of the 4 parameters

There are four parameters in the model: \( \{ k_1, K_d^{(B)}, K_d, k_{\text{off}} \} \). Our approach for determining these parameters is first to determine 4 quantities (equations) involving these parameters and then to solve for the 4 parameters from the 4 equations. Specifically, we construct 2 equations from the data for low dGTP concentrations, and construct 2 more equations from the data for high dGTP concentrations.

In the range of low dGTP concentrations, the mean dwell time given in (4) is approximately a linear function of \([\text{dGTP}]\). Mathematically when \( [dGTP]/K_d^{(B)} \ll 1 \), we expand (4) and neglect terms of the order \( O \left( (|dGTP|/K_d^{(B)})^2 \right) \) or higher order. We obtain

\[
\langle T \rangle = \frac{1}{k_1} + \frac{[dGTP]}{K_d} \frac{1}{k_1} + \frac{[dGTP]}{K_d^{(B)}} \left( 1 + O \left( \frac{[dGTP]}{K_d^{(B)}} \right) \right) \frac{1}{k_{\text{off}}}
\]

\[
\approx [dGTP] \left( \frac{1}{K_d k_1} + \frac{1}{K_d^{(B)} k_{\text{off}}} \right) + \frac{1}{k_1}
\]

We fit the observed mean dwell times in low dGTP concentrations to a linear function:

\[
f \left( [dGTP] \right) = s [dGTP] + y
\]

We use the least square formulation in linear scale to determine the optimal values for \( s \) and \( y \).

\[
(s_1, y_1) = \arg \min_{(s, y)} \sum_j \left[ \langle T \rangle_j - s [dGTP]_j - y \right]^2
\]

where \( \langle T \rangle_j \) is the observed mean dwell time at concentration \([dGTP]_j\). Once \( s_1 \) and \( y_1 \) are determined, we have two equations involving the 4 model parameters.

\[
\frac{1}{K_d k_1} + \frac{1}{K_d^{(B)} k_{\text{off}}} = s_1 \quad \frac{1}{k_1} = y_1
\]

(5)

In the range of high dGTP concentrations, the mean dwell time given in (4) is also approximately a linear function of \([dGTP]\) (but different from the one for low dGTP concentrations).
Mathematically when \( \frac{[dGTP]}{K_d^{(B)}} \gg 1 \), \( K_d^{(B)}/[dGTP] \) is small. We expand (4) and neglect terms of the order \( O \left( \frac{K_d^{(B)}}{[dGTP]} \right) \) or higher order. We obtain

\[
\langle T \rangle = \frac{1}{k_i} + \frac{[dGTP]}{K_d} \cdot \frac{1}{k_i} + \left( 1 + O \left( \frac{K_d^{(B)}}{[dGTP]} \right) \right) \cdot \frac{1}{k_{off}}
\]

\[
\approx \frac{[dGTP]}{K_d k_i} + \left( \frac{1}{k_i} + \frac{1}{k_{off}} \right)
\]

We fit the observed mean dwell times in high dGTP concentrations to a linear function:

\[
f \left( [dGTP] \right) = s [dGTP] + y
\]

For high dGTP concentrations, we use the least square formulation in logarithmic scale to determine the optimal values for \( s \) and \( y \):

\[
(s_2, y_2) = \arg \min_{(s, y)} \sum_j \left[ \log \left( \frac{\langle T \rangle_j}{y} \right) - \log \left( \frac{[dGTP]_j}{s} \right) \right]^2
\]

Here the selection of a log-scale distance function is for the purpose of minimizing the effect of statistical errors. When the data values are roughly of the same order of magnitude, a linear scale distance function is a good choice regardless of whether the fitting function is linear. When the data values span many orders of magnitude and are roughly uniformly distributed in a log-scale, a linear distance function will give a very large weight to the data point with the largest value, which is reasonable if the absolute statistical errors are roughly the same for all data points so the data point with the largest value has the smallest relative statistical error. In the case of measured sample mean dwell times, however, the absolute statistical error is proportional to the sample mean. Therefore, we select a log-scale distance function to average over the relative statistical error of all data points to minimize the effect of statistical errors.

Once \( s_2 \) and \( y_2 \) are determined, we obtain two more equations involving the 4 model parameters.

\[
\frac{1}{K_d k_i} = s_2, \quad \frac{1}{k_i} + \frac{1}{k_{off}} = y_2
\]

Solving for the 4 model parameters from equations (5) and (6), we obtain
\[ k_1 = \frac{1}{y_1}, \quad k_{\text{off}} = \frac{1}{y_2 - y_1}, \quad K_d = \frac{y_1}{s_2}, \quad K_{d^{(B)}} = \frac{y_2 - y_1}{s_1 - s_2} \]

3. Results of determining the 4 model parameters

In the range of low dGTP concentrations, we use the lowest 4 dGTP concentrations ([dGTP] = 0, 0.2, 0.4, 0.8 µM) to determine \( s_1 \) and \( y_1 \). We obtain

\[ s_1 = 10.4, \quad y_1 = 3.45 \]

The linear fitting for low dGTP concentration is shown in figure 4A in the text.

In the range of high dGTP concentration, we use the highest 6 dGTP concentrations ([dGTP] = 100, 180, 300, 1000, 3000, 10000 µM) to determine \( s_2 \) and \( y_2 \). We obtain

\[ s_2 = 0.459, \quad y_2 = 45.2 \]

The linear fitting for high dGTP concentration is shown in figure 4B in the text.

Using the values of \( s_1, y_1, s_2 \) and \( y_2 \), and using (7), we arrive at

\[ k_1 = \frac{1}{y_1} = 0.290 (\text{ms}^{-1}) = 290 \text{ s}^{-1} \]

\[ k_{\text{off}} = \frac{1}{y_2 - y_1} = 2.39 \times 10^{-2} (\text{ms}^{-1}) = 23.9 \text{ s}^{-1} \]

\[ K_d = \frac{y_1}{s_2} = 7.52 \mu\text{M} \]

\[ K_{d^{(B)}} = \frac{y_2 - y_1}{s_1 - s_2} = 4.20 \mu\text{M} \]

\[ k_{\text{on}} = \frac{k_{\text{off}}}{K_d} = 3.18 \times 10^{-3} \mu\text{M}^{-1} (\text{ms})^{-1} = 3.18 \mu\text{M}^{-1} \text{s}^{-1} \]

We also calculated 95% confidence intervals for these estimated parameter values. The calculation of 95% confidence intervals will be described in section 5 of this document.

The affinity of dGTP binding onto DNA-KF complexes in solution is estimated to be \( K_{d^{(B)}} = 4.20 \mu\text{M} \) while the affinity of dGTP binding onto the DNA-KF complex on the pore
(which is being pulled by the voltage against the pore) is estimated to be \( K_d = 7.52 \mu\text{M} \). The fact that \( K_d > K_d^{(b)} \) is consistent with our intuitions that the voltage pulling reduces the binding affinity. It may be debatable whether or not the difference is significant. Here we point out that our mathematical model is capable of accommodating this difference.

In the above, when we fit the data to a linear function at the low concentration end, the number of dGTP concentrations used in fitting is constrained by two factors. On one hand, we should use as many dGTP concentrations as possible in fitting to minimize the effect of statistical error. On the other hand, the largest dGTP concentration used should be sufficiently small so the systematic error in determining \( K_d \) is small. In section 5.4 of this document, we calculate and report the 95% confidence interval for each estimated parameter value. Notice that although 0.8 \( \mu\text{M} \) is about 20% of the estimated value of \( K_d^{(b)} = 4.20 \mu\text{M} \), the systematic error caused by including 0.8 mM is not significant in comparison with the statistical error as indicated by the 95% confidence interval (see section 5.4 of this document).

4. **Posterior analysis on the concentration of free dGTP in solution**

In experimental data, the dGTP concentration refers to the initial dGTP concentration with which we start the bulk solution along with 1 \( \mu\text{M} \) of DNA and 2 \( \mu\text{M} \) of KF. As the bulk solution relaxes to equilibrium, some dGTP molecules end up in DNA-KF-dGTP (ternary) complexes. What we really need in our modeling analysis is the concentration of free dGTP in solution at equilibrium. It is important to investigate whether or not the free dGTP concentration is significantly different from the initial dGTP concentration. For that purpose, we do a posterior analysis. Specifically, we start with the assumption that the free dGTP concentration is well approximated by the initial dGTP concentration given in experimental data. Under this assumption, we use the experimental data to determine the values of the 4 model parameters using the method described above. Once the values of the 4 model parameters are estimated, we use them to calculate how different these two dGTP concentrations are. If they are significantly different, then the assumption is incorrect. If these two dGTP concentrations are close to each other, then the assumption is, at least, self-consistent.
We first estimate, \( K_d^{(KF)} \), the affinity of KF binding onto DNA. When we start the bulk solution with 1 \( \mu \)M of DNA and 2 \( \mu \)M of KF in the absence of dGTP, we find that about 60% of DNA (0.6 \( \mu \)M) is in the DNA-KF form. This observation gives us

\[
\frac{0.6}{(1 - 0.6)} = \frac{2 - 0.6}{K_d^{(KF)}}
\]

which yields \( K_d^{(KF)} = 0.9333 \mu \)M. In the equation above, (1-0.6) = 0.4 \( \mu \)M is the concentration of free DNA (without KF bound) in solution and (2-0.6) = 1.4 \( \mu \)M is the concentration of free KF in solution. Before we proceed mathematically, we introduce several notations

\[
\begin{align*}
[dGTP]_0 & : \quad \text{the initial dGTP concentration} \\
[dGTP] & : \quad \text{the concentration of free dGTP in solution} \\
[KF]_0 = 2 \mu \text{M} & : \quad \text{the initial KF concentration} \\
[KF] & : \quad \text{the concentration of free KF in solution} \\
[DNA]_0 = 1 \mu \text{M} & : \quad \text{the initial DNA concentration} \\
[DNA] & : \quad \text{the concentration of free DNA in solution} \\
[DNA-KF] & : \quad \text{the concentration of DNA-KF (binary) complex in solution} \\
[DNA-KF-dGTP] & : \quad \text{the concentration of DNA-KF-dGTP (binary) complex in solution}
\end{align*}
\]

The mathematical goal here is to solve for [dGTP] for each input value of [dGTP]_0 so we can compare these two dGTP concentrations for a wide range of initial dGTP concentrations. The two binding affinities \( K_d^{(KF)} \) and \( K_d^{(B)} \) give us two equations:

\[
\begin{align*}
\frac{[\text{DNA-KF}]}{[\text{DNA}]} &= \frac{[\text{KF}]}{K_d^{(KF)}} \\
\frac{[\text{DNA-KF-DNA}]}{[\text{DNA-KF}]} &= \frac{[\text{dGTP}]}{K_d^{(B)}}
\end{align*}
\]

(7)

System (7) involves more than two unknowns. To solve system (7), we need to reduce the number of unknowns to two. Fortunately, these concentrations are constrained by the conservation of dGTP molecules, KF molecules and DNA molecules
\[
[dGTP]_0 = [dGTP] + [DNA-KF-dGTP]
2 = [KF] + [DNA-KF] + [DNA-KF-dGTP]
1 = [DNA] + [DNA-KF] + [DNA-KF-dGTP]
\]

Substituting these constraints into (7), we arrive at a nonlinear system that has only two unknowns, [DNA-KF] and [dGTP].

\[
\frac{[DNA-KF]}{1-[DNA-KF]+[dGTP]} = \frac{2-[DNA-KF]-[dGTP]_0+[dGTP]}{K_d^{(KF)}}
\]

For each input value of [dGTP]_0, we solve for [DNA-KF] and [dGTP] simultaneously from equation (8). The results are shown in Figure S3 below.

![Graph](image)

**Figure S3.** Free dGTP as a fraction of initial dGTP. The solid line shows the fraction as a function of initial dGTP concentration; the dashed line marks the level of 100%.

At low initial dGTP concentrations, the free dGTP concentration is about 88% of the initial dGTP concentration. As the initial dGTP concentration is increased, this percentage gradually increases and approaches 100% at high dGTP concentration. It is clear that over the full range of
initial dGTP concentrations, the initial dGTP concentration is a fairly good approximation to the concentration of free dGTP in solution at equilibrium.

5. Statistical analysis

5.1 Elimination of outliers

The first task in our statistical analysis of the observed dwell time samples is to get rid of outliers in the data. To detect outliers, we examine the mathematical behaviors of the dwell time (time to escape) distribution in a general escape problem. Consider a general escape problem described by a n-state kinetic model in which escape may occur from some of states. Let $p_j(t)$ denote the probability that the system is in state $j$ at time $t$. For mathematical convenience, we introduce the vector notation:

$$\tilde{p}(t) \equiv (p_1(t), p_2(t), \ldots, p_n(t))^T$$

The governing equation for $\tilde{p}(t)$ has the general form

$$\frac{d}{dt} \tilde{p}(t) = -A \tilde{p}(t) \quad (9)$$

where $A$ is an $n \times n$ matrix. The general solution of equation (9) can be expressed as

$$\tilde{p}(t) = \sum_{j=1}^{n} \exp(-r_j t) \tilde{v}^{(j)}$$

where $\{\tilde{v}^{(j)}, j = 1, 2, \ldots, n\}$ are constant vectors and $\{r_j, j = 1, 2, \ldots, n\}$ are eigenvalues of matrix $A$. Consider the complementary cumulative distribution function of the dwell time:

$$F_c(t) \equiv \sum_{j=1}^{n} p_j(t) = \sum_{j=1}^{n} c_j \exp(-r_j t) \quad (10)$$

In (10), each exponential term decays with different rate. For moderately large time, $F_c(t)$ is dominated by the term that has the smallest decay rate. Thus, for moderately large time, $F_c(t)$ is well approximated by a single exponential term:

$$F_c(t) = c \cdot \exp(-rt)$$

Or equivalently
\[
\log[F_C(t)] \approx -rt + \log(c)
\]

The slope, \(r\), can be estimated as

\[
r \approx \frac{\log[F_C(t_a)] - \log[F_C(t_b)]}{t_b - t_a}
\]

where \(t_a\) and \(t_b\) are two moderately large times. We introduce a new function

\[
G(t) \equiv \log[F_C(t)] + \frac{\log[F_C(t_a)] - \log[F_C(t_b)]}{t_b - t_a} \cdot t + \alpha
\]

It is clear that function \(G(t)\) should be approximately a constant for moderately large time. In the equation above, \(\alpha\) is a constant added to make \(\min G(t) = 1\) so we can use log scale for plotting \(G(t)\). This is numerically convenient since outliers may make the range of \(G(t)\) very large.

To detect outliers, we calculate \(G(t)\) from the observed dwell time samples. If the calculated \(G(t)\) deviates significantly from being a constant, then an outlier is detected at the location where the deviation from a constant level is most significant. Once detected, the outlier is deleted from the data and the process is repeated on the revised data (with detected outlier removed). Below we describe the specific procedure.

Consider a set of observed dwell time samples. We first sort the set into the ascending order.

\[
\{T_j, j = 1, 2, \ldots, m\}, \quad T_i \leq T_j \quad \text{for} \quad i < j
\]

**Step 1:** Form the grid on which we calculate \(F_C(t)\) and \(G(t)\).

We use middle points between adjacent samples as grid points:

\[
\{t_{j+1/2}, j = 1, 2, \ldots, m - 1\}, \quad t_{j+1/2} = \frac{T_j + T_{j+1}}{2}
\]

**Step 2:** Calculate \(F_C(t_{j+1/2})\).

\[
F_C(t_{j+1/2}) = \frac{m-j}{m}
\]

**Step 3:** Calculate \(G(t_{j+1/2})\).

We select \(t_a\) and \(t_b\) from the grid \(\{t_{j+1/2}, j = 1, 2, \ldots, m - 1\}\) such that
\[ F_c (t_a) \approx 16\%, \quad F_c (t_b) = 8\% \]

**Step 4:** Plot \( G(t_{j+1/2}) \) vs \( F(t_{j+1/2}) \) where \( F(t) \) is the cumulative distribution function

\[ F(t) = 1 - F_c (t) \]

The advantage of plotting \( G(t_{j+1/2}) \) vs \( F(t_{j+1/2}) \) is that the samples are uniformly distributed in the space of cumulative distribution \( F(t) \). We could plot \( G(t_{j+1/2}) \) vs \( t_{j+1/2} \). But the samples are very densely distributed for small \( t \) and very sparsely distributed for large \( t \).

The presence of an outlier at \( T_{j+1} \) increases \( F_c (t_{j+1/2}) \) and consequently increases \( G(t_{j+1/2}) \).

If the plot of \( G(t_{j+1/2}) \) vs \( F(t_{j+1/2}) \) is approximately a straight line, then no outlier is detected and we stop at here.

If the plot of \( G(t_{j+1/2}) \) vs \( F(t_{j+1/2}) \) deviates significantly from being a straight line, then outliers are present in the data. To detect outliers, we find the location of the maximum of \( \{ G(t_{j+1/2}), j = 1, 2, \ldots, m-1 \} \). Suppose the maximum is attained at \( j = d \). We identify \( T_{d+1} \) as an outlier, remove it from the data, and start again at Step 1 with the revised data.

Figure S4 below shows results of each iteration when applying this procedure to the data set for \([dGTP] = 0.2 \mu\text{M}\). For this data set, 7 outliers are detected and removed.
Figure S4. Results of each iteration when applying the procedure described above to eliminate outliers in the data set at [dGTP] = 0.2 µM
5.2 Calculation of key statistical quantities of the EBS dwell time distributions

After outliers are removed, we calculate 1) sample mean, 2) sample variance, 3) 95% confidence interval of each sample mean, 4) sample median, and 5) 95% confidence interval of each sample median.

The sample mean is

$$\langle T \rangle \approx \frac{1}{m} \sum_{j=1}^{m} T_j$$

The sample variance is

$$\text{var}(T) = \frac{1}{m-1} \sum_{j=1}^{m} (T_j - \langle T \rangle)^2$$

The 95% confidence interval of the sample mean is

$$\left( \langle T \rangle - R, \quad \langle T \rangle + R \right)$$

where the radius $R$ of the 95% confidence interval of the sample mean is given by

$$R \approx \frac{2}{\sqrt{m}} \sqrt{\text{var}(T)}$$

which is twice the standard deviation of the sample mean.

The sample median is calculated as the 50% percentile.

The 95% confidence interval of the sample median is estimated using the bootstrap re-sampling method (B. Efron and G. Gong, “A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation”, *The American Statistician*, Vol. 37 (1983), pp36-48.)

Table S1 below lists these key statistical quantities of the EBS (enzyme bound state) dwell time distributions for various initial dGTP (correct nucleotide) concentrations while the initial DNA concentration is 1 µM and the initial KF concentration is 2 µM.
Table S2 below lists these key statistical quantities of the EBS dwell time distributions for various initial dTTP (incorrect nucleotide) concentrations while the initial DNA concentration is 1 µM and the initial KF concentration is 2 µM.

Table S3 below lists these key statistical quantities of the EBS dwell time distributions for various initial KF concentrations while the initial DNA concentration is 1 µM and the initial dGTP concentration is zero.
<table>
<thead>
<tr>
<th>[dGTP] (µM)</th>
<th>Number of samples</th>
<th>Median (ms)</th>
<th>95% confidence interval of median (ms)</th>
<th>Mean (ms)</th>
<th>Standard Deviation (ms)</th>
<th>Radius of 95% confidence interval of mean (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>177</td>
<td>1.90</td>
<td>(1.6, 2.1)</td>
<td>3.05</td>
<td>3.857</td>
<td>0.580</td>
</tr>
<tr>
<td>0.2</td>
<td>105</td>
<td>2.4</td>
<td>(1.8, 3.4)</td>
<td>5.24</td>
<td>7.536</td>
<td>1.471</td>
</tr>
<tr>
<td>0.4</td>
<td>287</td>
<td>2.4</td>
<td>(2.0, 2.9)</td>
<td>8.84</td>
<td>17.58</td>
<td>2.08</td>
</tr>
<tr>
<td>0.8</td>
<td>269</td>
<td>3.0</td>
<td>(2.3, 3.6)</td>
<td>11.22</td>
<td>23.11</td>
<td>2.82</td>
</tr>
<tr>
<td>1.0</td>
<td>177</td>
<td>2.3</td>
<td>(1.9, 3.1)</td>
<td>8.94</td>
<td>18.69</td>
<td>2.81</td>
</tr>
<tr>
<td>1.3</td>
<td>337</td>
<td>3.4</td>
<td>(2.8, 4.1)</td>
<td>10.82</td>
<td>19.46</td>
<td>2.12</td>
</tr>
<tr>
<td>2</td>
<td>365</td>
<td>10.2</td>
<td>(7.9, 12.9)</td>
<td>26.55</td>
<td>41.07</td>
<td>4.3</td>
</tr>
<tr>
<td>2.5</td>
<td>483</td>
<td>5.9</td>
<td>(4.7, 7.2)</td>
<td>21.74</td>
<td>43.42</td>
<td>3.95</td>
</tr>
<tr>
<td>4</td>
<td>370</td>
<td>17.8</td>
<td>(13.5, 20.2)</td>
<td>42.9</td>
<td>72.96</td>
<td>7.59</td>
</tr>
<tr>
<td>8</td>
<td>653</td>
<td>25.0</td>
<td>(22.0, 28.6)</td>
<td>51.49</td>
<td>71.28</td>
<td>5.57</td>
</tr>
<tr>
<td>15</td>
<td>251</td>
<td>30.1</td>
<td>(25.0, 38.3)</td>
<td>61.35</td>
<td>82.95</td>
<td>10.47</td>
</tr>
<tr>
<td>30</td>
<td>236</td>
<td>29.1</td>
<td>(22.1, 35.9)</td>
<td>50.35</td>
<td>60.01</td>
<td>7.81</td>
</tr>
<tr>
<td>50</td>
<td>167</td>
<td>31.5</td>
<td>(22.7, 41.7)</td>
<td>57.8</td>
<td>77.17</td>
<td>11.94</td>
</tr>
<tr>
<td>100</td>
<td>323</td>
<td>51.6</td>
<td>(44.8, 62.5)</td>
<td>84.47</td>
<td>92.19</td>
<td>10.26</td>
</tr>
<tr>
<td>180</td>
<td>631</td>
<td>76.9</td>
<td>(67.9, 88.2)</td>
<td>144.9</td>
<td>193.7</td>
<td>15.52</td>
</tr>
<tr>
<td>300</td>
<td>121</td>
<td>224.8</td>
<td>(183.0, 290.0)</td>
<td>312.9</td>
<td>270.9</td>
<td>49.25</td>
</tr>
<tr>
<td>1000</td>
<td>125</td>
<td>314.5</td>
<td>(251.0, 410.0)</td>
<td>478.9</td>
<td>516.9</td>
<td>92.46</td>
</tr>
<tr>
<td>3000</td>
<td>89</td>
<td>965.1</td>
<td>(625.8, 1218)</td>
<td>1295.4</td>
<td>1251.0</td>
<td>265.3</td>
</tr>
<tr>
<td>10000</td>
<td>69</td>
<td>1981</td>
<td>(1429, 2525)</td>
<td>2699.7</td>
<td>2410.0</td>
<td>580.2</td>
</tr>
</tbody>
</table>

**Table S1.** Key statistical quantities of the EBS dwell time distributions for various initial dGTP (correct nucleotide) concentrations while the initial DNA concentration is 1 µM and the initial KF concentration is 2 µM.
<table>
<thead>
<tr>
<th>[dTTP] (µM)</th>
<th>Number of samples</th>
<th>Median (ms)</th>
<th>95% confidence interval of median (ms)</th>
<th>Mean (ms)</th>
<th>Standard Deviation (ms)²</th>
<th>Radius of 95% confidence interval of mean (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>177</td>
<td>1.90</td>
<td>(1.6, 2.1)</td>
<td>3.05</td>
<td>3.857</td>
<td>0.580</td>
</tr>
<tr>
<td>100</td>
<td>105</td>
<td>2.2</td>
<td>(1.6, 2.6)</td>
<td>3.54</td>
<td>4.40</td>
<td>0.858</td>
</tr>
<tr>
<td>500</td>
<td>189</td>
<td>2.4</td>
<td>(2.1, 2.9)</td>
<td>4.35</td>
<td>5.03</td>
<td>0.731</td>
</tr>
<tr>
<td>3000</td>
<td>162</td>
<td>2.85</td>
<td>(2.2, 3.7)</td>
<td>16.25</td>
<td>41.17</td>
<td>6.47</td>
</tr>
</tbody>
</table>

**Table S2.** Key statistical quantities of the EBS dwell time distributions for various initial dTTP (incorrect nucleotide) concentrations while the initial DNA concentration is 1 µM and the initial KF concentration is 2 µM

<table>
<thead>
<tr>
<th>[KF] (µM)</th>
<th>Number of samples</th>
<th>Median (ms)</th>
<th>95% confidence interval of median (ms)</th>
<th>Mean (ms)</th>
<th>Standard Deviation (ms)²</th>
<th>Radius of 95% confidence interval of mean (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>48</td>
<td>1.8</td>
<td>(1.3, 2.6)</td>
<td>2.563</td>
<td>2.454</td>
<td>0.709</td>
</tr>
<tr>
<td>0.5</td>
<td>138</td>
<td>2.5</td>
<td>(2.0, 2.65)</td>
<td>2.915</td>
<td>2.625</td>
<td>0.447</td>
</tr>
<tr>
<td>1.0</td>
<td>143</td>
<td>1.9</td>
<td>(1.6, 2.2)</td>
<td>2.424</td>
<td>2.008</td>
<td>0.336</td>
</tr>
<tr>
<td>1.5</td>
<td>200</td>
<td>2.0</td>
<td>(1.6, 2.3)</td>
<td>2.908</td>
<td>3.0</td>
<td>0.424</td>
</tr>
<tr>
<td>2.0</td>
<td>176</td>
<td>1.85</td>
<td>(1.6, 2.1)</td>
<td>2.875</td>
<td>3.208</td>
<td>0.484</td>
</tr>
<tr>
<td>3.0</td>
<td>69</td>
<td>2.2</td>
<td>(1.7, 3.0)</td>
<td>3.319</td>
<td>3.145</td>
<td>0.757</td>
</tr>
</tbody>
</table>

**Table S3.** Key statistical quantities of the EBS dwell time distributions for various initial KF concentrations in the absence of dGTP while the initial DNA concentration is 1 µM.
5.3 Calculation of 95% confidence intervals for model parameters

As we described in sections 2 and 3, parameters \( \{k_i, K^{(b)}_d, K_d, k_{off}\} \) are estimated by fitting the model to the low concentration data set and to the high concentration data set. We now discuss how we calculate 95% confidence intervals for these estimated parameter values.

First, we point out that in the calculation of parameter values we used only sample means of dwell time at various dGTP concentrations. At each dGTP concentration, the sample mean of dwell time is calculated from many experimentally observed dwell times. To facilitate the presentation, we introduce mathematical notation:

- \( D_{\text{Exp}} \): the experimental data set of sample means of dwell time
- \( P_{\text{Exp}} \): the vector of parameter values \( \{k_i, K^{(b)}_d, K_d, k_{off}\} \) estimated from \( D_{\text{Exp}} \)
- \( F \): the mathematical function that maps data to estimated parameter values

Mathematically, the process of estimating parameter values from data is written as

\[
P_{\text{Exp}} = F(D_{\text{Exp}})
\]

To calculate the uncertainty in the estimated parameter values, we generate artificial data sets of sample means of dwell time. Sample means of dwell time are random variables. At each dGTP concentration, the sample mean of dwell time is approximately a Gaussian with

- mean = experimentally observed sample mean
- variance = experimentally observed sample variance

We generate \( N = 50000 \) artificial data sets

\[
\{D_n, \quad n = 1, 2, \ldots, N\}
\]

From each artificial data set, we calculate the corresponding vector of estimated parameter values by fitting the model to the data

\[
\{P_n, \quad n = 1, 2, \ldots, N\} \quad \text{where} \quad P_n = F(D_n)
\]

For the \( i \)-th component of \( P \), the 95% confidence interval is approximated by \( [a_i, b_i] \) where

\[
a_i = 2.5\% \text{ percentile of } \{P_n(i), \quad n = 1, 2, \ldots, N\}
\]
\[ b_i = 97.5\% \text{ percentile of } \{P_n(i), \quad n = 1, 2, \ldots, N\} \]

Here we need to point out that the intervals constructed this way are not the true confidence intervals. They are actually the prediction intervals for the estimated parameter values when \( P_{\text{Exp}} \) is used as the true values of model parameters in generating artificial data sets. For a random artificial data set, the corresponding estimated parameter values are random variables. The prediction intervals tell us about the distribution of parameter values estimated from artificial data sets, which can be viewed as a measure on the uncertainty in estimated parameter values. Since the true confidence intervals are very difficult to determine, we use the prediction intervals to approximate the confidence intervals.

We use the procedure described above to calculate the 95\% confidence intervals for estimated values of model parameters. The results are listed below. For each model parameter, we display the estimated value followed by the 95\% confidence interval.

\[ y_1 = 3.45 \quad (2.49, 4.41) \]
\[ s_1 = 10.4 \quad (6.77, 14.1) \]
\[ y_2 = 45.2 \quad (33.6, 56.5) \]
\[ s_2 = 0.459 \quad (0.408, 0.507) \]
\[ k_i = 290 \text{ s}^{-1} \quad (227, 402) \]
\[ K_d = 7.52 \mu M \quad (6.81, 8.46) \]
\[ k_{\text{off}} = 23.9 \text{ s}^{-1} \quad (18.8, 33.1) \]
\[ K_{d_i}^{(n)} = 4.20 \mu M \quad (2.72, 6.89) \]
\[ k_{\text{on}} = 3.18 \mu M^{-1}\text{s}^{-1} \quad (2.29, 4.71) \]

### 5.4 Histograms of observed dwell times with 95\% confidence intervals

In this subsection, we discuss the calculation of 95\% confidence intervals for histograms. Specifically, we calculate the 95\% confidence interval for the observed number of samples in each individual bin.

Let \( N \) be the number of samples used in the histogram. We consider an arbitrary bin. Let \( p \) be the true underlying probability that a random sample is in the bin. Let \( n \) be the observed number of samples in the bin. Approximately, we have
\[ p \approx \frac{n}{N} \]

If the experiment is repeated, the number of samples in the bin is a random variable. Let \( B \) denote this random variable. \( B \) has a binomial distribution with number of trials \( = N \) and probability of success \( = p \approx \frac{n}{N} \). For the observed number of samples in the bin, the 95% confidence interval is approximated by \([a_i, b_i]\) where

\[ a_i = 2.5\% \text{ percentile of binomial distribution with number of trials} = N \]

and probability of success \(= \frac{n}{N} \).

\[ b_i = 97.5\% \text{ percentile of binomial distribution with number of trials} = N \]

and probability of success \(= \frac{n}{N} \).

Again, the interval constructed this way is not the true confidence interval. It is actually the prediction interval for the number of samples in the bin when \( \frac{n}{N} \) is used as the true probability of the bin in repeating experiments. The number of samples in the bin is a binomial random variable. The prediction interval tells us about the distribution of the number of samples in the bin, which can be viewed as a measure on the uncertainty in the observed number of samples in the bin. Since the true confidence interval is very difficult to determine, we use the prediction interval to approximate the confidence interval.

We use the procedure described above to calculate the 95% confidence intervals for individual bins in histograms. Below we display the histograms of log dwell time for various dGTP concentrations. In the histograms, error bars represent the 95% confidence intervals, and the thick black solid lines represent the probability densities of dwell time predicted from the mathematical model using the model parameters estimated in section 3.
Figure S5. Histograms of log dwell time for [dGTP] = 0, 0.2, 0.4, 0.8 µM. The error bars represent the 95% confidence intervals, and the thick black lines represent the probability densities of dwell time predicted from the mathematical model.
**Figure S6.** Histograms of log dwell time for [dGTP] = 1.3, 2.5, 4.0, 8.0 µM. The error bars represent the 95% confidence intervals, and the thick black lines represent the probability densities of dwell time predicted from the mathematical model.
Figure S7. Histograms of log dwell time for [dGTP] = 15, 30, 75, 180 µM. The error bars represent the 95% confidence intervals, and the thick black lines represent the probability densities of dwell time predicted from the mathematical model.
Figure S8. Histograms of log dwell time for [dGTP] = 300, 1000, 3000, 10000 µM. The error bars represent the 95% confidence intervals, and the thick black lines represent the probability densities of dwell time predicted from the mathematical model.
5.5 Comparison of histograms of observed dwell times and probability densities of dwell time predicted from the mathematical model

In Figures S5-S8 above, we displayed both the histograms of observed dwell times and the probability densities of dwell time predicted from the mathematical model.

Although we used only mean dwell times in model fitting, it is important to point out that our model is not based on the assumption that dwell time has a single exponential distribution. To the contrary, our model equation is a 2 x 2 ODE system. Since a 2 x 2 ODE system has two eigenvalues, the probability density of the dwell time predicted from the model has the form of a sum of two exponential terms, one corresponding to each eigenvalue. We solved the 2 x 2 ODE system semi-analytically and examined the behavior of the two exponential terms in the probability density of dwell time as dGTP concentration is varied. We found the following results.

1. At very low dGTP concentration ([dGTP] << $K_d^{(B)}$), the exponential term with the smaller eigenvalue has negligible weight so the dwell time distribution is a single exponential term with the larger eigenvalue.

2. In the range of intermediate low dGTP concentrations (from [dGTP] << $K_d^{(B)}$ to ([dGTP] ~ $K_d^{(B)}$), as the dGTP concentration is increased, the magnitudes of the two eigenvalues stay roughly the same while the weight shifts from the term with the larger eigenvalue gradually toward the term with the smaller eigenvalue.

3. In the range of intermediate high dGTP concentrations (from ([dGTP] ~ $K_d^{(B)}$ to [dGTP] >> $K_d^{(B)}$), as the dGTP concentration is increased, the weight shifts further away from the term with the larger eigenvalue toward the term with the smaller eigenvalue. At the same time the smaller eigenvalue becomes smaller as the dGTP concentration is increased.

4. At very high dGTP concentration ([dGTP] >> $K_d^{(B)}$), the exponential term with the larger eigenvalue has negligible weight so again the dwell distribution is a single exponential term with the smaller eigenvalue.

In particular, in the intermediate range of dGTP concentrations, the dwell time distribution exhibits two exponential terms. All these results predicted from the mathematical model are qualitatively consistent with the observed dwell time histograms (see Figures S5 – S8 above). To have a more meaningful quantitative comparison of the model prediction and data on the dwell
time distribution, we need to have much more data (and more accurate data), which will be carried out in a subsequent study.