Single-particle trajectories reveal two-state diffusion-kinetics of hOGG1 proteins on DNA

Citation: Vestergaard, Christian L et al. "Single-Particle Trajectories Reveal Two-State Diffusion-Kinetics of hOGG1 Proteins on DNA." Nucleic Acids Research 46, 5 (January 2018): 2446–2458 © 2018 The Author(s)

As Published: http://dx.doi.org/10.1093/NAR/GKY004

Publisher: Oxford University Press (OUP)

Persistent URL: http://hdl.handle.net/1721.1/117579

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

2446–2458 *Nucleic Acids Research, 2018, Vol. 46, No. 5* *Published online 18 January 2018*
doi: 10.1093/nar/gky004

Single-particle trajectories reveal two-state diffusion-kinetics of hOGG1 proteins on DNA

Christian L. Vestergaard[1,2,3,*], Paul C. Blainey[4] and Henrik Flyvbjerg[1]

[1]Department of Micro- and Nanotechnology, Technical University of Denmark, Kgs. Lyngby DK-2800, Denmark, [2]Decision and Bayesian Computation, Pasteur Institute, CNRS UMR 3571, 25–28 rue du Dr Roux, 75015 Paris, France, [3]Bioinformatics and Biostatistics Hub, C3BI, Pasteur Institute, CNRS USR 3756, 25–28 rue du Dr Roux, 75015 Paris, France and [4]MIT Department of Biological Engineering and Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA

Received May 31, 2017; Revised December 20, 2017; Editorial Decision December 30, 2017; Accepted January 05, 2018

ABSTRACT

We reanalyze trajectories of hOGG1 repair proteins diffusing on DNA. A previous analysis of these trajectories with the popular mean-squared-displacement approach revealed only simple diffusion. Here, a new optimal estimator of diffusion coefficients reveals two-state kinetics of the protein. A simple, solvable model, in which the protein randomly switches between a loosely bound, highly mobile state and a tightly bound, less mobile state is the simplest possible dynamic model consistent with the data. It yields accurate estimates of hOGG1's (i) diffusivity in each state, uncorrupted by experimental errors arising from shot noise, motion blur and thermal fluctuations of the DNA; (ii) rates of switching between states and (iii) rate of detachment from the DNA. The protein spends roughly equal time in each state. It detaches only from the loosely bound state, with a rate that depends on pH and the salt concentration in solution, while its rates for switching between states are insensitive to both. The diffusivity in the loosely bound state depends primarily on pH and is three to ten times higher than in the tightly bound state. We propose and discuss some new experiments that take full advantage of the new tools of analysis presented here.

INTRODUCTION

Molecular fluorescent labels and super-resolution microscopy allow us to track single biomolecules in cells (1). There, diffusion is ubiquitous, as many cellular processes rely on diffusion for transport (2). A precise understanding of such processes requires a precise determination of diffusion constants. Less than that may miss process-specific details by lumping them into one, simple diffusive process, we show below.

Recent examples of experimental measurements of diffusion in biological systems include supercoils on DNA (3), proteins on biopolymers such as DNA (4–9) or microtubules (10), on surfaces (11), in natural (12–14) and artificial (15) lipid membranes, in films (16) and inside cells (17,18), all recorded with time-lapse photography.

Diffusion of proteins on DNA captured the attention of biophysicists nearly half a century ago when it was observed that the Lac repressor was able to locate its cognate site *in vitro* even faster than theoretically predicted for 3D diffusion in bulk (19). That such rates can even be approached under conditions, where many protein molecules bind DNA nonspecifically is all the more impressive. Such rapid target-binding activity is explained by fast 1D sliding along the DNA in many transcription factor and DNA repair protein systems. But high-speed transport is not a sufficient explanation since these proteins must probe the DNA to identify their targets. For some proteins, target-containing DNA is only subtly different from non-specific DNA, making the search all the more difficult. This is particularly the case for human oxoguanine DNA glycosylase 1 (hOGG1) repair protein (20–22), the subject of the present study. hOGG1 scans DNA to identify oxidized guanine bases and thus must diffuse fast along nonspecific DNA and also recognize and bind stably to oxoguanine bases, which may require local conformational change.

The apparently dichotomous requirements of rapid transport on nonspecific DNA and effective probing/stable target binding through DNA shape-sensing and/or base-specific interactions (which almost certainly increases free energy barriers for translocation) is known as the 'speed-stability paradox'. Recently, it was shown that while the Lac repressor recognizes its cognate sites with low probability in a single pass, the redundancy of 1D diffusional search leads to a high probability of site recognition in the course of a single binding event (23).

*To whom correspondence should be addressed. Tel: +33 1 45 68 87 38; Email: cvestergaard@gmail.com
Present address: Christian L. Vestergaard, Decision and Bayesian Computation, Pasteur Institute, CNRS UMR 3571, 25–28 rue du Dr Roux, 75015 Paris, France.

One proposed solution to the speed-stability paradox is the ability of the searching protein to adopt multiple states: a fast-sliding 1D 'search' state that is minimally affected by the presence of targets interconverting with a slower-sliding 'recognition' state that recognizes targets with significant probability and has sacrificed speed-of-search in favor of the increased DNA interaction necessary for recognition (24).

It is a major technical challenge to determine the dynamics of search and recognition states quantitatively, as they are predicted to interconvert rapidly and stochastically to search DNA effectively. Consequently, the two states are not resolved, but averaged over, in standard ensemble biochemical assays.

Single-molecule-tracking assays promise to resolve individual biomolecule dynamics on DNA. Trajectories are mostly short, however, and position data are contaminated by multiple sources of experimental error. In particular: localization errors due to diffraction in microscope optics and limited numbers of recorded photons; motion blur due to particle movement during the camera exposure time; and thermal fluctuations of the DNA on which the particle diffuses. We need to account for these errors or risk severely biased results (25). The analysis is further complicated by the fact that we cannot simply average over multiple trajectories to reduce statistical error since we are interested in resolving individual molecular dynamics (26), and we can neither see nor model the moving substrate: It is invisible, and no solvable theory exists for its motion. Thus we face a sixtuple of challenges that reinforce each other: (i) resolve *individual* particle dynamics (ii) from mostly *short* trajectories (iii) recorded with *considerable localization errors and motion blur* (iv) on a *moving* substrate (v) that is *invisible* and (vi) its motion *uncharted* by theory.

Here, we show how optimal estimators that treat noise sources in the single-molecule tracking data rigorously enable identification and quantitative characterization of two distinct states of searching of hOGG1 molecules along non-specific DNA. We analyze the diffusion of individual hOGG1 proteins that were tracked *in vitro* while diffusing on λ DNA stretched in a shear flow over a cover slip (Figure 1A–E).

We characterize the motion of the invisible substrate using data for the motion of the protein. That motion is recorded in two dimensions, while the substrate, the DNA, is one-dimensional and fairly stretched. Using this, we overcome the absence of a specific theory with a generic yet precise phenomenology that all realistic theories must share.

Using this phenomenology, we have constructed optimal estimators of diffusion coefficients of particles diffusing on a fluctuating substrate, e.g. DNA (25). Applying these estimators here, we find diffusion constants that, together with the distribution of residence times of hOGG1 on DNA, point unambiguously to a two-state kinetics of hOGG1 on DNA.

We propose an analytically solvable model for the kinetics of hOGG1: a protein binds to DNA in a loosely bound state and switches stochastically between this state and another more tightly bound state, until it detaches again from the DNA, from the loosely bound state. We provide accurate estimates of the kinetic parameters of our model, showing that the loosely bound state has much higher diffusivity

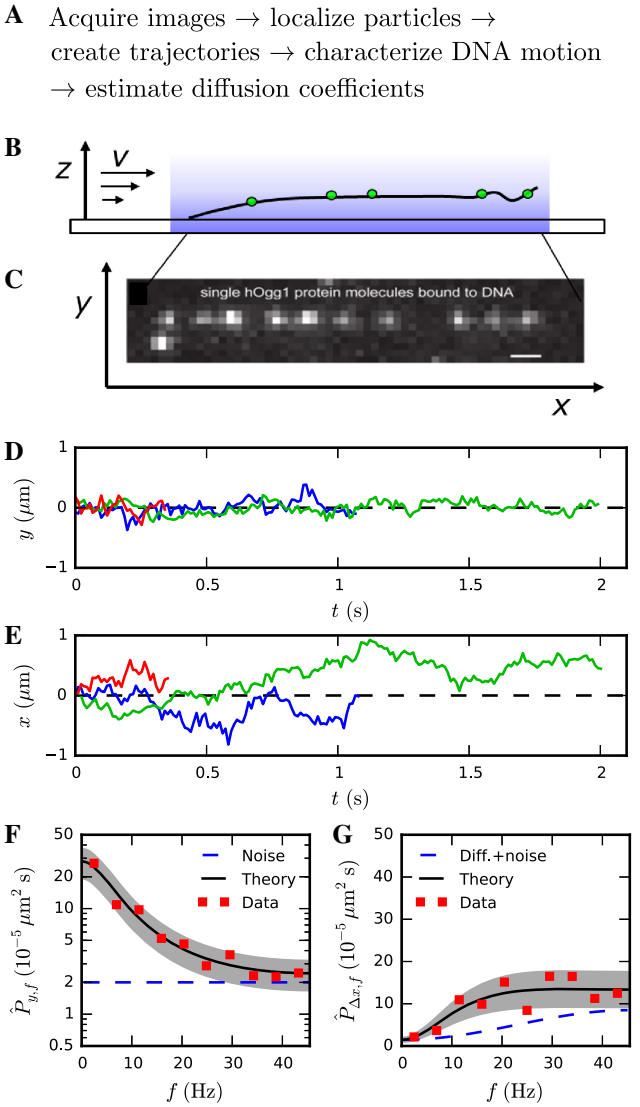

Figure 1. Experimental measurements of diffusion coefficients of hOGG1 proteins on flow-stretched λ DNA. (**A**) Workflow for estimation of diffusion coefficients from experimental tracking data of diffusing particles on DNA. (**B**) Experimental setup (not to scale): Proteins (green) on flow-stretched DNA (black) attached to cover slip at one end and fluctuating in a shear flow. (**C**) Image of several fluorescent hOGG1 molecules bound to and diffusing on a DNA molecule at higher density than during data acquisition for illustration (27). Scale-bar = 1 μm. (**D**) Transverse coordinates y of the trajectories of three hOGG1 proteins diffusing on DNA and recorded with time-lapse $\Delta t = 11$ ms. The mean residence time of proteins on DNA ranges from 50 to 500 ms, depending on solution conditions. (**E**) Longitudinal coordinates x of the same three proteins. (**F**) Periodogram $\hat{P}_y$ of the transverse coordinate of a protein diffusing on DNA. The transverse fluctuations fit a Lorentzian plus a constant. The corner frequency (3 dB frequency) f_c of the Lorentzian is 6.7 Hz. (**G**) Periodogram $\hat{P}_{\Delta x}$ of longitudinal displacements of the same protein. The periodogram of displacements Δx is used, because diffusion is an unbounded process, making the periodogram of x a bad statistics. The expected value of $\hat{P}_{\Delta x}$, the power spectrum $P_{\Delta x}$, is the sum of a diffusion term, a white-noise term, and a single Lorentzian term describing longitudinal DNA fluctuations. The corner frequency of these fluctuations is $f_{c,x} = 2f_c$, in agreement with our assumption that DNA fluctuations in the y- and z-directions contribute equally to DNA fluctuations in the x-direction. Shown values for data in F, G are block averages, each over 20 periodogram values, and the grey areas mark the 68% confidence interval (CI) for the block averages.

than the tightly bound state. We hypothesize that the loosely bound and highly mobile state allows the protein to efficiently cover the length of the DNA, i.e. acting as a 1D travel and 'search' state, while the tightly bound and less mobile state allows it to probe for and recognize oxidative damage to DNA bases.

The experiment's pH and salt concentration have been shown to affect the diffusivity and binding times of hOGG1 on DNA (27,28). Our estimates of the kinetic parameters of our model show that pH and salt concentration do not significantly influence the proteins' transition rates between the two states and thus do not influence the time it spends in each state within the range of conditions tested here. Instead, pH and salt concentration affect hOGG1's diffusion coefficient in the loosely bound state and its rate of detachment from the DNA from this state.

Some of the data analyzed here were previously analyzed using a standard method, a straight-line fit to an ensemble of mean squared displacements (MSDs) of trajectories (27). This previous analysis assumed that substrate motion did not contribute to observed longitudinal displacements and that all such displacements were described by a single diffusion coefficient. The present analysis thus demonstrates the usefulness of optimized statistical methods, such as the CVE and MLE (25). They make the analysis of individual trajectories possible and that is crucial to fully exploit the single-molecule resolution offered by single-molecule experiments.

MATERIALS AND METHODS

Experimental data

The experimental setup and data acquisition are described in detail in (27). It produced five different data sets consisting of time-series of (x, y)-coordinates of thousands of fluorescently marked hOGG1 proteins that diffused on single flow-stretched λ-phage DNA molecules (48.5 kb = 16 μm long) at a range of different pH-values, 6.6–7.8, and salt concentrations, 0.01–0.1 M. Each data-set consisted of hundreds to thousands of time-series recorded at the experimental conditions listed in Table 1. Three of the data-sets (corresponding to DNA molecules 2–5) were previously analyzed in (27), while two datasets (DNA molecules 1 and 6) have not been analyzed previously.

In each experiment, a DNA molecule (two in one experiment) was end-biotinylated and fused to a coverslip (Figure 1B). The DNA was stretched to ∼75% of its contour length by a shear flow (the flow speed at the DNA was 100 $\mu m\, s^{-1}$). If the free end of the DNA, where it tends to curl up, is excluded, the remaining DNA is stretched 90% (27). hOGG1 proteins fluorescently marked with Cy3B diffused on the DNA molecule and were filmed using total internal reflection microscopy (TIRF) by an EMCCD camera with pixel width corresponding to 250 nm. The proteins were tracked in the resulting movie until they detached from the DNA.

Trajectories of protein positions were estimated by fitting 2D Gaussians plus constant backgrounds to the recorded point-spread functions as described in (27). Trajectories were previously analyzed by fitting a straight line to the average over the MSDs of all trajectories in each data-set (27).

Here, we estimated diffusion coefficients, the variance of localization errors, and parameters characterizing the DNA's motion statistically from these same trajectories using the optimal MLE and CVE methods developed in (25), as described below and in Supplementary Section S3.

In Supplementary Section S3F we compare the CVE and MLE to MSD-based methods and to recent Bayesian methods for inferring multi-state diffusion.

A priori knowledge used to separate the protein's diffusion along DNA from DNA motion

For each tracked protein, we know its trajectory measured in lab coordinates, apart from localization errors and motion blur. It is its trajectory on the DNA, however, that relates to its diffusion coefficient on the DNA. The difference between the two trajectories is given by the motion of the invisible DNA in the lab. This may seem an impasse because the DNA's motion was not observed and we have no complete theory for its motion.

We do, however, have knowledge of general mathematical properties that any linear model of the DNA's motion must satisfy. This leads to a phenomenological model for the motion of local segments of the DNA, and that is sufficient, since any visiting protein visits only a local segment of the DNA. This model dictates a protocol for how to analyze trajectory data in order to separate the motion of a protein along DNA from the motion of the DNA in the lab and thus obtain accurate estimates of protein diffusion coefficients. We sketch this phenomenological model in the present subsection and give the protocol in the next subsection. Details are given in (25).

Worm-like Chain model in free-draining flow. The ultimate way to do describe the DNA's motion, is to solve a realistic dynamical model for it. This is impossible, however: The simplest realistic model would treat the DNA as a massless, semi-flexible, unstretchable fiber (the Kratky-Porod model, a.k.a. the *Worm-Like Chain* model) in a free-draining shear flow over the coverslip to which one end of the DNA is attached. The result is a non-linear partial differential equation for the over-damped thermal Brownian motion of the fiber in the flow. The boundary condition on that motion imposed by the coverslip adds to the intractability of this model by exact analytical means.

Spectral theory invoked. The non-linear nature of this simplest realistic model means that general results from mathematics' *spectral theory* do not apply directly to it. They may apply indirectly, however, e.g., if our theory is well approximated by a linearized version of it. We expect this to be the case since the DNA is stretched taut and its fluctuations consequently are relatively small.

So, despite the model's analytical intractability, qualitative mathematical results exist that may well apply to it. These results are of such a nature that experimental data will reveal whether they apply or not. Thus we need not argue their case theoretically. Right or wrong, our explanation does not affect the correctness of our ensuing data analysis, because that analysis relies solely on the phenomenological spectral theory described here and dictated by experimental data.

Table 1. Goodness-of-fit of the one-state and the two-state models. Column 1: Identity of DNA molecule(s). Column 2: pH-value of buffer. Column 3: Salt concentration of buffer. Column 4: Time lapse and shutter time of movie; the shutter time was set equal to the time lapse to maximize the number of recorded photons (45). Column 5: Number of time-series. Column 6: *P*-value (and number of degrees of freedom (dof) fitted) for the one-state model's exponential distribution to the distribution of measured residence times and its predicted constant average diffusion coefficient as function of residence time to measured diffusion coefficients. Column 7: Akaike weight wAIC for the one-state model. Column 8: *P*-value (and number of degrees of freedom (dof) fitted) for the two-state model's double-exponential distribution to measured residence times and its predicted average diffusion coefficient as function of residence time to experimental data. Column 9: Akaike weight wAIC for the two-state model; a more complex model that allows detachment from the tightly bound state shows no improvement of the fits, and is thus rejected by the Akaike information criterion (Supplementary Table S3). A low P-value means that the model can be rejected, while a high *P*-value means that the data supports the model. $P \geq 0.05$ means the model is supported by any meaningful significance level. Akaike weights measure the relative probability of the models given the experimental data (47)

1	2	3	4	5	6	7	8	9
		Data set			One-state model		Two-state model	
DNA no.	pH	[NaCl]	Δt	n	P (dof)	w_{AIC}	P (dof)	w_{AIC}
1	6.6	100 mM	25 ms	228	0.11 (11)	0.002	0.64 (8)	1.00
2 & 3	7.0	10 mM	28.5 ms	254	0.004 (12)	0.10	0.49 (9)	0.90
4	7.0	75 mM	7 ms	55	3×10^{-6} (10)	2×10^{-8}	0.15 (7)	1.00
5	7.5	10 mM	11 ms	246	6×10^{-10} (13)	2×10^{-12}	0.24 (10)	1.00
6	7.8	50 mM	15 ms	92	8×10^{-8} (12)	6×10^{-12}	0.61 (9)	1.00
All	–	–	–	875	$<10^{-16}$ (58)	$<10^{-16}$	0.38 (43)	1.00

In any linear theory, the fiber's motion can be written as a linear superposition of spatial eigen-modes. The amplitudes multiplying each spatial eigen-mode are time-dependent; each has its own thermal dynamics equal to that of a massively overdamped harmonic oscillator at finite temperature, and each mode's amplitude has its own characteristic relaxation time.

It is well known from spectral theory that the spectrum of eigen-values, here the relaxation rates of eigen-modes, is bounded from below and discrete when the system is *compact* in the mathematical sense of that word. DNA of finite extent is compact. This means that there is a slowest relaxation time. Since DNA is also one-dimensional, we expect the discrete spectral values to be well separated; i.e. we expect the next-slowest relaxation time to be many times faster than the slowest one. Thus, with the limited time-resolution of time-lapse recordings, we may only resolve the motion of the slowest or a few of the slowest spatial eigen-modes even if we could see and track the motion of the DNA. All higher modes show up in our measurements as uncorrelated (white) noise, and simply add to the localization error that we observe.

Thus the task of describing the DNA's invisible motion is simplified by realizing the manner in which this motion affects our data. Spectral theory facilitates this realization and formulates its result.

For illustration of spectral theory, we have analyzed a linear mean-field model for the DNA's motion that is analytically tractable and discuss two other linear, exactly solvable models for DNA as well (Supplementary Section S1). Their solutions show that the DNA's motion is composed of a sum of independent, orthogonal modes, each having its own characteristic relaxation time and its own characteristic mean amplitude. They also confirm that contributions from higher modes quickly become negligible: our mean-field model for DNA stretched by a shear flow shows that both the relaxation time (Supplementary Figure S1I and Supplementary Eq. (S31)) and the mean amplitude (Supplementary Figure S1J and Supplementary Eq. (S5)) of the second mode are six times lower than those of the lowest mode; our two other models, DNA stretched by a plug flow and by pulling at its ends, have second modes that have relaxation times and amplitudes that are 5.3 and 4 times lower than the respective values for their respective lowest modes (Supplementary Figures S1A, B, E and F).

Returning to our experiment, the body of *a priori* knowledge we have invoked here has reduced our task to a point where we can deduce the DNA's motion from the information we have at hand, when supplemented with yet another piece of *a priori* knowledge: two different length-scales occur in the experiment. The large ratio between them invites a highly simplifying approximation.

Separation of length scales. A hOGG1 protein diffuses only for a while (its residence time) on a flow-stretched DNA molecule (Figure 1B–E). Thus, its diffusion length while in residence is small (of the order of 1 μm or smaller) compared to the length of the DNA (16 μm). The DNA is stretched out in the x-direction and its dominant fluctuations are large-wave-length according to spectral theory. Consequently, the protein's diffusion on the DNA changes the protein's y-coordinate insignificantly compared to changes in the protein's y-coordinate due to the DNA's motion in the lab. Thus, the y-coordinate of a protein's trajectory is, effectively, the trajectory of a fixed, physical point on the DNA, e.g., the mean position, $\overline{s}$, of the protein, where s is distance along the DNA measured from its tethered end (e.g. in number of base-pairs). The y-coordinate's periodogram $\hat{P}_{y,f}$ (Figure 1F) bears this out, as discussed in the Results section below.

Tethered and unstretchable, the DNA's unobservable longitudinal motion follows from its observable transversal motion. Since the DNA is unstretchable, its longitudinal (x-direction) fluctuations are completely determined by its upstream fluctuations in the transverse y- and z- directions through the relation $x' = \sqrt{1-(y')^2-(z')^2}$, where the prime on a variable denotes the derivative with respect to s. Since the DNA is taut, this expression may be approximated by $x' \approx 1 - (y')^2/2 - (z')^2/2$. This shows that longitudinal

(x-direction) DNA fluctuations are determined by transverse fluctuations through a quadratic dependence. This implies that the power spectrum of $x(\bar{s})$ is a Lorentzian with a corner frequency that is twice that of the power spectrum of $y(\bar{s})$, $f_{c,x} = 2f_c$. Here we have assumed that the statistics of DNA motion is the same in the two transverse directions, y and z (Supplementary Sections S1 and S2).

We cannot observe $z(\bar{s})$, which is why we make this assumption, so we cannot verify the assumption directly. It is a reasonable assumption if the presence of the cover slip mainly affects the average value of $z(\bar{s})$ with little effect on fluctuations in $z(\bar{s})$ about this shifted average. Our mean-field model for the DNA's motion indicates that this is the case (Supplementary Section S1C), and data are consistent with this assumption (Figure 1G). Our protocol for data analysis exploits this to increase the precision of our statistical description of longitudinal DNA fluctuations based on the observed transverse motion.

Statistical independence of the three contributions to changes in the protein's x*-coordinate.* Finally, x-displacements of the protein, Δx_n, as measured in lab coordinates, are a sum of three statistically independent contributions: one from the protein's actual diffusion along the DNA, another from localization errors, and a third from motion of the DNA in the x-direction. Localization errors are by nature statistically independent of the two actual, physical displacements. The latter two are independent of each other because the separation of length scales makes the observed diffusion of the protein on the DNA independent of the DNA's motion and vice versa.

Consequently, the covariance $\langle \Delta x_m \Delta x_n \rangle$ of the measured displacements of a protein's x-coordinate is the sum of three contributions, one from each of those three statistically independent terms: the first, proportional to the protein's diffusion coefficient D, which we want to determine, another parameterized by the variance of localization errors, σ^2, which we will have to determine simultaneously with D if it is not known otherwise, and a third, the autocovariance of DNA fluctuations, $\mathcal{C}_{\Delta x}$, which we know, up to a prefactor, from substrate motion in the y-direction. Supplementary Equations (S39)–(S41) express this mathematically.

Supplementary Equations (S39)–(S40) show that the displacements of a protein's experimentally measured x-coordinate directly give estimates for D and σ^2 as solutions to two coupled linear equations: The necessary input is the covariance $\langle \Delta x_m \Delta x_n \rangle$ of the measured displacements and the covariances $\mathcal{C}_{\Delta x}$, which also are determined from experimental data.

Protocol for data analysis

In actual practice, we determine the values of D, σ^2, and the parameters of $\mathcal{C}_{\Delta x}$ simultaneously, using Maximum Likelihood Estimation (MLE) applied simultaneously to the periodogram $\hat{P}_{\Delta x,f}$ of the measured displacements along the x-coordinate, and the periodogram $\hat{P}_{y,f}$ of the y-coordinate of the recorded position (Supplementary Eqs. (S44) and (S45)).

This section walks through our protocol, motivating each step, while pointing to relevant sections of the Supplementary Information for details. Points 1–5 below describe our procedure for analyzing individual trajectories. Points 6–9 describe our procedure for investigating the results of this analysis and for comparing it to theoretical models for hOGG1's dynamics on DNA.

1. We discarded time-series that displayed anomalies or were too short to be tested reliably for anomalies (Supplementary Section S3A).
   Motivation: To ensure that the time-series we analyzed correspond to mobile DNA-bound proteins that experienced constant conditions along the DNA.
2. For each time-series longer than $N = 25$, we used MLE to estimate the protein's diffusion coefficient D, the variance σ^2 of localization errors, and the parameters $\boldsymbol{\phi} = (f_c, K_x, K_y)$ characterizing DNA fluctuations locally in 1 μm long patches along the DNA with the phenomenological model of substrate motion developed in (25) and sketched above (Supplementary Section S3C).
   Motivation: For long time-series, MLE is unbiased and hence can reliably estimate D and substrate motion simultaneously. So we do that for the relatively few long time-series we have in order to characterize the motion of the DNA locally, where each long time-series visited. We have enough long time-series to characterize the local motion of the entire stretch of DNA visited by the short time-series we analyse below. With the motion of the DNA thus described, we can correct the CVE for DNA motion below.
3. For time-series shorter than $N = 25$, we used the CVE corrected for substrate motion, using weighted averages of the parameters $\boldsymbol{\phi}$ that had been estimated using MLE for the several long time-series in the same patch (Supplementary Section S3D).
   Motivation: For time-series shorter than $N = 25$, the MLE may be significantly biased (Supplementary Figure S5). Thus, for such short time-series, recorded in a patch on the substrate that already had been characterized with the help of several long time-series, we used the CVE corrected for substrate motion.
4. For time-series longer than $N = 25$ in the same patch, we also estimated diffusion coefficients using the uncorrected CVE.
   Motivation: We compared these estimates to those obtained with MLE; their difference is an experimental estimate of the bias b_D of the CVE, which we compare to our theoretical prediction for this bias given by Equation (1) below. This step provides a consistency check of our procedure (Figure 2F and Supplementary Section S3E).
5. Diffusion coefficients for diffusion along the DNA were obtained by multiplying all estimates with the tortuosity of the DNA, which was constant, equal to 1.2, across the segment of the flow-stretched DNA that was used in our analysis (Supplementary Figure S1J and Supplementary Sections S2 and S3).
   Motivation: Estimated diffusion coefficients describe the projection on the x-axis of diffusion along the DNA. Since the DNA is stretched to 90% along the x-axis, diffusion coefficients for the proteins' motion along the DNA contour are $1/(0.9)^2 \approx 1.2$ times higher than those we measure for the motion projected along the x-axis.

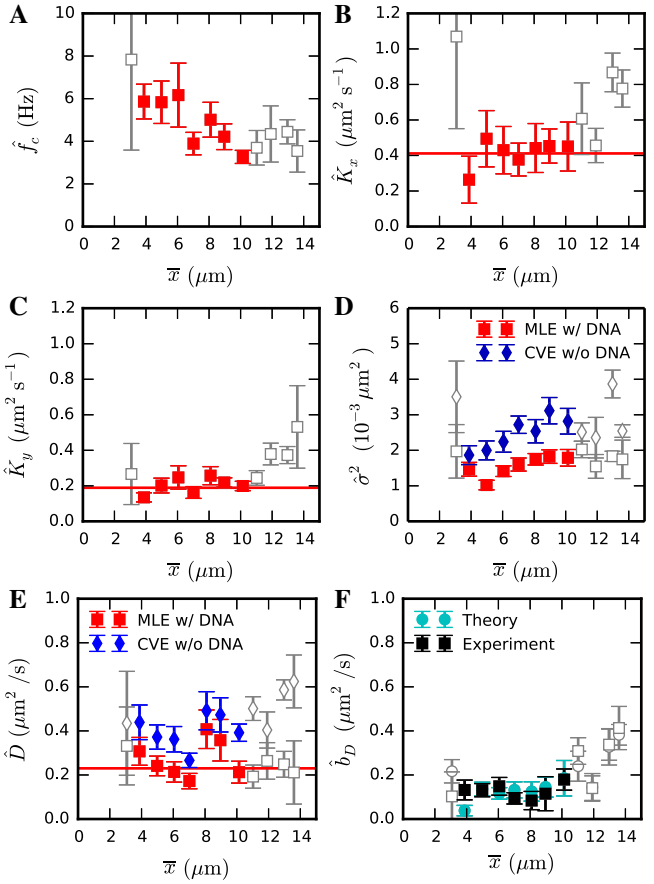

Figure 2. Estimated parameter values as functions of the protein's position $\overline{x}$ on DNA. Results for molecule no. 5 (pH 7.5 and [NaCl] = 0.01 M). Averages are weighted means and error bars are weighted s.e.m., except for estimates of diffusion coefficients; their averages are simple mean values. Data close to the DNA ends (open gray symbols) are excluded in the following analysis to avoid bias (see discussion in Supplementary Section S3A). (**A**) MLE of the corner frequency f_c of transverse (y-direction) DNA motion. Measured values vary by a factor two between the tethered and the free end. (The hypothesis that f_c is constant is refuted with $P = 6 \times 10^{-7}$.) The faster dynamics near the tethered end is due to larger tension in the DNA there. (The corner frequency of the DNA's longitudinal motion is twice the corner frequency of its transverse motion.) (**B**) The diffusivity $K_x(\overline{s})$ of longitudinal DNA fluctuations is constant along the DNA, corresponding to increasing amplitude in the downstream direction in consequence of decreasing tension; $\overline{K}_x = 0.41 \pm 0.03\,\mu$m/s ($P = 0.93$). (**C**) The diffusivity $K_y(\overline{s})$ of transverse DNA fluctuations is also constant along the DNA; $\overline{K}_y = 0.19 \pm 0.02\,\mu\text{m}^2/\text{s}$ ($P = 0.19$). (**D**) The CVE overestimates σ^2, the variance of localization errors, by almost a factor two because it does not account for DNA motion. The MLE of σ^2 increases slightly towards the free end. The assumption that it is constant has negligible support: $P = 0.003$. (**E**) Estimated diffusion coefficients. The uncorrected CVE overestimates diffusion coefficients significantly, because it does not account for DNA motion. The MLE does not, and shows that diffusion coefficients do not depend on the protein's position on the DNA ($P = 0.10$). (**F**) Experimental estimates of the bias b_D of the CVE, calculated as $\hat{D}_{\text{cve}} - \hat{D}_{\text{mle}}$, and theoretical estimates, calculated from Equation (1) using weighted means of estimates of f_c and K_x (Materials and Methods and Supplementary Section S3D). The theory generally agrees excellently with experiments. The bias increases near the DNA's free end, where DNA fluctuations are larger and slower.

Note that although this tortuosity factor is a source of error on estimated values for the diffusion coefficients, it does not affect their ratios. So it does not affect any of the conclusions that we draw from our analysis, such as the statistical support for proposed models. Nor does it affect the relative importance of bias due to substrate fluctuations. Optimally, the tortuosity could be measured directly, e.g. by using optical tweezers to stretch the DNA.

6. We plotted the estimated diffusion coefficients of proteins on DNA against the measured residence time of proteins, as scatter plots (Figure 3A–E) and as binned on the residence-time axis (Figure 3F–J). Histograms of measured residence times were also plotted to investigate the distribution of residence times (Figure 3K–O).
   Motivation: These plots were used, with standard statistical testing as described below, to compare experimental data with theoretical models for the underlying kinetics (Supplementary Section S4).
7. Maximum likelihood estimates of parameters in the two-state model presented in Figure 4B–G (Supplementary Section S4C) were obtained by a combined fit to the block-averaged estimates of diffusion coefficients (Figure 3F–J) and to the full distribution of residence times (Figure 3K–O).
   Motivation: The two-state model predicts correlated values for a protein's residence time and the protein's residence-time-averaged diffusion coefficient. Thus, combined simultaneous fitting of model predictions to experimental residence times as well as experimental diffusion coefficients (which are time-averages, each an average over one residency) is the logical and statistically optimal way to fit to these data.
8. We also fitted the two-state model directly to diffusion coefficient estimates obtained using the uncorrected CVE (Supplementary Figure S2 and Table S4).
   Motivation: This is another consistency test of our results: it leads to the same two-state model for hOGG1's kinetics, albeit with biased estimates of the diffusivities (Supplementary Table S4). It confirms that the observed correlation between diffusion coefficients and residence times is not a spurious effect induced by a possible bias in the MLE and the corrected CVE due to finite statistics.
9. The next four sections, two on tests and two on Monte Carlo simulations, are peripheral parts of our protocol.
   Motivation: One can reproduce our results without the next four sections, but cannot reproduce our confidence in our results without them. The confidence rests on our understanding of the steps described above as well as on the reflections over self-consistency, precision, and accuracy that the next four sections describe.

Testing for drift of DNA-attached proteins with surrounding flow

The estimators for the diffusion coefficient were derived under the assumption of negligible drift. To test for drift of the proteins along the DNA in response to the drag-force from the surrounding flow, we calculated their mean displacement per time-lapse (Supplementary Section S3A). There was no discernible drift, neither locally on the DNA (Sup-

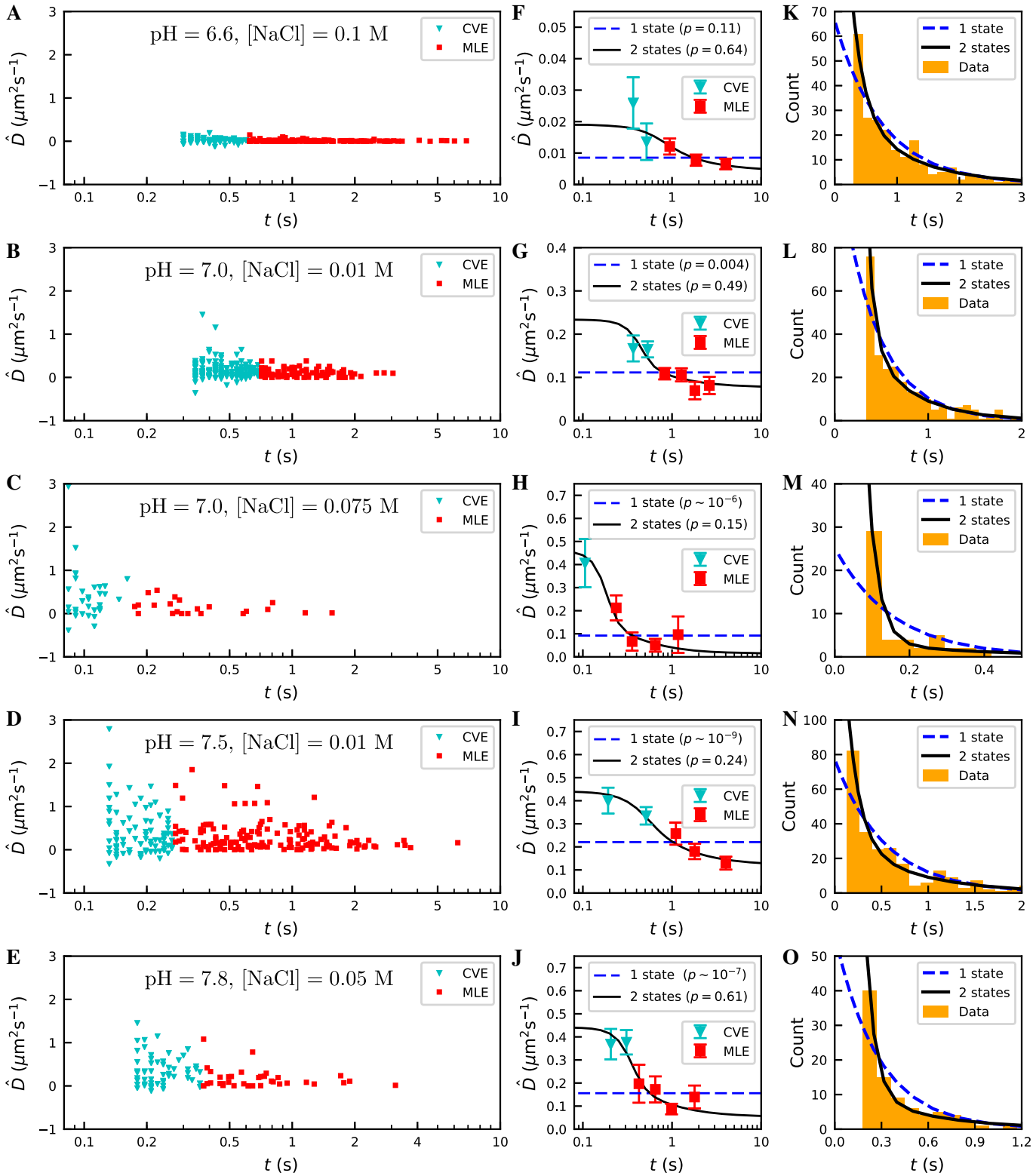

Figure 3. Estimated diffusion coefficients and residence times for proteins on DNA. Each row shows results for the experimental conditions listed in the top of the left panels. (**A–E**) Estimated diffusion coefficients $\hat{D}$ versus residence times for proteins on DNA. (**F–J**) Block averages of diffusion coefficient estimates in A–E binned on the time axis. (**A**, **F**, **K**) DNA molecule no. 1 (pH 6.6 and [NaCl] = 0.1 M); (**B**, **G**, **L**) DNA molecules nos. 2 and 3 (pH 7.0 and [NaCl] = 0.01 M); (**C**, **H**, **M**) DNA molecule no. 4 (pH 7.0 and [NaCl] = 0.075 M); (**D**, **I**, **N**) DNA molecule no. 5 (pH 7.5 and [NaCl] = 0.01 M); (**E**, **J**, **O**) DNA molecule no. 6 (pH 7.8 and [NaCl] = 0.05 M). A clear dependence on the residence time is seen in the measured diffusion coefficients, quite contrary to what one finds for a simple diffusion process. (**K–O**) Distribution of protein residence times on DNA. The distributions are not simple exponentials, so the rate of detachment of protein from DNA is not constant but decreases with the time bound. Dashed blue lines and full black lines in F–O mark combined ML fits of the one- and two-state models, respectively, to data in the second and third columns (*P*-values are given in legends and in Table 1, and estimated parameter values are given in Table 2).

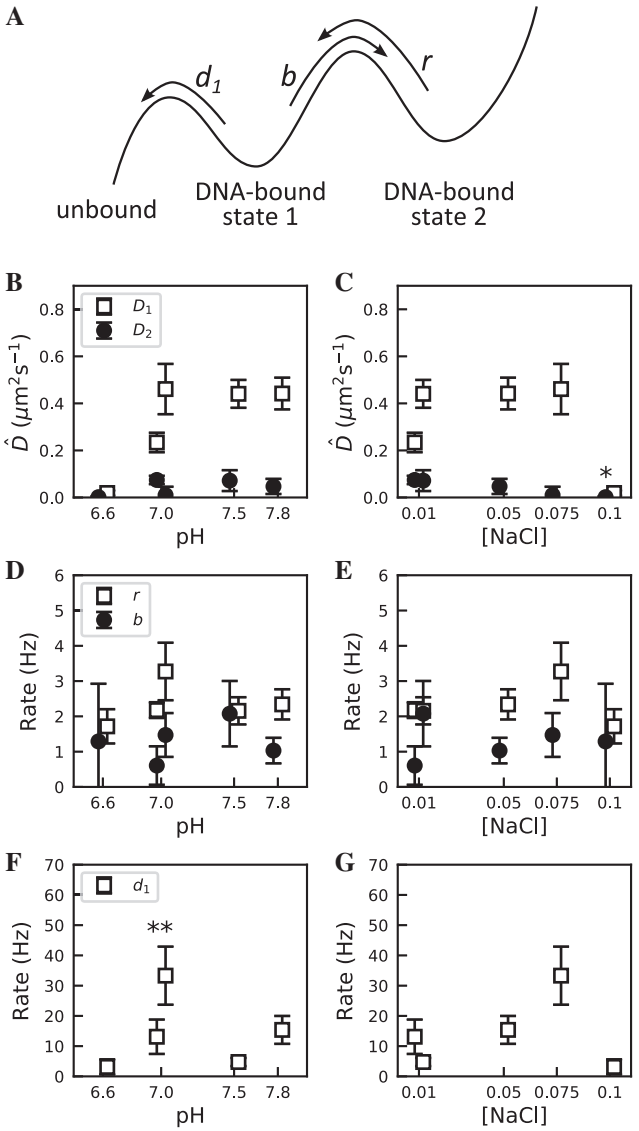

Figure 4. Two-state model for hOGG1's diffusion on DNA. (**A**) Schematic free energy landscape for the two-state model for diffusion on DNA. Maximum likelihood fits of this two-state model to data shown in Figure 3 give *P*-values which support the two-state model and Akaike weights which overwhelmingly favor the two-state model over the one-state model (Table 1). (**B**–**G**) Maximum likelihood estimates of parameters in the model as a function of pH-value (**B**, **D**, **F**) and salt concentration (**C**, **E**, **G**). (**B**, **C**) Diffusion coefficients in the loosely bound state, D_1, and in the tightly bound state, D_2. *The low value of D_1 found for [NaCl] = 0.1 M is explained by the low value of the pH (pH 6.6) for this DNA molecule. (**D**, **E**) Rates of transitions, b from the loosely to the tightly bound state and r for returning. These rates for changing between states do not show significant dependence on pH (D) or salt concentration (E). (**F**, **G**) Detachment rate d_1 from the loosely bound state. **Salt concentration is varied 8-fold between the two measurements at pH 7.0, which explains the difference in detachment rates.

plementary Figure S4D) nor on average (Supplementary Table S5).

Statistical testing

Theoretical models of the distribution of residence times were tested against experimental data using Pearson's chi-square goodness-of-fit test with each bin containing at least five observed counts, and the number of bins $n \geq 7$. For other tests we used a chi-squared test for variance to test whether averages of estimated parameters agreed with the expected values. Estimates were divided into m bins, and averages were calculated for each bin. The number of estimates in each bin was on average eleven or more to assure that averages were approximately Gaussian distributed. The one- and two-state models for hOGG1 on DNA predict both a distribution of protein residence times and a mean diffusion coefficient for a given residence time. They were fitted to both types of data in a combined fit. Consequently, the goodness-of-fit of each model was determined from a corresponding total chi-squared value: the sum of a Pearson's chi-squared value for its agreement with the distribution of residence times, and a variance-based chi-squared value for its agreement with measured diffusion coefficients as function of residence time. The total degrees of freedom is $n + m - p - 1$ for each model, where p is the number of parameters fitted.

Monte Carlo simulations of diffusion on a crowded DNA strand

In (27), experimental conditions were chosen that result in a low average density of proteins on the DNA (fewer than three hOGG1 molecules at a time). However, even if the TIRF setup limits bleaching of proteins before they bind to DNA, bleached or unlabeled proteins could also bind to DNA, so the number of proteins bound to DNA might have been higher than what we observed. Thus a protein may occasionally have prevented another protein from diffusing in a given direction. We therefore Monte Carlo simulated diffusing proteins at concentrations up to twenty times larger than in experiments. At these concentrations, estimated diffusion coefficients differ only negligibly from their values at zero concentration and can account neither for the observed dispersion in diffusion coefficients nor the correlation between residence times and diffusion coefficients that we see in experiments (Supplementary Table S6 and Supplementary Section S4A). It is easy to see why: Protein encounters are rare, and the rest of the time diffusion goes on as at zero concentration. This is in agreement with a recent study of the effects of crowding on single-file diffusion (29). Those authors introduced the order parameter $\tau_p = c^2 D \Delta t$, where c is the concentration of diffusers on the DNA. They concluded that for $\tau_p \ll 1$, the effect of crowding is negligible; for our experimental conditions we find $\tau_p \sim 10^{-5}$.

Monte Carlo simulations of random walkers in a quenched potential landscape

DNA is not a perfectly homogeneous medium since AT and CG base-pairs differ physically. To investigate whether this could cause our observed nontrivial distributions of residence times and diffusion coefficients, we Monte Carlo simulated the model proposed in (24), which assumes that proteins perform random walks in a quenched energy-landscape with uncorrelated and Gaussian distributed binding energies (Supplementary Figure S6A and Supplementary Section S4B). The time-lapse of our experimental

measurements is much longer than the molecular time-scale set by the mean time between consecutive steps in the random walk of a protein. Consequently, a protein diffusing on DNA effectively averages over the DNA's quenched energy-landscape during a single time-lapse. Thus, the Monte Carlo simulation shows no correlations between diffusion coefficients and residence times (Supplementary Figure S6C), and its distribution of residence times is a simple exponential (Supplementary Figure S6B), which means that detachment is a simple Poisson process with a constant rate of detachment. In summary, the model proposed in (24) predicts that the simple one-state model should explain observations. Since it does not, we have eliminated the model proposed in (24) as an explanation.

RESULTS

Crucial experimental result for the DNA's motion

The y-coordinate's periodogram $\hat{P}_{y,f}$ (30) (Figure 1F) fits a Lorentzian plus a constant (25). (This Lorentzian is aliased and low-pass filtered due to limited time-resolution and motion blur, respectively (25), Supplementary Eq. (S44).) This fit means that the y-motion of the tracked point on the DNA is that of a massively over-damped harmonic oscillator at finite temperature (31,32) tracked with a white-noise localization error. The only physical explanation of this observation is that locally on the DNA, at the segment whose y-coordinate was tracked, only a single mode contributes discernibly to the DNA's recorded motion, in agreement with spectral theory (Materials and Methods).

This is an extremely useful result. The DNA has very many degrees of freedom, but now we found phenomenologically that we can describe the DNA's motion locally with a *single* degree of freedom which, to boot, displays the simplest possible dynamics.

The mode with the slowest relaxation time is the dominant contributor to the motion of the DNA at most positions along the DNA. The other modes contribute, but with corner frequencies (= 3 dB frequencies) that increase and amplitudes that decrease so rapidly with mode-number that already the second mode is indistinguishable from white noise in our data: The Nyquist frequency $f_{\mathrm{Nyq}} = 1/(2\Delta t)$, where Δt is the time-lapse between measurements, is the maximal frequency that we can resolve in our time-lapse recorded data. It is 45 Hz in Figure 1F. Our mean field approximation makes the relaxation time of the second mode six times shorter than the relaxation time of the first mode (Supplementary Figure S1I). In Figure 1F, that factor six places the corner frequency of the second mode at 40 Hz, which is so close to the Nyquist frequency that our limited statistics cannot distinguish it from white noise, and hence also cannot distinguish higher modes. This, we believe, is why we see only the Lorentzian of a single mode, the first mode, in Figure 1F. The corner frequency of this Lorentzian is $f_c(\bar{s}) = 6.7$ Hz in Figure 1F, which is much smaller than the Nyquist frequency.

We consequently use this lowest mode to characterize the DNA's local transverse fluctuations. We know how the DNA's longitudinal motion follows from the transverse motion that we now have characterized (Materials and Methods and Supplementary Section S2). Thus, we now can separate the DNA's longitudinal fluctuations from the observed longitudinal motion of the protein, as described in the Materials and Methods section.

Characterizing the diffusion of hOGG1 on DNA

The key steps in our data analysis are outlined in Figure 1A and in the Materials and Methods section. The experimental setup and data samples are shown in Figure 1B–E. Power spectral analysis of the trajectories reveals that the correlation time τ of transverse DNA fluctuations is longer than the time-lapse between frames Δt (Figure 1F, G and Figure 2A)—compare $\tau = 1/(2\pi f_c) = 25$–50 ms to $\Delta t = 11$ ms. Diffusion coefficients (Figure 2E), localization error (Figure 2D), and DNA motion parameters (Figure 2A–C) are Maximum Likelihood estimated from time-series longer than $N = 25$ (Materials and Methods and Supplementary Section S3C). For time-series shorter than $N = 25$, estimates of parameters of DNA motion from long time-series are used to calculate and correct the bias of the covariance-based estimator (CVE) to obtain unbiased estimates of diffusion coefficients (Materials and Methods and Supplementary Section S3D).

Comparison between diffusion coefficients estimated with MLE and the uncorrected CVE shows that DNA motion causes a bias of up to $0.4\,\mu\mathrm{m}^2\ \mathrm{s}^{-1}$ (Figure 2E–F, and Materials and Methods and Supplementary Table S1), depending on the protein's position on the DNA and the time-lapse according to

$$b_D(\bar{s}) = \left(\frac{1 - e^{-4\pi f_c(\bar{s})\Delta t}}{4\pi f_c(\bar{s})\Delta t}\right)^3 K_x(\bar{s}), \qquad (1)$$

where $K_x(\bar{s})$ parametrizes the amplitude of longitudinal fluctuations of the point $\bar{s}$ on the DNA (25). Equation (1) shows that the bias caused by substrate fluctuations depends linearly on K_x, with a proportionality coefficient that is at most equal to unity (with equality when the relaxation time of longitudinal substrate fluctuations, $1/(4\pi f_c)$, is much smaller than Δt) and decreases with $f_c\Delta t$ as $1 - 6\pi f_c\Delta t$ when $f_c\Delta t$ is small and as $(4\pi f_c\Delta t)^{-3}$ when $f_c\Delta t$ is large. (For the trajectory analyzed in Figure 1F, G, the estimated bias is $\hat{b}_D(\bar{s}) \approx 0.3\,K_x(\bar{s}) \approx 0.2\,\mu\mathrm{m}^2\mathrm{s}^{-1}$, as compared to an estimated diffusion coefficient of $\hat{D} = 0.06\,\mu\mathrm{m}^2\mathrm{s}^{-1}$.) We measure diffusion coefficients in the range 0.01–$0.5\,\mu\mathrm{m}^2\ \mathrm{s}^{-1}$ (Figure 3F–J and $\overline{D}$ in Table 2), so the bias could, in the worst case, be many times larger than the actual value of the diffusion coefficient.

The bias of the uncorrected CVE does not depend on the proteins' residence time on DNA (Supplementary Figure S2 and (25)), and (unbiased) estimated diffusion coefficients do not depend on the proteins' position on the DNA (Figure 2E). However, the proteins' estimated diffusion coefficients show a much higher spread around their mean than what would be expected from pure statistical error, i.e. they are not consistent with the proteins having a single unique diffusion coefficient (Supplementary Table S2). The diffusion coefficients furthermore depend on the proteins' residence time on DNA (Figure 3A–J). This result invites close inspection of the residence times themselves. They are appealingly

Table 2. Estimated parameter values of the two-state model. Column 1: Identity of DNA molecule(s). Column 2: pH-value of buffer. Column 3: Salt concentration of buffer, [NaCl]. Column 4: Dissociation rate from the loosely bound state, d_1. Column 5: Rate for switching from the loosely to the tightly bound state, b. Column 6: Rate for switching from the tightly to the loosely bound state, r. Column 7: Diffusion coefficient in the loosely bound state, D_1. Column 8: Diffusion coefficient in the tightly bound state, D_2. Column 9: Mean fraction of time proteins spend in the loosely bound state, f_1. Column 10: Effective mean dissociation rate from DNA, $\overline{k_{\text{off}}}$. Column 11: Mean observed diffusivity $\overline{D}$. Measured values of $\overline{D}$ for DNA molecules 2, 3, 4, and 5 are smaller than those reported for corresponding samples in (27). The exclusion of trajectories shorter than $N_{\text{min}} = 12$ points that show higher average diffusivity, compared to $N_{\text{min}} = 5$ in (27), as well as correction for substrate fluctuations, account for this difference.

1	2	3	4	5	6	7	8	9	10	11
DNA no.	pH	[NaCl] (mM)	d_1 (Hz)	b (Hz)	r (Hz)	D_1 ($\mu m^2\ s^{-1}$)	D_2 ($\mu m^2\ s^{-1}$)	f_1	$\overline{k_{\text{off}}}$ (Hz)	$\overline{D}$ ($\mu m^2\ s^{-1}$)
1	6.6	100	3 ± 3	1.3 ± 1.6	1.7 ± 0.5	0.02 ± 0.01	0.001 ± 0.004	0.47 ± 0.23	2 ± 2	0.009 ± 0.002
2 & 3	7.0	10	13 ± 6	0.6 ± 0.5	2.2 ± 0.2	0.23 ± 0.04	0.07 ± 0.02	0.73 ± 0.17	10 ± 5	0.11 ± 0.01
4	7.0	75	33 ± 10	1.4 ± 0.6	3.3 ± 0.8	0.46 ± 0.11	0.01 ± 0.04	0.65 ± 0.11	23 ± 7	0.09 ± 0.04
5	7.5	10	5 ± 1	2.1 ± 0.9	2.2 ± 0.4	0.44 ± 0.06	0.07 ± 0.04	0.42 ± 0.09	2 ± 1	0.22 ± 0.05
6	7.8	50	15 ± 5	1.0 ± 0.4	2.3 ± 0.4	0.44 ± 0.07	0.05 ± 0.03	0.64 ± 0.09	11 ± 2	0.16 ± 0.05

simple data, as each residence time is just measured with a clock. Nevertheless, the distributions of the protein residence times are non-trivial: they do not fit the simple exponential distributions that would result from a fixed rate of detachment (Figure 3K–O). hOGG1 proteins on DNA must have more than one state of attachment according to this distribution of residence times.

The absence of drift in experiments (Materials and Methods and Supplementary Table S5) implies that the non-trivial distribution of hOGG1 residence times and their correlation with measured diffusion coefficients cannot be explained by intermittent excursions of hOGG1 into bulk, so-called 'hopping' (33) (Supplementary Section S4D). Monte Carlo simulations of diffusion on 'crowded' DNA indicate that the non-trivial dynamics of hOGG1 cannot be explained by crowding either (Materials and Methods and Supplementary Section S4A). Monte Carlo simulations of diffusion of a one-state protein on DNA with a random potential landscape indicate that such a landscape also cannot explain this non-trivial behavior (Materials and Methods and Supplementary Section S4B).

Two-state model for hOGG1's kinetics when bound to DNA

In order to explain the non-trivial distribution of residence times and their correlation with measured diffusion coefficients, we propose a minimal two-state model for hOGG1 on DNA: proteins bind to DNA in a loosely bound state (State 1); they switch stochastically to a more tightly bound state (State 2) with rate b and return to the loosely bound state with rate r; proteins detach from the loosely bound state with rate d_1; and proteins have diffusion coefficient D_1 and D_2 in the loosely and tightly bound state, respectively (Figure 4A). We derive analytical expressions for the distribution of protein residence times on DNA and for the mean value of a protein's diffusion coefficient on DNA as a function of its residence time (Supplementary Section S4C). This allows us to directly fit parameters of the two-state model to experimental data using maximum likelihood estimation (estimated parameter values are listed in Table 2). The model shows excellent agreement with experimental data (Figure 3F–O and Table 1). Proteins diffuse much faster in the loosely bound state than in the tightly bound state, and diffusion coefficients in the loosely bound state depend highly on pH (Figure 4B), while diffusivity does not depend on salt concentration (Figure 4C). Taken with the observation that the diffusivity of the H270A hOGG1 mutant protein is pH-insensitive, this suggests that hOGG1's His270 residue is important for diffusion in the loosely bound state (27). Proteins spend slightly more time loosely bound than tightly bound (Figure 4D, E and Table 2), and they make only few transitions between states during typical residence times on DNA (compare b and r in Figure 4D, E to d_1 in Figure 4F, G).

DISCUSSION

At the resolution available experimentally, the thermal motion of a stretched DNA molecule is accurately described locally as a single thermally driven overdamped harmonic oscillator. This simple theory works because higher order modes of the DNA's motion are too fast and too small in amplitude to be resolved. Using this theory, we characterized the fluctuations of the DNA strands from the longer time-series of proteins diffusing on them. This allowed us to predict the bias caused by DNA fluctuations in covariance-based estimates of diffusion coefficients. With this bias removed, we estimated diffusion coefficients with accuracy and optimal precision for single hOGG1 repair proteins diffusing on λ DNA even from short time-series. Our increased precision revealed a two-state kinetics in hOGG1's diffusion on DNA, which was missed by the cruder but common method based on the ensemble-averaged mean squared displacement, and allowed us to accurately estimate the proteins' diffusivity in each state as well as the rates of switching between states.

A two-state kinetics was earlier proposed to explain the motion of Proliferating Cell Nuclear Antigen (PCNA) on DNA (8). This study, however, relied on (i) PCNA undergoing rotation-coupled sliding in one state and rotation-uncoupled linear diffusion in another state, where the protein topologically entraps the substrate, (ii) elaborate experiments using labels of different sizes, and (iii) inference of the existence of the two states of motion from indirect evidence. Another recent study found evidence of two-state kinetics in TALE proteins' diffusion along DNA (9). The study relied on subjective visual identification of periods of low diffusivity from the recorded time-series: a method which (i) can

only be applied when switching between states is so slow that such periods are long enough to be identified visually in a noisy time-series and (ii) provides no rigorous statistical evidence for the inferred model. Importantly, the above approaches did not allow quantitative estimation of the diffusivity in the two states.

Using optimal estimation of individual diffusion coefficients, we have shown the existence of two-state kinetics in the diffusion of hOGG1 on DNA directly from single-molecule tracking data from a single experiment. We have derived analytical expressions for the distribution of the proteins' residence times on DNA and their expected diffusion coefficient as function of their residence time, enabling rigorous statistical testing and model selection, even when state changes are too short to be detected visually, and we have provided accurate estimates of all of the kinetic parameters of the two-state model. A plausible physical mechanism for this two-state kinetics is that hOGG1 undergoes conformal change between a highly mobile state, where it is loosely bound to the DNA, and a more tightly bound, less mobile state that allows greater engagement with the DNA bases to detect oxidative damage. Our analysis showed that salt concentration in solution solely affected hOGG1's detachment rate from its loosely bound state and not rates for switching between states for the range of conditions tested. This indicates that electrostatic screening due to free ions does not affect hOGG1 dynamics in the tightly bound state. Furthermore, pH affected hOGG1's diffusivity in the loosely bound state only. This suggests that the pH-sensitive His270 residue, which was previously shown to play a role in hOGG1's diffusion along DNA (27), is sensitive to solution pH and/or affects sliding in the loosely bound state only, while it is not involved in facilitating changes in the protein's state of binding.

Here, we demonstrated the ability to identify and characterize multiple dynamic states of searching protein molecules from single-molecule tracking data, which will enable refinement of theoretical models explaining how the speed-stability paradox is overcome by hOGG1. The speed-stability paradox is a general physical problem in protein-DNA interaction, and we predict that many proteins may use conformationally and dynamically distinct binding modes to solve the paradox. Resolving dynamic states of DNA-bound proteins from single-molecule diffusion data could be broadly useful as an approach for generating targets for structural studies of each bound state and functional studies characterizing the role of each state in search and/or target binding.

Our results point to sliding as the dominant contributor to hOGG1's 1D motion and search for target sites along DNA on the mesoscopic scale, in agreement with other single-molecule studies (7,27), but here shown to consist of two distinct sliding modes. Conversely, ensemble studies indicate much shorter mean sliding lengths, of the order of several to tens of base-pairs; much shorter binding times, on the micro to millisecond timescale (34–38); and a relatively greater contribution of hopping (34,35,37). Much of these differences can be attributed to differences between salt concentrations used in ensemble studies (typically high, comparable to intracellular levels) and single-molecule studies (typically 5- to 20-fold lower). The archetypical sensitivity of nucleic acid proteins' sequence-nonspecific DNA binding to salt concentration results in a strong shift of the binding equilibrium toward the unbound state and an increased rate of unbinding from the DNA with rising salt concentrations.

Numerical simulations of the Lac repressor's target search have shown that if the protein follows a helical path along the DNA, e.g., the phosphate backbone or the major or minor groove, short-range hopping does not contribute significantly to target search *in vivo* due to geometrical constraints on rebinding (33). This picture agrees with our conclusion that hopping does not contribute to hOGG1's motion along stretched DNA *in vitro* based on drift analysis ((27) and here), largely salt concentration independent diffusion constants ((27) and here), and strong rotation coupling (7), suggesting persistant sliding on DNA along a helical path for hOGG1. 'Molecular clock' ensemble measurements, however, indicate that hOGG1 can overcome obstacles on the DNA phosphate backbone and transfer between strands without dissociating to bulk (34). Groove-focused sliding in the loosely bound state may explain hOGG1's rotation-coupled sliding together with an ability to transfer between strands and negotiate backbone-bound obstacles without fully escaping from DNA.

Dynamic NMR spectroscopy measurements have revealed that non-specifically bound human uracil DNA glycosylase (hUNG) undergoes conformational dynamics in the millisecond range (39). Others have suggested that the millisecond timescale of hUNG's conformational dynamics are tuned to the timescale of DNA breathing (i.e. bubble formation and coalescence) (39,40). This is two orders of magnitude faster than the rates for hOGG1's switching between the sliding states measured here. It would be interesting to perform spectroscopy measurements to see whether the rates of protein conformational changes of non-specifically bound hOGG1 correspond to the rates of switching between states observed here.

We finally conclude that more single-molecule tracking measurements should be done, with hOGG1 on DNA and with other diffusers. New experimental protocols ease large-scale single-molecule data collection (41), and optimized tools of analysis are now available and promise novel insights, as demonstrated here with old, suboptimal data. The increased resolution and much higher throughput will enable rigorous exploration of a larger range of pH and salt concentrations and may help integrate results from single-molecule and ensemble experiments.

Our results indicate that one may increase the precision of estimates many times by using an experimental setup that applies constant tension along the entire substrate; $\sim$1 pN would be preferred. One configuration capable of this is a dual optical trap setup (42). Dual optical traps give precise control of the extension of the substrate and can measure DNA fluctuations directly via fluctuations of the two trapped beads holding the DNA. Moreover, modeling of the DNA's motion is simple in this case; a linear model applies and is fully solvable. This has the very practical consequence that recordings of the motions of the two beads and transverse motion of the DNA measured at various points along the DNA all can be combined to determine the few parameters of the linear model that describes the DNA's motion, as done in (43) for a similar, less simple setup. The

model will thus be over-determined and hence can be falsified, and it enables direct and rigorous analysis of time-series that describe diffusion over length scales that are of the order of the DNA's extent. Also, the local tortuosity (or, equivalently, the degree of stretching) of the DNA can be calculated from the fitted model (43), so absolute values of diffusion constants can be measured with accuracy and precision. With dual traps one can furthermore investigate rigorously how stretching of the substrate affects kinetic parameters, providing an experimental means to verify numerical predictions for the influence of DNA dynamics on sliding (44) and to link single-molecule measurements made on stretched DNA to ensemble measurements made on DNA in free configuration.

Flow-stretch assays like the one studied here and in (41) trade the precision of optical tweezers for a simpler experimental setup and the possibility to record on multiple DNA molecules in parallel, providing much more data. Here we recommend to use a strong flow to stretch the DNA strands in order to drive their fluctuations as fast as possible relative to the time-lapse of recordings, while assuring that this does not lead to drift of the attached proteins. Fast fluctuations will ensure that the phenomenological single-mode theory for the DNA's motion is valid and increases the precision of estimated diffusion coefficients (25). This is all the more important when one wants to choose the time-lapse as small as possible in order optimize the number of recorded positions (45). If the fluctuations are fast enough, they may even be absorbed in the localization error term and do not need to be explicitly accounted for, leading to a simple, and more precise, estimation procedure—this should be checked by plotting the periodograms of the recorded time-series.

Finally, tracking with MINFLUX promises to increase the acquisition rate of recordings substantially without sacrificing localization precision (46). This would facilitate the separation of DNA fluctuations from diffusive motion and improve the precision of inferred model parameters, and would make it possible to probe faster dynamics of single proteins sliding on DNA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The experimental data analyzed here were produced in the lab of Xiaoliang Sunney Xie in the Department of Chemistry and Chemical Biology at Harvard University.

FUNDING

National Institutes of Health [ST32 GM07598-25 to P.C.B.] through the Harvard University's Molecular, Cellular and Chemical Biology training program; Burroughs Wellcome Fund through a Career Award at the Scientific Interface (to P.C.B); Human Frontier Science Program Research [GP0054/2009-C to H.F.]. Funding for open access charge: Human Frontier Science Program; Burroughs Wellcome Fund.

Conflict of interest statement. None declared.

REFERENCES

1. Brameshuber,M. and Schütz,G.J. (2012) Detection and quantification of biomolecular association in living cells using single-molecule microscopy. *Methods Enzymol.*, **505**, 159–186.
2. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1994) *Molecular Biology of the Cell*, 3rd edn. Garland Publishing, Inc., pp. 95–563.
3. van Loenhout,M.T.J., de Grunt,M.V. and Dekker,C. (2012) Dynamics of DNA supercoils. *Science*, **338**, 94–97.
4. Wang,Y.M., Austin,R.H. and Cox,E.C. (2006) Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett.*, **97**, 048302.
5. Granéli,A., Yeykal,C.C., Robertson,R.B. and Greene,E.C. (2006) Long-distance lateral diffusion of human Rad51 on double-stranded DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 1221–1226.
6. Tafvizi,A., Mirny,L.A. and van Oijen,A.M. (2011) Dancing on DNA: kinetic aspects of search processes on DNA. *Chemphyschem*, **12**, 1481–1489.
7. Blainey,P.C., Luo,G., Kou,S.C., Mangel,W.F., Verdine,G.L., Bagchi,B. and Xie,X.S. (2009) Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.*, **16**, 1224–1229.
8. Kochaniak,A.B., Habuchi,S., Loparo,J.J., Chang,D.J., Cimprich,K.A., Walter,J.C. and van Oijen,A.M. (2009) Proliferating cell nuclear antigen uses two distinct modes to move along DNA. *J. Biol. Chem.*, **284**, 17700–17710.
9. Cuculis,L., Abil,Z., Zhao,H. and Schroeder,C.M. (2016) Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.*, **10**, 8277.
10. Helenius,J., Brouhard,G., Kalaidzidis,Y., Diez,S. and Howard,J. (2006) The depolymerizing kinesin MCAK uses lattice diffusion to rapidly target microtubule ends. *Nature*, **441**, 115–119.
11. Rocha,S., Hutchison,J.A., Peneva,K., Herrmann,A., Müllen,K., Skjøt,M., Jørgensen,C.I., Svendsen,A., De Schryver,F.C., Hofkens,J. and Uji-i,H. (2009) Linking phospholipase mobility to activity by single-molecule wide-field microscopy. *Chemphyschem*, **10**, 151–161.
12. Wieser,S. and Schütz,G.J. (2008) Tracking single molecules in the live cell plasma membrane—Do's and Don't's. *Methods*, **46**, 131–140.
13. Vrljic,M., Nishimura,S.Y., Brasselet,S., Moerner,W.E. and McConnell,H.M. (2002) Translational diffusion of individual class II MHC membrane proteins in cells. *Biophys. J.*, **83**, 2681–2692.
14. Sergé,A., Bertaux,N., Rigneault,H. and Marguet,D. (2008) Dynamic multiple-target tracing to probe spatiotemporal cartography of cell membranes. *Nat. Methods*, **5**, 687–694.
15. Renner,M., Domanov,Y., Sandrin,F., Izeddin,I., Bassereau,P. and Triller,A. (2011) Lateral diffusion on tubular membranes: quantification of measurements bias. *PLoS ONE*, **6**, e25731.
16. Sharonov,A., Bandichhor,R., Burgess,K., Petrescu,A.D., Schroeder,F., Kier,A.B. and Hochstrasser,R.M. (2008) Lipid diffusion from single molecules of a labeled protein undergoing dynamic association with giant unilamellar vesicles and supported bilayers. *Langmuir*, **24**, 844–850.
17. Persson,F., Lindén,M., Unoson,C. and Elf,J. (2013) Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods*, **10**, 265–269.
18. Smith,M.B., Karatekin,E., Gohlke,A., Mizuno,H., Watanabe,N. and Vavylonis,D. (2011) Interactive, computer-assisted tracking of speckle trajectories in fluorescence microscopy: application to actin polymerization and membrane fusion. *Biophys. J.*, **101**, 1794–1804.
19. Riggs,A.D., Bourgeois,S. and Cohn,M. (1970) The lac represser-operator interaction: III. Kinetic studies. *J. Mol. Biol.*, **53**, 401–417.
20. Radicella,J.P., Dherin,C., Desmaze,C., Fox,M.S. and Boiteux,S. (1997) Cloning and characterization of hOGG1, a human homolog of the OGG1 gene of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 8010–8015.
21. Bauer,N.C., Corbett,A.H. and Doetsch,P.W. (2015) The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res.*, **43**, 10083–10101.
22. Banerjee,A., Yang,W., Karplus,M. and Verdine,G. (2005) Structure of a repair enzyme interrogating undamaged DNA elucidates recognition of damaged DNA. *Nature*, **434**, 612.

23. Hammar,P., Leroy,P., Mahmutovic,A., Marklund,E., Berg,O. and Elf,J. (2012) The lac repressor displays facilitated diffusion in living cells. *Science*, **336**, 1595–1598.
24. Slutsky,M. and Mirny,L.A. (2004) Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.*, **87**, 4021–4035.
25. Vestergaard,C.L., Blainey,P.C. and Flyvbjerg,H. (2014) Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys. Rev. E*, **89**, 022726.
26. Vestergaard,C.L., Pedersen,J.N., Mortensen,K.I. and Flyvbjerg,H. (2015) Estimation of motility parameters from trajectory data. *Eur. Phys. J. Special Topics*, **224**, 1151–1168.
27. Blainey,P.C., van Oijen,A.M., Banerjee,A., Verdine,G.L. and Xie,X.S. (2006) A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 5752–5757.
28. Cravens,S.L. and Stivers,J.T. (2016) Comparative effects of ions, molecular crowding, and bulk DNA on the damage search mechanisms of hOGG1 and hUNG. *Biochemistry*, **55**, 5230–5242.
29. Koslover,E.F., Díaz de la Rosa,M. and Spakowitz,A.J. (2017) Crowding and hopping in a protein's diffusive transport on DNA. *J. Phys. A: Math. Theor.*, **50**, 074005.
30. An experimental power spectrum is also called a *periodogram*. When it is, *power spectrum* may refer to the periodogram's expected value or a mathematical function that describes this expected value, e.g., a Lorentzian. This is also called the *theoretical power spectrum*. In that vein, periodogram is synonymous with *experimental power spectrum*, and *power spectrum* is short for theoretical power spectrum.
31. Berg-Sørensen,K. and Flyvbjerg,H. (2004) Power spectrum analysis for optical tweezers. *Rev. Sci. Instrum.*, **75**, 594–612.
32. Nørrelykke,S.F. and Flyvbjerg,H. (2011) Harmonic oscillator in heat bath: exact simulation of time-lapse-recorded data, exact analytical benchmark statistics. *Phys. Rev. E*, **83**, 041103.
33. Tabaka,M., Burdzy,K. and Hołyst,R. (2015) Method for the analysis of contribution of sliding and hopping to a facilitated diffusion of DNA-binding protein: Application to in vivo data. *Phys. Rev. E*, **92**, 022721.
34. Rowland,M.M., Schonhoft,J.D., McKibbin,P.L., David,S.S. and Stivers,J.T. (2014) Microscopic mechanism of DNA damage searching by hOGG1. *Nucleic Acids Res.*, **42**, 9295–9303.
35. Schonhoft,J.D. and Stivers,J.T. (2012) Timing facilitated site transfer of an enzyme on DNA. *Nat. Chem. Biol.*, **8**, 205–210.
36. Hedglin,M. and O'Brien,P.J. (2010) Hopping enables a DNA repair glycosylase to search both strands and bypass a bound protein. *ACS Chem. Biol.*, **5**, 427–436.
37. Hedglin,M., Zhang,Y. and O'Brien,P.J. (2014) Probing the DNA structural requirements for facilitated diffusion. *Biochemistry*, **54**, 557–566.
38. Schonhoft,J.D. and Stivers,J.T. (2013) DNA translocation by human uracil DNA glycosylase: the case of single-stranded DNA and clustered uracils. *Biochemistry*, **52**, 2536–2544.
39. Friedman,J.I., Majumdar,A. and Stivers,J.T. (2009) Nontarget DNA binding shapes the dynamic landscape for enzymatic recognition of DNA damage. *Nucleic Acids Res.*, **37**, 3493–3500.
40. Schonhoft,J.D., Kosowicz,J.G. and Stivers,J.T. (2013) DNA translocation by human uracil DNA glycosylase: role of DNA phosphate charge. *Biochemistry*, **52**, 2526–2535.
41. Xiong,K. and Blainey,P.C. (2017) A simple, robust, and high throughput single molecule flow stretching assay implementation for studying transport of molecules along DNA. *J. Viz. Exp.*, **128**, e55923.
42. Harada,Y., Funatsu,T., Murakami,K., Nonoyama,Y., Ishihama,A. and Yanagida,T. (1999) Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys. J.*, **76**, 709–715.
43. Pedersen,J.N., Marie,R., Kristensen,A. and Flyvbjerg,H. (2016) How to determine local stretching and tension in a flow-stretched DNA molecule. *Phys. Rev. E*, **93**, 042405.
44. Mondal,A. and Bhattacherjee,A. (2015) Searching target sites on DNA by proteins: Role of DNA dynamics under confinement. *Nucleic Acids Res.*, **43**, 9176–9186.
45. Vestergaard,C.L. (2016) Optimizing experimental parameters for tracking of diffusing particles. *Phys. Rev. E*, **94**, 022401.
46. Balzarotti,F., Eilers,Y., Gwosch,K.C., Gynnøa,A.H., Westphal,V., Stefani,F.D., Elf,J. and Hell,S.W. (2017) Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science*, **355**, 606–612.
47. Burnham,K.P. and Anderson,D.R. (2002) *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd edn, Springer.