Sequence-dependent sliding kinetics of p53

Leith, Jason S.; Tafvizi, Anahita; Huang, Fang; Uspal, William E.; Doyle, Patrick S.; Fersht, Alan R.; Mirny, Leonid A.; van Oijen, Antonius; Hammes, Gordon G.

Published in:
Proceedings of the National Academy of Sciences of the United States of America

DOI:
10.1073/pnas.1120452109

Publication date:
2012

Supporting Information

Leith et al. 10.1073/pnas.1120452109

SI Text

Materials and Data Acquisition. Because only relative movements of proteins on the contour of DNA of trajectories were recorded in our earlier studies (1), we collected new data for this study that allowed us to obtain information on the position of the protein with respect to the underlying DNA sequence.

Before labeled p53 was introduced to the flow cell, a 0.016% suspension of biotinylated beads were flowed in with a concentration and incubation time such that 2–5 beads appeared in each field of view, following Elenko (2). When the flow cell had been studded with beads, movies were taken of p53 sliding on flow-stretched λ-phage DNA, flowing 100 μL/ min through a flow cell 2 mm wide, 100 μm tall, and 36 mm long p53 sliding buffer consisted of 20 mM HEPES (equilibrated to pH 7.9 with NaOH), 150 mM KCl, 0.5 mM EDTA, 2 mM MgCl2, 0.25 mg/mL BSA, and 2.5 mM DTT. p53 concentrations were between 50 and 150 pM. Fig. 2*B* shows a kymogram of a single p53 molecule sliding on DNA. At the end of the experiment, DNA was visualized with Sytox Orange (Invitrogen) and movies of it taken.

The beads were present at fixed locations for both the protein movies and the DNA movies, which allowed p53 trajectories to be aligned to positions on DNA (*SI Text*, *Data Analysis*) despite stage drift. This alignment additionally required lower concentrations of DNA than in earlier work, because the DNA could not be so dense as to prevent us from assigning protein particles to a distinct DNA molecule. In the previous work, stained DNA illuminated nearly the entire field of view; in the current work, DNA concentration was lowered to approximately 40 DNA molecules per field of view.

A further difference from our earlier work is that, in this study we did not impose artificial minimum trajectory lengths or durations. All trajectories that our tracking scripts identified are included in the present work's analysis, excepting those of particles that we identified as being on the flow cell surface rather than bound to DNA. Our criterion for being stuck to the surface was having an end-to-end displacement less than half a pixel (85 nm). This distance is nearly the standard deviation in the frame-to-frame displacement of the quantum dot (QD) nearest the tether (368 bp, or 100 nm assuming the DNA is 80% stretched), and so a particle bound to DNA but not sliding on the DNA contour would be expected to move this distance within a single frame. The particles eliminated by this cutoff were clearly distinguishable from particles bound to DNA, as can be seen in Fig. 3*B*. The cutoff of 85 nm corresponds to an end-to-end squared displacement of 10^5 bp^2, which is where the first gray circle representing the QD nearest the tether lies. The share of particles with end-to-end squared displacements vanishes as the cutoff is approached from above, making us confident that we did not discard particles bound to DNA but perhaps trapped at a near-cognate site or otherwise immobile.

Because Brownian dynamics simulations (*SI Text*, *Data Analysis*) show no part of the DNA molecule stretched beyond 90% of its contour length, we conclude that the stretching is largely entropic rather than enthalpic—that is, bond lengths and angles and local conformations are negligibly affected by the buffer flow—we assume that the dependence of the stretching in the DNA as a function of position does not appreciably affect the sliding kinetics or binding thermodynamics of p53 to the DNA.

Data Analysis. To map the spatial positions in our movies of protein particles to positions on the contour of DNA, Brownian dynamics simulations of DNA as a tethered polymer in shear flow were performed to determine the degree of compression in the DNA as a function of the distance along the contour from the tether (3). Integrating and inverting this function yields a function that transforms positions in recorded images to positions along the sequence of λ-phage DNA.

For each p53 trajectory mapped to the DNA contour, a drift rate, v, and diffusion coefficient, D, were determined using maximum likelihood estimation (MLE). Assuming that a particle's displacement due to drift is independent of its displacement due to diffusion, and that the particle's displacements are all independent, the MLEs for a particle's v and D in the absence of DNA fluctuations are derived as follows:

$$p(\Delta x; v, D) = \exp\left(\frac{-(\Delta x - v\Delta t)^2}{4D\Delta t}\right)(4\pi D\Delta t)^{-1/2}$$

$$L(\Delta x_1, \ldots, x_n | v, D) = \exp\left(\sum_i^n \frac{-(\Delta x_{i,p} - v\Delta t_i)^2}{4D\Delta t_i}\right)\prod_i^n (4\pi D\Delta t_i)^{-1/2}$$

$$\log L = -\sum_i^n \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{4D\Delta t_i} - \frac{1}{2}\sum_i^n \log(4\pi D\Delta t_i), \quad \textbf{[S1]}$$

where $\Delta x_{i,p}$ is displacement i of the protein on DNA, which takes place over the duration Δt_i. Taking the partial derivative of L with respect to the drift rate, v, and setting the result equal to zero,

$$0 = \frac{\partial \log L}{\partial v} = \sum_i^n \frac{2\Delta x_{i,p}\Delta t_i - 2v\Delta t_i^2}{4D\Delta t_i}$$

$$= \sum_i^n (\Delta x_{i,p} - v\Delta t_i)$$

$$v = \frac{\sum_i^n \Delta x_{i,p}}{\sum_i^n \Delta t_i}. \quad \textbf{[S2]}$$

Here, the index i is over the largest non-overlapping set of $\frac{\Delta x_{i,p}}{\Delta t_i}$, which are the final and initial frames of each trajectory j, so:

$$v = \frac{\sum_j^{\text{all traj.}} x_{j,\text{final}} - x_{j,\text{initial}}}{\sum_j^{\text{all traj.}} t_{j,\text{final}} - t_{j,\text{initial}}}. \quad \textbf{[S3]}$$

We now take the partial derivative with respect to the diffusion coefficient, D, and equate to zero:

$$0 = \frac{\partial \log L}{\partial D} = \sum_i^n \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{4D^2\Delta t_i} - \frac{1}{2}\sum_i^n \frac{1}{D}$$

$$= \sum_i^n \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{\Delta t_i} - 2nD$$

$$D = \frac{1}{2}\frac{1}{n}\sum_i^n \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{\Delta t_i}$$

$$D = \frac{1}{2}\frac{1}{n}\sum_i^n \frac{\Delta x_{i,p}^2 - 2\Delta x_{i,p} v\Delta t_i + v^2\Delta t_i^2}{\Delta t_i}. \quad \textbf{[S4]}$$

The observed displacements, Δx_i, are in fact the sum of displacement from protein diffusion, $\Delta x_{i,p}$, and displacement from DNA fluctuations, $\Delta x_{i,d}$. Substituting $\Delta x_{i,p}$ with $\Delta x_i - \Delta x_{i,d}$ in Eqs. **S2** and **S4**, and substituting $\Delta x_{i,p}^2$ with $\Delta x_i^2 - 2\Delta x_{i,p} - \Delta x_{i,b}^2$ in Eq. **S4**, yields the following:

$$v = \frac{\sum_i^n \Delta x_i - \Delta x_{i,d}}{\sum_i^n \Delta t_i} \quad \textbf{[S5]}$$

and

$$D = \frac{1}{2}\frac{1}{n}\sum_i^n \frac{\Delta x_i^2 - 2\Delta x_{p,i}\Delta x_{d,i} - \Delta x_{d,i}^2 - 2\Delta x_i v\Delta t_i + 2\Delta x_{d,i} v\Delta t_i + v^2\Delta t_i^2}{\Delta t_i}. \quad \textbf{[S6]}$$

Separating the terms under the sum in the expression for v gives

$$v = \frac{\sum_i^n \Delta x_i}{\sum_i^n \Delta t_i} - \frac{\sum_i^n \Delta x_{i,d}}{\sum_i^n \Delta t_i}. \quad \textbf{[S7]}$$

The second term in Eq. **S7** vanishes because the displacements due to DNA fluctuations, $\Delta x_{d,i}$, have mean zero, and so the drift rate is simply that given in Eq. **S2**. In the equation for D **[S6]**, the DNA displacements are likewise independent of the protein displacements, $\Delta x_{p,i}$, and the drift, $v\Delta t_i$, so the sums of the cross terms $2\Delta x_{p,i}\Delta x_{d,i}$ and $2\Delta x_{d,i} v\Delta t$ also go to zero. Eliminating these terms and separating into four remaining sums yields

$$D = \frac{1}{2}\left(\frac{1}{n}\sum_i^n \frac{\Delta x_i^2}{\Delta t_i} - \frac{1}{n}\sum_i^n \frac{\Delta x_{i,d}^2}{\Delta t_i} - \frac{1}{n}\sum_i^n \frac{2\Delta x_i v\Delta t_i}{\Delta t_i} + \frac{1}{n}\sum_i^n \frac{v^2\Delta t_i^2}{\Delta t_i}\right). \quad \textbf{[S8]}$$

This is equivalent to Eq. **2**. The third and fourth terms are known from the estimate of v in Eq. **S2** and from observed Δx_i and Δt_i. The second term in Eq. **S8** is equivalent to

$$\frac{1}{2}\frac{1}{n}\sum_i^n \frac{\Delta x_{i,d}^2}{\Delta t_i} = \frac{1}{2}\frac{1}{n}\sum_{\Delta t} n_{\Delta t}\frac{\langle \Delta x_d^2\rangle}{\Delta t}, \quad \textbf{[S9]}$$

where $n_{\Delta t}$ is the number of displacements with duration Δt in the trajectory, and Δx_d are displacements owing to DNA fluctuation. Trajectories of DNA fluctuations were measured in previous work (4) by examining the trajectories of QDs covalently attached to λ-phage DNA at known positions. The expression in Eq. **S9** is thus the expected contribution of DNA fluctuations to the apparent diffusion of the protein (*SI Text*, *Interpolations of DNA-Fluctuation Variance and Distributions*), the effect of which on our data and estimated diffusion coefficients is shown in Fig. S3.

Once a p53 particle's diffusion coefficient had been determined, the diffusion coefficient was assigned to every midpoint of the particle's trajectory's displacements (Fig. 3*C*, dots). Data from the third of the DNA farthest from the tether was discarded owing to the large amplitude of DNA fluctuations beyond that point. The DNA was divided into segments with a width chosen equal to the mean end-to-end distance of remaining trajectories, approximately 2.9 kb (Fig. 3*C*, dashed lines). The mean of all the diffusion coefficients assigned to positions within each segment was calculated and then compared with the predicted diffusion coefficient based on theoretical energy landscapes. This method is equivalent to calculating the mean of D over all particles contributing to a segment, weighted by the number of displacements each particle contributed.

Significance and Consistency of Experimental Results. Because the observed variation in D_{expt} among segments is not dramatic, we thought it especially important to see whether this variation was significant. If some segments truly have more rugged energy landscapes than others, then we expect lower-D particles to be especially likely to be found in certain segments and higher-D particles especially likely to be found in others. Randomizing the assignment of D to particles would desegregate particles with different diffusivities, and so the difference between segments should decrease. We performed 1,000 such randomizations, and for each pair of segments i and j, we calculated the absolute log-ratio of the two segments' diffusion coefficients, $|\log(D_i/D_j)|$, as determined in *Methods*, *Data analysis* in the main text. Pairs of segments with significantly different D_{expt} should only rarely have randomized absolute log-ratios greater than the absolute log-ratio for data where particles were assigned their observed D rather than the D from another, randomly selected particle. We found that 11 in 36 pairs had unrandomized absolute log-ratios greater than all but 5% of the absolute log-ratios from the shuffled data ($p < \alpha = 0.05$), and that 6 of these pairs had unrandomized absolute log-ratios greater than all but 1% of the absolute log-ratios from the shuffled data ($p < \alpha = .01$). The p-values of the pairs' absolute log-ratios in D are shown in Fig. S1.

Having a D_{expt} that differed significantly from other segments' D_{expt} was imperfectly correlated with having an extreme D_{expt}. Segments 8 and 11, centered 20.4 kb and 29.3 kb from the tether, had nearly identical D_{expt} : 1.322×10^6 bp^2/s and 1.319×10^6 bp^2/s, respectively. Yet segment 8 differed significantly from 3 out of 8 other segments at $\alpha = 0.01$ and from another segment at $\alpha = 0.05$, while segment 11 differed significantly from no other segment. We are unsure of why this is the case; we conjecture that it might owe to segment 8's D_{expt} deriving from 68 particles while segment 11's D_{expt} derives from 54, and so segment 8's D_{expt} when computed with randomly reassigned Ds from other segments will average over more particles' Ds and thus be less likely to take on a value far enough from the mean D of all particles to produce absolute log-ratios with other segments that exceed the absolute log-ratios between segment 8 and other segments without random reassignment. If this is correct, then observing more particles would help us resolve differences between more segment pairs than we currently can.

Although 11 in 36 segment pairs differed significantly in their D_{expt}, most did not. This owes in part to many pairs having predicted landscapes of similar ruggedness and thus similar expected D_{expt}. That these pairs of segments have D_{expt} values that cannot be solidly differentiated accords with our main result: Segments with similar theoretical D/D_0 are expected to have similar D_{expt}. Additionally, there is substantial individuality among the particles, as can be seen in Fig. 3*C*. To assess the effect on the range of Ds among the particles in a segment on the segment's D_{expt}, we employed a bootstrapping procedure. For each segment, we discarded from each particle's set of displacements those displacements that fell outside the segment. If N particles contributed to the segment, we sampled with replacement N times, with sampling probability for a particle proportional to the number

of displacements contributed by the particle. We performed 10,000 such resamples for each segment. The distributions of the resulting D_{expt}s are shown in Fig. S4*A*.

To get a sense of the sufficiency of the statistics we collected, we performed an identical bootstrapping procedure, but sampling from the N particles in a segment only $N/2$ times. This simulates having collected only half the data we did. The greater range in resampled D_{expt} appears as cyan error bars in Fig. 3. That halving the data noticeably widens the bootstrap uncertainty estimate suggests, similarly to what was mentioned above, that more particles would allow us to sharpen our estimates of D_{expt}.

In addition to examining heterogeneity among particles' observed D, we grouped our data according to data-collection session (a morning or an afternoon) in order to assess consistency across time and p53 aliquots. We used the same bootstrapping method described earlier in this section, but with sampling on the level of batches rather than particles. Fig. S4*B* shows the distributions of the bootstrapped D_{expt} estimates. For some segments, the distribution is wider than is the corresponding distribution for bootstrapped D_{expt} based on particle rather than batch resampling, which may mean that the differences between a few batches and average behavior owe to inherent variation in the batch. On the other hand, some segments are underrepresented in some batches, and on the level of a batch may have insufficient averaging, giving an aberrant D for that particular batch-segment pair. Additionally, the smaller number of batches (10) than particles per segment ($60 \sim 70$) causes some of the distributions in Fig. S4*B* to not be bell-shaped.

An additional combinatorial test we performed was to randomize the location of particle trajectories on the DNA. If all the particles were undergoing uniform, position-independent Brownian motion, scrambling positional information would be expected to have little effect on the variation in D_{expt} among segments. Results from five such randomizations are shown in Fig. S5. As can be seen, randomization reduces the variation in D_{expt}.

For every segment, the number of particles, the number of displacements, the estimates of D_{expt} as determined in *Methods*, *Data analysis* in the main text, and the weighted standard deviation of particles' D are shown in Fig. 5*E*. As can be seen, D_{expt} for a notional DNA segment is obtained from averaging over $60 \sim 70$ actual p53–DNA complexes. Many particles contribute to multiple segments on the same DNA molecule (for instance, in Fig. 3*A*). Most DNA molecules contribute only a single trajectory or two distant trajectories, and therefore we cannot assess the extent to which different p53 particles in the same actual segment of a single DNA molecule behave uniformly relative to each other. Despite this, the large number of molecules contributing to each ~2.9-kb division of λ-phage DNA allows us to determine the diffusive properties of p53 particles in the aggregate within those divisions. This segment width is much larger than the error in uncertainty in particle position assignment (~500 bp, from measurements of DNA fluctuations alone; see *SI Text*, *Interpolations of DNA-Fluctuation Variance and Distributions*), and we take this error into account when predicting segments' aggregate D_{expt}.

Alternative Data Analysis. The estimation of particle diffusivity in *SI Text*, *Data Analysis* as well as in the main text treats each p53 particle as if it were undergoing normal diffusion, with a constant diffusion coefficient D. This D aggregates the base-pair–level nonuniformity of the energy landscape experienced by a particle. We also analyzed our data using a treatment that does not attempt to assign particles a diffusion coefficient, but rather calculates the variance in all displacements by all particles in a segment.

In this treatment, we fragmented trajectories wherever they crossed segment boundaries. Then, for each segment, every displacement within a trajectory or fragment was corrected for drift and normalized by dividing it by the square-root of its corresponding duration [**S11**]. A trajectory fragment for which we recorded N frames would have $N-1$ displacements between adjacent frames, $N-2$ displacements between frames with one intervening frame, etc., to 1 displacement with $N-1$ intervening frames. These corrected and normalized displacements were then fit to a Gaussian distribution, and the variance of the fit distribution taken as the estimate of the segment's diffusivity, comparable to twice a diffusion coefficient, $2D$.

$$\text{drift rate} = \frac{\sum_i^{\text{all traj.}} x_{i,\text{final}} - x_{i,\text{initial}}}{\sum_i^{\text{all traj.}} t_{i,\text{final}} - t_{i,\text{initial}}} \qquad \textbf{[S10]}$$

normalized displacements

$$= \left\{ \frac{(x_{i,n} - x_{i,m}) - \text{drift rate} \cdot (t_{i,n} - t_{i,m})}{\sqrt{t_{i,n} - t_{i,m}}} \;:\; n > m; i \text{ over all traj.} \right\} \qquad \textbf{[S11]}$$

$$\left\{ \frac{\Delta x_j}{\sqrt{\Delta t_j}} \right\} \sim N(\mu, 2D), \quad j \text{ over all normalized displacements.} \qquad \textbf{[S12]}$$

The index i is over trajectories and fragments; indices n and m are over frames within a trajectory or fragment; and index j is over all normalized displacements. Each normalized displacement is the sum of a Gaussian random variable owing to diffusion along the DNA with mean zero and variance $2D\Delta t$, and another Gaussian random variable owing to fluctuations of the DNA molecule on which the proteins are diffusing. To account for the increase in apparent diffusion coefficient due to DNA fluctuations we determined the DNA's longitudinal mean squared displacement (MSD) as a function of time window Δt, as discussed in *SI Text*, *Interpolations of DNA-Fluctuation Variance and Distributions*.

After $2D$ was determined for each segment according to Eqs. **S10–S12**, the average MSD_{DNA} over all displacements for that segment was estimated to be the share of $2D$ owing to DNA fluctuations, and was subtracted:

$$2D_{\text{protein}} = 2D_{\text{apparent}} - \frac{1}{n}\sum_j^n \text{MSD}_{\text{DNA}}(\Delta t_j), \qquad \textbf{[S13]}$$

j over all normalized displacements.

We compare the sequence-dependent diffusivity using this method to that discussed in the main text and *SI Text*, *Data Analysis* in Fig. S6. The alternative diffusivity, $D_{\text{alt}} = \frac{1}{2}2D$, correlates better with theoretical D/D_0 ($r = 0.931$) than does the diffusion coefficient D_{expt} using the MLE-based approach ($r = 0.810$). We nonetheless chose to present our results using D_{expt}, as D_{alt} is less rigorously theorized, and does not allow us to report diffusion coefficients for individual particles.

Prediction of Energy Landscape and Local Diffusion Coefficients. Our work bears some similarity to a single-molecule study by Harada et al. (5) that found a dependence in the dissociation kinetics of RNA polymerase from λ-phage DNA both on GC content and on the presence or absence of known promoters or promoter-like sequences. We took the additional steps, however, of quantifying the match between every site on the λ genome and our sequence of interest, using a position weight matrix (PWM), as well as quantifying the correlation between an energy landscape based on the scored genome and the observed kinetics of the protein.

We built an effective predicted landscape $U(x)$ as follows. Every position on λ DNA was scored according to a PWM for a single dimer, based on a list of known p53 binding sites. The PWM we used closely resembled those derived from six other lists based on a variety of experimental techniques (Fig. S2*B*). As discussed in *Methods* in the main text, the differences between scores are assumed to be proportional to differences between corresponding half-site energies:

$$E_{\mathrm{R}}(x) - E_{\mathrm{S}} = c(PWM(x) - PWM_{\mathrm{S}}), \qquad \textbf{[S14]}$$

where $PWM(x)$ is the score for position x, and PWM_{S} is the score corresponding to binding energy in the **S** mode. Thus, in the event that a site scores equal to the reference score, the specific and nonspecific binding energies for p53 to that site will be equal. We chose a value for PWM_{S} based on studies of eukaryotic transcription factor binding energies on defective versions of their consensus sequences (6). It was observed that for all the transcription factors studied, binding weakened as the consensus sites were mutated to contain one and then two mismatches (equivalent to four bits), but then became no weaker with further mutations. We therefore chose a nonspecific reference score equal to the score of the best-scoring half-site minus four bits. Varying PWM_{S} by a bit in either direction had little effect on our results. The choice of a four-bit threshold receives some additional justification from fluorescence-recovery-after-photobleaching measurements of p53 and two other eukaryotic transcription factors that found all three transcription factors' search dynamics to be similar (7).

The remaining unknown in Eq. **S14** is the proportionality constant c that relates score to energy. Dissociation constants for p53 binding to the left-hand Mdm2 half-site as well as to random DNA are available from biochemical measurements (8). At our experimental conditions, p53 favors the Mdm2 half-site by a factor of 47 (8), and so for this half-site, we estimate $E_{\mathrm{R}}(x) - E_{\mathrm{S}} = \log(47)\ k_{\mathrm{B}}T = 3.9\ k_{\mathrm{B}}T$. Substituting this value into the left-hand side of Eq. **S14**, and the site's PWM score minus PWM_{S} into the right-hand side gives a value for c of $0.97\ k_{\mathrm{B}}T$/nat or $0.67\ k_{\mathrm{B}}T$/bit.

At any site x, the protein may bind in four distinct modes owing to the left and right dimers being able each to bind in either mode: (*i*) both dimers in **S**; (*ii*) left dimer in **S**, right dimer in **R**; (*iii*) left dimer in **R**, right dimer in **S**; and (*iv*) both dimers in **R** (Fig. S2*A*). The statistical weight of a site x is thus the sum of the Boltzmann factors corresponding to each of the four modes:

$$w(x) = e^{-2E_{\mathrm{S}}} + e^{-(E_{\mathrm{S}}+E_{\mathrm{R}}(x+\Delta))} + e^{-(E_{\mathrm{R}}(x)+E_{\mathrm{S}})} + e^{-(E_{\mathrm{R}}(x)+E_{\mathrm{R}}(x+\Delta)+\varepsilon)}. \qquad \textbf{[S15]}$$

The constant ε is a cooperativity term representing additional binding energy when both dimers are bound in specific mode. Its value was determined from Eq. **S15** by substituting in energies for the left-hand and right-hand sites of the Mdm2 promoter as determined by Eq. **S14** and our PWM scoring, and substituting experimental values for the K_{d} of the full Mdm2 site relative to the K_{d} for a random sequence. From this, we find $\varepsilon = -1.39\ k_{\mathrm{B}}T$, the negative sign indicating that the energy of a protein on a full-site that binds both component half-sites in specific mode is $1.39\ k_{\mathrm{B}}T$ lower than it would be absent any cooperativity.

A small (~10%) proportion of known p53-binding sites include a gap of 1–14 bp between half-sites. To allow gapped full-sites to be treated as such in our predicted energy landscape, $E_{\mathrm{R}}(x+\Delta)$ at each binding site was assigned as follows:

$$E_{\mathrm{R}}(x+\Delta) = \min_i(E_{\mathrm{R}}(x+\Delta_0+i) - c\log(f_i/f_0)); \quad i = 0, \ldots, 14, \qquad \textbf{[S16]}$$

where Δ_0 is the length of a half-site, 10 bp, and thus the separation between half-site start positions in the absence of a gap. The index i is over gaps of length 0 to 14, and f_i is the frequency of gaps of length i in the dataset used to build the PWM. The second term under the minimum accounts for the suboptimal binding conformation the protein must adopt when binding to half-sites separated by a gap. As $f_{i>0} < f_0$, gapped full-sites suffer an energy penalty, while full-sites with zero gap suffer none.

Setting the energy scale such that $E_{\mathrm{S}} \equiv 0$, Eq. **S15** becomes

$$w(x) = 1 + e^{-E_{\mathrm{R}}(x+\Delta)} + e^{-E_{\mathrm{R}}(x)} + e^{-(E_{\mathrm{R}}(x)+E_{\mathrm{R}}(x+\Delta)+\varepsilon)}. \qquad \textbf{[S17]}$$

A single-mode model would not include nonspecific binding and thus omit all but the final term in Eq. **S17**, and a model that disallowed hemi-specific binding would omit the middle two terms. From this function of the statistical weights across all positions, we may treat p53 as interacting with DNA on a "golf-course landscape," the energy at position x of which is equal to the negative logarithm of $w(x)$:

$$U(x) = -\log w(x). \qquad \textbf{[S18]}$$

We used the resulting effective landscape to calculate D_{theo}. We segmented the landscape at the same positions as we did the experimental data, and for each segment predicted the diminution in diffusion coefficient owing to sequence-specific binding by estimating the mean ratio of the time during a visit to the segment that the protein spends sliding on DNA, t_s, versus the total time that it spends on DNA:

$$\frac{D}{D_0} = \left\langle \frac{\Delta x^2/2t_{\mathrm{total}}}{\Delta x^2/2t_s} \right\rangle = \left\langle \frac{t_s}{t_{\mathrm{total}}} \right\rangle, \qquad \textbf{[S19]}$$

where D_0 is diffusion coefficient in the absence of sequence-specific binding; i.e., D on a completely smooth landscape, without an **R** mode. The ratio t_s/t_{total} for a trajectory **x** is

$$\frac{t_s}{t_{\mathrm{total}}} = \frac{\sum_i^{\mathbf{x}} \exp(-2E_S)}{\sum_i^{\mathbf{x}} \exp(-U(x_i))}, \qquad \textbf{[S20]}$$

where $U(x_i)$ is the effective energy at site x_i, which is the ith site visited in trajectory **x**. If the transition state for translocating between two sites is constant across all sites—equivalent to assuming that for any position on DNA, p53's microscopic rates to step left and right are equal or that traps are isolated—then averaging over trajectories results in a uniform distribution of visits to all sites in a given segment, and

$$\left\langle \frac{t_s}{t_{\mathrm{total}}} \right\rangle = \frac{n\exp(-2E_{\mathrm{S}})}{\sum_x^n \exp(-U(x))}, \qquad \textbf{[S21]}$$

where n is the number of sites in the segment. The right-hand side of Eq. **S21** consists entirely of constants, and E_{S} is defined to be zero, so

$$\frac{D}{D_0} = \frac{1}{\frac{1}{n}\sum_x^n \exp(-U(x))}; \qquad \textbf{[S22]}$$

that is, the diffusion coefficient is diminished by a factor equal to the average of e raised to the effective energy in the segment. Because p53's half-site-binding sequence logo is not perfectly

palindromic, $\exp(-U(x))$ was taken to be the mean for the forward and reverse strands.

Experimental D_{expt} was compared with predicted D/D_0 by calculating Pearson's correlation coefficient r_{expt} for the two quantities over all the segments. Assessment of statistical significance was made using the permutation test described in *Methods*, *Prediction of Diffusion Coefficients*. Owing to the 10-bp half-site *PWM* having the bulk of its information content in two nucleotides three positions apart, permuting the *PWM* is not a viable control, as $10-3=7$out of $10^2=100$ permuted PWM_S will closely resemble the original *PWM*. We thus chose to permute the scores of the positions on λ DNA rather than permuting the *PWM*. Each permutation of scores gives rise to a permuted $ER(x)$ and thus a control landscape $U(x)$ and corresponding control D/D_0. To obtain *p*-values, we calculated an r_{ctl} between each control D/D_0 and D_{expt}. Reported p is the proportion of r_{ctl} equaling or exceeding r_{expt}.

Nonspecific Binding in Model Parametrization. To parametrize our scored λ genome into an energy landscape, we used dissociation constants from in vitro affinity assays of p53 and 30-bp oligonucleotides bearing full-sites, half-sites, and random DNA (8). Because p53's binding site is 20-bp long, it is possible that one or more noncognate sites are available for p53 to bind to on either side of the full- and half-sites. Indeed, oligonucleotides of only 26 bp have been used to study binding between p53 and its cognate sites (9), so it is not improbable that a 30-bp oligonucleotide can accommodate p53 binding at least four noncognate sites. If this is the case, then the apparent preference of p53 for half-site 30-mers relative to random 30-mers, of approximately a factor of 8, reflects a true preference for a single half-site over a single random site of 35:

$$\frac{n\exp(-E_n)+\exp(-E_h)}{n\exp(-E_n)}=x_{hn}\qquad\frac{\exp(-E_h)}{\exp(-E_n)}=n(x_{hn}-1), \tag{S23}$$

where n is the number of sites available on the oligonucleotide for binding, including the cognate site, E_h and E_n are half-site and noncognate binding energies, respectively, and x_{hn} is the apparent factor by which p53 prefers to bind the half-site in hemi-specific mode relative to noncognate DNA in nonspecific mode. For values of $n=5$ and $x_h=8$, the true preference for half-sites is approximately four-and-a-half times greater than the apparent preference, corresponding to an energy difference of 1.5 k_BT.

This energy difference is reflected in a greater value for the proportionality constant c relating the score of a site to its energy. With available binding sites flanking the cognate site, $c=0.97\ k_BT$/nat, while with four sites on either side ($n=5$ in Eq. **S24**), it increases to 1.37 k_BT/nat. This has the concomitant effect of raising the energy of cooperativity between specific-mode binding in the two dimers (that is, raising the energy of the fully specifically-bound state) from $\varepsilon=-1.39\ k_BT$ to $+0.19\ k_BT$; that is, specific binding becomes weakly anticooperative. The increase in c amounts to a more rugged landscape, with deeper wells at half- and full-sites, while the decrease in ε causes full-site binding to become weaker. The information content of the p53 sequence logo is such that these two effects are similar in magnitude and opposite in sign, and thus largely cancel each other out. For a pair of adjacent half-sites that each score a typical 4 bits better than the score corresponding to nonspecific binding, s_0, the energy for fully specific binding, which is the dominant form of binding on such a site, equals $2\cdot(\log(2)\text{nat/bit})\cdot 4\text{ bits}\cdot 0.97\ k_BT/\text{nat}+1.39\ k_BT=6.8\ k_BT$ in the absence of available flanking sites, and $2\cdot(\log(2)\text{nat/bit})\cdot 4\text{ bits}\cdot 1.37\ k_BT/\text{nat}-0.19\ k_BT=7.4\ k_BT$. We presented results assuming no flanking sites, but the landscapes based on the availability of 4 flanking sites are very similar in the predicted local diffusion coefficients they produce: Both have a correlation coefficient of 0.81 with experimental D.

A similar treatment for the true preference of a dimeric DNA-binding protein for binding a full-site in full-specific mode relative to a noncognate site in nonspecific mode, $\exp(-2E_h-\epsilon)/\exp(-E_n)$, as a function of the apparent preference, denoted x_{fn}, follows:

$$\frac{n\exp(-E_n)+2\exp(-E_h)+\exp(-2E_h-\epsilon)}{n\exp(-E_n)}=x_{fn}.$$

Rearranging and substituting in Eq. **S24**,

$$\frac{n\exp(-E_n)+2n(x_{hn}-1)\exp(-E_n)+\exp(-2E_h-\varepsilon)}{n\exp(-E_n)}=x_{fn}$$

$$\frac{\exp(-2E_h-\varepsilon)}{\exp(-E_n)}=n(x_{fn}-2x_{hn}+1). \tag{S24}$$

Although nonspecific binding to the oligonucleotides did not turn out to affect our results substantially, this owes to an accident of the parameters relevant to our system. Nonspecific binding of proteins to specific probes receives little attention, and yet is necessary to consider when making accurate estimates of binding preferences.

Interpolations of DNA-Fluctuation Variance and Distributions. We used our data from earlier work (4) of QDs covalently attached to positions on λ-phage DNA one-third and two-thirds the distance from the tether to estimate the mean apparent diffusivity owing to DNA fluctuations, $\langle\Delta x_d^2\rangle$, in Eq. **S9**. $\langle\Delta x_d^2\rangle$ at position x along the contour is expected to fluctuate according to a polynomial in x with nonzero linear and quartic coefficients (10). For all time windows Δt up to a maximum of two seconds, we fit these coefficients to the observed variance in displacement of the QDs at $x=1/3L$ and $x=2/3L$ ($L=$ the contour length of λ DNA), and an assumed zero-variance point at the tether, between frames separated by Δt to arrive at an expression for $\langle\Delta x_d(\Delta t)^2\rangle$:

$$\langle\Delta x_d^2(\Delta t)\rangle=a_1(\Delta t)\cdot x+a_4(\Delta t)\cdot x^4. \tag{S25}$$

The same QD data was used to correct estimates of D/D_0 for the uncertainty in the assignment of experimental displacements to segments owing to DNA fluctuations. We determined for each segment's D/D_0 the proportion α of the apparent population of the segment s that can be expected to originate in fact from neighboring segments $s-1$ to the left and $s+1$ to the right:

$$\frac{D_{\text{corrected}}}{D_0}[s]=(1-\alpha_{-1}-\alpha_{+1})\frac{D}{D_0}[s]+\alpha_{-1}\frac{D}{D_0}[s-1]+\alpha_{+1}\frac{D}{D_0}[s+1] \tag{S26}$$

$$\alpha_{\Delta s}=\int_{-w/2}^{+w/2}Q(x|s+\Delta s)*\frac{1}{w}\mathrm{d}x. \tag{S27}$$

The variable s identifies the segment whose D/D_0 is estimated; $\alpha_{\pm1}$ is the contribution to a segment's observed population of neighboring segments $s=\pm1$. The integral is over all base pairs in the indicated segment. $Q(x|s)$ is the distribution of longitudinal DNA displacements from equilibrium for segment s, normalized such that $\int_0^\infty Q(x|s)\mathrm{d}x=1$, which we obtained from the same QD measurements used to correct experimental D for DNA fluctuations. We assumed that the density of data giving rise to observed diffusion coefficients in each segment was uniform within that

segment, and so convolved the distributions of the quantum dots displacements with a uniform distribution the width of a segment, $1/w$. It is worth remarking that the distribution of DNA displacements, Q, is itself a function of distance from the tether, so the convolution kernel widens as it moves farther from the tether.

To determine the distribution $Q(x|s)$ used in Eq. **S27**, we constructed sample distributions of the position of the QDs at 1/3 and 2/3 the length of the DNA from the tether, about their mean positions. The variances of these distributions were used to find the coefficients of a similar polynomial as the one in Eq. **S25**. Interpolated distributions consisted of a linear combination of the two closest experimental QD distributions, including a zero-variance delta distribution assumed for the tether point, such that the variance of the interpolated distribution at a position s equaled the fitted polynomial evaluated at that position:

$$Q(x|s) = \begin{cases} b_s Q(x|0) + (1-b_s)Q(x|\frac{1}{3}L) & 0 < s \le \frac{1}{3}L \\ b_s Q(x|\frac{1}{3}L) + (1-b_s)Q(x|\frac{2}{3}L) & \frac{1}{3}L \le s < \frac{2}{3}L \end{cases} \quad \textbf{[S28]}$$

$$\mathrm{Var}(Q(x|s)) = a_1 s + a_4 s^4. \quad \textbf{[S29]}$$

The QD measurements were also used to add noise to simulations, which was then subtracted out using an identical procedure as described in *SI Text*, *Data Analysis*.

Control for Specific Binding. To verify that p53 can recognize its cognate sites in our experimental conditions, we synthesized a DNA construct to which we expected p53 to bind specifically. In brief, we cloned into the pET-28b plasmid a 36-bp insert containing the p21 5' site, p53's strongest known functional binding element (GAACATGTCCCAACATGTTG), as well as two sites absent from pET-28b recognized by the nicking endonuclease *Nt.BspQI*. After extracting DNA from the transformed cells, we nicked the plasmid, treated it with an excess of a biotinylated oligonucleotide equivalent to the nicked segment, and used rolling-circle amplification with the T7 DNA replisome to produce long (>100 kb) DNA constructs with a p53 RE repeated every 5,380 nucleotides (the length of the plasmid, minus the fragment lost during the double digest, plus our insert). The resulting constructs were treated with *Escherichia coli* DNA polymerase I and T4 ligase to join Okazaki fragments.

We repeated our experiments with the same biochemical and imaging conditions, but using this synthetic construct instead of λ-phage DNA. We found p53 to bind nonuniformly to DNA; rather, we observed a periodicity in its binding, with a period corresponding to the expected separation between instances of the binding site, 5,380 bp (Fig. S8). While the binding profile is enriched for particles spaced by integer multiples of 5,380 bp, we found there to be a distribution of distances, which can be attributed to the concentration of p53 used in the experiment being sufficiently high to have multiple particles bind within some 5,380-bp segments, and thus locally saturate the p21 5' sites. We intend to explore additional properties of p53 binding on this and other engineered constructs in a future publication.

1. Tafvizi A, et al. (2008) Tumor suppressor p53 slides on DNA with low friction and high stability. *Biophys J* 95:L01–L03.
2. Elenko MP, Szostak JW, van Oijen AM (2010) Single-molecule binding experiments on long time scales. *Rev Sci Instrum* 81:083705.
3. Doyle PS, Underhill PT (2005) *Handbook of Materials Modeling*, ed Yip S (Springer, The Netherlands) 2619–2630.
4. Tafvizi A, Huang F, Fersht AR, Mirny LA, van Oijen AM (2011) A single-molecule characterization of p53 search on DNA. *Proc Natl Acad Sci USA* 108:563–568.
5. Harada Y, et al. (1999) Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys J* 76:709–715.
6. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315:233–237.
7. Mueller F, Wach P, McNally JG (2008) Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. *Biophys J* 94:3323–3339.
8. Weinberg RL, Veprintsev DB, Fersht AR (2004) Cooperative binding of tetrameric p53 to DNA. *J Mol Biol* 341:1145–1159.
9. Rajagopalan S, Huang F, Fersht AR (2011) Single-molecule characterization of oligomerization kinetics and equilibria of the tumor suppressor p53. *Nucleic Acids Res* 39:2294–2303.
10. Underhill PT, Doyle PS (2004) On the coarse-graining of polymers into bead-spring chains. *J Nonnewton Fluid Mech* 122:3–31.

A

Segment #	3	4	5	6	7	8	9	10	11
3	0	0.1484	0.0371	0.1144	0.2046	0.2609	0.2753	0.1092	0.2589
4	0.1484	0	0.1113	0.2628	0.3530	0.4093	0.4237	0.2576	0.4072
5	0.0371	0.1113	0	0.1515	0.2418	0.2980	0.3124	0.1463	0.2960
6	0.1144	0.2628	0.1515	0	0.0902	0.1465	0.1609	0.0052	0.1445
7	0.2046	0.3530	0.2418	0.0902	0.	0.0563	0.0707	0.0954	0.0542
8	0.2609	0.4093	0.2980	0.1465	0.0563	0	0.0144	0.1517	0.0020
9	0.2753	0.4237	0.3124	0.1609	0.0707	0.0144	0	0.1661	0.0165
10	0.1092	0.2576	0.1463	0.0052	0.0954	0.1517	0.1661	0	0.1496
11	0.2589	0.4072	0.2960	0.1445	0.0542	0.0020	0.0165	0.1496	0

B

Segment #	3	4	5	6	7	8	9	10	11
3	1	0.424	0.667	0.015	0	0	0	0.291	0.167
4	0.424	1	0.254	0.088	0.064	0.039	0.042	0.274	0.140
5	0.667	0.254	1	0.049	0.042	0.012	0.018	0.383	0.173
6	0.015	0.088	0.049	1	0.073	0	0.006	0.962	0.450
7	0	0.064	0.042	0.073	1	0.386	0	0.252	0.771
8	0	0.039	0.012	0	0.386	1	0.805	0.292	0.992
9	0	0.042	0.018	0.006	0	0.805	1	0.083	0.937
10	0.291	0.274	0.383	0.962	0.252	0.292	0.083	1	0.246
11	0.167	0.140	0.173	0.450	0.771	0.992	0.937	0.246	1

C

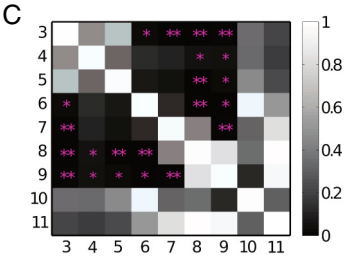

Fig. S1. Differences in D_{expt} among the segment pairs and assessment of significance. (*A*) For each pair of segments *i* and *j*, we show the absolute log-ratios, $|\log(D_i/D_j)|$, between the pairs' D_{expt}. (*B*) *p*-values for absolute log-ratios between segments' D_{expt} shown in *A*. All particles' *D*s were randomly reassigned to another particle, and D_{expt} was calculated for each segment using these reassigned particle *D*s. One thousand such rounds were performed. Table entries are the proportion of reassignment rounds in which the absolute log-ratio of D_{expt} between a pair of segments was greater than or equal to the absolute log-ratio for the unreassigned data. (*C*) Graphical depiction of information in *B*. $p < 0.05$ denoted by *; $p < 0.01$ denoted by **.

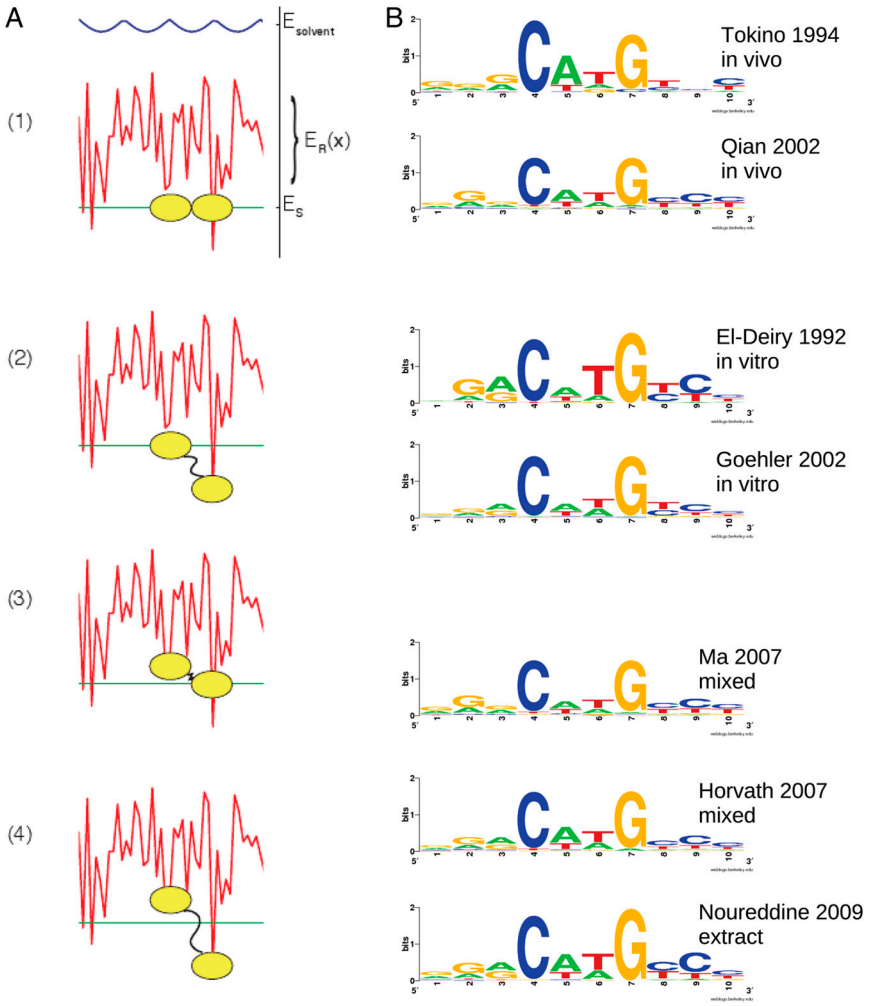

Fig. S2. (*A*) Four modes of binding: (*i*) fully nonspecific; (*ii*) first dimer nonspecific, second dimer specific; (*iii*) first dimer specific, second dimer non-specific; (*iv*) fully specific. The energy at a position *x* in the golf-course landscape is equal to the negative logarithm of the sum of the statistical weights of these four modes. (*B*) Sequence logos of the p53 half-site from a variety of position weight matrices (1–7).

1 Tokino T, et al. (1994) p53 tagged sites from human genomic DNA. *Hum Mol Genet* 3:1537–1542.
2 Qian H, Wang T, Naumovski L, Lopez CD, Brachmann RK (2002) Groups of p53 target genes involved in specific p53 downstream effects cluster into different classes of DNA binding sites. *Oncogene* 21:7901–7911.
3 El-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B (1992) Definition of a consensus binding site for p53. *Nat Genet* 1:45–49.
4 Göhler T, et al. (2002) Specific interaction of p53 with target binding sites is determined by DNA conformation and is regulated by the C-terminal domain. *J Biol Chem* 277:41192–41203.
5 Ma B, Pan Y, Zheng J, Levine AJ, Nussinov R (2007) Sequence analysis of p53 response-elements suggests multiple binding modes of the p53 tetramer to DNA targets. *Nucleic Acids Res* 35:2986–3001.
6 Horvath MM, Wang X, Resnick MA, Bell DA (2007) Divergent evolution of human p53 binding sites: Cell cycle versus apoptosis. *PLoS Genet* 3:e127.
7 Noureddine MA, et al. (2009) Probing the functional impact of sequence variation on p53-DNA interactions using a novel microsphere assay for protein-DNA binding with human cell extracts. *PLoS Genet* 5:e1000462.

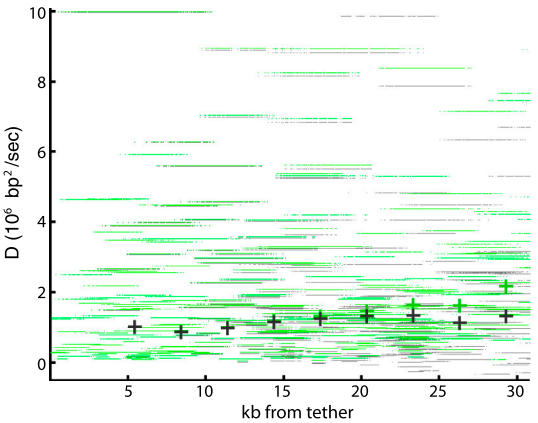

Fig. S3. Effect of DNA fluctuations on diffusion coefficients. Each particle is represented by dots in a horizontal line, as in Fig. 3*C* (main text). Data in Fig. 3*C* is reproduced in gray. As in Fig. 3*C*, each dot represents the center of a displacement, positioned along the horizontal axis according to the location on DNA of the displacement, and positioned vertically according to its parent trajectory's diffusion coefficient. Gray dots correspond to data after correction for DNA fluctuations, with segments' average diffusivity denoted by gray crosses. Green dots correspond to data before apparent diffusivity owing DNA fluctuations subtracted out per Eqs. **2** and **S9**, with the average diffusivity in a segment denoted by green crosses.

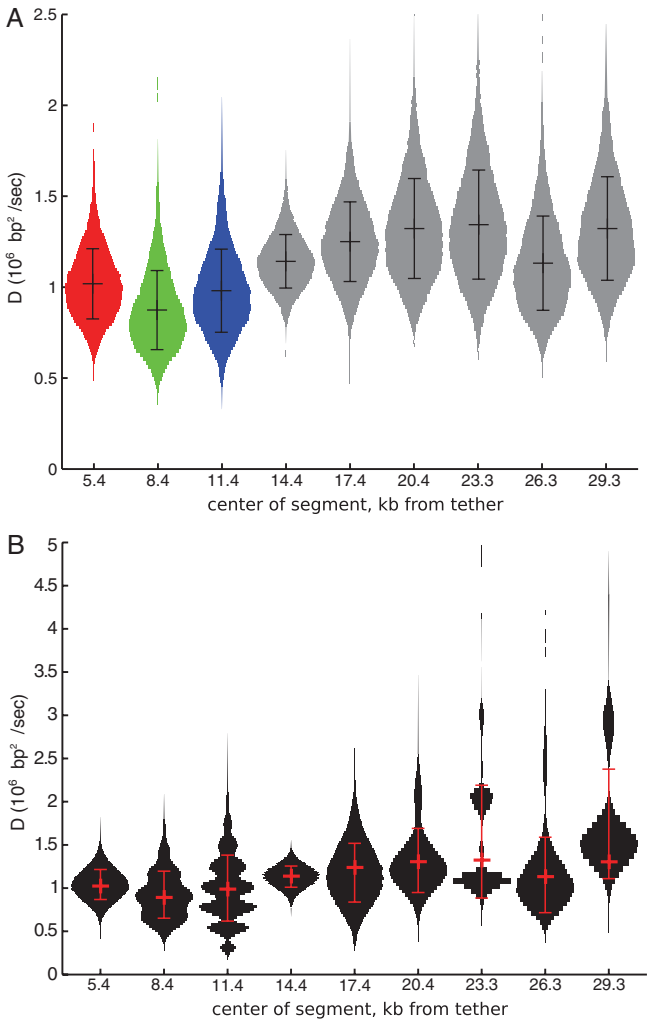

Fig. S4. (*A*) Violin plot of segments' D_{expt} from 10,000 bootstrap resamplings of particles contributing to the segments' D_{expt}. Means and standard deviations of resampled D_{expt} shown as black error bars. (*B*) Violin plot of segments' D_{expt} from 10,000 bootstrap resamplings of data-collection batches. Non-resampled means shown as thick red crosses (same position as bars in Fig. 3*D*), with standard deviations of resampled D_{expt} shown as red error bars.

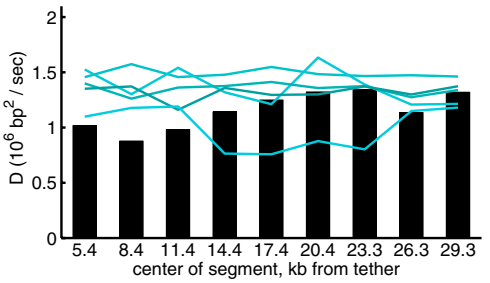

Fig. S5. Comparison of calculations of experimental D_{expt} (black bars) with D_{expt} obtained after randomizing particle positions on DNA (cyan traces).

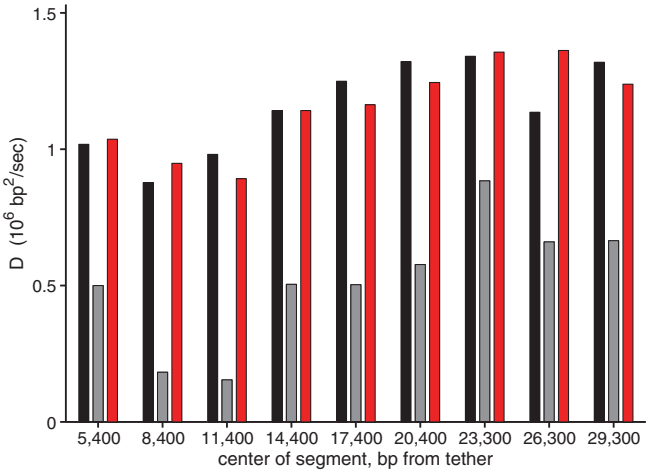

Fig. S6. Comparison of diffusion coefficients determined using the alternative method described in *SI Text, Alternative Data Analysis*. Red and black bars are identical to those in Fig. 5. Gray bars are half 2*D*, called D_{alt}, as determined by finding the variance of the fitted Gaussian distribution of normalized displacements in a segment. The correlation coefficient, *r*, between theoretical D/D_0 and D_{alt} is 0.931; and between the MLE-based D_{expt} and D_{alt} is 0.831.

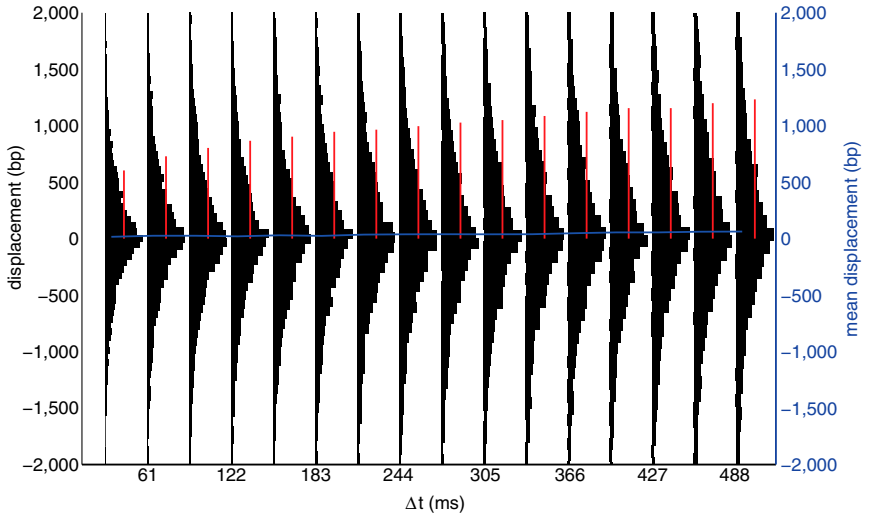

Fig. S7. Distribution of displacements across all analyzed segments, as a function of time window Δt. The red bars indicate the standard deviation of the distributions, and the blue trace the mean. As can be seen, the mean displacement is nearly zero, though close inspection will reveal that it increases approximately linearly with time, as is expected from hydrodynamic drag.

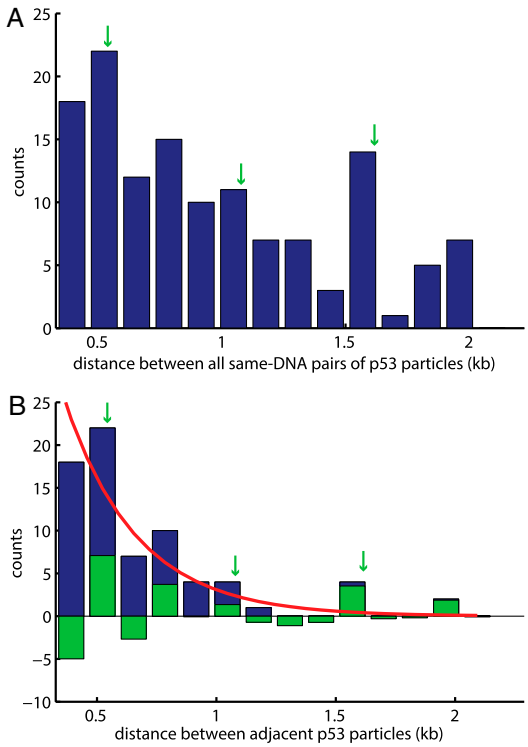

Fig. S8. Verification of p53's ability to recognize target sites. We synthesized DNA constructs consisting of several (>10) repeats of a ~5.4-kb plasmid (*SI Text*, *Control for Specific Binding*) into which we cloned the p21 5′ binding site for p53, immobilized them in our flow cell, and treated them with fluorescently labeled p53. For 35 DNA molecules, we recorded the distances between all pairs of particles on the construct, the aggregate distribution of which is shown as blue bars (bin width = 2 pixels) in *A*. Green arrows indicate distances corresponding to integer multiples of the plasmid length (i.e., the spacing between known p53 Res) and are found at positions locally enriched for p53 binding. We also present distances only between adjacent particles on the same construct as blue bars in *B*. In the absence of preferential binding, nearest neighbors would have an exponential distribution, a fit of our data to which is shown as a red trace. The difference between the fit and the measured distribution is shown as green bars. Green arrows are identical to those in *A*.