Sequence specificity in DNA binding is mainly governed by association

Emil Marklund[1], Guanzhong Mao[1#], Jinwen Yuan[1#], Spartak Zikrin[1], Eldar Abdurakhmanov[2], Sebastian Deindl[1*], Johan Elf[1*]

Affiliations:

[1]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Box 596, 75124, Uppsala, Sweden

[2]Drug Discovery and Development Platform, Science for Life Laboratory, Department of Chemistry - BMC, Uppsala University, Box 576, 751 23 Uppsala, Sweden

*Correspondence to: sebastian.deindl@icm.uu.se, johan.elf@icm.uu.se

[#] Equal contribution

Abstract: Sequence-specific binding of proteins to DNA is essential for accessing genetic information. Here, we derive a model that predicts an anti-correlation between the macroscopic association and dissociation rates of the DNA-binding protein. We test the model for thousands of different *lac* operator sequences using a protein binding microarray and by observing kinetics for individual *lac* repressor molecules in single-molecule experiments. We find that sequence specificity is mainly governed by the efficiency with which the protein recognizes different targets. The variation in probability to recognize different targets is at least 1.7 times larger than the variation in microscopic dissociation rates. Modulating the rate of binding instead of the rate

of dissociation, effectively reduces the risk of the protein being retained on non-target sequences while searching.

One Sentence Summary:

Association and dissociation rates are anti-correlated for reactions that include a nonspecific probing step.

Main Text:

Sequence-specific recognition and binding of DNA target sites by proteins such as polymerases, DNA-modifying enzymes, and transcription factors are essential for gene expression and regulation across all kingdoms of life (*1*). The textbook explanation for this sequence dependence of binding posits that favorable hydrogen bonding interactions between the protein and particular DNA sequences result in prolonged binding times (*2*). Consequently, the rate of protein dissociation would depend on the DNA sequence, while the association rate would be invariant with respect to sequence. Indeed, the rate of protein association with DNA has often been assumed to be sequence-independent (*3–6*). However, single-molecule measurements have shown that when a protein scans the DNA for binding sites, the association rate does depend on the sequence *(7)*, and that different target sequences can be bypassed with distinct probabilities (*8*). These differences have been ascribed to differences in the probability of recognition when the protein is centered on the target sequence (*7*). It is however unknown if the specific retention of a given sequence is chiefly caused by differences in the probability of recognition or in the

rate of dissociation. In fact, the physical constraints on the rate constants are unknown beyond the fact that the ratio of association and dissociation rates is necessarily dictated by the free energy difference between the free and bound states.

To explore what limits the association and dissociation rates, we considered the theoretical standard model (*9*), according to which a protein has a nonspecific testing mode where it is bound nonspecifically to DNA (Fig. 1A). In the testing state, where the protein can slide into the target sequence through nonspecific interactions, the protein can either specifically bind the target with probability p_{tot}, or dissociate into solution with probability 1-p_{tot}. When the association process is modeled as a three-state (specifically bound, nonspecifically bound, and dissociated) continuous-time Markov chain, the effective macroscopic target association and dissociation rate constants (k_a and k_d) relate to each other according to (see Supplementary Text for derivation)

$$k_a = k_{on,max} - \frac{k_{on,max}}{k_{off,\mu}} k_d \quad , \tag{1}$$

where $k_{on,max}$ is the association rate constant given by a searching protein that binds the target upon every nonspecific encounter (p_{tot}=1), and $k_{off,\mu}$ is the rate of microscopic dissociation from the bound state into the nonspecifically-bound searching mode. This equation implies that the macroscopic association and dissociation rates are inherently coupled, and linearly anti-correlated if binding sites exhibit identical microscopic dissociation rates, since $k_{on,max}$ does not depend on the specific sequence. The linear relationship between k_a and k_d described by Eq. 1 is implicitly parameterized by the probability, p_{tot}, of binding rather than dissociating from the

nonspecifically bound state, such that an increase in p_{tot} causes an increase in k_a, and a corresponding decrease in k_d (Fig. 1B). The anti-correlation can be intuitively understood by acknowledging that a decrease in the number of target site encounters required for successful binding must, in turn, result in a corresponding increase in the number of dissociation attempts needed for the macroscopic dissociation from the target (Fig. 1C). Importantly, Eq. 1 makes it possible to access the microscopic parameters $k_{off,\mu}$ and p_{tot} from macroscopically measurable parameters, such as k_a and k_d. In Fig. 1D, we show predictions for the distributions of k_a and k_d that would be observed experimentally for different binding sequences when $k_{off,\mu}$ and p_{tot} are varied in different ways (see also Methods). Three scenarios that yield the same range of $K_d = k_d/k_a$ are simulated by (1) varying mainly $k_{off,\mu}$, (2) varying $k_{off,\mu}$and p_{tot} to the same extent, or by (3) varying mainly p_{tot}. Notably, all three scenarios give distinct distributions in (k_a,k_d)-space (Fig. 1D). The linear anti-correlation between k_a and k_d is observed only in the scenario when p_{tot} varies to a larger extent than $k_{off,\mu}$.

To experimentally test if there is anti-correlation between association and dissociation rates, we measured the kinetics by which a prototypical DNA-binding protein, the *lac* repressor (LacI), binds to different operator sequences. In order to directly compare the rates for the association to and dissociation from different operators under identical experimental conditions, we used a protein binding microarray (PBM) (*10*) with 2479 different operator sequences that are mutated versions of the natural O_1 and O_2 as well as the artificially strong O_{sym} sequences. PBMs are normally used to study equilibrium binding, but by mounting the array in a flow cell on the microscope (Fig. S1A), we were able to monitor the binding and unbinding kinetics of

fluorescent LacI-Cy3 in real time (Fig. 2A, B). The Cy3 label at a site distal to the DNA binding domain has previously been shown to affect neither the specific nor the nonspecific DNA binding (*8*) (Labeling efficiency: 84.5 %; see also Supplementary Text and Table S1). Since it is impossible to measure a dissociation rate when the fluorescence signal at equilibrium is not substantially higher than background, weak target sequences showed bad reproducibility for individual sequences in repeat experiments (grey points in Fig. 2C, D). In the remainder of our analysis we therefore focused on operators where the fluorescence signal at equilibrium was > 3% of the signal for O_{sym}. For these operators the measurements of both association and dissociation rates were reproducible in repeat experiments (cyan points in Fig. 2C, see also Fig. S1B). Moreover, equilibrium dissociation constants K_D estimated using $K_d = k_d/k_a$, *versus* K_D estimated from the fluorescence values at equilibrium (see Methods), show excellent agreement (Fig. S1C). In a plot of the association *versus* dissociation rates for all operators, we observed an anti-correlation (Fig. 2D), which implies that the microscopic rate of binding p_{tot} is different for different operators. To quantify the relative importance of p_{tot} and $k_{off,\mu}$ for the binding strength, we compute which range of p_{tot} and $k_{off,\mu}$ values would give rise to the observed spread in (k_a,k_d)-space. According to this analysis, the ratio of variation in p_{tot} to variation in $k_{off,\mu}$ was 1.70 ± 0.20, depending on the noise floor reached in different experiments (Fig 2D and S1B). A larger measurement noise limits the range of p_{tot} and at the same time increases the estimated variation in $k_{off,\mu}$. For this reason, the measured ratio represents a lower limit, implying that variation in the binding probability is the major source of variation in binding strength.

Although PBMs have been shown to enable accurate measurements of relative binding kinetics, the absolute rates are not expected to be identical to those measured by other methods (*11*). We therefore also calculated $k_{off,\mu}$ for the different O_{sym}, O_1, O_2, O_3 operators (Fig. 3A) from *in vivo* estimates of k_a and k_d ((*7*, *12–14*), Fig. 3B). $k_{on,max}$ has been measured *in vivo* (*7*) (see Methods), and $k_{off,\mu}$ can therefore be calculated as the only unknown in Eq. 1 for each operator. The large error in the k_d estimate for O_3 (68% CI: [-0.08,0.24] s^{-1}) renders it impossible to determine how similar the $k_{off,\mu}$ for O_3 (68% CI: [-0.08,0.29] s^{-1}) actually is in relation to the other operators. However, consistent with our *in vitro* PBM experiments, the $k_{off,\mu}$ estimates obtained from *in vivo* data are similar for the rest of the operators (Fig. 3C). For example, even though the K_D value of O_2 exceeds that of O_1 by more than 4-fold, and that of O_{sym} by 20-fold, these operators exhibit a similar $k_{off,\mu} \approx 0.006 s^{-1}$ *in vivo*, as they all fall on the same k_a *versus* k_d line in Fig. 3B. In terms of a binding energy diagram, similar $k_{off,\mu}$ for different operators suggests that LacI binding dynamics can be described with one kinetic barrier, the height of which differs for different operators; i.e., it is more favorable for LacI to bind to certain operators than to others when sliding by, but the rate of escaping from the specifically-bound state does not change much with sequence (Fig. 3D). For the in vivo data the ratio of variation in p_{tot} to variation in $k_{off,\mu}$ is [0.67, 3.01] (68% CI), where the large error in the estimate comes from the inaccuracies in the single molecule *in vivo* measurements. This also means that we can not statistically exclude the possibility that $k_{off,\mu}$ and p_{tot} have a similar contribution to binding based on these data alone.

To change association and dissociation rates in a manner that is orthogonal to changing the operator sequence, we next performed single-molecule measurements where we varied the salt

concentration in experiments with a single operator (Fig. 4A). Changes in salt concentration are expected to affect the time that LacI spends nonspecifically bound to DNA while sliding along it (*9*, *15*). This, in turn, would change the number of operator encounters per nonspecific association, such that p_{tot} is expected to increase with decreasing salt concentrations. To probe this, we surface-immobilized a Cy5-labeled DNA construct containing a natural *lacO* operator site (O_1) and used total-internal-reflection fluorescence microscopy to monitor individual DNA molecules (Fig. 4; see also Fig. S2). Upon addition of fluorescent LacI, we monitored the appearance and disappearance of well-defined spots with co-localized fluorescence emission from both Cy3 and Cy5 (Fig. 4B). Few DNA molecules featured co-localized LacI-Cy3 spots in control experiments with Cy5-labeled DNA constructs lacking an operator site (11% and 3% at 1 and 100 mM NaCl, in contrast to >65%, >60% and >20% at 1, 100 and 200 mM NaCl for DNA with an O_1 operator; Fig. S2A), indicating that the Cy3 spots represent complexes of LacI-Cy3 specifically bound to the operator with only a minor contribution from nonspecific binding of LacI-Cy3 to DNA or to the surface. We note that the measured k_a values should be interpreted as being merely proportional to the true bimolecular association rate constants, since the exact concentration of active LacI needed for normalization can vary between salt titration repeats due to differences in the extent of protein surface adsorption, protein stability, etc. Nevertheless, we obtain an anti-correlated relationship between the measured k_a and k_d, with an estimated ratio of variation in p_{tot} to variation in $k_{off,\mu}$ of 1.65 ± 0.19 (Fig. 4C, D). To independently corroborate the dependence on salt concentration, we used surface plasmon resonance (SPR) to measure k_a and k_d for surface-immobilized O_1 operators. We found an anti-correlation between k_a and k_d with this

measurement technique as well, with an estimated ratio of variation in p_{tot} to variation in $k_{off,\mu}$ of 1.97 ± 0.07 (Fig. S3).

As an extension, our theoretical framework can be used to dissect the binding path for proteins with more complex, sequential binding mechanisms, considering that mutations along the binding pathway can be seen as energetic barriers for binding. Accordingly, one would first mutate a binding sequence in several different ways, measure the resulting macroscopic rates k_a and k_d, and then determine which sector of the (k_a,k_d)-space the different mutations fall into (Fig. S4A, B). Assuming that the native sequence has the highest k_a value and that mutations introduce a rate-limiting step, the sectors will be ordered according to the position of the mutations along the reaction pathway. Thus, rate-limiting steps closer to the bound state will result in fewer rebinding events, leading to an increase in k_d for the same value of k_a. In the case of LacI (Fig. 2, 3), the mutations fall on one line corresponding to one common rate limiting step. In the supplementary materials, we apply this method to high-throughput association and dissociation data available for dCas9 binding to off-target, mismatched mutants ((*16*), Fig. S4, 5). As expected, mutations related to the same step in the reaction pathway fall into sectors of the (k_a,k_d)-space, and the order of the sectors corresponds to the previously known reaction path.

In conclusion, the efficiency of target-site recognition is not only crucial for determining protein-DNA association rates, but also plays an equally important role in determining how long proteins remain bound to their targets. In the case of the *lac* repressor, we have shown that the efficiency of target-site recognition (p_{tot}) - and *not* how long the protein remains in the bound state - is the main determinant for binding strength observed for different sequences. This

behavior may represent an evolutionary adaptation to facilitate fast search by minimizing the risk of the protein being retained on sequences that resemble *bona fide* operators. We note that the measurements and models for LacI in this work all consider the non-induced, allolactose analogue-free repressor, which is the conformation of LacI that is capable of binding to DNA with sequence specificity and high affinity (*17*, *18*). Adding an inducer pushes the system into a new steady state where the k_d/k_a-ratio is very high. Moreover, earlier work (*19*) indicates that the induced repressor spends a non-negligible time in a non-specific testing state before binding the inducer. What this implies on the microscopic level, i.e. if the inducer exerts its effect mainly *via* a change in $k_{on,\mu}$ or $k_{off,\mu}$, is not clear from the current study. However, considering that dissociation is very fast at high concentrations of inducer (*14*), changes in $k_{off,\mu}$ are expected.

The coupling between association and dissociation rates proposed in this work holds for all bimolecular association-dissociation processes adhering to detailed balance, where a step of rapid testing for molecular recognition precedes the strong binding of a target. Indeed, the anti-correlated relationship between association and dissociation has been observed previously for numerous other systems when perturbing the sequence or salt condition (*16*, *20–22*). We therefore believe that our theoretical result is very likely to be generally applicable to a wide range of kinetic systems in addition to the ones investigated here, including processes that do not involve protein-DNA interactions.

Acknowledgments: We thank Otto Berg, Måns Ehrenberg, Helena Danielson, Jakub Wiktor, Malin Lüking, David Fange, Irmeli Barkefors, and Daniel Jones for discussions.

Funding: KAW (2016.0077 & 2019.0439 to JE; 2019.0306 to SD), VR (2016-06213 to JE, 2020-06459 to EM), ERC (StG, 714068 to SD; AdG, 885360 to JE); **Author contributions:** JE and EM conceived the study; EM derived models and equations; SD, EM, and MG designed single-molecule experiments; MG performed single-molecule experiments; EM analyzed single-molecule data; JY, EM and JE designed PBM experiments; JY performed PBM experiments; JY and SZ analyzed PBM experiments; EM and EA designed, EA performed, and EM analyzed SPR experiments; EM, JE, and SD interpreted results and wrote the paper, with input from all authors; **Competing interests:** Authors declare no competing interests; **Data and materials availability:** All raw data and analysis codes is available at https://doi.org/10.17044/scilifelab.17099687 (*23*).

List of Supplementary Materials:

References 24-38 are only cited in the Supplementary MaterialsMaterials and Methods

Supplementary Text

Figures S1 to S5

Table S1

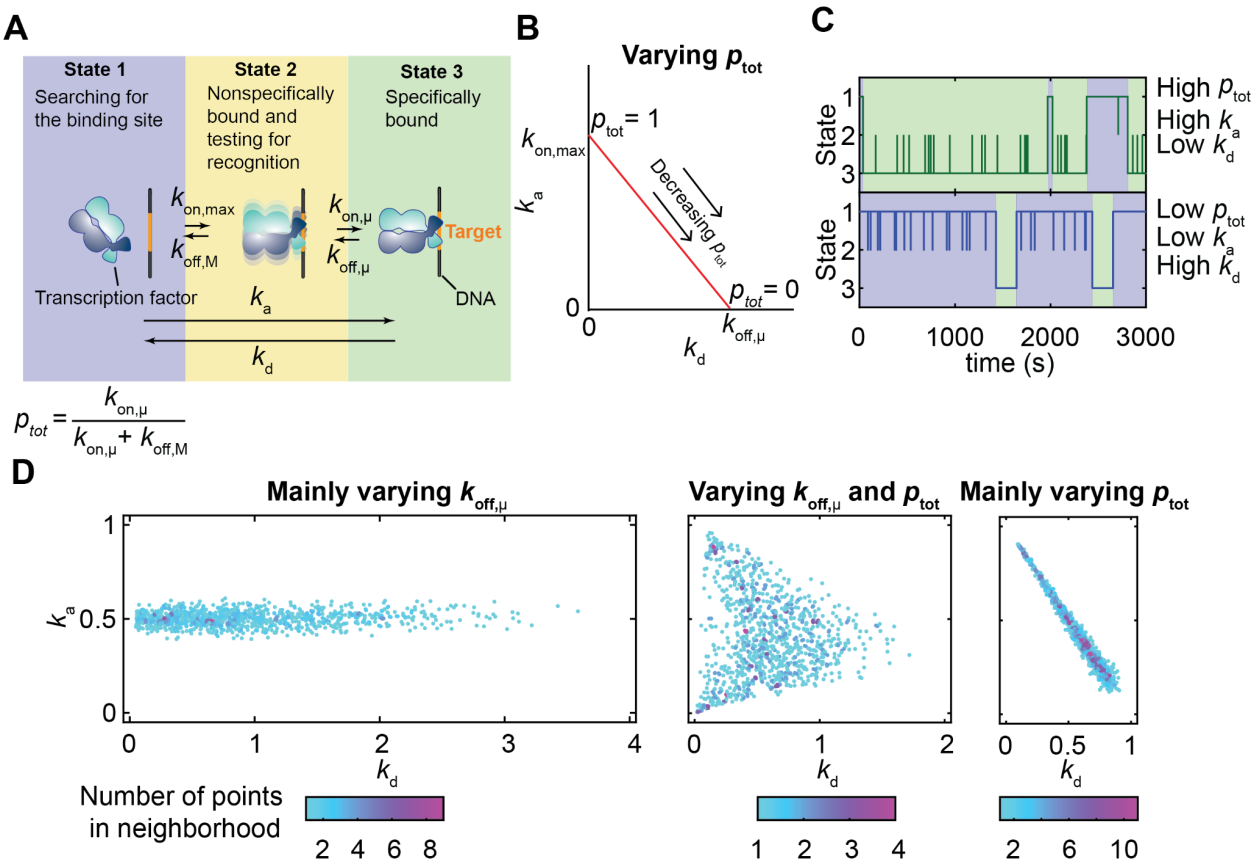

Fig. 1. Bimolecular association and dissociation rates are inherently coupled due to target-site probing. (**A**) Schematic of the kinetic model describing protein-DNA binding. (**B**) The effective rate constants for the association to (k_a) and dissociation from (k_d) the target site are coupled according to Eq. 1. This relationship becomes anti-correlated and linear when $k_{off,\mu}$ is constant and p_{tot} changes (red line). (**C**) Example traces from stochastic simulations sampling the association, dissociation, and nonspecific binding with target-site probing. When p_{tot} is high (top), the search time becomes short ($1/k_a$, blue areas) and the binding time long ($1/k_d$, green areas). When p_{tot} is low (bottom), the search time becomes long and the binding times short. (**D**) Effect on k_a and k_d when varying: $k_{off,\mu}$ 10 times more than p_{tot} (left), $k_{off,\mu}$ and p_{tot} to the same

extent (center), and p_{tot} 10 times more than $k_{off,\mu}$ (right), in simulations of the model. Each plot contains 1000 points, where each point represents one target site with a randomly sampled ($k_{off,\mu}$, p_{tot})

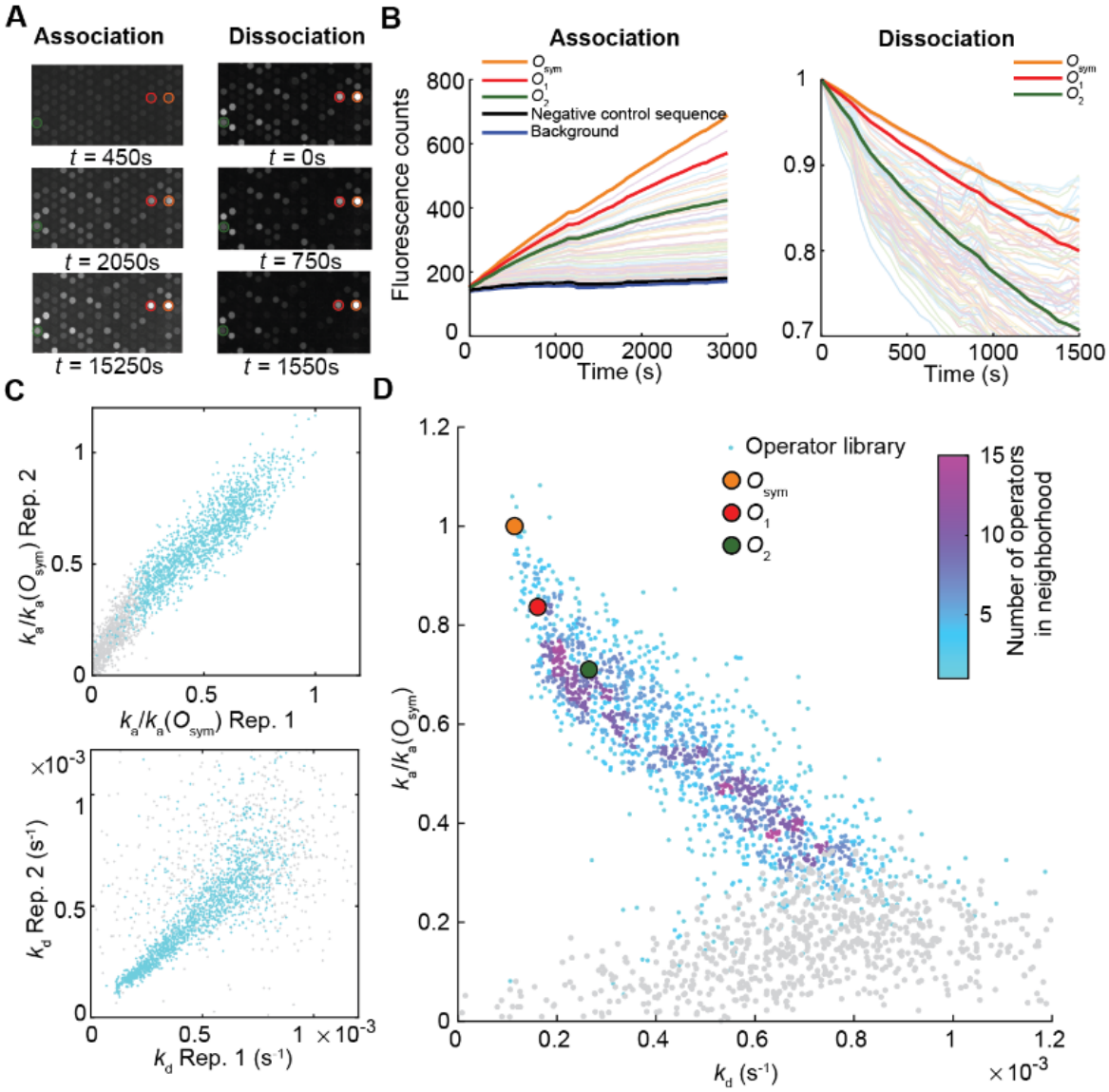

Fig. 2. Kinetic measurements with LacI on a protein binding microarray. (**A**) Example images taken during association (left) and dissociation (right) of LacI-Cy3 to spots on the DNA microarray. Orange circle: O_{sym} operator, Red circle: O_1 operator, Green circle: O_2 operator. O_3 is not present on the array, since initial experiments with this operator showed no binding over background. (**B**) Association and dissociation curves for the O_{sym}, O_1, and O_2 operators (thick lines), and 100 examples of their mutants (faint colors). The dissociation curves for each operator were normalized to the fluorescence count for the first frame of the corresponding dissociation movie (**C**) Reproducibility of k_a (top) and k_d (bottom) for individual operators (cyan points) between replicates. The k_a values are all normalized to the k_a value of O_{sym} in that replicate. Sequences associated with weak binding (fluorescence signal at equilibrium < 3% of signal for O_{sym}; grey points). (**D**) Measured association and dissociation rates for wild-type operators (circles) and their single and double mutants (points colored by operator density in that (k_a,k_d) neighborhood). (Grey points, same as in C). The value for each operator is a mean from 2 PBM replicate experiments, see S1B for data from individual replicates.

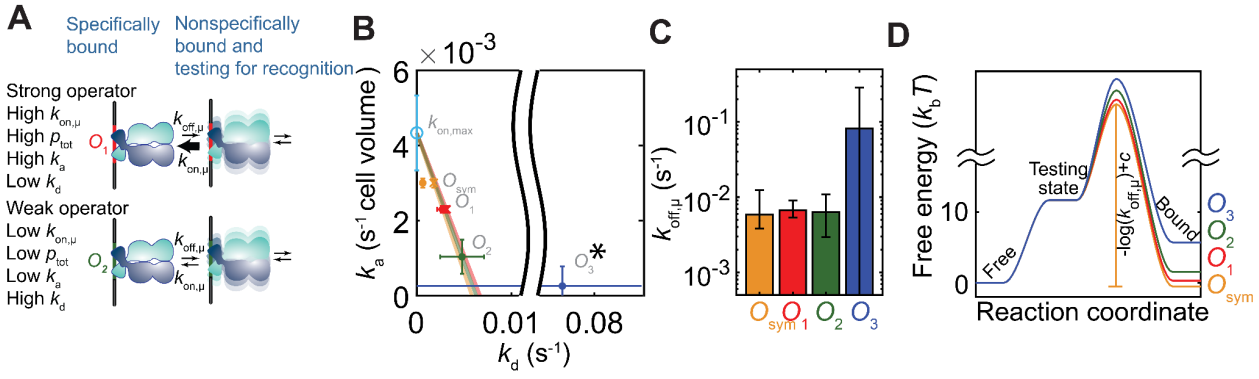

Fig. 3. Analysis of *in vivo* binding kinetics. (**A**) Predicted effect on the association and dissociation rates if the $k_{on,\mu}$ values were different but $k_{off,\mu}$ identical for the different operators. (**B**) Experimental single-molecule target-site association rate constants k_a (*7*) plotted against the dissociation rate constants k_d for the different *lac* operators. For the crosses, k_d was directly measured by single-molecule imaging (*14*). For the dots, k_d was calculated as $K_D \times k_a$, where the equilibrium constant K_D was measured via the repression ratio of gene expression (*12*, *13*). Cyan circle, measured $k_{on,max}$ (see Methods). Colored lines, $k_{off,\mu}$-lines, found by evaluating Eq. 1, for individual operators. *Due to the large error in the k_d estimate for O_3 (68% CI: [-0.08,0.24] s^{-1}), the $k_{off,\mu}$-line for O_3 is not shown. Error bars are standard errors, obtained by propagating experimental errors. (**C**) Microscopic dissociation rates $k_{off,\mu}$ for the different operators, estimated from the *in vivo* data using Eq. 1. Error bars are 68 % confidence intervals, obtained by propagating experimental errors. (**D**) Energy landscapes (a putative rather than a true reaction coordinate is shown) for the transition from free (State 1) to bound (State 3) states for the different operators, as determined by the measured K_D and $k_{off,\mu}$ values. The activation energy on the transition path between the testing state and bound state is not uniquely determined, but the differences in activation energies between the different operators are. The activation energy is equal to $-\log(k_{off,\mu}) + c$, where c is the same constant for all operators (see Supplementary Text).

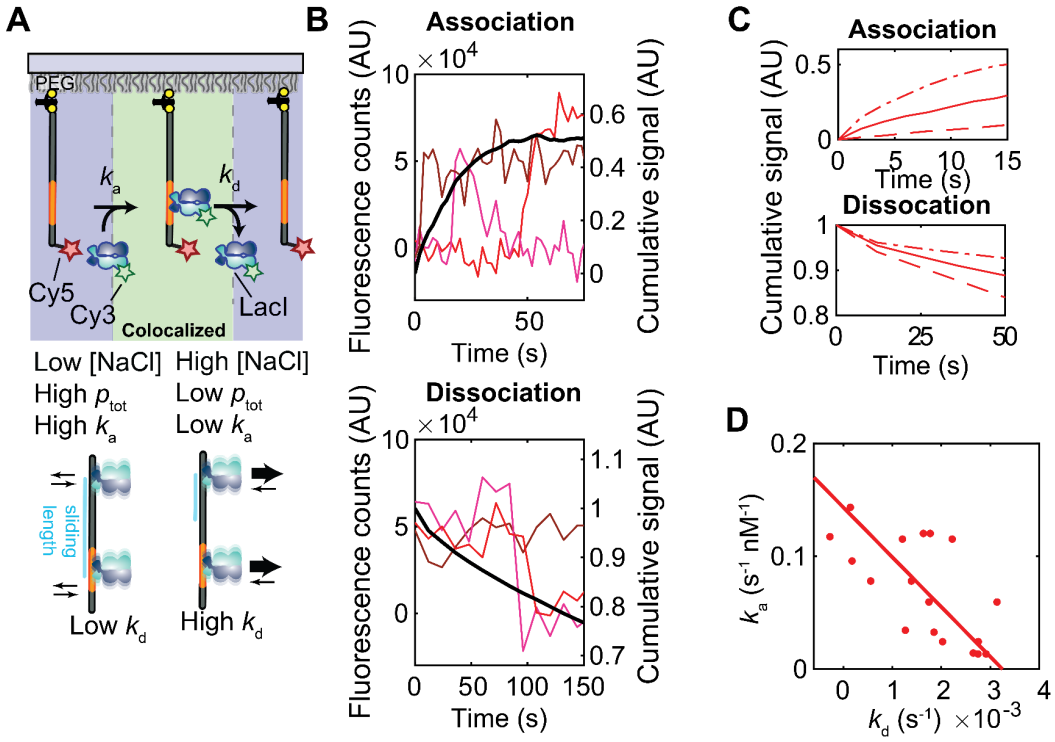

Fig. 4. Effect of changing the salt concentration. (**A**) Single-molecule colocalization measurements detect association and dissociation for LacI binding to its operators (top) and the predicted effect on association and dissociation rates of changing the salt concentration (bottom). (**B**) Example single-molecule traces showing binding to and unbinding from the O_1 operator at 100 mM NaCl (colored lines) and the normalized association and dissociation curves (black lines) obtained after summing 649 and 777 traces for the association and dissociation experiment, respectively. a.u., arbitrary units. (**C**) Normalized association and dissociation curves for O_1 binding at 1 mM (dashed dotted), 100 mM (solid), and 200 mM NaCl (dashed). (**D**) Measured k_a and k_d values for O_1 binding at different salt concentrations, and a fit to Eq. 1 (red line). The salt concentrations used for the different experiments are in the range 1-250 mM supplemented NaCl (Fig. S2A-D).

Supplementary Materials for

Sequence specificity in DNA binding is mainly governed by association

Emil Marklund, Guanzhong Mao, Jinwen Yuan, Spartak Zikrin, Eldar Abdurakhmanov, Sebastian Deindl, Johan Elf

Correspondence to: sebastian.deindl@icm.uu.se, johan.elf@icm.uu.se

This PDF file includes:

Material and Methods

Sampling association and dissociation rates while varying model parameters

The three simulated scenarios of (k_a,k_d) distributions in Fig. 1D contain sampled points with the same $K_D = k_d/k_a$ values for every scenario, sampled from a normal distribution with average 0.5 and standard deviation 2, that was truncated at 0.1 and re-normalized, so that 0.1 was the lowest possible K_D value that could be obtained. $k_{on,max} = 1$ in all simulated scenarios. For the first scenario p_{tot} was sampled from a normal distribution with average 0.5, that was truncated at 0 and 1 and re-normalized, while $k_{off,\mu}$ was set from the target site probing model according to (see Supplementary Text for derivation)

$$k_{off,\mu} = K_D k_{on,max} p_{tot}/(1 - p_{tot}). \quad (2)$$

The standard deviation of the p_{tot} distribution was set to get the desired ratio = 0.1 between the variation in p_{tot} and $k_{off,\mu}$, where variation in each variable was measured as the width of the sampled 68 % central probability mass, divided by the mean of the samples in the 68% central probability mass. In the second and third scenario $k_{off,\mu}$ was sampled from a normal distribution with average 1, that was truncated at 0.01 and re-normalized, while p_{tot} was set from the target site probing model according to (see Supplementary Text for derivation)

$$p_{tot} = k_{off,\mu}/(K_D k_{on,max} + k_{off,\mu}). \quad (3)$$

The standard deviation of the $k_{off,\mu}$ distribution was set to get the desired ratio between the variation in p_{tot} and $k_{off,\mu}$ (1 and 10), where variation in each variable was measured the same way as in the first and second scenario.

Expression, purification and fluorophore labelling of the *lac* repressor

Cy3 labelled *lac* repressor dimer (LacI-Cy3) was prepared according to previously published methods (*8*, *24*), with a Cy3 introduced distal from the DNA binding domain of LacI. Briefly, the protein contains a C-terminal 6xHis-tag for affinity purification purposes, and the C-terminal tetramerization domain has been removed. A cysteine for labeling was introduced at amino acid position 312. All cysteines found in the wild-type protein were removed from the sequence, except for the solvent-excluded cysteine in the monomer-monomer interface required to maintain an intact dimer (*24*).

Preparation of protein binding microarrays

The operator library was designed and ordered on 8x15K comparative genomic hybridization (CGH) DNA microarrays (Agilent). 14262 DNA spots (including repeats), with 2479 unique sequences, were synthesised on each microarray. For O_1, all possible single and double mutants were present on the array. For O_{sym} and O_2, all possible single mutants, and all double mutants where G was changed to C, C was changed to G, A was changed to T, and T was changed to A, were present on the array. O_{sym}, O_1, O_2 and the negative control sequence without an operator occurred at 23 different repeat spots on each array. Each sequence mutant of the operators occurred at 5 or 6 different repeat spots on each array. Agilent microarray double stranding reactions were performed in the same conditions as in PBM protocols (*10*). Two microarrays were simultaneously probed in one flow cell, as shown in Fig. S1A. In initial experiments with O_3, DNA spots on the microarray (not shown) gave a signal similar to the background level, and O_3 and its mutants were therefore not included in the final microarray design.

Microscopy of protein binding microarrays

Microscopy was performed with a Ti2-E (Nikon) inverted microscope equipped with TU Plan

Fluor 5x objective (Nikon), Spectra III (Lumencor) fluorescent light source, a Sona 4.2B-11 sCMOS camera 360 (Andor), and μManager (*25*). The light source was triggered by the camera at each image acquisition. The filter sets used were, for Cy3 (LacI): excitation filter: FF01-543/22 (Semrock), emission filter: FF01-586/20 (Semrock), dichroic mirror: FF562-Di03 (Semrock). When detecting LacI in the Cy3 channel an exposure time of 800 ms was used. The flow cell was scanned as 45 images with ~18% overlap, taking 57s for each cycle. Each location on the PBM was thus imaged every 57 seconds in the Cy3 channel for the first 60 scans of association experiments. After that, the flow cell was imaged every 5 min. In dissociation experiments, the flow cell was imaged every 57 seconds, if longer dissociation time were persuaded, after 60 scans, the flow cell was also imaged every 5 min after this time point. Before the association experiment, the flow cell were incubated with Imaging Buffer (Potassium phosphate 10mM pH 7, EDTA 0.1mM, Glycerol 5%, NaCl 20mM, 1mM 2-Mercaptoethanol, 0.5mg/ml BSA, 0.3ug/ml salmon testes DNA) for 30 min as described in *(8)* and *(10)*. For association experiments, 1.7nM of labelled LacI-Cy3 dimer in Imaging Buffer was flowed in with 1.8 ml/min flow rate and imaging in the Cy3 channel started when LacI-Cy3 was introduced. After four to five hours of association, dissociation experiments were started by flowing in Imaging Buffer without LacI-Cy3 with the same flow rate as in association.

Analysis of kinetic protein binding microarray measurements

The data were analyzed by extracting the Cy3 fluorescence signal of each DNA spot in each flow cell frame. A *flow cell frame* was generated by merging 45 images of different locations of the two PBMs present in the flow chamber. The initial flow cell frame was imaged before introduction of LacI and was subtracted from the subsequent frames to reduce background noise.

Cy3 fluorescence intensity of DNA spots were extracted by aligning a binary mask with the array geometry to the flow cell frame. This mask was generated through the physical information about microarray offered by Agilent. The Cy3 fluorescence intensity of each spot was assigned with the corresponding DNA sequence through Agilent's production information of the microarray. PBM regions that were damaged or distorted due to scratches on the microarray surface, or covered by tubing regions *etc.,* were excluded manually. This removed a fraction of DNA microarray spots, but each sequence mutant still occurred at least 5 times in the first replicate, and 3 times in the second replicate experiment. The median number of DNA spots per sequence was 10 in the first replicate, and 7 in the second replicate. The fluorescent counts for each DNA spot in each merged flow cell frame were calculated as the difference between its raw fluorescence signal and the mean negative control sequence fluorescence value in the same merged *flow cell frame*, so the fluorescence signal given from background binding was removed. The negative control sequence spots all contained the same non-operator DNA sequence (see Table S1), and had a low mean fluorescence level very similar to the mean fluorescence of the glass surface background adjacent to each spot (Fig. 2B), indicating that no binding occurred at nonspecific DNA. The acquisition of association kinetics started just after LacI were introduced into the flow cell and ended when the buffer without LacI was flowed in, which defined the start of the dissociation measurement. The signal value in association and dissociation curves, for each operator and time point, was obtained by calculating the 15 % trimmed mean of the fluorescence signal of all DNA spots containing the operator, and subtracting with the corresponding 15 % trimmed mean of the non-operator DNA sequence spots. In the example curves in Fig. 2B the signal for non-operator DNA sequence spots were not subtracted.

Non-normalized association rate constants k_a of each operator was taken as the slope of the initial 115s - 746s in the first replicate, and 171s - 800s in the second replicate (avoiding initial noisy images taken between 0~200s), of the association curve. The normalized dissociation curves were calculated by dividing by the fluorescence value at the first dissociation frame. The dissociation rate k_d of each operator was taken as the negative of the slope of the initial 600s of the normalized dissociation curve. All association rates are reported relative to the association rate of O_{sym}, ($k_a/k_a(O_{sym})$). For each operator, standard errors of k_d and $k_a/k_a(O_{sym})$ were obtained by bootstrapping, implemented by re-sampling the DNA spots containing the operator 500 times, and re-running the analysis on these samples. Each standard error was calculated as the standard deviation of the bootstrap samples, with exclusion of outlier bootstrap samples that were more than three scaled median absolute deviations from the median. Mean standard errors for all operators with a fluorescence signal at equilibrium that were at least 3% of the corresponding signal for O_{sym} are reported for each PBM replicate in Fig. S1B. Two relative K_D estimates were calculated from the data. First, from the initial slopes of the binding curves as $K_D = k_d/k_a$. Second, from the fluorescence signal at equilibrium, at the end of the association experiment, as

$$\frac{K_{D,i}}{K_{D,Osym}} = \left(\frac{f_i}{f_{Osym}}\right)^{-1}, \quad (4)$$

where f_i is the fluorescence signal at equilibrium for operator i. Eq. 4 is only valid under the assumption that the DNA spots on the microarray are far from saturated with LacI at equilibrium, also for the strongest operators. We reasoned that this assumption is most likely true because the PBM experiments were performed with a high concentration of nonspecific salmon testes DNA in the buffer, sequestering most of the LacI, and making the free LacI concentration in solution

low. The excellent agreement between the two different K_D measures, also for the strongest operators, validates the assumption (Fig. S1C).

Estimating the variation in model parameters from PBM data

Model parameters p_{tot} and $k_{off,\mu}$ were estimated for each DNA operator by evaluating Eq. 2 and 3, using the experimentally measured k_a and k_d, and $K_D = k_d/k_a$. Operators with a fluorescence signal at equilibrium that were at least 3% of the corresponding signal for O_{sym} (colored points in S1B) were included in this analysis. $k_{on,max}$ was estimated as the k_a-intercept of a linear regression of the k_a *versus* k_d data, where the operators with the 5 % highest k_a values were included in the regression. After evaluating Eq. 2 and 3 for each operator, variation in p_{tot} and $k_{off,\mu}$ was estimated as the width of the sampled 68 % central probability mass for each variable, divided by the mean of the samples within the 68 % central probability mass. We note that the estimated ratio between variation in p_{tot} and $k_{off,\mu}$ is a lower limit, since the dynamic range of the experiment limits us from measuring k_a and k_d for weak operators. This makes it impossible to measure p_{tot} values lower than some limit, depending on the level of experimental noise in that particular experiment . The larger the measurement noise is in a replicate of the PBM experiment, the smaller is the range of p_{tot} values that can be measured, and the larger is the estimated variation in $k_{off,\mu}$. This is consistent with the fact that the estimated ratio of variation in p_{tot} to variation in $k_{off,\mu}$ is largest for the PBM replicate with the lowest standard errors in the k_a and k_d estimates for individual operators in the same experiment (Fig. S1B). The estimated ratio of variation in p_{tot} to variation in $k_{off,\mu}$ in this experiment is 1.9, which is our best estimate on the lower limit of the true ratio. In the main text the ratio of variation in p_{tot} to variation in $k_{off,\mu}$ is reported as the mean ± standard error of the mean from the two PBM repeats.

DNA constructs for single-molecule fluorescence co-localization measurements

Double-stranded DNA constructs that contained operator sites as indicated, a backbone-incorporated Cy5 fluorophore attached to position 5 of a dT base via a 6-carbon linker (Integrated DNA Technologies), and an end-positioned biotin moiety were generated by annealing and ligating a set of overlapping, complementary oligonucleotides. High-performance liquid chromatography (HPLC)-purified oligonucleotides were mixed at equimolar concentrations in 50 mM Tris pH 8.0, 100 mM KCl, 1 mM EDTA, annealed with a temperature ramp (95–3°C), ligated with T4 DNA ligase (New England Biolabs), and purified by polyacrylamide gel electrophoresis (PAGE). Successful ligation was confirmed by denaturing PAGE.

Single-molecule colocalization microscopy

Biotinylated and fluorophore-labeled DNA constructs were surface-anchored onto PEG-coated quartz microscope slides through biotin-streptavidin linkage (*26*, *27*). Cy3 and Cy5 dyes were excited with 532 nm Nd:YAG and 638 nm diode lasers, respectively, and fluorescence emissions from the two fluorophores were detected using a custom-built prism-based TIRF microscope, filtered with ZET532NF (Chroma) and NF03-642E (Semrock) notch filters, spectrally separated by 635 nm (T635lpxr) and 760 nm (T760lpxr) dichroic mirrors (Chroma), and imaged onto the separate regions of an Andor iXon Ultra 888 electron multiplying charge-coupled device (EMCCD) camera. Imaging was carried out in imaging buffer containing 20 mM K_2HPO4:KH_2PO4 pH 7.4, 1 mM β-Mercaptoethanol, 0.05 mM EDTA, 100 μg/ml acetylated BSA (Promega), 10% (v/v) glycerol, 10% (w/v) glucose, 0.01% Tween 20 (v/v), 2 mM Trolox, an enzymatic oxygen scavenging system (composed of 800 μg/ml glucose oxidase and 50 μg/ml

catalase), as well as 1-300 mM NaCl (as indicated). LacI was introduced by infusing the sample chamber with imaging buffer supplemented with 0.5 nM LacI using a syringe pump (Harvard Apparatus). During image acquisition, a laser exposure time of 1 s was used. A frame rate of 0.5 Hz was used when detecting association and measuring k_a. Twelve and four association experiments were performed for O_1 and non-operator DNA respectively. For six of the O_1 experiments an additional movie at 1/12 Hz was captured directly after the 1/6 Hz movie. For these experiments, k_d values were estimated individually from the 1/12 Hz and 1/6 Hz movie, while the same 0.5 Hz movies were used for estimating the corresponding k_a. This resulted in a total of 18 k_d measurements (Fig. S2C) and twelve k_a measurements (Fig. S2D) for O_1 at different salt concentrations. The data were collected in two salt titrations during two different days.

Analysis of single-molecule colocalization measurements

The data were analyzed by summing the Cy3 fluorescence intensities within a 7 x 7-pixel square for each frame and each Cy5 dot. An à trous wavelet decomposition was used for dot detection in the Cy5 images (*28*). Dots were detected through scale-dependent standard deviation thresholding in the second wavelet plane with a threshold of three standard deviations, where the standard deviation was estimated by the median absolute deviation method (*29*). Dot centers were localized by calculating the weighted centroid from the pixel regions obtained from dot detection. Background intensities for pixels were estimated by a 2D moving average of each Cy3-fluorescence image, with exclusion of outliers by assigning them lower weights when calculating the average (*30*), so as not to include pixels corresponding to fluorescent dots when calculating the moving average. The fluorescence counts for each trace and frame were calculated as the difference between the raw fluorescence signal and the local background in the

Cy3 image. Cumulative fluorescence curves were obtained by aligning and summing regions of single-molecule traces that had a current startpoint of the region corresponding to either low (association) or high (dissociation) fluorescence (Fig. 4B) values. For the association curves, the trace regions being aligned and summed over started in a three-frame window just after LacI was introduced into the flow channel, and ended at the end of the movie. For the dissociation curves, the trace regions being aligned and summed over started at any point in time more than 200 s before the end of the movie, and ended 200 s after the start point. Fluorescence traces were classified as 'low fluorescence' or 'high fluorescence' if the current count was below 20,000 or above 35,000, respectively. A threshold was set such that all counts above 50,000 were set to this value. A binding event was detected when a trace had a fluorescence count higher than 50,000 in a 12-s moving-average window, and only such traces were included when calculating the cumulative binding curves. For each experiment, the association and dissociation rates were estimated from the initial slopes of the cumulative curves. The frequency of binding to the glass surface at non-DNA locations was estimated from the non-operator experiments, by uniformly sampling 1300 locations from the non-DNA containing pixels in the experiment, and estimating how many binding events occurred at these locations. A pixel was defined as being non-DNA containing if it had a time averaged Cy5 fluorescence lower than 1200. To account for photobleaching, we performed calibration dissociation experiments at different laser exposure times (fractional laser on times compared to the frame rate), and subtracted the constant contribution due to photobleaching from each k_d estimate (Fig. S2E). $k_{on,max}$ for each salt titration was estimated by fitting Eq. 1 to the experimental data by minimizing the sum of squared deviations between the model predicted (k_a,k_d) and the experimentally obtained (k_a,k_d). p_{tot} and

$k_{off,\mu}$ were estimated for each salt concentration by solving Eq. 1 and 16 (see Supplementary Text) for these variables. The $k_{off,\mu}$ estimate is only informative when the (k_a,k_d) for this data point is substantially different from the k_a-intercept. Therefore data points with k_d close to 0 and k_a close to $k_{on,max}$ were excluded when estimating the variation in p_{tot} and $k_{off,\mu}$. This was implemented by excluding data points having a k_a value larger than 95% of $k_{on,max}$. Variation in p_{tot} and $k_{off,\mu}$ was estimated as the standard deviation over the data points divided by the mean over the data points, for each variable. In the main text the ratio of variation in p_{tot} to variation in $k_{off,\mu}$ is reported as the mean ± standard error of the mean from the two salt titrations. The entire analysis pipeline was validated by performing stochastic simulations of binding and dissociation and simulated microscopy (*31*). This analysis method returned essentially the same association and dissociation rates as those that were put into the simulations (Fig. S2F-H). The simulations were performed with a number of DNA molecules that matches the average number of surface-immobilized DNA molecules found per field of view in the experiments (1400 Cy5 dots), and with imaging conditions mimicking the experiments in terms of level of background and fluorescence counts when LacI was bound. A frame rate of 0.5 Hz was used when simulating microscopy images for k_a estimation, and a frame rate of 1/6 Hz was used when simulating microscopy images for k_d estimation.

Kinetic measurements using surface plasmon resonance biosensor

High-performance liquid chromatography (HPLC)-purified oligonucleotides were mixed with 20% molar excess of the non-biotinylated strand, to a final concentration of 30 μM of the biotinylated strand, in 10 mM Tris pH 7.5, 50 mM NaCl, 1 mM EDTA and annealed with a

temperature ramp (95–25°C). The kinetic experiments were performed using a Biacore T200 instrument and SA biosensor chips (Cytiva, Uppsala, Sweden) at 25 ºC. The SA biosensor chip was conditioned with three pulse injections of 1 M NaCl/50 mM NaOH solution prior DNA immobilizations. DNA constructs were diluted to 40 nM in the running buffer without NaCl (10 mM K_2HPO_4:KH_2PO_4 pH 7.4, 1 mM β-mercaptoethanol, 0.1 mM EDTA, 5% (v/v) glycerol and 0.05% Tween-20) and injected at the flow rate of 2 µl/min, reaching a typical immobilization level of 30-50 response units. After immobilization of DNA constructs, unlabeled LacI was injected over the surface in 10 increasing concentrations, ranging from 1 nM to 500 nM, at a flow rate of 90 µl/min in the running buffer with four different salt concentrations (150, 200, 250 and 300 mM NaCl). The data of an association phase was collected for 120 sec and a dissociation phase for 300 sec. After each dissociation phase the surface was regenerated by an injection of 2 M NaCl/2 M $MgCl_2$ solution for 60 sec at a flow rate of 30 µl/min. Sensorgrams were double-referenced by subtracting the signals from a reference surface and the signal from two blank injections. k_a and k_d were estimated by fitting equations corresponding to a reversible, one-step, 1:1 binding to these background subtracted sensorgrams. The first 1.5 seconds, and the last second was excluded from the association curves, and the first second was excluded from the dissociation curves, to exclude parts of the data where LacI injection was switched between being on or off. For the association part of the sensorgram at LacI concentration *i*, the fitted equation is

$$R_{a,i}(t) = \frac{R_{max} k_{a,i} [LacI]_i}{k_{a,i}[LacI]_i + k_d} \left(1 - exp(-(k_{a,i}[LacI]_i + k_d)(t - t_0))\right), \quad (5)$$

For the dissociation part of the sensorgram the fitted equation is

$$R_{d,i}(t) = R_{0,i}exp(-k_d t), \quad (6)$$

where the $R_{0,i}$ parameters were locked and set to have the final value of the corresponding association data curve. The parameters ($k_{a,1}$,...$k_{a,10}$,k_d, R_{max}, t_0) were fitted to the sensorgrams of the 10 LacI concentrations simultaneously, by minimizing the sum of squared deviations between the experimental data and Eq. 5 and 6. An initial fit was performed to obtain R_{max} from the fit to the lowest salt concentration data (150 mM NaCl), and R_{max} was then locked to this value for all salt concentrations in that salt titration. ($k_{a,1}$,...$k_{a,10}$,k_d, t_0) was then fitted to the sensorgrams of all salt concentrations individually, and the reported k_a was obtained from the estimated $k_{a,i}$ (different LacI concentrations) as the slope of a linear regression of $[LacI]_i k_{a,i}$ as a function of $[LacI]_i$. Since sensorgrams at the highest LacI concentrations reached steady-state very fast and give no information about the association rate, $[LacI]k_{a,i}$ that were larger than 0.5 s^{-1} (binds in ~2s) were excluded from the linear regression. The time delay t_0 was bound in the fit to be between -1 and 1s, and handles the scenario if LacI injection did not occur exactly at the time denoted as 0s by the instrument. k_a and k_d were also estimated by another method by using the initial slopes (see PBM analysis method) during the first 2 seconds of the association and dissociation curves. Variation in p_{tot} and $k_{off,\mu}$ was estimated in the same way as for the single-molecule colocalization data, with the k_a and k_d obtained from the exponential 1:1 binding model fit. In the main text the ratio of variation in p_{tot} to variation in $k_{off,\mu}$ is reported as the mean ± standard error of the mean from the two salt titrations.

Analysis and regression of *in vivo* association and dissociation measurements

Measured values and associated errors for association rates (*7*), dissociation rates (*14*), and equilibrium constants (*12*, *13*) for *lac* repressor binding to different operators were obtained from their respective references. In (*12*, *13*) equilibrium data was measured as the fold-change of repression, and is reported in binding energies of the repressor to its operator. With the model used in (*13*), these binding energies can be recalculated to equilibrium constants via

$$K_D = N_{genome}/exp(-\Delta\varepsilon), \tag{7}$$

where $N_{\text{genome}} = 5\times10^6$ base pairs is the size of the *E. coli* genome, and $\Delta\varepsilon$ is the binding energy to the operator as defined in (*13*). In (*12*) and (*13*), the fold-change is measured for the LacI tetramer and dimer, respectively. The model fit of the data in (*12*) is shown to perfectly describe the data in (*13*), demonstrating that these constructs have identical binding energies. Thus, we can use the values of the binding energies and the associated errors from (*12*) when estimating K_D for the LacI dimer from Eq. 7. Association and dissociation rates are here reported in units per cell volume, and since ~ 4 *lac* repressor molecules were searching for the target sites in (*7*), all association rates were divided by 4. The association rates used here are taken from the additive model fit of all the data in (*7*), that is, they are extracted from the strains JE13, JE12, JE117, JE118, JE116, JE101 and JE104 containing different combinations of the O_{sym}, O_1, O_2, and O_3 operators. When estimating $k_{\text{off},\mu}$ and p_{tot} we used the k_{d} values measured directly for O_{sym} and O_1 (*14*). Since k_{d} has not been measured directly for O_2 and O_3 we instead calculated k_{d} as $K_{\text{D}}\times k_a$,

where the K_D values were taken from (*12*). The reported values for $k_{off,\mu}$ were obtained by evaluating Eq. 1 with the experimentally estimated values for (k_a,k_d,$k_{on,,max}$), where $k_{on,max}$= $k_a(O_1)/p_{tot}(O_1)$, where $k_a(O_1)$ and $p_{tot}(O_1)$ are the values reported for O_1 in (*7*). p_{tot} were estimated for each operator by taking $p_{tot}= k_a/k_{on,max}$ (see Eq. 16, Supplementary Text). Variation in p_{tot} and $k_{off,\mu}$ was estimated as the standard deviation over the operators divided by the mean over the operators, for each variable. Since the $k_{off,\mu}$ estimate for O_3 is essentially uninformative (68% CI: [-0.08,0.29] s^{-1}) we did not include this data point when estimating the variation in p_{tot} and $k_{off,\mu.}$ Error bars for $k_{off,\mu}$ and the variation in p_{tot} and $k_{off,\mu}$ are 68 % confidence intervals obtained by resampling (k_a,k_d, $k_{on,,max}$) with the errors reported in (*7*, *12–14*), while assuming that the errors for all reported values are normally distributed.

Analysis and regression of dCas9 association and dissociation measurements

All data were taken from (*16*). Dissociation rate constants were taken from the dataset with chase, and association rate constants were obtained after combining the 1 nM and 10 nM datasets as described in (*16*). The colored lines in Fig. S4A, B were acquired by fitting Eq. 1 to the experimental data by minimizing the sum of squared deviations between the model predicted (k_a,k_d) and the experimentally obtained (k_a,k_d), while constraining the line to go through the data point corresponding to the perfectly complementary sequence. In Fig. S4C colored lines are representations of a global fit of a model with an 8-state sequential recognition to all data (See Supplementary Text and Fig. S5A).

Supplementary Text

Derivation of Equation 1 and the coupling between macroscopic association and dissociation

The mean first passage times for transitions between the different states of a continuous time Markov chain are given by

$$t_{ij} = t_{i,k\neq i} + \sum_{k\neq j} p_{i,k} t_{k,j} \quad , \tag{8}$$

where i, j and k are state indices, $t_{i,j}$ is the mean first passage time to transition from state i to state j, $t_{i,k\neq i}$ is the mean time to exit from state i into any of the other states in the model, and $p_{i,j}$ is the probability to transition to state j given that the model is currently in state i.

We now consider the three-state model shown in Fig. 1A of the main text and evaluate Eq. 8 for all possible transitions in the model, which gives two linear systems of equations

$$\begin{pmatrix} -1 & 1 \\ 1-p_{2,3} & -1 \end{pmatrix} \begin{pmatrix} t_{1,3} \\ t_{2,3} \end{pmatrix} = \begin{pmatrix} -t_{1,k\neq 1} \\ -t_{2,k\neq 2} \end{pmatrix} \qquad \text{(for } j=3\text{)} \tag{9}$$

and

$$\begin{pmatrix} -1 & p_{2,3} \\ 1 & -1 \end{pmatrix} \begin{pmatrix} t_{2,1} \\ t_{3,1} \end{pmatrix} = \begin{pmatrix} -t_{2,k\neq 2} \\ -t_{3,k\neq 3} \end{pmatrix} \qquad \text{(for } j=1\text{)}, \tag{10}$$

where we have used the fact that $p_{2,1} = 1 - p_{2,3}$. We now solve Eq. 9 and 10 for $1/t_{1,3}$ and $1/t_{3,1}$, which are the sought macroscopic association and dissociation rates. We then obtain

$$k_a[R] = \frac{1}{t_{1,3}} = \frac{p_{2,3}}{t_{1,k\neq1} + t_{2,k\neq2}} \tag{11}$$

and

$$k_d = \frac{1}{t_{3,1}} = \frac{1 - p_{2,3}}{t_{3,k\neq3} + t_{2,k\neq2}}\ , \tag{12}$$

where [R] is the concentration of the searching protein. We now solve for $p_{2,3}$ (= p_{tot}) in both Eq. 11 and 12, and equate the resulting expressions, which gives

$$k_a[R](t_{1,k\neq1} + t_{2,k\neq2}) = 1 - k_d(t_{3,k\neq3} + t_{2,k\neq2}). \tag{13}$$

We then assume that $t_{2,k\neq2} << t_{1,k\neq1}$ and that $t_{2,k\neq2} << t_{3,k\neq3}$, i.e. that the time spent nonspecifically bound is much shorter than both the time spent specifically bound, and the time spent searching elsewhere (dissociated from the relevant DNA region). We believe that this is a reasonable assumption, which is true for most search processes where a searcher is looking for a rare target among many false decoys. With this assumption, the $t_{2,k\neq2}$ terms in Eq. 13 are negligible. We now solve for $k_a[R]$, which gives

$$k_a[R] = \frac{1}{t_{1,k\neq1}} - \frac{t_{3,k\neq3}}{t_{1,k\neq1}}k_d\ . \tag{14}$$

Using the same notations as in the main text, where $t_{1,k\neq1} = 1/\ k_{on,max}[R]$ and $t_{3,k\neq3} = 1/\ k_{off,\mu}$, yields the final expression,

$$k_a = k_{on,max} - \frac{k_{on,max}}{k_{off,\mu}} k_d \quad . \tag{15}$$

Two useful equations are obtained when we invoke the assumptions $t_{2,k\neq2} << t_{1,k\neq1}$ and $t_{2,k\neq2} << t_{3,k\neq3}$ for Eq. 11 and 12, and write them with the same notations as in the main text, giving

$$k_a = k_{on,max} p_{tot} \tag{16}$$

and

$$k_d = k_{off,\mu}(1 - p_{tot}). \tag{17}$$

The equilibrium constant K_D can then be written as a function the microscopic model parameters by dividing Eq. 17 with Eq. 16,

$$K_D = \frac{k_{off,\mu}(1 - p_{tot})}{k_{on,max} p_{tot}} \quad . \tag{18}$$

Solving Eq. 18 for p_{tot} and $k_{off,\mu}$ gives Eq. 2 and 3, which are used to generate the simulated (k_a,k_d) distributions in Fig. 1D (see also Methods).

The limitations, simplifications and assumptions used in the presented model follows directly from the math that is used in the derivation, and here we state these limitations explicitly: First, the model requires that the molecules have a nonspecific first interaction with an association rate constant of $k_{on,max}$. Such a nonspecific first interaction is indeed expected for many macromolecules. Second, the model assumes that a second interaction (with probability p_{tot}) has to occur for reaching a specifically bound state. Third, the reaction system must obey detailed

balance. In the context of the model, this implies that the reactions do not consume energy and are not allosterically modulated. Fourth, there is a separation of time scales between the times spent in the probing, bound, and macroscopically dissociated states, where the probing state is sufficiently short-lived to render the simultaneous probing of two different molecules extremely unlikely. Finally, the system has no memory beyond the instantaneous configuration, i.e., all microscopic internal states are rapidly equilibrated in relation to the state transitions. When these constraints are strictly satisfied, the model will provide a good approximation in many cases.

Extension of the model to handle multiple states of testing

The modeling framework can also be extended and used to determine the reaction mechanisms of binding and unbinding, if presented with measured kinetic rates of many binding site mutants. To achieve this, we extended the model to feature $N = n + 2$ number of states, and n number of testing states that the protein has to proceed through sequentially to reach the specifically bound state, and where each testing state represents a group of nucleotides in the target sequence. The generalized model is thus

$$S_1 \underset{1-p_{2,3}}{\overset{k_{on,max}}{\rightleftharpoons}} S_2 \underset{1-p_{3,4}}{\overset{p_{2,3}}{\rightleftharpoons}} \cdots \underset{1-p_{i,i+1}}{\overset{p_{i-1,i}}{\rightleftharpoons}} S_i \underset{1-p_{i+1,i+2}}{\overset{p_{i,i+1}}{\rightleftharpoons}} \cdots \underset{1-p_{N-1,N}}{\overset{p_{N-2,N-1}}{\rightleftharpoons}} S_{N-1} \underset{k_{off,\mu}}{\overset{p_{N-1,N}}{\rightleftharpoons}} S_N \,. \tag{19}$$

The model is parameterized by two rates $k_{on,max}$ and $k_{off,\mu}$, defined by the diffusion-limited association time and the time spent specifically bound, along with n probabilities $p_{i,i+1}$, defining how likely it is for the protein to transition from one testing state to the next state in the

sequential binding. The two linear equation systems obtained after evaluating Eq. 8 for the state transitions in the model are now

$$\begin{pmatrix} -1 & 1 & 0 & 0 & 0\cdots \\ 1-p_{2,3} & -1 & p_{2,3} & 0 & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 1-p_{i,i+1} & -1 & p_{i,i+1} & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 0 & 1-p_{N-2,N-1} & -1 & p_{N-2,N-1} \\ 0\cdots & 0 & 0 & 1-p_{N-1,N} & -1 \end{pmatrix} \begin{pmatrix} t_{1,N} \\ t_{2,N} \\ \vdots \\ t_{i,N} \\ \vdots \\ t_{N-2,N} \\ t_{N-1,N} \end{pmatrix} = \begin{pmatrix} -t_{1,k\neq 1} \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

(20)

and

$$\begin{pmatrix} -1 & p_{2,3} & 0 & 0 & 0\cdots \\ 1-p_{3,4} & -1 & p_{3,4} & 0 & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 1-p_{i,i+1} & -1 & p_{i,i+1} & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 0 & 1-p_{N-2,N-1} & -1 & p_{N-1,N} \\ 0\cdots & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} t_{2,1} \\ t_{3,1} \\ \vdots \\ t_{i,1} \\ \vdots \\ t_{N-1,1} \\ t_{N,1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ -t_{N,k\neq N} \end{pmatrix},$$

(21)

where we have used the assumption that the time spent in the nonspecific testing states is negligible, i.e. that $t_{i,k\neq i} = 0$ for all $i \in [2,N\text{-}1]$, and note that $t_{1,N} = 1/k_a[R]$, $t_{N,1} = 1/k_d$, $t_{1,k\neq 1} = 1/k_{on,max}[R]$ and $t_{N,k\neq N} = 1/k_{off,\mu}$. We now consider how changes in one specific $p_{i,i+1}$ parameter will change the k_a and k_d obtained from solving Eq. 20 and 21, while keeping the rest of the model parameters constant. Since the time spent in any of the testing states is assumed to be 0,

effectively the model is always in state 1 (free protein) or state N (bound protein). Hence, the passage times for reaching state i from state 1 must be exponentially distributed, with an average passage time of $(1/k_{on,max}[R])P_{1,i}$, where $P_{1,i}$ is the probability that the model will reach state i given that the model has just left state 1, before returning to state 1 again. Similarly, the passage times required to go from state N to state i must also be exponentially distributed, with an average passage time of $(1/k_{off,\mu})P_{N,i}$, where $P_{N,i}$ is the probability that the model will reach state i given that the model has just left state N, before returning to state N again. If we then view states $\{1, 2,...i\text{-}1\}$ as one state, state i as one state, and states $\{i\text{+}1,i\text{+}2,...N\}$ as one state, this new model describes a three-state continuous time Markov chain, which we have shown to have k_a and k_d coupled by Eq. 1. In this case, when considering the effect of changing one $p_{i,i+1}$ parameter, we can rewrite Eq. 1 as

$$k_a = k_{on,max}P_{1,i} - \frac{k_{on,max}P_{1,i}}{k_{off,\mu}P_{N,i}}k_d, \tag{22}$$

where $P_{1,i}$ and $P_{N,i}$ are functions of $p_{j,k}$, where $j \neq i$ and $k \neq i$+1. With $k_{on,max,i-1} = k_{on,max}P_{1,i}$ and $k_{off,\mu,i-1} = k_{off,\mu}P_{N,i}$, we obtain the same notation as used in Fig. S4 and S5, where $k_{off,\mu,i-1}$ is the effective rate of transitioning to state i from state N. Using this notation in Eq. 22 gives an equation of the familiar form

$$k_a = k_{on,max,i-1} - \frac{k_{on,max,i-1}}{k_{off,\mu,i-1}}k_d. \tag{23}$$

The model described by Eq. 20 and 21 can be used to show effects of mutations in certain sequence regions (certain $p_{i,i+1}$) on association and dissociation, which can be compared with

experimental rates to deduce which nucleotides the protein contacts first, second, and so on in the recognition process of binding. Since $k_{\mathrm{off},\mu,i-1} = k_{\mathrm{off},\mu} P_{\mathrm{N},i}$ becomes smaller and smaller with decreasing state indices i, mutations and decreases in a $p_{i,i+1}$ close to the free state (i=1) results in k_a decrease while k_d remains more constant (mostly k_a modulation, low k_d-intercept i.e. $k_{\mathrm{off},\mu,i-1}$). Correspondingly, mutations and decreases in a $p_{i,i+1}$ close to the bound state results in k_d decrease while k_a remains more constant (mostly k_d modulation, high k_d-intercept i.e. $k_{\mathrm{off},\mu,i-1}$). To demonstrate how this can be used to deduce a reaction pathway, we have applied the model on high-throughput data of association and dissociation available for dCas9 binding to off-target, mismatch mutants ((*16*), Fig. S4C). dCas9 is guided by an RNA (gRNA) when it binds DNA, with a reaction coordinate for the testing of recognition that is already well-established (*32–34*). Cas9 first detects a protospacer adjacent motif (PAM, NGG sequence), which is followed by DNA melting and gRNA-DNA hybridization, were the gRNA binds the DNA by base pairing in a sequential manner starting from a seed sequence, and then continuing hybridisation at base pairs more distal from the seed. If this sequential binding is completely memoryless, DNA binding sites with mismatches corresponding to gRNA region j will have the same microscopic dissociation rate ($k_{\mathrm{off},\mu,j}$) for the transition from a completely hybridized and specifically bound gRNA, to a melted region j. Just as our theory predicts according to Eq. 23, DNA binding sites with mutations in the same DNA region have association rates that are anti-correlated to the dissociation rates (Fig. S4C), where the k_d-intercept of each k_a *versus* k_d line is the effective microscopic dissociation rate $k_{\mathrm{off},\mu,j}$ associated with each DNA region. When mismatches are present in the seed sequence, the slope of the k_a *versus* k_d line line is steep (low $k_{\mathrm{off},\mu,j}$ and mostly k_a modulation, Fig. S4A, B). The further away from the seed sequence the mismatches are, the

flatter the slope becomes (higher $k_{off,\mu,j}$ and more k_d modulation Fig. S4A, B). Ordering the gRNA regions by increasing $k_{off,\mu,j}$ gives us the reaction pathway for dCas9 to go from the free to bound state, which is exactly the order and pathway expected for dCas9 (Fig. S4).

Next, given the pathway of the binding, we can fit the parameters of the model to the experimental data. We have done this with eight-step sequential recognition, that is with eight regions in the on-target DNA and one $p_{i,i+1}$ parameter fitted for each unique sequence region, by solving Eq. 20 and 21 for $1/t_{1,N}$ and $1/t_{N,1}$ and choosing the parameters ($k_{on,max}$,$k_{off,\mu,n}$, $p_{2,3}$...) so that the sum of squared deviations between the model predicted (k_a,k_d) and the experimentally obtained (k_a,k_d) is minimized. In total the model has 199 parameters, which are fitted to 2586 data values. With this model we can then show the predicted effect of mutations for the individual DNA regions (colored lines in Fig. S4C, varying one $p_{i,i+1}$ for each line). The model fit captures the large-scale changes and sequence-dependent coupling between k_a (Fig. S5B) and k_d (Fig. S5C) for single-and double-mismatch mutants, also when the model is trained with only half of the measured k_a and k_d values, while being tested on the other half of the dataset (Fig. S5D). We note that the model is simplistic, since it assumes that the testing of recognition is infinitely fast. Since dCas9 is observed to bind more off-target sites than Cas9 can cleave with high efficiency (*33*), a more realistic model would be one where the time that Cas9 actually spends in the testing state is considered (*35*). This discrepancy is a likely reason for why our simple model does not capture all of the data variance for single-and double-mismatch mutants (Fig. S4C and S5D), and performs poorly in terms of predicting rates for triple mutants when trained on single and double mutants (Fig. S5E).

Measurements of LacI operator binding in relation to previous work

In one of our previous papers, we performed salt titrations with LacI labeled with bifunctional rhodamine (LacI-R) and detected the binding of O_1 operators via single-molecule FRET (*8*). From the titration end points of these measurements we obtained K_D, $k_{a,obs}$ and $k_{d,obs}$ estimates of 0.0974 ± 0.0005 nM, 0.0033 ± 0.0003 $s^{-1}nM^{-1}$ and 0.0062 ± 0.0005 s^{-1} at 1 mM supplemented NaCl, or of 3.4 ± 0.6 nM, 0.0009 ± 0.0001 $s^{-1}nM^{-1}$ and 0.0089 ± 0.0007 s^{-1} at 80 mM supplemented NaCl. We note that these previous measurements, when compared with the measurements in this current work, were performed in a baseline imaging buffer with substantially higher ionic strength (10% glucose, 10% glycerol, 1mM NaCl, 0.05mM EDTA, 0.01% Tween 20, 0.1mg/ml BSA, 1mM 2-Mercaptoethanol, 2 mM Trolox, and 100mM K_2HPO4:KH_2PO_4 pH 7.4 in the previous work versus 20mM K_2HPO_4:KH_2PO_4 pH 7.4 in this current work), which makes it most suitable to compare the results in this current work with measurements from the previous work that were obtained with ~100 mM lower concentrations of supplemented NaCl. The earlier estimates above should therefore be compared with the K_D, k_a, and k_d values for LacI-Cy3 at 100 mM or 200 mM supplemented NaCl in this work (0.025 ± 0.028 nM, 0.088 ± 0.029 $s^{-1}nM^{-1}$ and 0.0014 ± 0.0017 s^{-1} or 0.13 ± 0.07 nM, 0.023 ± 0.0096 $s^{-1}nM^{-1}$, and 0.0023 ± 0.0004 s^{-1}, respectively). Furthermore, the previous work uses a LacI construct labeled at a different location, with which rates were measured using a different method. Most likely, the discrepancy between estimates from the current and earlier work is predominantly caused by the difference in how association events are detected. For the single-molecule FRET assay, a higher number of false negative (i.e. non detected) association events is expected, since the associating protein and the fluorophore must adopt a specific

conformation capable of producing an interpretable FRET signal. In our single-molecule colocalization measurements, the only requirement for the detection of an association event is that the protein must contain an intact fluorophore label. Effectively, when normalizing the association rate to the concentration of labeled protein, single-molecule FRET is expected to yield lower apparent association rate constants compared to the single-molecule colocalization measurements, just as observed when comparing the measurements for LacI-R with those for LacI-Cy3. We note that our previous work focused on measuring intramolecular properties of target search when LacI scans the DNA via 1D diffusion, and that the absolute value of the apparent intermolecular association rate constant $k_{a,obs}$ reported there does not influence any of the conclusions. Prior to our work, the K_D for the full length LacI tetramer binding to a 80bp DNA fragment with the O_1 operator has been estimated via nitrocellulose filter binding to be 0.16 nM at 50 mM KCl, and 0.43 nM at 100 mM KCl (*36*). Taken all together, we believe that the discrepancy in K_D estimates is in line with what can be expected from measurements with different protein constructs and batches that were carried out in different buffers and with different experimental methods.

k_a and k_d estimates for O_1 obtained via single molecule colocalization (Fig. 4 and S2) and the biosensor SPR (Fig. S3) differ around one order of magnitude. When comparing measurements with the same reactant pair, different biosensor methods typically give one order of magnitude difference in absolute determination of rate constants (*37*). The difference in absolute values for k_a and k_d compared to single molecule measurements is in this context not surprising.

Energy landscape for reaching the specific bound state via the putative reaction coordinate

To estimate and draw the energy landscapes *in vivo* for the different operators in Fig. 3D, we first consider the free energy difference between the free (state 1) and bound (state 3) state. Since the time spent in the testing state is much shorter than the time spent in the free and bound states, this free energy difference is directly given by the measured K_D values according to

$$\Delta G_{3\to 1} = -log(K_D), \tag{24}$$

with the energy difference given in k_bT units. Next, we consider the free energy difference between the free state and the testing state, which is the same for all operators. This free energy difference is defined as

$$\Delta G_{1\to 2} = -log(\frac{P_{testing}}{P_{free}}), \tag{25}$$

where $P_{testing}$ is the probability that the protein is testing for recognition within one sliding length from the operator, and P_{free} is the probability that the protein is searching somewhere else in the cell at any given time point. This probability can be calculated according to

$$P_{free} = P_{free,1D} + P_{free,3D}, \tag{26}$$

where $P_{free,1D}$ is the probability that the protein is bound nonspecifically to DNA somewhere else in the genome of the cell, and $P_{free,3D}$ is the probability that the protein is dissociated from DNA and is searching in the cytoplasm. The fraction of time f that the protein spends nonspecifically bound to DNA when searching is thus defined as

$$f = \frac{P_{free,1D}}{P_{free,1D} + P_{free,3D}}. \quad (27)$$

After combining Eqs. 25-27 we obtain

$$\Delta G_{1\to 2} = log(\frac{P_{free,1D}}{P_{testing}} + \frac{1-f}{f}\frac{P_{free,1D}}{P_{testing}}) \quad (28)$$

where the ratio $P_{\text{free,1D}}/P_{\text{testing}}$ can be calculated from the size of the genome N_{genome} and the average sliding length N_{sliding} as

$$\frac{P_{free,1D}}{P_{testing}} = \frac{N_{genome}}{N_{sliding}}. \quad (29)$$

With $N_{\text{genome}} = 5\times10^6$ base pairs, $N_{\text{sliding}} = 45$ base pairs (*7*), and $f = 0.9$ (*38*) we obtain the free energy difference shown in Fig. 3D.

The relative difference in activation energy on the transition path between the testing state (state 2) and bound state (state 3) for the different operators can be calculated from the measured $k_{\text{off},\mu}$ values. To achieve this, we model the transition state that the protein has to go through to reach the bound state from the testing state as a distinct species. This modeling scheme is similar to what is used in transition state theory (TST), but here we model the transition state as a true equilibrated Markovian state, instead of as a quasi-equilibrated state as is done in TST. The model is thus

$$S_2 \underset{1-p_{\ddagger,3}}{\overset{k_{2,\ddagger}}{\rightleftharpoons}} S_{\ddagger} \underset{k_{3,\ddagger}}{\overset{p_{\ddagger,3}}{\rightleftharpoons}} S_3 , \tag{30}$$

where S_2 is the testing state, $S_{\ddagger}$ is the transition state, and S_3 is the bound state. As we have shown, Eq. 12 describes how the effective transition rate from the last state to the first state in this type of model can be calculated. This effective rate is now $k_{off,\mu}$, and with the notations used in Eq. 30, Eq. 12 is for this model

$$k_{off,\mu} = \frac{1 - p_{\ddagger,3}}{\frac{1}{k_{3,\ddagger}} + t^{\ddagger}} , \tag{31}$$

where $t^{\ddagger}$ is the average time that the model spends in the transition state. Furthermore, the free energy difference between the bound state and the transition state is defined as

$$\Delta G_{3\to\ddagger} = -log(\frac{k_{3,\ddagger}}{k_{\ddagger,3}}) = -log(\frac{k_{3,\ddagger}}{p_{\ddagger,3}\frac{1}{t^{\ddagger}}}) . \tag{32}$$

Solving for $k_{3,\ddagger}$ in Eq. 31 and putting this into Eq. 32 gives

$$\Delta G_{3\to\ddagger} = -log(k_{off,\mu}) + log(1 - p_{\ddagger,3} - k_{off,\mu}t^{\ddagger}) + log(\frac{p_{\ddagger,3}}{t^{\ddagger}}) . \tag{33}$$

When $1/k_{off,\mu} >> t^{\ddagger}$, i.e. when the time spent in the transition state is very small, Eq. 33 can be written as

$$\Delta G_{3\to\ddagger} = -log(k_{off,\mu}) + c , \tag{34}$$

where

$$c = log(1 - p_{\ddagger,3}) + log(\frac{p_{\ddagger,3}}{t^{\ddagger}}) \quad (35)$$

If we now assume that $t^{\ddagger}$ and $p_{\ddagger,3}$ are the same for all operators, the difference in $\Delta G_{3\to\ddagger}$ for different operators, i.e. the difference in energy barrier between the bound state and the transition state for different operators in Fig. 3D, is given by the difference in -log($k_{off,\mu}$) for the different operators, where a fast $k_{off,\mu}$ gives a low energy barrier, and a slow $k_{off,\mu}$ gives a high energy barrier.

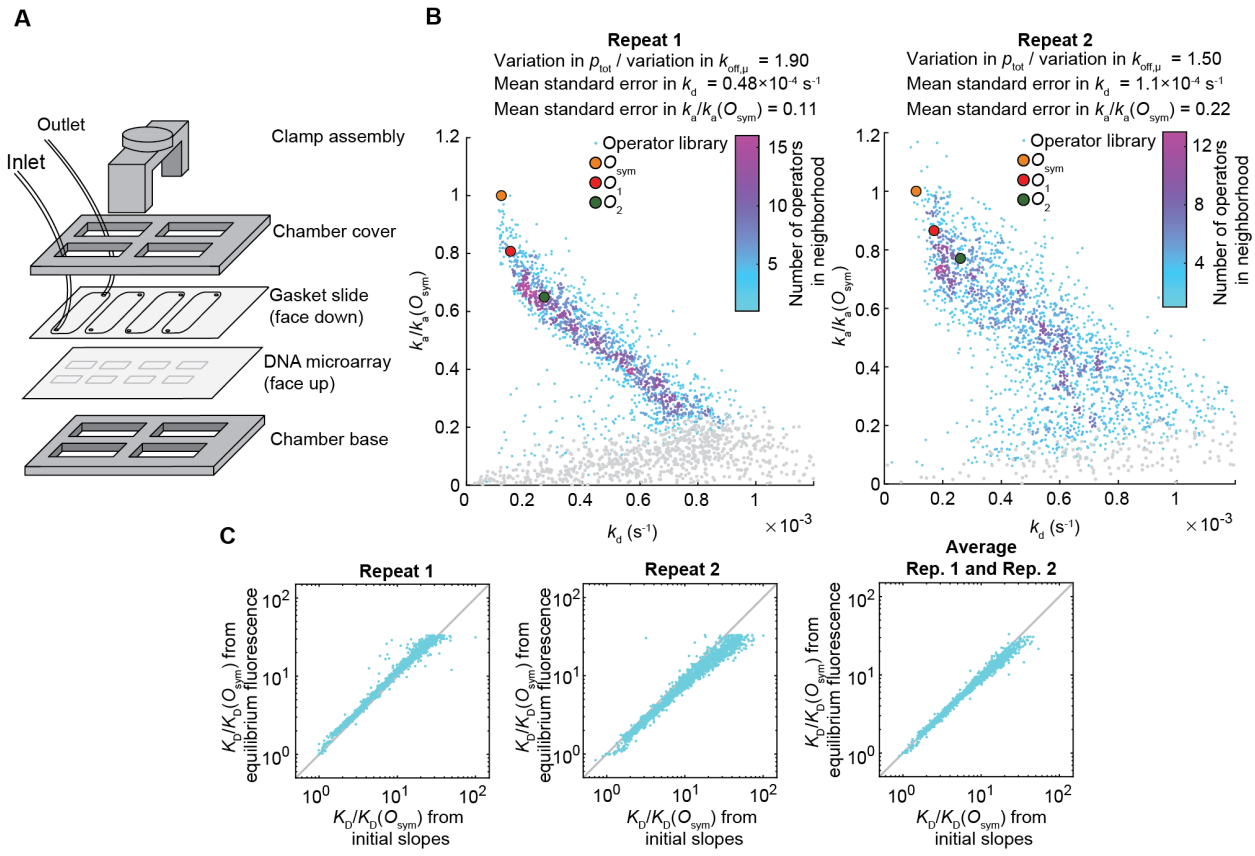

Fig. S1. Kinetic measurements on a protein binding microarray (**A**) Schematic of flow cell assembly for the kinetic measurements on a protein binding microarray. Clamp assembly and chamber base come from Agilent G2534A Microarray Hybridization Chamber. Chamber cover is customized with frosted stainless steel, having the same length, width and height as the chamber cover from Agilent G2534A Microarray Hybridization Chamber, but carved with four same size squares as in the chamber base. Each chamber on the gasket slide is drilled with two diagonal 1mm diameter holes and connected with 1mm tubings. Figure is not drawn to scale. (**B**) Measured association and dissociation rates in replicate 1 and 2 for wild-type operators (circles) and their single and double mutants (points colored by operator density in that (k_a,k_d) neighborhood). Sequences with weak binding (fluorescence signal at equilibrium < 3% of signal

for O_{sym}; grey points). (**C**) Relative K_D estimates for the operators on the PBM, estimated using two different methods.

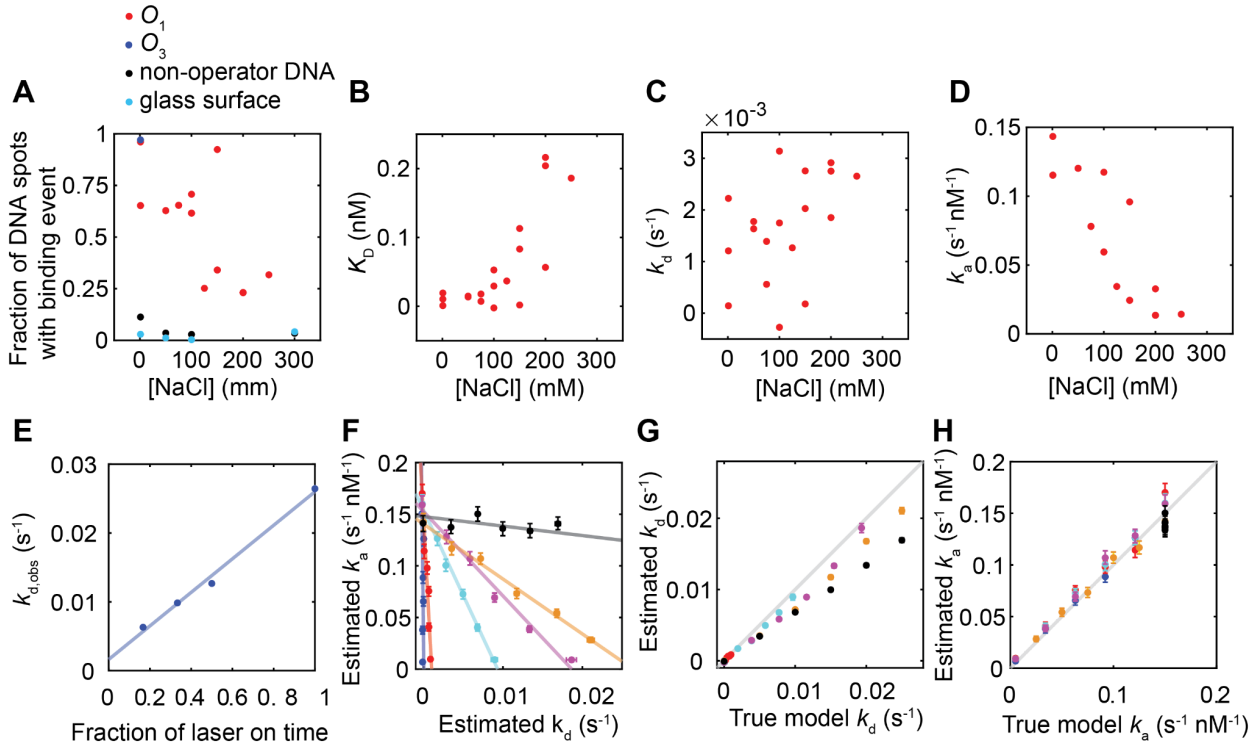

Fig. S2. Single-molecule colocalization measurements and simulations. (**A**) The fraction of spots that had at least one LacI binding event, as a function of the salt concentration of the experiment, for different DNA constructs, and non-DNA locations representing nonspecific binding to the glass surface. K_D (**B**) k_d (**C**) and k_a (**D**) for the O_1 operator as a function of the salt concentration of the experiment. (**E**) Observed dissociation rates $k_{d,obs}$, obtained as the initial slopes of the dissociation curves, plotted against the fractional exposure time f used in the experiment. $k_{d,obs}$ was measured for O_3 dissociation at 1 mM NaCl, with a laser exposure time of 1 s, and frame rates of 1,0.5, 1/3 and 1/6 Hz (blue points). The blue line is the best fit to the equation $k_{d,obs} = k_d + f\,k_{bleach}$, so that $f\,k_{bleach}$ from the fit can be subtracted from $k_{d,obs}$ to obtain k_d for each measurement. (**F**) Estimated k_a and k_d for simulated data. (**G**) True and estimated k_d values

from simulated data. (**H**) True and estimated k_a values from simulated data. All error bars are 68 % confidence intervals obtained by bootstrapping the simulated fluorescence traces.

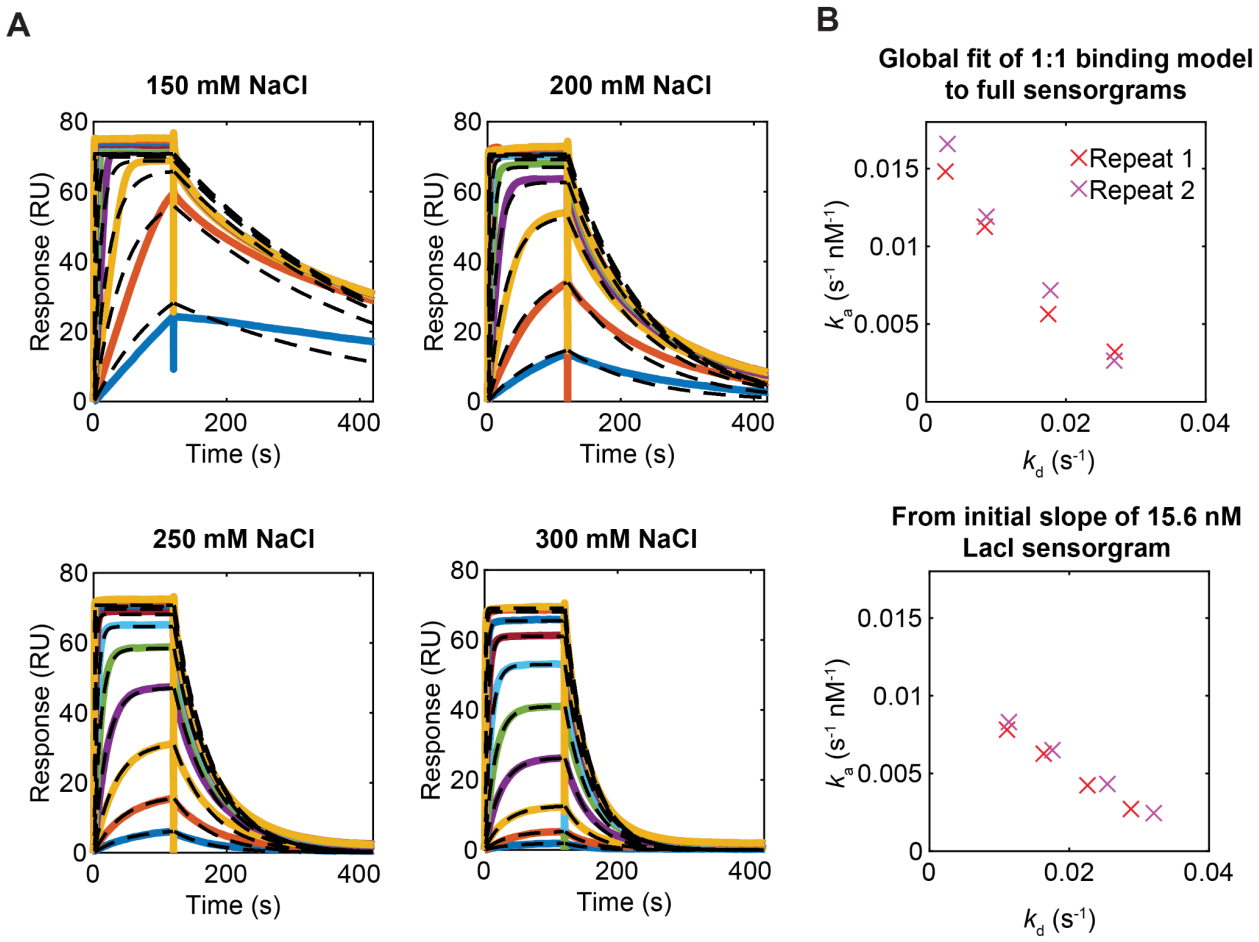

Fig. S3. Surface plasmon resonance measurements. (**A**) Sensorgrams for one of the replicates of O_1 binding and unbinding, at 10 different LacI concentrations between 1 and 500 nM (coloured lines), and global fits for each salt concentration of a 1:1 binding model (dashed lines). Sensorgrams from experiments with a non-operator construct (not shown) reached steady-state at less than 3 response units, for all protein and salt concentrations in both replicates. (**B**) k_a and k_d for LacI binding O_1 DNA at 150, 200, 250 and 300 mM NaCl, estimated using a global fit of the full sensorgrams to a 1:1 binding model (top) , and using the initial slope of the 15.6 nM LacI sensorgram (bottom).

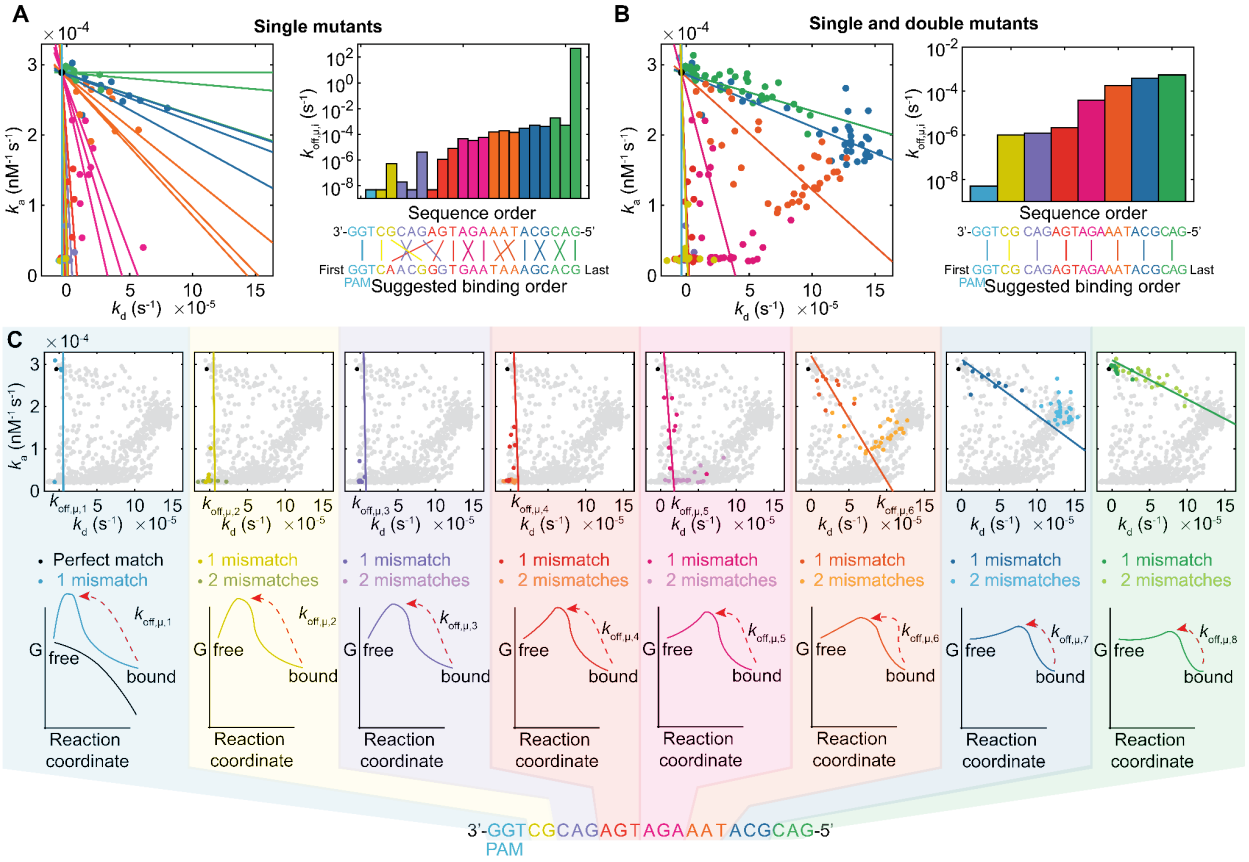

Fig. S4. The target-site recognition model describes the coupling between dCas9 off-target binding and unbinding. Triplets along the sequence are indicated by different colors. (**A**) Measured association and dissociation rates for dCas9 and different off-target sequences (single-site mutant data from (*16*)) (left). Linear fits to the data for each mutated base yield the estimated effective microscopic dissociation rate $k_{\mathrm{off},\mu,i}$ (i.e. k_d-intercept) associated with each nucleotide, indicating the binding pathway (right). (**B**) Measured association and dissociation rates for dCas9 and different off-target sequences with single and double mismatches (*16*) (left), and estimated effective microscopic dissociation rate $k_{\mathrm{off},\mu,i}$ associated with each 3-nucleotide DNA region, indicating the binding pathway (right). (**C**) Measured association and dissociation

rates for dCas9 and different off-target sequences (grey points; single and double mismatch mutant data from (*16*)), where data for mismatches occuring in specific gRNA regions are highlighted in colors. In (**A**) and (**B**), colored lines are individual fits of Eq. 1 to the single, or the single and double mismatch data, respectively. In (**C**), colored lines are representations of a global fit of a model with an 8-state sequential recognition to all the data. Here each line shows the model-predicted effect of mutations in one specific gRNA triplet region (see Supplementary Text and Fig. S5). Note that some of the data points in the plots are not highlighted in colors. These data points represent sequences with mutations in two different gRNA regions. In the right panels of (**A**) and (**B**) estimated $k_{\mathrm{off},\mu,\mathrm{i}}$ values smaller than 5×10^{-9} s^{-1} have been rounded up to this value.

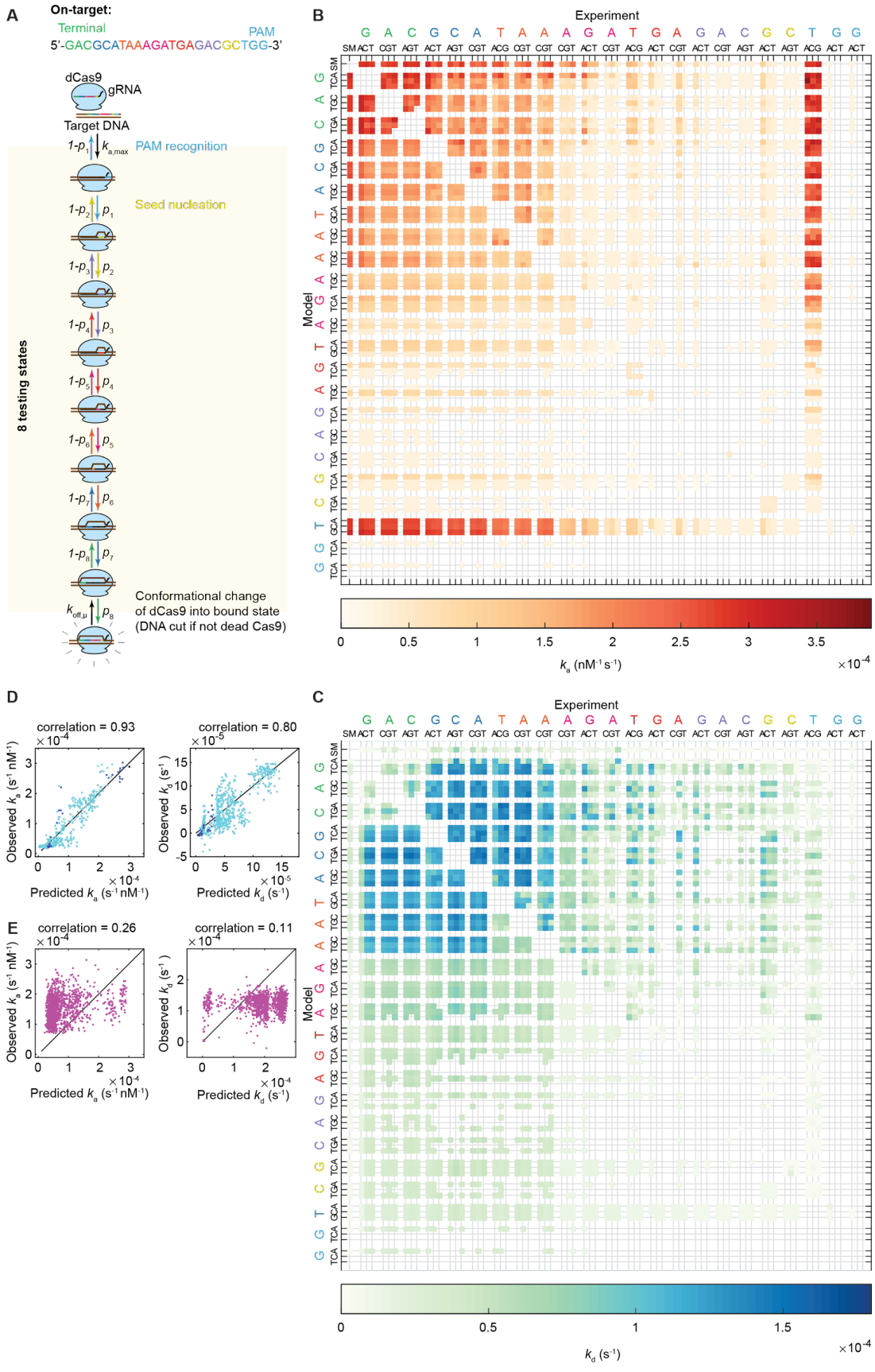
A
On-target:
Terminal
PAM
5'-GACGCATAAAGATGAGACGCTGG-3'
dCas9
gRNA
Target DNA
PAM recognition
Seed nucleation
8 testing states
Conformational change
of dCas9 into bound state
(DNA cut if not dead Cas9)
B
Experiment
Model
k_a (nM^{-1}s^{-1})
C
Experiment
Model
k_d (s^{-1})
D
correlation = 0.93
correlation = 0.80
Observed k_a (s^{-1} nM^{-1})
Predicted k_a (s^{-1} nM^{-1})
Observed k_d (s^{-1})
Predicted k_d (s^{-1})
E
correlation = 0.26
correlation = 0.11

Fig. S5. Model fit to dCas9 off-target binding and unbinding data. (**A**), Cartoon model diagram. Measured (above diagonal) and model-predicted (below diagonal) association (**B**) and dissociation (**C**) rates for single and double off-target mutants, when simultaneously fitting association and dissociation rates. The on-target, all single mutants, and all double mutants except the ones containing a mutation in the degenerate PAM (T in TGG) were used in the training data set. (**D**) Correlation between the model-predicted and observed association (left) and dissociation (right) rates when half of the data set (random sample) in (**B**,**C**) was used for training, and the other half of the data set was used for testing (points in the plots), for the on-target (black), single (blue) and double mutants (cyan). (**E**), Correlation between the predicted and observed association (left) and dissociation (right) for triple mutants when the same data set as in (**B**,**C**) is used for training the model.

Table S1. Sequence of LacI and DNA constructs. For LacI-Cy3, the cysteine introduced for labeling is marked in red. For the DNA constructs, modifications are reported using Integrated DNA Technologies (IDT) nomenclature. LacI operator sites are highlighted in orange.

name	Sequence
LacI-Cy3	MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAEL NYIPNRVAQQLAGKQSLLIGVATSSLALHAPSQIVAAIKSRADQLGAS VVVSMVERSGVEAAKAAVHNLLAQRVSGLIINYPLDDQDAIAVEAA ATNVPALFLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQQIALLAGP LSSVSARLRLAGWHKYLTRNQIQPIAEREGDWSAMSGFQQTMQML NEGIVPTAMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDS SCYIPPLTTIKQDFRLLGQTSVDRLLQLSQGQCVKGNQLLPVSLVKRK TTLAPNTQTHHHHHH
O_{sym} construct for PBM	5’-CAACTAAGCAGCTAATTGTGAGCGCTCACAATTAGGGTCTGTGT TCCGTTGTCCGTGCTG-3’
O_1 construct for PBM	5’-AACTAAGCAGCTAATTGTGAGCGGATAACAATTAGGGTCTGTGT TCCGTTGTCCGTGCTG-3’
O_2 construct for PBM	5’-AACTAAGCAGCTGGTTGTTACTCGCTCACATTTAGGGTCTGTGT TCCGTTGTCCGTGCTG-3’

Non-operator construct for PBM	5’-AACTAAGCAGCTATGAGGCATGGAATCCTGGCTAGGGTCTGTG TTCCGTTGTCCGTGCTG-3’
O_1 construct for SPR	Top strand: 5’-/5BioTinTEG/AACTAAGCAGCTAATTGTTATCCGCTCACAATTAG GGTCTGTGTTCCGTTGTCCGTGCTG-3’ Bottom strand: 5’-CAGCACGGACAACGGAACACAGACCCTAATTGTGAGCGGATAA CAATTAGCTGCTTAGTT-3’
Non-operator construct for SPR	Top strand: 5’-/5BioTinTEG/AACTAAGCAGCTATGAGGCATGGAATCCTGGCTAG GGTCTGTGTTCCGTTGTCCGTGCTG-3’ Bottom strand: 5’-CAGCACGGACAACGGAACACAGACCCTAGCCAGGATTCCATGC CTCATAGCTGCTTAGTT-3’

O_1 construct for single-molecule experiments	Top strand: 5’-/5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGG ATTGC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACT CTCAATCGCAATTGTTATCCGCTCACAATTCCGAAAGCCT-3’ Bottom strand: 5’-AGGCT/iCy5/TCGGAATTGTGAGCGGATAACAATTGCGAATGAGA GTCT TGCGTTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATC CTTGGACTTGACACGCCCAAAACTTGAAGTACGA-3’

O_3 construct for single-molecule experiments	Top strand: 5'-/5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGG ATTGC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACT CTCAATCGCGGCAGTGAGCGCAACGCAATTCCGAAAGCCT-3' Bottom strand: 5'-AGGCT/iCy5/TCGGAATTGCGTTGCGCTCACTGCCGCGAATGAGA GTCT TGCGTTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATC CTTGGACTTGACACGCCCAAAACTTGAAGTACGA-3'
Non-operator construct for single-molecule experiments	Top strand: 5'-/5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGG ATTGC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACT CTCA /iCy5/TCGCGATTGCAGCTCGAAGCAGCATCCGAAAGCC-3'

	Bottom strand: 5’-GGCTTTCGGATGCTGCTTCGAGCTGCAATCGCGAATGAGAGTCT TGCG TTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATCCTTG GACTTGACACGCCCAAAACTTGAAGTACGA-3’

References and Notes:

1. W. Gilbert, B. Müller-Hill, The lac operator is DNA. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 2415–2421 (1967).

2. R. Milo, R. Phillips, *Cell Biology by the Numbers* (Garland Science, 2015).

3. A. Grönlund, P. Lötstedt, J. Elf, Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nat. Commun.* **4**, 1864 (2013).

4. D. L. Jones, R. C. Brewster, R. Phillips, Promoter architecture dictates cell-to-cell variability in gene expression. *Science*. **346**, 1533–1536 (2014).

5. M. Z. Ali, V. Parisutham, S. Choubey, R. C. Brewster, Inherent regulatory asymmetry emanating from network architecture in a prevalent autoregulatory motif. *eLife*. **9** (2020), , doi:10.7554/elife.56517.

6. M. Morrison, M. Razo-Mejia, R. Phillips, Reconciling kinetic and thermodynamic models of bacterial transcription. *PLoS Comput. Biol.* **17**, e1008572 (2021).

7. P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, J. Elf, The lac Repressor Displays Facilitated Diffusion in Living Cells. *Science*. **336** (2012), pp. 1595–1598.

8. E. Marklund, B. van Oosten, G. Mao, E. Amselem, K. Kipper, A. Sabantsev, A. Emmerich, D. Globisch, X. Zheng, L. C. Lehmann, O. G. Berg, M. Johansson, J. Elf, S. Deindl, DNA surface exploration and operator bypassing during target search. *Nature*. **583** (2020), pp. 858–861.

9. O. G. Berg, R. B. Winter, P. H. Von Hippel, Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*. **20** (1981), pp. 6929–6948.

10. M. F. Berger, M. L. Bulyk, Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* **4**, 393–411 (2009).

11. T. Siggers, M. H. Duyzend, J. Reddy, S. Khan, M. L. Bulyk, Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 555 (2011).

12. H. G. Garcia, R. Phillips, Quantitative dissection of the simple repression input–output function. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12173–12178 (2011).

13. R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, R. Phillips, The transcription factor titration effect dictates level of gene expression. *Cell*. **156**, 1312–1323

(2014).

14. P. Hammar, M. Walldén, D. Fange, F. Persson, O. Baltekin, G. Ullman, P. Leroy, J. Elf, Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. *Nat. Genet.* **46**, 405–408 (2014).

15. P. C. Blainey, A. M. van Oijen, A. Banerjee, G. L. Verdine, X. S. Xie, 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 (2006).

16. E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna, W. J. Greenleaf, High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5461–5466 (2017).

17. M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, P. Lu, Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*. **271**, 1247–1254 (1996).

18. J. Chen, S. Alberti, K. S. Matthews, Wild-type operator binding and altered cooperativity for inducer binding of lac repressor dimer mutant R3. *J. Biol. Chem.* **269**, 12482–12487 (1994).

19. S. L. Laiken, C. A. Gross, P. H. Von Hippel, Equilibrium and kinetic studies of Escherichia coli lac repressor-inducer interactions. *J. Mol. Biol.* **66**, 143–155 (1972).

20. A. Poddar, M. S. Azam, T. Kayikcioglu, M. Bobrovskyy, J. Zhang, X. Ma, P. Labhsetwar, J. Fei, D. Singh, Z. Luthey-Schulten, C. K. Vanderpool, T. Ha, Effects of individual base-pairs on in vivo target search and destruction kinetics of bacterial small RNA. *Nat. Commun.* **12**, 874 (2021).

21. N. F. Dupuis, E. D. Holmstrom, D. J. Nesbitt, Single-molecule kinetics reveal cation-promoted DNA duplex formation through ordering of single-stranded helices. *Biophys. J.* **105**, 756–766 (2013).

22. S. Bonilla, C. Limouse, N. Bisaria, M. Gebala, H. Mabuchi, D. Herschlag, Single-Molecule Fluorescence Reveals Commonalities and Distinctions among Natural and in Vitro-Selected RNA Tertiary Motifs in a Multistep Folding Pathway. *J. Am. Chem. Soc.* **139**, 18576–18589 (2017).

23. E. Marklund, G. Mao, J. Yuan, S. Zikrin, E. Abdurakhmanov, S. Deindl, J. Elf, Data and code for: Sequence specificity in DNA binding is mainly governed by association. 10.17044/scilifelab.1709968.

24. K. Kipper, N. Eremina, E. Marklund, S. Tubasum, G. Mao, L. C. Lehmann, J. Elf, S. Deindl, Structure-guided approach to site-specific fluorophore labeling of the lac repressor

LacI. *PLoS One*. **13**, e0198416 (2018).

25. A. D. Edelstein, M. A. Tsuchida, N. Amodaj, H. Pinkard, R. D. Vale, N. Stuurman, Advanced methods of microscope control using μManager software. *J Biol Methods*. **1** (2014), doi:10.14440/jbm.2014.36.

26. S. Deindl, X. Zhuang, Monitoring conformational dynamics with single-molecule fluorescence energy transfer: applications in nucleosome remodeling. *Methods Enzymol.* **513**, 59–86 (2012).

27. A. Sabantsev, R. F. Levendosky, X. Zhuang, G. D. Bowman, S. Deindl, Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling. *Nat. Commun.* **10**, 1720 (2019).

28. J.-C. Olivo-Marin, Extraction of spots in biological images using multiscale products. *Pattern Recognit.* **35**, 1989–1996 (2002).

29. B. M. Sadler, A. Swami, Analysis of multiscale products for step detection and estimation. *IEEE Trans. Inf. Theory*. **45**, 1043–1051 (1999).

30. D. Garcia, Robust smoothing of gridded data in one and higher dimensions with missing values. *Comput. Stat. Data Anal.* **54**, 1167–1178 (2010).

31. M. Lindén, V. Ćurić, A. Boucharin, D. Fange, J. Elf, Simulated single molecule microscopy with SMeagol. *Bioinformatics*. **32**, 2394–2395 (2016).

32. S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*. **507**, 62–67 (2014).

33. X. Wu, D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu, D. B. Dadon, A. W. Cheng, A. E. Trevino, S. Konermann, S. Chen, R. Jaenisch, F. Zhang, P. A. Sharp, Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).

34. M. D. Szczelkun, M. S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, R. Seidel, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9798–9803 (2014).

35. M. Klein, B. Eslami-Mossallam, D. G. Arroyo, M. Depken, Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Rep.* **22**, 1413–1423 (2018).

36. R. B. Winter, P. H. von Hippel, Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor--operator interaction: equilibrium measurements. *Biochemistry*. **20**, 6948–6960 (1981).

37. D. Yang, A. Singh, H. Wu, R. Kroe-Barrett, Comparison of biosensor platforms in the

evaluation of high affinity antibody-antigen binding kinetics. *Anal. Biochem.* **508**, 78–96 (2016).

38. J. Elf, G.-W. Li, X. S. Xie, Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*. **316**, 1191–1194 (2007).