Analyzing the Dynamics of Single TBP-DNA-NC2 Complexes Using Hidden Markov Models

Nawid Zarrabi,[1,2] Peter Schluesche,[3] Michael Meisterernst,[4,5] Michael Börsch,[1,2] and Don C. Lamb[3,*]
[1]Physikalisches Institut, University of Stuttgart, Stuttgart, Baden-Württemberg, Germany; [2]Single-Molecule Microscopy Group, Jena University Hospital, Jena, Thuringia, Germany; [3]Department Chemie, Center for Nano Science, Center for Integrated Protein Science, and Nanosystems Initiative München, Ludwig-Maximilians-Universität Munich, Munich, Bavaria, Germany; [4]GSF-National Research Center for Environment and Health, Gene Expression, Munich, Bavaria, Germany; and [5]Institute of Molecular Tumor Biology, Faculty of Medicine, University of Muenster, Muenster, North Rhine-Westphalia, Germany

ABSTRACT Single-pair Förster resonance energy transfer (spFRET) has become an important tool for investigating conformational dynamics in biological systems. To extract dynamic information from the spFRET traces measured with total internal reflection fluorescence microscopy, we extended the hidden Markov model (HMM) approach. In our extended HMM analysis, we incorporated the photon-shot noise from camera-based systems into the HMM. Thus, the variance in Förster resonance energy transfer (FRET) efficiency of the various states, which is typically a fitted parameter, is explicitly included in the analysis estimated from the number of detected photons. It is also possible to include an additional broadening of the FRET state, which would then only reflect the inherent flexibility of the dynamic biological systems. This approach is useful when comparing the dynamics of individual molecules for which the total intensities vary significantly. We used spFRET with the extended HMM analysis to investigate the dynamics of TATA-box-binding protein (TBP) on promoter DNA in the presence of negative cofactor 2 (NC2). We compared the dynamics of two promoters as well as DNAs of different length and labeling location. For the adenovirus major late promoter, four FRET states were observed; three states correspond to different conformations of the DNA in the TBP-DNA-NC2 complex and a four-state model in which the complex has shifted along the DNA. The HMM analysis revealed that the states are connected via a linear, four-well model. For the H2B promoter, more complex dynamics were observed. By clustering the FRET states detected with the HMM analysis, we could compare the general dynamics observed for the two promoter sequences. We observed that the dynamics from a stretched DNA conformation to a bent conformation for the two promoters were similar, whereas the bent conformation of the TBP-DNA-NC2 complex for the H2B promoter is approximately three times more stable than for the adenovirus major late promoter.

INTRODUCTION

Protein biosynthesis begins with DNA transcription and RNA translation. Many regulatory and accessory factors exist to control the early steps during DNA transcription (1). For genes with TATA-box promoter sites in eukaryotic cells, the first step in DNA transcription is binding of the TATA-box-binding protein (TBP) (2) to the core promoter TATA boxes. This step is accompanied by deformation of the DNA strand, resulting in an 80° bend (3–9). This conformation change is believed to lead to the recruitment of additional general transcription factors (TFs) that form the preinitiation complex (10–12). In eukaryotic cells, positive cofactors play the major role in regulation of the DNA transcription process (13,14), whereas negative cofactors can sterically occlude association of other general TFs, which intermits the preinitiation complex formation and leads to repression of transcription. Some proteins, such as the evolutionarily conserved negative cofactor 2 (NC2) protein complex (15–17), have the capability to both suppress and enhance gene expression (18–22). Recent studies show that, in addition to steric interactions, dynamics also play an important role when investigating the interaction of TFs on DNA (23).

From x-ray crystallography experiments, it is known that NC2 forms a ringlike structure with TBP around the DNA (24), which can delocalize from TATA without leaving the DNA strand (23). Assuming that the formation of the TBP-NC2 subcomplex loosens the TBP-DNA interaction, the DNA is expected to relax into its original linear configuration. This stretched DNA conformation enables the TBP-NC2 subcomplex to move away from TATA and slide along the DNA strand. Using single-pair Förster resonance energy

Submitted June 5, 2018, and accepted for publication November 12, 2018.
*Correspondence: d.lamb@lmu.de
Editor: Keir Neuman.
https://doi.org/10.1016/j.bpj.2018.11.015

transfer (spFRET), we could directly visualize the conformational fluctuations of the TBP-NC2-DNA complex as well as movement of the TBP-NC2 complex along the DNA upon the binding of NC2 (23).

A wealth of information regarding the dynamics of the biomolecular system is buried within the spFRET traces. A detailed analysis can yield information regarding the number of states involved, which states can interconvert, and the transition rates between the individual states. One objective approach to extract this information from the spFRET data is the hidden Markov model (HMM) analysis. HMM was initially developed for speech-recognition algorithms but since then has been applied to many different fields and has become an important tool for analyzing spFRET data (25–37). A Markov model assumes discrete states with instantaneous transitions between the different states. In measurements with limited signal/noise ratio, the actual states become "hidden" because of the noise. The probability of measuring a particular value for a given state becomes distributed and, for the case of spFRET measurements, the distributions often overlap when multiple Förster resonance energy transfer (FRET) states are present. The power of the HMM analysis is that it can deal with overlapping probability distributions functions and, by optimizing a whole spFRET trace or family of traces, can reliably assign values to the hidden states. The HMM approach has been combined with maximal likelihood algorithms (25,26,29,31,37–39), variational Bayesian techniques (40), and empirical Bayesian methods (41).

One of the drawbacks of current HMM methods is that they assume a constant noise value for each state. However, the noise in spFRET traces is not necessarily constant. For example, the donor molecule can undergo partial quenching during the measurement. More importantly, a global HMM analysis is often desirable, but the total measured intensities of the donor and acceptor fluorophores and hence the signal/noise ratio will vary for different molecules. The signal/noise ratio for the individual molecules can be extracted from the raw data by estimating the total number of detected photons (37). With this approach, the shot noise does not need to be added as a parameter to the HMM analysis. To account for the diverse total intensities, particularly for a global analysis of hundreds of traces, we changed from the estimators commonly used in spFRET experiments, the donor and acceptor intensities, to the total intensity and proximity ratio (28).

In this work, we use the above-developed HMM analysis to investigate the number of conformations and the dynamics of the conformational changes induced by the formation of the TBP-NC2 subcomplex on DNA. SpFRET experiments were performed on immobilized molecules using total internal reflection fluorescence (TIRF) microscopy with a time resolution of 5 ms. Based on the new estimators, the HMM-assigned states could be determined and related to conformations of the TBP-NC2 subcomplex on the adenovirus major late (AdML) promoter sequence. Four states were observable. Three states corresponded to different conformations of the TBP-NC2-DNA complex with sharply bent DNA, partially bent DNA, and extended DNA. The fourth state is attributed to motion of the TBP-NC2 complex along the DNA. We also measured the dynamics of TBP-NC2 on the H2B promoter site, which revealed much richer dynamics. A comparison between the two promoter sites indicated that the dynamics were much more prevalent on the major late promoter site because of the lower stability of the bent conformation.

HMM

General introduction

A Markov model is described by a discrete number of states, q_i (where $i = 1...Q$), that the system can adopt. The system undergoes transitions between the different states, and the probability of a transition is constant, independent of the previous transitions. Hence, the dwell time in each state can be described by an exponential distribution. For a Q-state system, there are $Q \times (Q-1)$ independent transition probabilities k_{ij} of going from state i to state j, and together, they form the transition probability matrix $\boldsymbol{K}$. A schematic of a three-state Markov model is shown in Fig. 1 *a*. Typically, one is interested in which state the system is in as a function of time as well as the transition probabilities between states (the *upper sequence* in Fig. 1 *b*). In a hidden Markov system, the states themselves are no longer directly observable but are hidden within the noise of the system (the *lower sequence* in Fig. 1 *b*). The measured observable, $\mathbf{x}_t$, depends on the state the system is in (i.e. q_i) but its exact value will vary because of random noise. Thus, it is no longer possible to unequivocally back-assign the states q from $\mathbf{x}_t$.

The goal of an HMM analysis is to infer from the trajectory of observables (42) all system parameters of the underlying HMM. When the noise of the system in the different states is known, the probability density function for possible values of $\mathbf{x}_t$ given that the system is in state q_i can be calculated and is referred to as the emission function $f_i(\mathbf{x}|q_i)$. For example, the emission functions of a three-state HMM shown are given in Fig. 1 *c*. A measured value of 0.35 is possible from all three states, but the probability of it arising from state 2 is much higher than that of state 1 or 3. Using the emission functions, we can estimate the most likely HMM that describes the measured time series.

The log-likelihood function, log L

The key tool used to determine the most probable set of system parameter values from the observable data is the log-likelihood function, log L. The likelihood function, L, calculates the probability of measuring the measured data set from the given set of parameters and is given by the following:

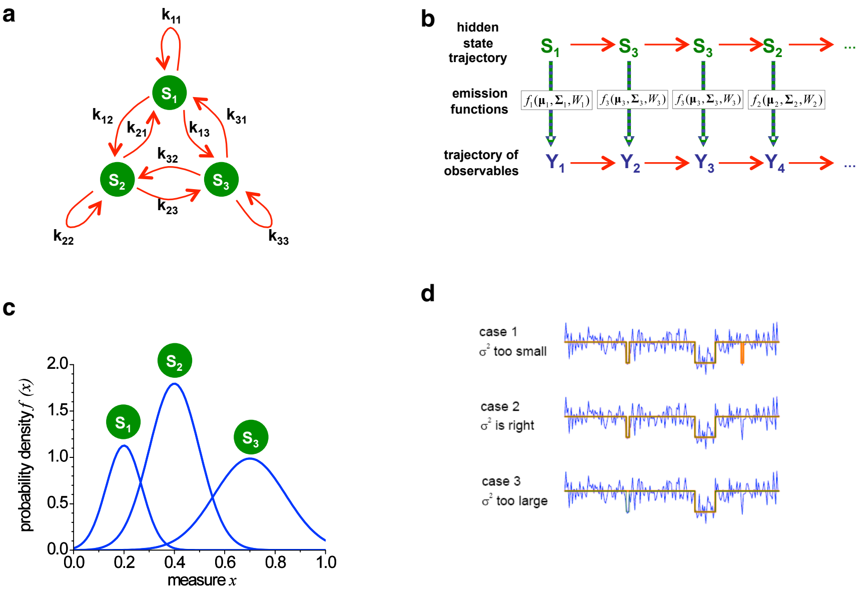
a
k11
k12
k21
k13
k31
k32
k23
k22
k33
S1
S2
S3
b
hidden state trajectory
S1
S3
S3
S2
...
emission functions
$f_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, W_1)$
$f_3(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, W_3)$
$f_3(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, W_3)$
$f_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, W_2)$
trajectory of observables
Y1
Y2
Y3
Y4
...
c
probability density f (x)
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
measure x
S1
S2
S3
d
case 1
σ² too small
case 2
σ² is right
case 3
σ² too large

FIGURE 1 Principles of a hidden Markov model (HMM). (*a*) A scheme of a Markov model with three states (S1, S2, and S3) and their corresponding transition probabilities k_{ij} is shown. (*b*) A possible realization of a Markov chain is shown. For a series of states, the model emits an observable x_i based upon the actual state S_i and the probability density function f_i. This yields a trajectory of observables. In an HMM, the states themselves are not directly determinable because of noise and are hence hidden. The HMM analysis finds the parameters of the emission functions that best describe the data. From the given parameters, the Viterbi path is determined, which gives the most likely trajectory of hidden states that describes the observed data. (*c*) Three Gaussian emission functions with different means, variances, and occurrences are shown ($\mu_q = \{0.2, 0.4, 0.7\}$, $\sigma_q = \{0.05, 0.1, 0.2\}$, $W_q = \{0.20, 0.45, 0.35\}$). (*d*) The influence of the variance on the state assignments extracted during an HMM analysis of an spFRET trace is depicted. When the variance is too small, the HMM introduces too many transitions, which yields a reduction in the duration of the individual states (case 1). When the variance is too large, state transitions are missed, which yields exaggerated state durations (case 3). When the variance models the broadness of a state correctly, the state durations are consistent with the transition probabilities (case 2). To see this figure in color, go online.

$$L = p\left(\{\mathbf{x}\} \mid \{\mu_q, \sigma_q, w_{q,t}\}\right) = \prod_{t=1}^{T} \prod_{q=1}^{Q} \left[f_q\left(\mathbf{x}_t \mid \mu_q, \sigma_q\right)\right]^{w_{q,t}}, \tag{1}$$

where μ_q, σ_q^2, and $w_{q,t}$, are the mean FRET value of state q, its covariance, and the probability of the data point x_t corresponding to q, respectively. T denotes the number of data points in the measured trajectory. To avoid underflow errors during determination of the likelihood function, it is advantageous to calculate the logarithm of the likelihood function:

$$\log L = \sum_{t=1}^{T} \sum_{q=1}^{Q} w_{q,t} \log\left(f_q\left(x_t \middle| \boldsymbol{\mu}_\mathbf{q}, \boldsymbol{\sigma}_\mathbf{q}\right)\right). \tag{2}$$

Because the logarithm is a monotonic increasing function, maximization of the log-likelihood function is equivalent to finding the maximum of the likelihood function. By defining the log-likelihood function, determination of the best parameter set of a given model for producing a given data set is reduced to an optimization problem. Because the log-likelihood function can depend on several parameters, a multidimensional optimization algorithm needs to be used (25,43). An algorithm that has been shown to converge rapidly is the forward-backward algorithm (38,44), which we have implemented in our approach.

Once the optimal model parameter values are obtained, the hidden-state trajectory itself can be reconstructed by the Viterbi algorithm (45), which assigns every time-binned data point to its most likely state. The main tasks in applying HMMs to spFRET data are now to choose the appropriate emission functions and derive the analytical estimators for the parameter determination (25,26,46).

Estimators for the emission functions

Often, Gaussian distributions are used as emission functions to model the probability density function of a state. The parameter estimators for the mean, the covariance, and the fraction of time spent in the q state is given by the following:

$$\text{mean}: \widehat{\mu}_q = \frac{\sum_t w_{q,t} \mathbf{x}_t}{\sum_t w_{q,t}}, \tag{3}$$

$$\text{covariance matrix}: \sigma_q^2 = \frac{\sum_t w_{q,t} \mathbf{x}_t^2}{\sum_t w_{q,t}} - \widehat{\mu}_q^2, \text{ and} \tag{4}$$

$$\text{fraction}: W_q = \frac{1}{T} \sum_{t=1}^{T} w_{q,t}. \tag{5}$$

$w_{q,t}$ is called the "responsibility matrix" or the "posterior probabilities" and depends in turn on the model parameters:

$$w_{q,t} = \frac{W_q \; f_q\left(x_t \middle| \boldsymbol{\mu}_\mathbf{q}, \boldsymbol{\sigma}_\mathbf{q}\right)}{\sum\limits_{q=1}^{Q} W_q \; f_q\left(x_t \middle| \boldsymbol{\mu}_\mathbf{q}, \boldsymbol{\sigma}_\mathbf{q}\right)}. \tag{6}$$

Because Eq. 3, 4, 5, and 6 are interdependent, the parameters cannot be determined directly but have to be refined iteratively.

Incorporation of the transition matrix

With the estimators above, the likelihood is increased by optimizing the parameters of the emission functions. The assignments of the data points to the hidden states are optimized by tuning the posterior probabilities. These posterior probabilities are connected to the emission functions of their preceding and subsequent states by the transition probabilities. Briefly, we calculate the probability of being in state i at the time step t, $\alpha_t(i)$, as the product of the transition probability of going from state j to state i, the probability of being in state j at time step $t-1$, and the emission function of state i (forward estimate):

$$\alpha_t(i) = \sum_{j=1}^{Q} \alpha_q(j)\; k_{ij} f_i(x_t). \tag{7}$$

Hence, $\alpha_t(j)$ can be iteratively determined. Likewise, we can calculate the probability of being in state j at the time step t, $\beta_t(j)$, by calculating backward from the end of the trace (backward estimate):

$$\beta_t(i) = \sum_{j=1}^{Q} \beta_{t+1}(j)\; k_{ij} f_j(x_t). \tag{8}$$

The total probability of being in state i at the time step t is then given by the following:

$$w_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{Q} \alpha_t(j)\beta_t(j)}. \tag{9}$$

From the forward and backward estimates, we can also determine an estimate for the transition probability matrix:

$$\widehat{k}_{ij} = \frac{\sum_{t=1}^{T} \alpha_t(i) k_{ij} f_j(x_{t+1}) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)}. \tag{10}$$

To begin the analysis, initial estimates for the parameters (i.e., μ_q, σ_q, and $\boldsymbol{K}$) are entered. From the initial values, a first likelihood value and posterior probabilities are estimated. The parameters are then adjusted to maximize the log likelihood. For optimization, we used the forward-backward algorithm, which is an implementation of an "expectation-maximization algorithm" (47,48). More detailed introductions to HMMs can be found in (44,49,50).

Observables in single-molecule FRET data

Everything discussed up to this point is independent of the type of data analyzed using HMM. In this work, we apply an HMM analysis to spFRET experiments on TBP (from *Saccharomyces cerevisiae*) interacting with DNA. TBP was labeled with the donor fluorophore, and DNA was labeled with the acceptor fluorophore (Fig. 2 *a*). A schematic of a single-molecule experiment with TBP bound to DNA immobilized on a PEGylated surface is shown in Fig. 2 *b*. In spFRET experiments, the fluorescence intensities of two fluorophores, the donor and acceptor molecules, are measured as a function of time (Fig. 2 *c*). The proximity ratio, E_{PR}, which is related to the FRET efficiency, contains information regarding the separation of the two fluorophores and hence information about the conformation of the complex. It can be calculated directly from the experimentally accessible fluorescence intensity traces of the donor and acceptor molecules, I_D and I_A, respectively, using Eq. 11:

$$E_{PR} = \frac{I_A}{I_D + I_A}. \tag{11}$$

E_{PR} provides information regarding the interfluorophore distance, and the total intensity

$$I_T = I_D + I_A \tag{12}$$

provides information on the accuracy of the measured proximity ratio. To convert the proximity ratio into FRET efficiency, differences in the detection efficiencies η of both detection channels as well as unequal fluorescence quantum yields ϕ of the fluorophores need to be accounted for. When the detection correction factor γ is known, the FRET efficiency E_{FRET} is given by the following:

$$E_{FRET} = \frac{1}{\gamma\left(E_{PR}^{-1} - 1\right) + 1} \text{ where } \gamma = \frac{\phi_A\ \eta_A}{\phi_D\ \eta_D}. \tag{13}$$

In general, we transform the variables I_D and I_A into a new pair of variables, E_{PR} and I_T. When performing spFRET experiments using single-photon counting detection, a Poisson distribution will describe both I_D and I_A. For moderate count rates (greater than ~50 photons per time bin), the Poisson distribution can be well approximated by a Gaussian distribution. The mean and variance of the Gaussian distribution are set equal to the mean (which is also the variance) of the Poisson distribution for the respective channel. The probability distribution function (pdf) for the total fluorescence intensity is then also approximated by a Gaussian distribution with a maximal value of I_T, and the variance is given by the following:

$$\sigma_T^2 = \sigma_D^2 + \sigma_A^2 = I_D + I_A = I_T. \tag{14}$$

The proximity ratio, E_{PR}, can also be approximated by a Gaussian distribution (see Supporting Materials and Methods), yielding Eq. 15 for the mean and Eq. 16 for the variance:

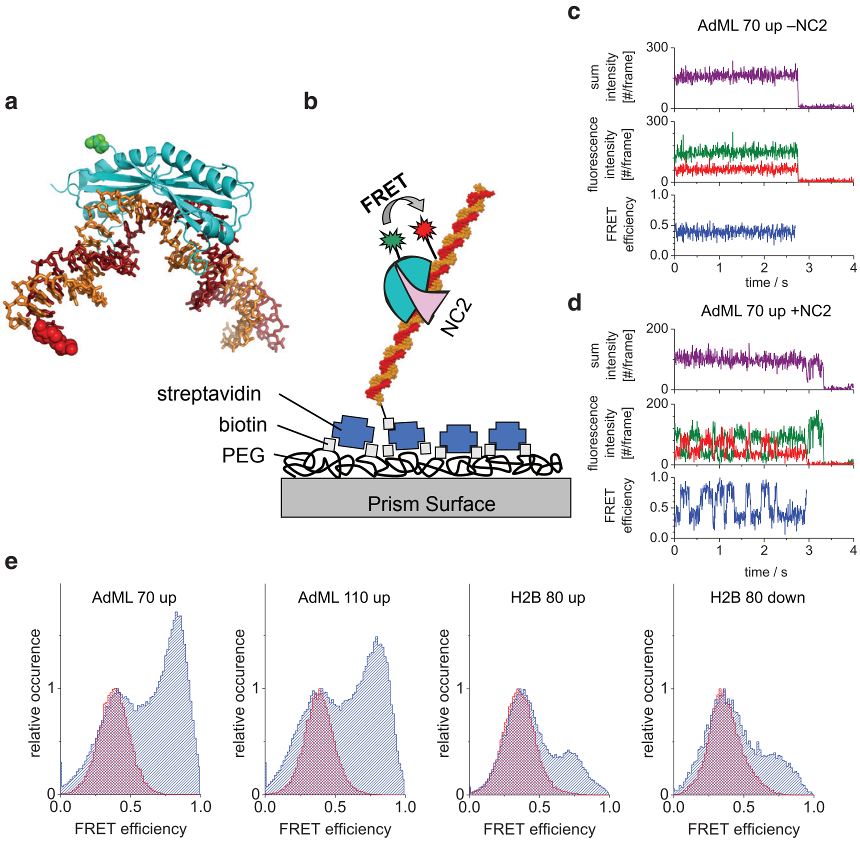

FIGURE 2 SpFRET data of TBP-DNA in the absence and presence of NC2. (*a*) The crystal structure of TBP-DNA complex (Protein Data Bank: 1RM1) is shown. The labeling positions of the TBP (*green*) and DNA (*red*) are shown as spheres. For display purposes, the DNA has been extended downstream. (*b*) A scheme of the TBP-NC2-DNA complex immobilized to a functionalized, PEGylated glass surface over a streptavidin-biotin linkage is shown. TBP is specifically labeled with a donor fluorophore, and the DNA is specifically labeled with the acceptor fluorophore. The sample is excited by the evanescent wave generated at the prism/buffer interface by total internal reflection. (*c* and *d*) Single-molecule FRET trajectories from TBP bound to 70-bp DNA containing the AdML TATA box are plotted in the (*c*) absence and (*d*) presence of NC2 (*purple*: the scaled total intensity, $I_T^{'} = \gamma I_D + I_A$, where γ is the detection correction factor; *green*: donor intensity; *red*: acceptor intensity; and *blue*: FRET efficiency). In the absence of NC2, a stable conformation of the TBP-DNA complex is observed, resulting in a constant FRET efficiency of ~0.40. After the addition of NC2, the FRET efficiency fluctuates between different states. (*e*) Histograms of the frame-by-frame FRET efficiencies from the complete data set of the four different constructs measured are shown. Red histograms represent before the addition of NC2, whereas blue histograms represent after the addition of NC2.

$$\mu_{E_{PR}} = \frac{\mu_{I_A}}{\mu_{I_D} + \mu_{I_A}} = \frac{\mu_{I_A}}{\mu_{I_T}} \quad \text{and} \qquad (15)$$

$$\sigma_{E_{PR}} = \frac{\mu_{E_{PR}}\left(1 - \mu_{E_{PR}}\right)}{\mu_I}. \qquad (16)$$

The expectation value of the total intensity appears in the denominator of the variance, indicating that the total fluorescence intensity is a direct measure for the accuracy of the determined apparent FRET efficiency. The limited number of photons per time bin reduces the accuracy with which the FRET efficiency can be estimated. Hence, the corresponding emission function becomes broader because of shot noise, which can be quantified by Eq. 16. As long as the total fluorescence intensity is constant in time, the shot-noise broadening of the emission functions is also constant and can be included in the time-invariant covariance matrix σ_i provided by the classical HMM approach.

Often, single-molecule experiments are performed using wide-field illumination (typically with TIRF excitation) and a charged-coupled device (CCD) as a detector as we used in our studies of the dynamics of TBP-DNA complexes upon the binding of NC2 (23). Such an approach has the advantage that many molecules can be investigated simultaneously. However, CCD cameras are not photon-counting devices. With proper calibration, camera counts can be converted into an approximate number of detected photons (see Supporting Materials and Methods). However, the influence of additional noise sources needs to be considered. For the electron-multiplying CCD (EMCCD) in our setup, the variance in the fluorescence intensity is increased by a factor of two over the shot noise, as has been discussed in detail elsewhere (51,52). This uncertainty can be included in the variance of the FRET efficiency, which can still be approximated, in this case, by a Gaussian distribution (see Supporting Materials and Methods).

When performing a global analysis on a collection of spFRET traces, the difference in the variances for the individual molecules as well as time-dependent changes in the total fluorescence intensity during a time trace needs to be accounted for. An incorrect variance for an HMM state will lead to errors in the recognition of transitions in the spFRET data. When the variance of the HMM is too small, noise fluctuations will also be incorrectly recognized as transitions (Fig. 1 *d*). Similarly, when the variance of the HMM is too large, rapid transitions between states will be ignored. Therefore, we have extended the standard HMM approach by introducing time-dependent weights to the classical parameter estimators, assuming that the degree of broadening is provided by the measured total fluorescence intensity I_T.

Weighted HMMs

Incorporating photon-counting statistics as well as other known noise sources into the HMM analysis allows for a more accurate determination of the dynamics measured using spFRET. As the total intensity of an spFRET signal can drop because of partial, dynamic quenching of the donor, the variance of the data used in the emission functions needs to be variable on a frame-by-frame basis. Because this is not reasonable, we used the information available from photon statistics to estimate the broadening of FRET levels due to shot noise, and this can be done in a frame-wise manner.

Shot-noise broadening of FRET levels

A stable conformational state of a protein can be described well by a single pdf with two parameters: a mean and a variance. The variance includes the inherent amount of fluctuations within this conformation and should be independent of the measurement method. In addition, the pdf of spFRET values is broadened by the limited number of detected photons, which is often the dominating factor. Hence, it is necessary to combine the inherent uncertainty of the spFRET state with a second pdf that accounts for the uncertainty of the measurement. The pdfs for FRET efficiencies derived from shot-noise broadened fluorescence count rates follow a β-function (53–55). However, for count rates typically obtained in single-molecule experiments, the β-function can be well approximated by a Gaussian distribution (25,53).

Weighted maximal likelihood estimators

When both factors contributing to the pdfs are Gaussian functions, the resulting emission function is again a Gaussian distribution with a mean and variance that are given by the sum of mean and variances of individual Gaussian distributions, respectively. Therefore, the classical HMM approach can be used by adding the variance of the broadened data point $\sigma_{x_t}^2$ to the inherent variance of the FRET state, σ_q^2:

$$\begin{aligned}\tilde{f}_q\left(x_t \middle| \mu_q, \sigma_q^2, \sigma_{x_t}^2\right) &= f_q\left(x_t \middle| \mu_q, \sigma_q^2 + \sigma_{x_t}^2\right) \\ &= \frac{1}{\sqrt{2\pi\left(\sigma_q^2 + \sigma_{x_t}^2\right)}} \\ &\quad \times exp\left(-\frac{\left(x_t - \mu_q\right)^2}{2\left(\sigma_q^2 + \sigma_{x_t}^2\right)}\right).\end{aligned} \tag{17}$$

Accordingly, it is possible to account for the changes in the variances caused by the high diversity in the total fluorescence intensity of different molecules while maintaining a constant variance for the FRET efficiency of the state itself. The respective log-likelihood function log L_q for one state yields the following form and is the basis for obtaining the new estimator functions:

$$\begin{aligned}L_q\left(\mu_q, \sigma_q^2, \sigma_{x_t}^2, w_{qt} \middle| \{x_t\}\right) = &-\frac{N}{2} ln(2\pi) \\ &- \sum_{t=1}^{T}\left(\frac{w_{qt}\left(x_t - \mu_q\right)^2}{2\left(\sigma_q^2 + \sigma_{x_t}^2\right)}\right. \\ &\left. + \frac{1}{2} ln\left(\sigma_q^2 + \sigma_{x_t}^2\right)\right).\end{aligned} \tag{18}$$

To derive initial expressions for the estimators of the HMM parameters, we set the derivate of the log-likelihood function, with respect to the desired parameters, equal to zero. Solving these equations for the parameters leads directly to the expressions for the estimators.

The newly introduced shot-noise variance leads to an additional weighting and hence to an expansion of the classical estimator. For the estimator of the mean value, the new parameter $\sigma_{x_t}^2$ appears as an additional weighting factor for the observable x and cannot be eliminated:

$$\widehat{\mu}_q = \frac{\sum_{t=1}^{T} \frac{w_{qt} x_t}{\sigma_q^2 + \sigma_{x_t}^2}}{\sum_{t=1}^{T} \frac{w_{qt}}{\sigma_q^2 + \sigma_{x_t}^2}}. \tag{19}$$

As a consequence, solving the derivative of the likelihood function with respect to the variance $\widehat{\sigma}_q^2$ for zero, given by:

$$\sum_{t=1}^{T} \frac{w_{qt}\left(\left(x_t - \mu_q\right)^2 - \left(\widehat{\sigma}_q^2 + \sigma_{x_t}^2\right)\right)}{\left(\widehat{\sigma}_q^2 + \sigma_{x_t}^2\right)^2} = 0, \tag{20}$$

is no longer solvable analytically. Fortunately, there are powerful methods that quickly find zeros for a one-dimensional function so that the computational time is only moderately increased (56).

MATERIALS AND METHODS

Sample preparation and labeling

A recombinant mutant of TBP from *S. cerevisiae* with a single cysteine introduced at position 61 was covalently labeled with the donor fluorophore, Atto532 (57,58). As shown previously, fluorescent labeling of the protein did not affect the functionality of the TBP (57,58). The FRET acceptor Atto647 was attached to a DNA sequence containing a TATA box for TBP binding and a biotin anchor for the immobilization on a streptavidin-coated quartz-prism surface (Fig. 2, *a* and *b*). The FRET signal from individual TBP-DNA complexes is measured until one of the fluorophores photobleaches (typically ~1 s, with a 5-ms time resolution). All samples were preincubated with the general initiation factor TFIIA recombinant homolog TOA to ensure the formation of stable TBP-DNA complexes forming the ~80° DNA bend (59) and proper orientation (60). Fluorescence intensities of both the donor and acceptor fluorophores were recorded simultaneously by an EMCCD camera (DV887-BV iXon$^+$; Andor Technology, Belfast, Northern Ireland) on our custom-build TIRF setup for single-molecule FRET. Details of the biochemical procedures and experimental conditions have been described previously (23).

For the studies reported here, a total of four different DNAs (summarized in Fig. 3) were investigated. Two double-stranded DNAs contained the AdML promoter and were fluorescently labeled 11 basepairs (bp) upstream from the TATA box with the acceptor molecule (Atto647). They differed both in length (70 vs. 110 bps) and in the position of the attachment point to the surface. In addition, two double-stranded DNAs (80 bp in length) containing the TATA box from the H2B-J promoter were investigated. One of the H2B-J constructs was labeled 12 bp upstream from the TATA box, whereas the other DNA strand was labeled 13 bp downstream from the start of the TATA box.

The Förster radius for the dye pair used is $R_0 = 6.0$ nm (according to supplier), making it very sensitive to conformational fluctuations of the DNA and movement of the TBP-NC2 complex along the DNA. Between 103 and 431 TBP-DNA complexes were analyzed bound to the AdML promoters, and 55–279 complexes were measured bound to the H2B-J promoters.

EMCCD shot-noise corrections for HMM

To estimate the number of detected photons from the EMCCD measurements, the output of the camera has to be corrected for offset, gain, and the analog-to-digital conversion factor of the camera. In addition to the shot noise, which follows a Poisson distribution, other noise factors from the camera need to be incorporated. A detailed description of noise coming from EMCCD cameras can be found in (51,52). The most important source of additional noise comes from the on-chip gain of the EMCCD, which broadens the variance of the intensity by a factor of two (or the SD by $\sqrt{2}$). Further information is given in the Supporting Materials and Methods. Background correction will also increase the uncertainty of the measurement of the signal. However, because of the large number of pixels used to determine the level of the background signal in the vicinity of each individual complex, the additional uncertainty due to the background correction is negligible.

HMM analysis

Two variations of the HMM analysis were performed. In the first case, molecule-wise, each trace was fitted individually with up to 10 different states. In the second case, a global fit was performed in which the same FRET states and transition rates were used to fit the entire data set. For the global analysis, analyses were performed with different HMMs containing 1 to 10 hidden states. For the HMM analysis, we used the MATLAB toolbox of Murphy (61), which includes the forward-backward and Viterbi algorithm and supports HMM with mixtures of Gaussian outputs. Details of the analysis using different numbers of FRET states are given in the Supporting Materials and Methods.

RESULTS

Monte Carlo test of the new estimators

Before analyzing the experimental spFRET data, we tested the reliability of our extended hidden Markov approach in handling additional shot-noise broadening by performing Monte Carlo simulations. Based on a normal distribution of FRET efficiencies representing a stable conformational state, a random trajectory with a length of 50,000 data points was created to mimic low inherent fluctuations of the FRET efficiency due to the flexibility of the observed protein. Mean total fluorescence intensities and average FRET efficiencies were introduced, and donor and acceptor fluorescence intensity trajectories were determined. Every point of this trajectory pair was finally replaced by a stretched Poissonian random number, taking the original value as its mean and a stretch factor of 2 to incorporate the additional noise component generated by the gain of an EMCCD camera (51,52).

The results of the simulations are summarized in Fig. S1. The estimators extracted the time-dependent shot noise from the data and resolved the correct mean values and inherent variances even at very low photon count rates. The Gaussian approximation of the β-function was performed such that their mean values and therefore their maximal likelihood estimators were identical. Slight deviations were observed for the estimation of the variance at very low count rates and broad distributions. For comparison, the inherent amount of fluctuations of the experimental single-molecule FRET data had an SD of ~0.1 with count

DNA 1 (AdML 70 bp upstream):

```
5′-GCCACGTGACCGGGTGTTCCTGAAGGGGGGCTATAAAAGGGGGTGGGGGCGCGTTCGTCCTCACTCTCTT-Biotin
3′-CGGTGCACTGGCCCACAAGGACTTCCCCCCGATATTTTCCCCCACCCCCGCGCAAGCAGGAGTGAGAGAA
```

DNA 2 (AdML 110 bp upstream):

```
Biotin-5′-GCCACGTGACCGGGTGTTCCTGAAGGGGGGCTATAAAAGGGGGTGGGGGCGCGTTCGTCCTCACTCTCTT...
       3′-CGGTGCACTGGCCCACAAGGACTTCCCCCCGATATTTTCCCCCACCCCCGCGCAAGCAGGAGTGAGAGAA...
                                     ...CCGCATCGCTGTCTGCGAGGGCCAGCTGTTGGGGTGAGTA-3′
                                     ...GGCGTAGCGACAGACGCTCCCGGTCGACAACCCCACTCAT-5′
```

DNA 3 (H2B-J 80 bp upstream):

```
Biotin-5′-CTTCACCTTATTTGCATAAGCGATTCTATATAAAAGCGCCTTGTCATACCCTGCTCACGCTGTTTTTCCTTTTCGTTGGC-3′
       3′-GAAGTGGAATAAACGTATTCGCTAAGATATATTTTCGCGGAACATGATGGGACGAGTGCGACAAAAAGGAAAAGCAACCG-5′
```

DNA 4 (H2B-J 80 bp downstream):

```
Biotin-5′-CTTCACCTTATTTGCATAAGCGATTCTATATAAAAGCGCCTTGTCATACCCTGCTCACGCTGTTTTTCCTTTTCGTTGGC
       3′-GAAGTGGAATAAACGTATTCGCTAAGATATATTTTCGCGGAACATGATGGGACGAGTGCGACAAAAAGGAAAAGCAACCG
```

FIGURE 3 DNA sequences used in this study. DNA 1 and DNA 2 contain the adenovirus major later promoter sequence and DNA 3 and DNA 4 contain the H2B-J promoter sequence. The TATA box is highlighted in grey and the thymine base to which the acceptor fluorophore was attached is highlighted in red. The location of the biotin tag used to immobilize the DNA for the spFRET experiments is also shown. To see this figure in color, go online.

rates of more than 50 counts per data point. The success of the new estimators, even at low signal/noise values, is notable because the shapes of the simulated distributions deviate strongly under those conditions.

Comparison to HMM without incorporation of photon counts

To compare our extended version of the HMM analysis with the standard approach (i.e., without explicit incorporation of the camera noise), we performed a simulation of a Markov model using four states. Details are given in the Supporting Materials and Methods and Table S1. An HMM analysis was performed, once with the width as a single free parameter and one in which the inherent noise of the FRET state due to shot noise and camera noise was calculated for each data point. After the learning algorithm determined the hidden Markov parameters, each frame of the simulated data was assigned with the best-fitting state by the Viterbi path, with the help of the learned parameters. A comparison of the two analyses is given in Table S1. Incorporation of the camera noise into the analysis slightly improves the already high accuracy, dropping the fraction of frames that are incorrectly assigned from 4 to 3%. Both approaches find the values of the four FRET states with high accuracy. The widths of the FRET states returned by the two HMM analyses are not comparable because the standard HMM is fitting the camera noise (the major contribution to the noise in this simulation), whereas the inherent noise of the FRET states is reliably returned with the new HMM model. Interestingly, there is a difference in the dwell times returned from the two different approaches. By incorporating the camera noise directly in the analysis, the probability of noise being misinterpreted as a transition decreases, yielding more accurate rates.

SpFRET measurements of TBP-DNA in the absence and presence of NC2

Having verified the reliability of our HMM analysis on simulated data, we now apply it to real data. Fig. 2 shows results from spFRET experiments on TBP-DNA complexes in the absence and presence of NC2. As observed in previous experiments (23), spFRET measurements typically showed a constant ("steady state") FRET efficiency with E_{PR} ~0.4 of the TBP-DNA complex before addition of NC2, demonstrating the stable binding of TBP to the TATA box (Fig. 2 *c*) for the 70-bp AdML promoter DNA. After addition of NC2, the complex becomes dynamic, and fluctuations between distinct FRET states are observed. A typical FRET trace is shown in Fig. 2 *d* for the 70-bp AdML promoter DNA. A histogram of the FRET efficiency for the individual frames of the 70-bp AdML sample (frame-wise histogram) is shown in Fig. 2 *e* before (*red*) and after (*blue*) the addition of NC2.

One of the advantages of the modified HMM that we present here is its ability to account for the shot noise within the spFRET data. Fig. S2 shows histograms of the total intensity (in photons) per frame for the different measurements. The measured total intensities varied by more than a factor of four, from 50 photons per frame to more than 200 photons per frame. This broad distribution of intensities indicates the heterogeneities of single-molecule experiments and how important it is to correct for shot noise when performing a global analysis. This can be circumvented by using an intensity window for selection of traces to be analyzed further. However, variations in the total intensity can also happen within an spFRET trace, for example, when the donor molecule is partially quenched. Because the proximity ratio is calculated from the ratio of intensity in the acceptor channel to the total intensity within a frame, donor quenching will not strongly influence the calculated proximity ratio, but the uncertainty will be increased. Such an example is shown in Fig. S3. Because the uncertainty due to shot noise is determined frame by frame, the modified HMM is able to assign a constant low-FRET state during transient quenching of the donor, although the FRET signal strongly fluctuates. Whether photophysics of the acceptor is leading to apparent fluctuations in FRET efficiency can be monitored using millisecond alternating-laser excitation (62). Because acceptor blinking was not typically observed for these constructs (23), we forwent alternating-laser excitation measurements and opted for high data collection rates to improve the kinetics analysis.

To test how well the noise characterization of the camera explains our data, we have plotted the mean and variance of the donor and acceptor signals in Fig. S4 for TBP-DNA (AdML promoter 70-bp DNA) in the absence and presence of NC2. The theoretical expectations, assuming Poissonian statistics for photon counting corrected for the additional EMCCD noise, are plotted as lines. In the absence of NC2, the experimental data are well described by the theoretical expression, indicating that the conformation of the TBP-DNA complexes with respect to the given labeling positions is static, and the shot-noise calculations of the measurement uncertainty are appropriate. In contrast, the individual traces demonstrate higher variances than expected from the shot noise alone because of the dynamics in the presence of NC2.

Dynamics of the FRET-labeled TBP-NC2 complex, two-state model

Histograms of the frame-wise spFRET efficiency for the measured complexes are shown in Fig. 2 *e*. In the presence of NC2 (shown in *blue*), two populations are observable with FRET efficiencies of ~0.40 and ~0.80. The subpopulation with a FRET efficiency of 0.40 is similar in conformation to TBP-DNA in the absence of NC2, whereas the higher FRET state is attributed to a conformational change of the DNA in the TBP-NC2-DNA complex (23). Fitting the data using a two-state HMM model, we obtained FRET values and equilibrium coefficients, K, of 0.39, 0.78, and $K_{40/80} = 1.32$ for the 70-bp DNA construct and 0.36, 0.75, and $K_{40/80} = 1.21$ for the 110-bp DNA construct. The results are in excellent agreement with each other, indicating that the dynamics are independent of the length of the DNA and do not depend on which end of the DNA is anchored to the surface.

The FRET histogram for the H2B-J promoter upstream-labeled construct after addition of NC2 shows clearly different dynamics. The same two dominant FRET subpopulations are observed, but the original TBP-DNA conformation is more stable and populated much more often. The results of the two-state HMM yielded FRET values and equilibrium coefficients of 0.36, 0.71, and $K_{40/70} = 0.52$ for the upstream- labeled H2B-J promoter. Interestingly, measurements with the downstream-labeled H2B-J promoter also show fluctuations in the FRET signal from ~0.4 to higher FRET values. The HMM analysis yielded 0.33, 0.68, and $K_{40/70} = 0.40$. The fact that FRET increases upon addition of NC2 for both the upstream- and downstream-labeled constructs confirms that the observed dynamics are coming from conformational changes of the DNA and not from motion of the TBP-NC2 complex along the DNA.

For a traditional spFRET analysis, the two FRET states would be fitted with Gaussian distributions, the FRET values would be given by the peak values, and the equilibrium coefficient would be given by the respective areas. For the AdML promoter constructs, these lead to FRET values of 0.47, 0.83, and an equilibrium coefficient of $K_{40/80} = 0.71$ for the 70-bp DNA (Fig. 4 *a*) and 0.43, 0.80, and $K_{40/80} = 0.71$ for the 110-bp DNA. The Gaussian fits are consistent between the two AdML DNAs again verifying that the DNAs have similar dynamics and that the experiments are reproducible. The FRET values extracted from the peak of the Gaussian distributions approximately agree with the HMM data, but the equilibrium coefficients are significantly different. This is due to the width of the Gaussian distributions in which a significant population of the 0.40 FRET state overlaps with the 0.80 FRET state. Because the HMM uses not only the FRET values but also the time information in the spFRET traces, it is capable of distinguishing overlapping populations. The difficulty in this case is that the two-state model is an oversimplification of the actual dynamics. This leads to an increase in the width of the 0.40 FRET population and hence incorrect results when fitting with two Gaussians.

Interestingly, for the H2B-J promoter DNAs in which the 0.40 FRET state is more prevalent, the results from Gaussian fits are more comparable with the HMM analysis. The FRET values and equilibrium coefficients are 0.39, 0.73, and $K_{40/70} = 0.35$ (Fig. 4 *b*) and 0.37, 0.73, and $K_{40/70} = 0.37$ for the upstream and downstream labels,

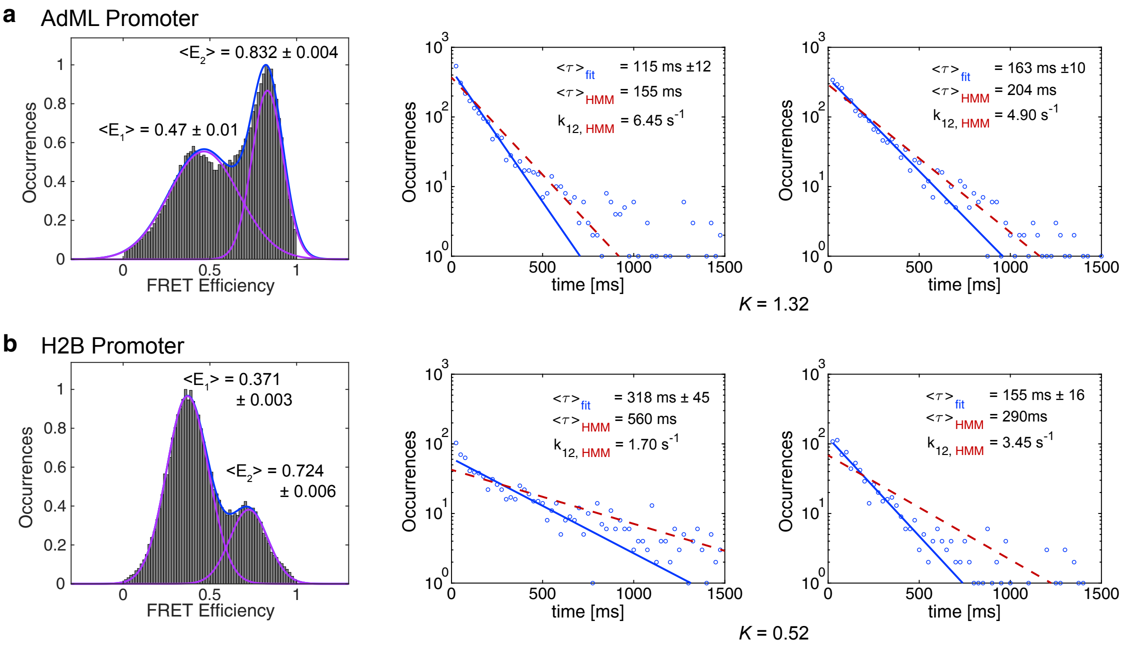

FIGURE 4 Two-well analysis of the TBP-DNA-NC2 dynamics. The results of a two-well analysis of the TBP-DNA-NC2 data are shown for the (*a*) 70-bp upstream-labeled AdML promoter DNA (DNA 1) and (*b*) the 80-bp upstream-labeled H2B promoter DNA (DNA 3). Left panels: Frame-wise spFRET histograms are shown for the respective measurements. Two dominant populations are observed, and the equilibrium coefficient can be estimated by fitting to the two populations to Gaussian distributions. Middle and right panels: from the two-state HMM analysis, the duration of each state before undergoing a transition was calculated from the Viterbi path determined for each trace from the parameters of the global HMM. The survival probability was fit to an exponential function (*solid line*, *red*). In addition, the rates from the HMM are shown (*dashed line*, *blue*). The given equilibrium coefficients are determined from the ratio of the rates determined from the HMM analysis. To see this figure in color, go online.

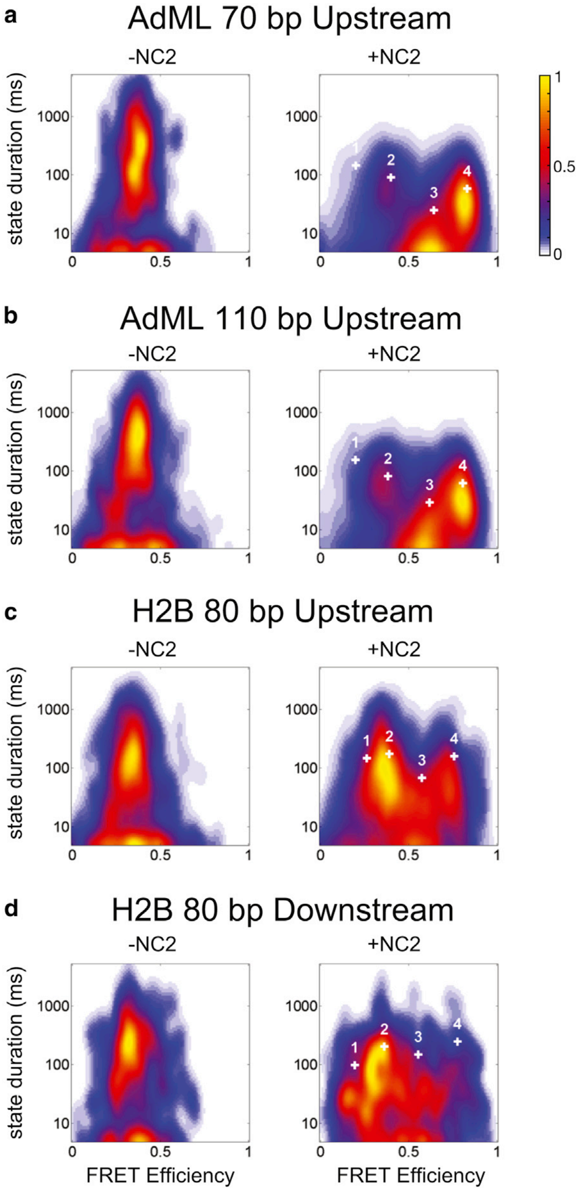

FIGURE 5 Molecule-wise HMM analysis. Displayed are 2D histograms of the FRET efficiency of the states derived by the molecule-wise HMMs versus the length of time the complex resided in that state before transitioning to another state or photobleaching. The plots are shown for TBP bound to a (*a*) 70-bp upstream-labeled DNA containing the AdML TATA box (DNA 1), (*b*) 110-bp upstream-labeled DNA containing the AdML TATA box (DNA 2), (*c*) an 80-bp upstream-labeled DNA containing the H2B TATA box (DNA 3), and (*d*) an 80-bp downstream-labeled DNA containing the H2B TATA box (DNA 4). Results for TBP-DNA in the absence of NC2 are shown in the left panels, and results for TBP-DNA in the presence of NC2 are shown in the right panels. The white plus signs show the FRET efficiencies returned from the global four-well HMM analysis. The histograms are normalized to the maximal number of transitions. The color scale is shown in the upper right corner.

respectively. However, the broader population is still overestimated by the Gaussian analysis. This indicates the advantages of using an HMM analysis even for a simple evaluation.

Determination of the number of states

Depending on the quality of the data and in how much detail one wishes to analyze the data, HMM can be used to investigate how many FRET states are present. Ideally, it would be convenient when the HMM analysis would directly yield how many significantly different FRET populations are present in the data. Unfortunately, the log likelihood, the Bayesian information parameter, and reduced χ^2 calculations of the Viterbi path were insufficient in unambiguously determining the optimal number of states necessary to describe the data (see Supporting Materials and Methods; Fig. S5). One approach we found that worked well was to perform a cluster analysis of the molecule-by-molecule HMM results of the data in which each trace was optimized independently with up to 10 available FRET states. The number of 10 states was sufficient to allocate all rarely appearing intermediate states in each molecule. The unassigned states were not occupied and did not disturb the analysis. Similar states extracted because of overfitting will merge into a single state when plotting the data in histograms. The results are summarized as two-dimensional (2D) histograms according to their mean FRET efficiencies and average duration within the conformation in Fig. 5 for all four DNA constructs in the absence and presence of NC2. Each transition is presented by a Gaussian with a width of ~1% of the image size. Four states are clearly observable for the AdML promoter, whereas the presence of additional minor states is observable for the H2B promoter. In addition to the HMM analysis, we visually inspected the individual spFRET traces to verify that the results from the HMM correspond to distinctly observable FRET states in a single trace. The advantage of the molecule-wise HMM analysis approach is that the number of states does not have to be known in advance. However, when the number of states is known, a global analysis also worked well.

In the absence of NC2, one dominant population at E_{PR} ~0.40 is observed. The average duration of the molecule in this configuration is given by photobleaching. A small amount of dynamics is observed between FRET states at 20 and 40% FRET efficiency, indicating that a minority of complexes exhibits dynamics before the addition of NC2. Measurements with the H2B-J promoter DNA in the absence of NC2 revealed a higher fraction of dynamic complexes than for the AdML promoter containing DNA (Fig. 5).

Upon the addition of NC2, four subpopulations are observable for complexes containing the AdML promoter site, having FRET efficiencies of ~0.20, ~0.40, ~0.60, and ~0.80% (Fig. 5). Visual inspection of the FRET traces of individual complexes often showed transitions between all four of these FRET states. When performing a global HMM analysis with the AdML-promoter complexes, four states were sufficient to completely describe the dynamics we measured. The three-state model does not find the

FRET state at $E_{FRET} = 0.2$, which is clearly visible in the spFRET traces, and models containing a higher number of states always have the four major states at ~0.20, 0.40, 0.64, and 0.83.

Analysis of the H2B-J promoter labeled upstream showed that the 0.40 FRET efficiency state is strongly populated, with fluctuations to a state with an efficiency value of 0.75 FRET. The 2D HMM histogram also shows transitions to other states with FRET values of 0.20, 0.50, 0.65, and 0.90. Hence, at least six FRET subpopulations are present for the H2B-J promoter. This is also confirmed by visual inspection of the individual traces. Analysis of the downstream-labeled H2B-J promoter DNA revealed at least seven states. As the number of states differs between the upstream and downstream label, FRET subpopulations are distinguishable with the downstream labeled that cannot be distinguished with the upstream label. Hence, an accurate mapping of the FRET states between the upstream and downstream label is currently not possible.

Dynamics of the FRET-labeled TBP-NC2 complex, four-state model

For a more detailed analysis of the dynamics of the TBP-DNA-NC2 complex, we analyzed the AdML promoter data using a four-state HMM. A four-state model is justified because all four states are directly observed in individual spFRET traces. Fig. 6 *a* shows a FRET trajectory and the corresponding optimized Viterbi path for a single TBP-NC2 complex on the 70-bp DNA promoter AdML. In the first second of this time trajectory, the FRET efficiency was high ($E_{FRET} = 0.83$), with short excursions to $E_{FRET} = 0.64$ and $E_{FRET} = 0.40$. After 1.2 s, the TBP-NC2-DNA complex switched to a low FRET conformation ($E_{FRET} = 0.20$) for ~1 s before returning to the 0.40 FRET efficiency conformation. At the end of the trace, the complex oscillates between E_{FRET} 0.40 and higher FRET states. From the HMM analysis, we do not only get the values of the different FRET states but also the intrinsic width of the FRET state beyond shot-noise broadening (Table S3). For the two AdML promoters, the widths determined for the different FRET states are very similar, validating the reproducibility of the analysis. An inherent broadening of the FRET distribution is always seen in single-molecule FRET experiments and most likely corresponds to slight heterogeneities in the structure and dynamics of the linkers used to attach the fluorophores to the protein. The FRET states of $E_{FRET} = 0.40$, 0.64, and 0.83 have widths of 0.07, 0.07, and 0.04 FRET efficiency values corresponding to structural heterogeneities of 3, 3, and 2 Å, respectively. These widths are relatively small corresponding to well-defined conformations. The inherent width of the $E_{FRET} = 0.20$ state, though similar in value (± 0.08), indicates a larger conformational heterogeneity of ±5–8 Å because of the lower sensitivity of FRET at the extremities due to the R^6 dependence of the FRET efficiency.

In addition to the average FRET values and intrinsic width of the FRET states, the HMM analysis also provides the transition rates between the different states. From the optimized Viterbi path for each spFRET trace, we can extract the lifetime distribution for each state (Fig. 6 *c*). The 0.83 and 0.64 FRET efficiency states have short average lifetimes of 82 and 34 ms, respectively, whereas the lower FRET efficiency states are more stable with lifetimes of 112 and 175 ms for $E_{FRET} = 0.40$ and $E_{FRET} = 0.20$, respectively. With the exception of the intermediate $E_{FRET} = 0.64$ state, all conformations can be described reasonably well with a monoexponential lifetime. For the $E_{FRET} = 0.64$ state, at least two components are visible. This may be due to a structural change in which a transition to the high FRET state becomes faster. This would explain the change in the dynamics observed between the beginning and end of the spFRET trace in Fig. 6 *a*. However, this is purely speculative. For the HMM analysis, exponential rates are assumed, but the model can explain the current data reasonably well even though the rates for the $E_{FRET} = 0.64$ state are nonexponential.

To look into the details of the kinetics, we generate a FRET transition density plot (TDP). To do this, we took the results from the global HMM analysis and calculated the optimized Viterbi path for each spFRET trace. From the Viterbi path, we determined the average FRET value for each transition and the survival probability of the different states (Fig. 6, *b* and *c*). Each transition was plotted as a Gaussian with a width of ~1% in FRET efficiency. In the TDP, only 8 of the theoretically possible 12 state transitions are populated. The most frequent transitions are between the 0.64 and 0.83 FRET states and between the 0.40 and 0.64 FRET states. Transitions to the 0.20 FRET efficiency state are rare and occur exclusively through the 0.40 FRET state. A few direct transitions are observed between the 0.40 and 0.83 FRET states. However, because of the fast fluctuations between the $E = 0.64$ and $E = 0.83$ FRET states, it is possible that a two-step transition between $E = 0.40$ and $E = 0.64$ and then on to the $E = 0.83$ FRET efficiency state occurs, which is detected as only a single step in the HMM analysis. From the rates, we can estimate how often such a transition would be missed. Assuming that a minimal dwell time of 5 ms (one frame from the data) is necessary for the HMM to detect the intermediate state, ~18% of the transitions from $E = 0.40$ to 0.064 to 0.83 would be detected as a straight transition to $E = 0.83$ (Fig. S6). This corresponds well to the relative amplitude of what is observed in the TDP in Fig. 6 *b*. When we set the transition rates between $E = 0.40$ and $\mathrm{E} = 0.83$ to zero and reanalyze the data, we did not detect a significant difference in the results of the HMM (Table S5). Hence, the HMM detects a direct transition between the 0.40 and 0.83 FRET states although the transition went through the 0.64 FRET state. Thus, the TDP indicates that the dynamics can be explained by a

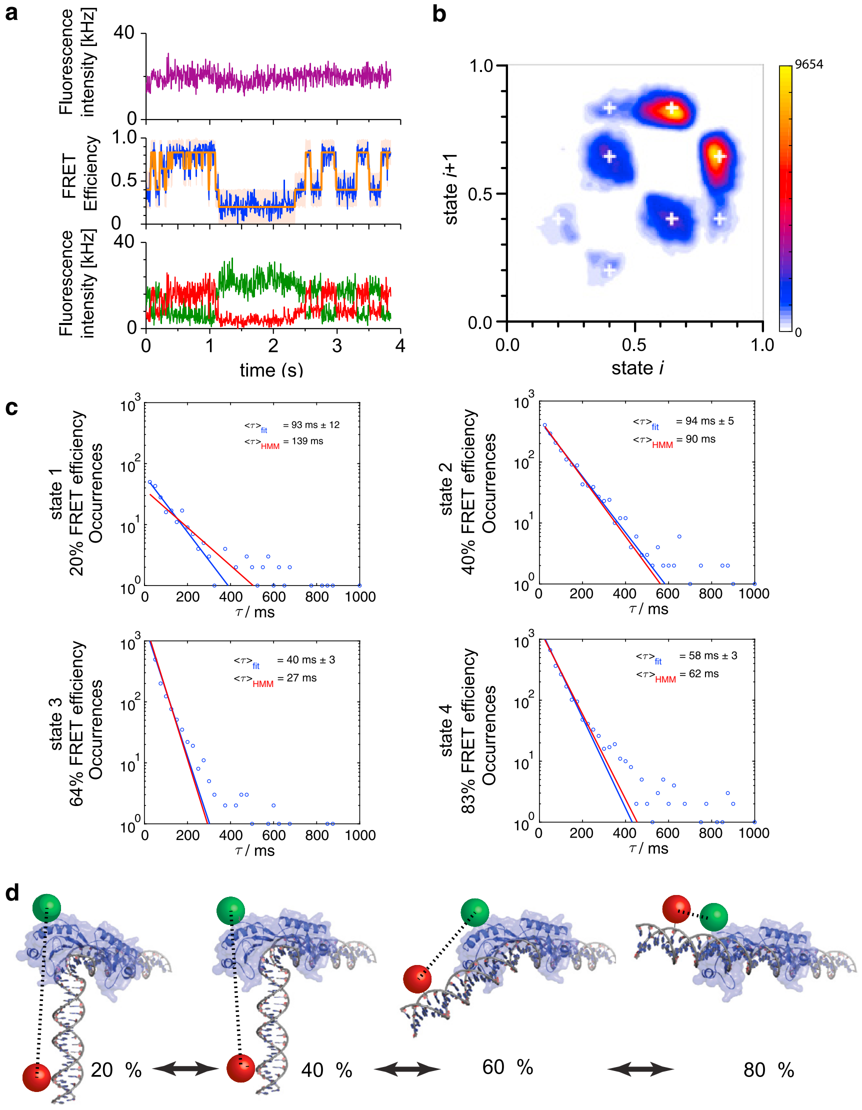

FIGURE 6 Four-well analysis of the TBP-DNA-NC2 dynamics. (*a*) A single-molecule FRET trajectory from a TBP-DNA-NC2 complex in which the 70-bp upstream-labeled DNA contains the AdML TATA box is plotted showing four different conformations with FRET efficiencies of 0.20, 0.40, 0.64, and 0.83 (*purple*: total intensity; *green*: donor intensity; *red*: acceptor intensity; and *blue*: FRET efficiency). The optimized Viterbi path derived using parameters from the global four-state HMM is overlaid in orange. (*b*) The TDP from the HMM analysis of the 70-bp AdML DNA is shown. The number of observed transitions is represented via the color code, with rare transitions shown in blue and more frequent transitions shown in yellow. The color scale bar is shown on the right with a maximal number of transitions at 9654. The FRET states returned from a global four-well HMM analysis are shown as white plus signs. Fast fluctuations between the high FRET state at ~0.83 and a short intermediate state at ~0.64 are observed. Transitions from 0.40 to 0.64 also occur frequently, whereas transitions between 0.40 and 0.83 occur over the short-lived intermediate FRET state with an efficiency of 0.64. The FRET state at 0.20 is connected only with the 0.40 FRET state. (*c*) The survival probabilities (*circles*), results from the global HMM values (*red line*) from an exponential fit to the data (*blue line*) are shown for the four FRET conformations. The survival probabilities can be reasonably described by a single exponential. (*d*) A schematic of the connections between the different conformations observed in the spFRET experiments on the AdML promoter complexes is shown. From the TDP, it is clear that the four different conformations are connected by a linear four-well model. The 0.40 FRET state is the initial, bent structure of TBP-DNA. The 0.60 and 0.80 FRET states represent conformational changes in the DNA in the TBP-DNA-NC2 complex, whereas the 0.20 FRET state represents motion of TBP-NC2 along the DNA. NC2 is not displayed for clarity.

linear four-well model (Fig. 6 *d*). The dynamics observed with the 70-bp DNA and 110-bp DNA strands containing the AdML-promoter TATA box are identical within experimental error (Fig. S7). This again suggests that neither the length of the DNA nor the attachment point of the DNA to the surface influences the dynamics.

In contrast to the AdML promoter DNA, the H2B-J promoter DNAs showed more complex dynamics. For the H2B-J promoter labeled upstream, at least six states are observable, and at least seven states are observable for the downstream-labeled construct (Fig. 5). One beautiful aspect of the HMM approach is that one has a complete description of all the states and all the transitions. Hence, for comparison, more complex models can be reduced using either an HMM with fewer states or by clustering results together. To quantify the difference between the AdML and H2B-J promoters, we approximated the H2B-J promoter with a global four-state model. The optimized Vitrebi paths for the global HMM were calculated for the individual traces, and the TDP was generated by calculating the average FRET efficiency for each state and plotting it as a Gaussian with a width of ~1%. The transition matrix for the HMM analysis is shown in Fig. S7. As for the AdML promoter, transitions to the low FRET state occur only through the 0.39 FRET subpopulation. From the 0.39 FRET state, transitions to all other states are observable.

For analysis purposes, we consider the following transitions: the transition between the low FRET state and initial FRET conformation ($k_{1 \leftrightarrow 2}$), the transition between the initial FRET conformation and the higher FRET states ($k_{2 \leftrightarrow 3+4}$), and the transition between the intermediate

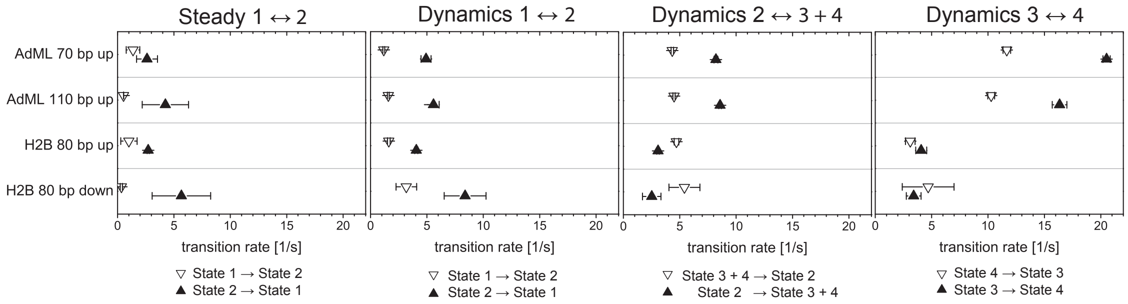

FIGURE 7 Comparison of the kinetics for the different constructs. Resulting rate constants of the four different sample preparations are shown. ($k_{1\leftrightarrow 2}$) The rate constants for transitions to a possible alternative TBP-binding site ($E_{\mathrm{FRET}} = 0.20$) in the absence (*left panel*) and presence (*middle left panel*) of NC2 are shown. The transition rate from $E_{\mathrm{FRET}} = 0.40$ to 0.20 is between 2 and 8 s^{-1} (*arrow up*) and the return the rate lies between 0.5 and 3.5 s^{-1} (*arrow down*). The affinity to this binding site appears to be neither dependent on the promoter nor influenced by the addition of NC2. ($k_{2\leftrightarrow 3+4}$) The rate for transitions between the $E_{\mathrm{FRET}} = 0.40$ and higher FRET states are shown (*middle right panel*). The transition rate from the bent conformation to the higher FRET conformations (*arrow up*) is a factor of ~3 higher for the AdML promoter than for the H2B promoter, whereas the relaxation rate for returning to the bent 40% FRET conformation (*arrow down*) was similar for all four samples. ($k_{3\leftrightarrow 4}$) The transition rates between the intermediate and high FRET states are shown. The transition rates from the intermediate state to high FRET conformation (*arrow up*) and back (*arrow down*) were faster for the AdML promoter than for the H2B promoter. The error bars are the standard error of the mean (SEM) determined empirically by splitting the data set into four equal subsets (two subsets in the case of transitions in the absence of NC2 due to limited statistics) and calculating the mean and the standard error of the mean (SEM) of the different subsets.

and high FRET states ($k_{3\leftrightarrow 4}$). The forward and backward transition rates for the transitions defined above are summarized in Fig. 7 for all four complexes investigated. In the absence of NC2 (*left panel*), only very few transitions between the ~0.40 and ~0.20 FRET efficiency states ($k_{1\leftrightarrow 2}$) were observed. The rates were similar to what was observed in the presence of NC2. For the upstream labels, the transition rates $k_{2\rightarrow 1}$ exhibited promoter-independent values of ~1.6 s^{-1}. The backward transition rates $k_{1\rightarrow 2}$ were 5–7 s^{-1} for the constructs labeled upstream. The transition rate from the low FRET state to the 0.40 FRET state was somewhat faster in the downstream-labeled construct, which may indicate that we are sensitive to slightly different motions with this construct. The dynamics of the conformational changes in the DNA are given by the fluctuations between the initial FRET state (bent DNA conformation) and the higher FRET states (highly bent conformations), $k_{2\leftrightarrow 3+4}$. Interestingly, the rate of unbending was independent of the promoter site (~4.5 s^{-1}), whereas the bending rate was a factor of 2.7 faster for the AdML promoter as for the H2B-J promoter (~8.4 vs. 3.1 s^{-1}). Hence, the initial TBP-DNA conformation is more stable for the H2B-J promoter, and the AdML promoter exhibits more pronounced dynamics. Also, the transition rates between different highly bent conformations were faster for the AdML promoter DNA. $k_{3\rightarrow 4}$ and $k_{4\rightarrow 3}$ were ~18.5 and 11 s^{-1}, respectively, for the AdML promoter DNA and ~4 and ~4 s^{-1} for the H2B-J promoter DNA. The results of the six-state HMM model of upstream-labeled H2B-J promoter showed slower transition rates between the various states in comparison to the $k_{3\leftrightarrow 4}$ values for the AdML promoter DNA above. This indicates that the slower dynamics of the H2B-J promoter is an attribute of the promoter and not the analysis.

DISCUSSION

We have presented a modification of the HMM approach for the analysis of spFRET data. By incorporating the photon-counting statistics into the analysis, the noise factor in the HMM analysis is reduced to the actual noise of the biological system. Thus, a global analysis is possible even when the total intensity coming from the single-molecule traces and hence the shot-noise contribution to the spFRET traces differ by severalfold (Fig. S2). In many cases, the shot noise is the dominant component to the noise of the measurement, and an HMM can be used without an additional factor to account for the noise. The number of fitted parameters is then reduced, making the analysis more robust. In addition, the shot noise is calculated on a frame-by-frame basis, making it possible to account for frames in which partial quenching of the donor molecule is observed via the HMM analysis (Fig. S3). We verified the performance of the HMM analysis using simulations. The algorithm was found to be robust over a large range of count rates and noise contributions.

We utilized the newly adapted HMM approach to analyze the dynamics of eukaryotic TBP-DNA complexes in the presence and absence of the TF NC2. From the framewise spFRET histograms, at least two populations were clearly visible (Fig. 2 *e*). However, a Gaussian fit to the spFRET histograms gave results that deviated significantly from the HMM analysis, suggesting that additional states are present. By using a cluster analysis of the molecule-by-molecule HMM analysis along with visual inspection of the different FRET trajectories, we could distinguish up to four different FRET states for the AdML promoter at FRET efficiencies of {0.20, 0.40, 0.64, 0.83}, and more states were detectable for the H2B-J promoter. For the sake of comparison, we analyzed both promoters using a linear four-well model (Fig. 6 *d*).

The ~0.40 FRET efficiency state is observed as a static state in the absence of NC2 and is also always observed in the dynamic traces after the addition of NC2 for all promoters. This suggests that the 0.40 FRET efficiency state corresponds to a conformation of the TBP-NC2-DNA complex similar to that of the TBP-DNA complex, which has a strongly bent DNA conformation (Fig. 2 *a*). The low FRET state (~0.20) is observable for both upstream-labeled promoters in the absence and presence of NC2. According to the crystal structure of the TBP-NC2-DNA construct and the positions of the labels, it is not possible to explain the 0.20 FRET efficiency with only a conformational change in the DNA (23). Furthermore, this state is only accessible through the 0.40 FRET state. This would suggest that the TBP-NC2 complex may slide along the DNA. The 0.20 FRET efficiency state is also observed as a minor conformation for static traces measured in the absence of NC2 (23). Hence, the low FRET state most likely represents an alternative binding position of the TBP and TBP-NC2 complex on the DNA. Assuming that the conformation of the TBP-NC2-DNA complex is similar for the 0.40 and 0.20 FRET efficiency states, the 0.20 FRET efficiency state would correspond to a shift of the TBP-NC2 duplex by ~4 bp. Transitions between the 0.20 and 0.40 FRET states are also observed in a small fraction of complexes in the absence of NC2 and were observed during the binding of TBP to the DNA (60). The higher FRET states (0.64 and 0.83 FRET values for the AdML promoter) are attributed to conformation changes in the DNA. This conclusion is supported by the experiments with the H2B-J promoter labeled in the upstream and downstream locations. Both constructs show transitions from the initial FRET conformation in the absence of NC2 (between 0.35 and 0.40) to higher FRET states. If the TBP-NC2 complex were to move along the DNA without a conformational change of the DNA, one would expect that when the signal of the upstream FRET pair increases, the downstream FRET signal must decrease and vice versa. Even when accounting for the three-dimensional motion of TBP along the minor groove of the DNA, it is not possible to explain an increase in FRET efficiency for both upstream and downstream labels by complex motion alone. Hence, we conclude that the DNA is changing conformation in the higher FRET states. Most likely, this is due to the interactions between NC2 and TBP decreasing the strength of the TBP-DNA interactions, allowing the highly kinked DNA to relax. As we see two intermediates for the AdML promoter, this could represent the removal of one or both of the "phenylalanine stirrups" that bind to the minor grove and kink the DNA. Theoretically, with a global analysis of both upstream- and downstream-labeled constructs, it should be possible to gain insight into the conformational changes occurring in the TBP-DNA complex. However, because of the complex nature of the dynamics of the H2B-J-promoter site and the different number of FRET states observed for the two constructs, a mapping of the FRET states between the two constructs was not possible.

The results for four-state HMM of the two AdML constructs were very similar, indicating that neither the length of the DNA nor the location of the biotin used to immobilize the complex on the surface influences the FRET values and dynamics. In addition, these experiments provide a measure of the reproducibility and accuracy of the HMM analysis. The TDP plots for the AdML promoter (Fig. S7) show that the four states are connected via a linear four-well model (Fig. 6 *d*). Hence, only adjacent states can interconvert. The low FRET state can only exchange with the initial bent conformation (0.40 FRET state). From the initial conformation, the complex can also fluctuate to the intermediate FRET state (0.64) in which the DNA is more stretched. From there, it can either return to the initial conformation or undergo quick fluctuations between the two high-FRET conformations (0.64 and 0.83). The kinetics of the fluctuations between the 0.64 and 0.83 FRET values in the AdML promoter determined from the Viterbi paths is ~18 s^{-1} for $k_{3 \rightarrow 4}$ and $k_{4 \rightarrow 3} = \sim 11\ s^{-1}$ (Fig. 7), and a similar ratio is obtained directly from the HMM (Table S5). The TDP for the upstream-labeled H2B-J promoter is similar to what we determined for the AdML promoter, with the exception that transitions are directly observed from the initial FRET conformation (0.39) to the highest FRET state (0.76) (Fig. S7). That the TDP pattern depends on the DNA sequence used supports our interpretation of the higher FRET states as different DNA conformations.

A comparison of the dynamics for two different promoter sites yields interesting molecular insights into the TBP-NC2-DNA complex. The quantitative difference between the two promoter sites comes from the difference in transition rates from the bent conformation to the stretched conformation, $k_{2 \rightarrow 3+4}$. However, the transition rates from the stretched conformation to the bent conformation, $k_{2 \leftarrow 3+4}$, are the same for both promoters. Hence, the promoter has a direct influence on the observed kinetics, with the bent conformation being more stable by a factor of three for the H2B-J promoter.

This HMM-based FRET analysis can be expanded further to freely diffusing proteins, when surface attachment is not suitable (28,37,63). For example, the membrane protein F_oF_1-ATP synthase was reconstituted into lipid vesicles of 100–150 nm diameter, and a pH gradient was established across the membrane for the catalytic synthesis of ATP (64). Monitoring the rotary subunit movement in single F_oF_1-ATP synthase was achieved in real time by single-molecule FRET using confocal excitation schemes (65–67). In this case, the overall fluorescence intensity varies strongly over millisecond time intervals within the time trajectory of each single protein because of the arbitrary diffusion pathways through the Gaussian-shaped excitation and detection volume. In addition, using fluorescent proteins as genetically fused labels to the enzyme results in lower photon count rates for FRET analyses of F_oF_1-ATP synthase (28,68). Identifying conformations and dwell times in these FRET trajectories by manual assignment is a time-consuming process and remains questionable, especially for those parts of the photon bursts with low-fluorescence sum intensity. Hence, applying our weighted HMM approach using the counting statistics for each data point in the photon burst provides an unbiased and robust methodology for the analysis of conformational dynamics of freely diffusing biomolecules, even for systems in which little a priori information about the likely structures or conformations and the reaction pathways is known.

One additional possibility with our expanded HMM approach is to include an additional variance for each FRET state that represents the residual broadening of the FRET efficiency beyond the shot-noise limit. As shown in Fig. S4, dynamics lead to a higher variance in the FRET signal beyond what is expected from photon statistics. This is also true when dynamics occur within a FRET state. The expanded analysis we describe allows one to explicitly account for the photon statistics, and any additional broadening can then be assigned to heterogeneities within the different states, yielding how well defined or dynamic a particular conformation is. For TBP/NC2 bound to the AdML promoter containing DNA studied here, the FRET states of 0.40, 0.64, and 0.83 showed well-defined conformations, whereas the FRET state at 0.20 showed more intrinsic heterogeneity.

CONCLUSIONS

The number of states and the corresponding transition rate constants for TBP-DNA in the presence and absence of NC2 were analyzed with an extended HMM approach. For the HMM analysis, it was necessary to determine the appropriate accuracy of the fluorescence intensities measured using EMCCD cameras. The developed HMM method was capable of acting directly on the time trajectory of the FRET efficiency without being affected by the variations in total fluorescence intensities from complex to complex.

Experiments on DNA with the same promoter and labeled at different positions confirmed that the higher FRET states observed are due to conformational changes in the DNA and not the motion of the complex along the DNA. Experiments with the AdML promoter with DNA of different lengths and attachment points to the surface verified that the observed dynamics do not depend on the length of the DNA construct nor on which end is attached to the surface.

Using the extended HMM approach, we could determine that the AdML promoter has four distinct conformations and extract the transition matrix for the different states. Transitions to the low FRET state can only be reached through the bent FRET conformation. The dynamics of the H2B-J promoter are more complex but could be simplified into a four-state model for comparison with the AdML promoter. The difference in the dynamics between these two promoter sites is due to the higher stability of the bent conformation of the DNA for the H2B-J promoter. Hence, the dynamics are qualitatively similar for the different promoter sites, but the stronger promoter site, AdML, shows faster dynamics.

SUPPORTING MATERIAL

Supporting Materials and Methods, seven figures, and five tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)31279-7.

AUTHOR CONTRIBUTIONS

N.Z. developed the analysis method and performed the simulations. P.S. performed the spFRET measurements. Both N.Z. and P.S. analyzed of the data. M.M. provided the protein. M.B. and D.C.L. oversaw the project. D.C.L. and N.Z. wrote the manuscript with contributions from all authors.

ACKNOWLEDGMENTS

We thank Christine Goebel for excellent technical assistance and Dr. Sushi Madhira and Dr. Evelyn Plötz for helpful and critical discussions regarding HMM and TBP. We also thank Dr. Evelyn Plötz and Simon Wanninger for assistance with the figures and Dr. Richard Börner for constructive comments on the manuscript.

We gratefully acknowledge the financial support by the Deutsche Forschungsgemeinschaft to D.C.L. and M.M. (SFB646; projects A2 and B11) and the Ludwig-Maximilians-Universität through the Center for NanoScience and the BioImaging Network. N.Z. was supported by the Deutsche Forschungsgemeinschaft (grant BO1891/10-1 to M.B.).

REFERENCES

1. Sims, R. J., III, S. S. Mandal, and D. Reinberg. 2004. Recent highlights of RNA-polymerase-II-mediated transcription. *Curr. Opin. Cell Biol.* 16:263–271.
2. Burley, S. K. 1996. The TATA box binding protein. *Curr. Opin. Struct. Biol.* 6:69–75.
3. Horikoshi, M., C. Bertuccioli, …, R. G. Roeder. 1992. Transcription factor TFIID induces DNA bending upon binding to the TATA element. *Proc. Natl. Acad. Sci. USA.* 89:1060–1064.
4. Nikolov, D. B., H. Chen, …, S. K. Burley. 1996. Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci. USA.* 93:4862–4867.
5. Werner, M. H., A. M. Gronenborn, and G. M. Clore. 1996. Intercalation, DNA kinking, and the control of transcription. *Science.* 271:778–784.
6. Pardo, L., M. Campillo, …, H. Weinstein. 2000. Binding mechanisms of TATA box-binding proteins: DNA kinking is stabilized by specific hydrogen bonds. *Biophys. J.* 78:1988–1996.
7. Kim, J. L., D. B. Nikolov, and S. K. Burley. 1993. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature.* 365:520–527.
8. Kim, J. L., and S. K. Burley. 1994. 1.9 A resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat. Struct. Biol.* 1:638–653.

9. Kim, Y., J. H. Geiger, …, P. B. Sigler. 1993. Crystal structure of a yeast TBP/TATA-box complex. *Nature*. 365:512–520.
10. Alberts, B. J. A., J. Lewis, …, P. Walter. 2002. Molecular Biology of the Cell, Fourth Edition. Garland Pub., New York.
11. Nikolov, D. B., and S. K. Burley. 1997. RNA polymerase II transcription initiation: a structural view. *Proc. Natl. Acad. Sci. USA*. 94:15–22.
12. Roeder, R. G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* 21:327–335.
13. Kaiser, K., and M. Meisterernst. 1996. The human general co-factors. *Trends Biochem. Sci.* 21:342–345.
14. Roeder, R. G. 2005. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Lett.* 579:909–915.
15. Meisterernst, M., and R. G. Roeder. 1991. Family of proteins that interact with TFIID and regulate promoter activity. *Cell*. 67:557–567.
16. Iratni, R., Y. T. Yan, …, M. M. Shen. 2002. Inhibition of excess nodal signaling during mouse gastrulation by the transcriptional corepressor DRAP1. *Science*. 298:1996–1999.
17. Prelich, G., and F. Winston. 1993. Mutations that suppress the deletion of an upstream activating sequence in yeast: involvement of a protein kinase and histone H3 in repressing transcription in vivo. *Genetics*. 135:665–676.
18. Gadbois, E. L., D. M. Chao, …, R. A. Young. 1997. Functional antagonism between RNA polymerase II holoenzyme and global negative regulator NC2 in vivo. *Proc. Natl. Acad. Sci. USA*. 94:3145–3150.
19. Xie, J., M. Collart, …, M. Meisterernst. 2000. A single point mutation in TFIIA suppresses NC2 requirement in vivo. *EMBO J.* 19:672–682.
20. Chitikila, C., K. L. Huisinga, …, B. F. Pugh. 2002. Interplay of TBP inhibitors in global transcriptional control. *Mol. Cell*. 10:871–882.
21. Geisberg, J. V., F. C. Holstege, …, K. Struhl. 2001. Yeast NC2 associates with the RNA polymerase II preinitiation complex and selectively affects transcription in vivo. *Mol. Cell. Biol.* 21:2736–2742.
22. Klejman, M. P., L. A. Pereira, …, H. T. Timmers. 2004. NC2alpha interacts with BTAF1 and stimulates its ATP-dependent association with TATA-binding protein. *Mol. Cell. Biol.* 24:10072–10082.
23. Schluesche, P., G. Stelzer, …, M. Meisterernst. 2007. NC2 mobilizes TBP on core promoter TATA boxes. *Nat. Struct. Mol. Biol.* 14:1196–1201.
24. Kamada, K., F. Shu, …, S. K. Burley. 2001. Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. *Cell*. 106:71–81.
25. McKinney, S. A., C. Joo, and T. Ha. 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* 91:1941–1951.
26. Messina, T. C., H. Kim, …, D. S. Talaga. 2006. Hidden Markov model analysis of multichromophore photobleaching. *J. Phys. Chem. B*. 110:16366–16376.
27. Bronson, J. E., J. Fei, …, C. H. Wiggins. 2009. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* 97:3196–3205.
28. Düser, M. G., N. Zarrabi, …, M. Börsch. 2009. 36 degrees step size of proton-driven c-ring rotation in FoF1-ATP synthase. *EMBO J.* 28:2689–2696.
29. Lee, T. H. 2009. Extracting kinetics information from single-molecule fluorescence resonance energy transfer data using hidden markov models. *J. Phys. Chem. B*. 113:11535–11542.
30. Bronson, J. E., J. M. Hofman, …, C. H. Wiggins. 2010. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics*. 11 (*Suppl 8*):S2.
31. Liu, Y., J. Park, …, T. Ha. 2010. A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B*. 114:5386–5403.
32. Taylor, J. N., D. E. Makarov, and C. F. Landes. 2010. Denoising single-molecule FRET trajectories with wavelets and Bayesian inference. *Biophys. J.* 98:164–173.
33. Pirchi, M., G. Ziv, …, G. Haran. 2011. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat. Commun.* 2:493.
34. Uphoff, S., K. Gryte, …, A. N. Kapanidis. 2011. Improved temporal resolution and linked hidden Markov modeling for switchable single-molecule FRET. *Chemphyschem*. 12:571–579.
35. Greenfeld, M., D. S. Pavlichin, …, D. Herschlag. 2012. Single molecule analysis research tool (SMART): an integrated approach for analyzing single molecule data. *PLoS One*. 7:e30024.
36. Okamoto, K., and Y. Sako. 2012. Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.* 103:1315–1324.
37. Zarrabi, N., S. Ernst, …, M. Börsch. 2014. Analyzing conformational dynamics of single P-glycoprotein transporters by Förster resonance energy transfer using hidden Markov models. *Methods*. 66:168–179.
38. Baum, L. E., T. Petrie, …, N. Weiss. 1970. A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41:164–171.
39. Schmid, S., M. Götz, and T. Hugel. 2016. Single-molecule analysis beyond dwell times: demonstration and assessment in and out of equilibrium. *Biophys. J.* 111:1375–1384.
40. Jordan, M. I., Z. Ghahramani, …, L. K. Saul. 1999. An introduction to variational methods for graphical models. *Mach. Learn.* 37:183–233.
41. van de Meent, J. W., J. E. Bronson, …, R. L. Gonzalez, Jr. 2014. Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophys. J.* 106:1327–1337.
42. Auble, D. T. 2009. The dynamic personality of TATA-binding protein. *Trends Biochem. Sci.* 34:49–52.
43. Press, W. H., S. A. Teukolsky, …, B. P. Flannery. 1992. Numerical Recipes the Art of Scientific Computing. Cambridge Univ. Press, New York.
44. Rabiner, L. R. 1989. A tutorial on hidden markov-models and selected applications in speech recognition. *Proc. IEEE*. 77:257–286.
45. Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*. 13:260–269.
46. Andrec, M., R. M. Levy, and D. S. Talaga. 2003. Direct determination of kinetic rates from single-molecule photon arrival trajectories using hidden Markov models. *J. Phys. Chem. A*. 107:7454–7464.
47. Press, W. H., S. A. Teukolsky, …, B. P. Flannery. 2007. Numerical Recipes the Art of Scientific Computing, Third Edition. Cambridge Univ. Press, New York.
48. Davison, A. C. 2003. Statistical Models. Cambridge University Press, Cambridge, UK.
49. Bilmes, J. A. 2006. What HMMs can do. *IEICE Trans. Inf. Syst.* E89D:869–891.
50. Talaga, D. S. 2007. COCIS: markov processes in single molecule fluorescence. *Curr. Opin. Colloid Interface Sci.* 12:285–296.
51. Hirsch, M., R. J. Wareham, …, D. J. Rolfe. 2013. A stochastic model for electron multiplication charge-coupled devices–from theory to practice. *PLoS One*. 8:e53671.
52. Börner, R., D. Kowerko, …, R. K. O. Sigel. 2018. Simulations of camera-based single-molecule fluorescence experiments. *PLoS One*. 13:e0195277.
53. Dahan, M., A. A. Deniz, …, S. Weiss. 1999. Ratiometric measurement and identification of single diffusing molecules. *Chem. Phys.* 247:85–106.
54. Gopich, I., and A. Szabo. 2005. Theory of photon statistics in single-molecule Förster resonance energy transfer. *J. Chem. Phys.* 122:14707.
55. Nir, E., X. Michalet, …, S. Weiss. 2006. Shot-noise limited single-molecule FRET histograms: comparison between theory and experiments. *J. Phys. Chem. B*. 110:22103–22124.
56. Forsythe, G. E., M. A. Malcolm, and C. B. Moler. 1977. Computer Methods for Mathematical Computations. Prentice-Hall, Englewook Cliffs, NJ.

57. Banik, U., J. M. Beechem, …, P. A. Weil. 2001. Fluorescence-based analyses of the effects of full-length recombinant TAF130p on the interaction of TATA box-binding protein with TATA box DNA. *J. Biol. Chem.* 276:49100–49109.

58. Gumbs, O. H., A. M. Campbell, and P. A. Weil. 2003. High-affinity DNA binding by a Mot1p-TBP complex: implications for TAF-independent transcription. *EMBO J.* 22:3131–3141.

59. Bleichenbacher, M., S. Tan, and T. J. Richmond. 2003. Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. *J. Mol. Biol.* 332:783–793.

60. Schluesche, P., G. Heiss, …, D. C. Lamb. 2008. Dynamics of TBP binding to the TATA box. *In* Single Molecule Spectroscopy and Imaging. J. Enderlein, Z. K. Gryczynski, and R. Erdmann, eds. SPIE:6862E-1–6862E-8.

61. Murphy, K. P. 1998. Hidden Markov Model (HMM) Toolbox for Matlab.

62. Kapanidis, A. N., T. A. Laurence, …, S. Weiss. 2005. Alternating-laser excitation of single molecules. *Acc. Chem. Res.* 38:523–533.

63. Zarrabi, N., M. G. Düser, …, M. Börsch. 2007. Detecting substeps in the rotary motors of F0F1-ATP synthase by hidden Markov models. *Proc. SPIE.* 6444:64440E.

64. Börsch, M., and P. Gräber. 2005. Subunit movement in individual H+-ATP synthases during ATP synthesis and hydrolysis revealed by fluorescence resonance energy transfer. *Biochem. Soc. Trans.* 33:878–882.

65. Börsch, M., M. Diez, …, P. Gräber. 2002. Stepwise rotation of the gamma-subunit of EF(0)F(1)-ATP synthase observed by intramolecular single-molecule fluorescence resonance energy transfer. *FEBS Lett.* 527:147–152.

66. Diez, M., B. Zimmermann, …, P. Gräber. 2004. Proton-powered subunit rotation in single membrane-bound F_0F_1-ATP synthase. *Nat. Struct. Mol. Biol.* 11:135–141.

67. Zimmermann, B., M. Diez, …, M. Börsch. 2005. Movements of the epsilon-subunit during catalysis and activation in single membrane-bound H(+)-ATP synthase. *EMBO J.* 24:2053–2063.

68. Düser, M. G., Y. Bi, …, M. Börsch. 2008. The proton-translocating a subunit of F0F1-ATP synthase is allocated asymmetrically to the peripheral stalk. *J. Biol. Chem.* 283:33602–33610.

Biophysical Journal, Volume *115*

Supplemental Information

Analyzing the Dynamics of Single TBP-DNA-NC2 Complexes Using Hidden Markov Models

Nawid Zarrabi, Peter Schluesche, Michael Meisterernst, Michael Börsch, and Don C. Lamb

Supporting Information

"Analyzing the Dynamics of Single TBP-DNA-NC2 complexes by Hidden Markov Models"

Nawid Zarrabi[1], Peter Schluesche[2], Micheal Meisterernst[3,4], Micheal Börsch[1], Don C. Lamb[2]

1) 3. Physikalisches Institut, Universität Stuttgart, Pfaffenwaldring 57, 70550 Stuttgart, Germany.

2) Department Chemie, Center for Nano Science (CENS), Center for Integrated Protein Science (CIPSM) and Nanosystems Initiative München (NIM), Ludwig Maximilians-Universität Munich, Butenandstr. 5-13, 81377 Munich, Germany

3) National Research Center for Environment and Health - GSF, Gene Expression, Marchionini-Str. 25, 81377 Munich, Germany

4) Institute for Tumor Biology, Department of Medicine, University of Muenster, 48149 Muenster

Table of contents **Page**

Estimation of the Number of Detected Photons and the Uncertainty using an EMCCD detector

To calculate the uncertainty in the measured fluorescence intensities in the single-pair FRET traces, it is necessary to know the number of photons that have been detected. At the low intensities available in single molecule experiments, the uncertainty is dominated by shot-noise. Shot-noise is Poisson distributed and hence, when using photon-counting detectors, the uncertainty (i.e. standard deviation) is given by the square root of the number of detected photons.

Although current EMCCD cameras detect single photons, the readout is not given in photons and additional noise-sources are present. Each detected photon generates an electron in the CCD chip. However, to readout the number of electrons, the charge has to be shifted and is often amplified. The number of detected photons can be estimated from the camera counts using:

$$N_{\text{Photons}} = \frac{N_{\text{Camera Counts}} \times N_{e^-/\text{Camera Counts}}}{Gain} \tag{S1}$$

where the number of electrons per camera count is given for each individual camera by the manufacture and the *Gain* is either provided directly, as in the case of a calibrated linear gain, or is determined independently. In our analysis, photons detected from the same molecule on different pixels are summed together to give an estimate of the total intensity of the molecule in the individual frame. The total fluorescence signal is coming from the molecule of interest and from the background. The background intensity is calculated from the pixels surrounding the individual molecules. The average background intensity is determined and subtracted from the total intensity of the individual molecule. Usually, one can select a large enough region of pixels for calculating the average background intensity that no additional error is brought in via background subtraction. When the background intensity changes slowly with time, the background can also be averaged with a sliding time window to provide a very accurate average value. Thus, the uncertainty in the fluorescence signal is given by the shot noise, i.e. the total number of detected photons (signal plus background). The fluorescence signal and uncertainty in each frame is then given by:

$$\begin{aligned} I_D &= N_D - I_D^{background}; \qquad \sigma_D^2 = N_D \\ I_A &= N_A - I_A^{background} - I_D^{crosstalk}; \quad \sigma_A^2 = N_A \end{aligned} \tag{S2}$$

where N_i represents the total number of detected photons in the *i*th channel and I_D and I_A the corrected intensities for the donor and acceptor channels respectively. We neglect direct excitation of the acceptor molecule with donor excitation, although it could be easily incorporated if desired. Additional noise sources include thermal noise or dark counts, clock-induced charge, multiplicative noise and readout noise. See references (1, 2) for a detailed description of camera noise. For our EMCCD camera (DV887-BV iXon$^+$, Andor Technology), the thermal noise and clock-induced charge are negligible. In addition, the readout noise is significantly smaller than one photon after amplification. If one could directly determine the number of electrons per pixel, then a CCD camera could be used as a photon-counting device. However, due to the different noise sources involved in reading the output of the camera, it is not possible to uniquely determine the number of detected photons. Therefore, we approximate the number of detected photons and thereby estimate the uncertainty of the measurement.

Comparison of HMM analyses with and without incorporation of the camera noise.

We simulated 20 single molecule FRET trajectories using a four-state Markov model. The parameters for the four-states were chosen with FRET values of {0.25, 0.45, 0.65, 0.85}, a variance and dwell time for each state of 0.002 and 100 ms, and an equal probability for transitions between all other states. The twenty state sequences (i.e. 20 molecules) were generated each with a length of 800 data points (one data point representing a 5 ms frame). Starting from the state sequence, the fluorescence signals were created. Each molecule was assigned an average total intensity from a random number between 10 and 110 counts per frame to generate bright and dark molecules. The average intensity of the donor and acceptor signal were calculated from the FRET state occupied at that time point. Noise was than generated by producing independent random numbers with a Gaussian distribution with a variance given by mean intensity (i.e. the variance of a Poissonian distribution) multiplied by a factor of 2 (i.e. to simulate the addition noise of the EMCCD-technology). This leads to 20 pairs of donor and acceptor traces of 800 frames duration each, which were than analyzed with the different HMM approaches. The results are summarized in Supporting Table S1.

Model selection criteria

For all four different promotors, the loglikelihood values increased with rising model complexity (Figure S5, first row). Every additional state for the parameter room led to a better description of the data and resulted in a higher loglikelihood value. To handle this problem, the Bayesian Information Criterion (BIC) has been established by Schwarz in 1978 to give a Bayesian argument for model selections by accounting for the model complexity (3). The BIC, which is still under scientific debate (4), is defined as:

$$BIC = -2\ln L + \theta \ln T \tag{S3}$$

with ln *L,* loglikelihood of the data given the complete set of free parameters; *T,* the number of data points, and θ, the number of free parameters.

The second term is called the "penalty term", because it increases with a rising number of states and, therefore, acts against the first term, which decreases with expanded models. In principle, this leads to a minimum of the BIC for a certain number of free parameters indicating an appropriate choice of the number of hidden states. However, in our case, the BIC always decreased with an increasing number of states (see Figure S5, second row) implying at least ten different hidden states.

Alternatively, we obtained a reasonable model size by investigating the variances of the steps delivered by the model by calculated the χ^2 values. The correct number of states should yield similar results for the steady traces (without NC2) and dynamic case (in the presence of NC2) (Figure 3, third row):

$$\chi^2 = \frac{1}{T}\sum_{step}\sum_{t\in T_{step}}(x_t - \mu_{step})^2, \text{ with } \mu_{step} = \frac{1}{T_{step}}\sum_{t\in T_{step}} x_t \qquad \text{(S4)}$$

For the static FRET efficiency case in the absence of NC2, all ten hidden Markov models were expected to predominantly mark one step per molecule. Calculating the variances for these steps defined a lower bound, which was basically independent of the model. In the dynamic case of TBP-NC2-DNA, only suitable models were capable to assign the hidden states correctly. Incorrectly assigned steps included either missing jumps in the FRET efficiency, leading to raised variances, or falsely divided steps that fit the noise, resulting in underestimated variances. The results are shown in Figure S5 for the four different promotors investigated, both before (static in red) and after (dynamic in blue) addition of NC2. The χ^2-curves of the dynamic cases suggested that already a two-state model should resolve the main behavior of the TBP-NC2-complex, i.e. jumps between the steady FRET-level (~0.4) and a second higher FRET-level (~0.8). Additional HMM states refined this picture and resolved more and more short-living intermediate states around ~0.65 between the two major states. A fourth additional low-FRET-state was found at E_{FRET} ~0.2. We assumed that a sufficient model complexity should be achieved when the χ^2 values of the dynamic cases reached the χ^2 value for the static conditions. Adding more than four states did not lead to any significant improvement of the χ^2 values for the AdML promoter and provided already reasonable information for the promoter H2B (Figure S5). As a result, a global HMM with a minimal number of four states was chosen to describe the main features of the TBP-NC2-DNA dynamics.

The validity of the usual BIC analysis can be increased by taking into account a potential deviance of the gamma-factors between the molecules. Therefore, an additional BIC was calculated from the modified χ^2-value (Eq. S4) from the Loglikelihood:

$$\mathrm{BIC} = T\log(\chi^2) + \theta\log(T) \tag{S5}$$

where T denotes the total number of analyzed frames, χ^2 is the modified residual sum of squares according to Eq. S4, and θ counts the number of free parameters.

Like the usual BIC, the first summand decreases with increasing model complexity whereas the second increases acting like a "penalty term". Both summands together should indicate the right model size by a minimum of the BIC value. The result is shown in Figure S5, forth row. The usual BIC decreases without interruption with increasing model complexity whereas the decrease of the modified BIC stagnates for all samples at four hidden states.

The gain in validity of the χ^2-BIC is given by the modified calculation of the χ^2-value. The χ^2-value usually describes the squared difference between the fit and the data. Note, that the fit here was replaced by the mean value of a marked step instead of using the resulting FRET-efficiency of the corresponding hidden state. Small deviances in the gamma-factors across the molecules yield to shifts of the FRET efficiencies of all hidden states. This usually increases the mismatch between the model and the data resulting in higher χ^2-value or lower loglikelihood values. With the modified χ^2-value, the influence of state shifts can be suppressed.

In summary, the best model should have a maximum likelihood and a minimal BIC value. However, the likelihood always increased, as more states were included in the model. The usual BIC (second row) does not show any minimum, neither for the steady (red) nor for the dynamic (blue) FRET-trajectories. This suggests that the penalty term in the BIC for the complexity of the model was too small. For both the χ^2-value and the χ2-BIC, the static traces (without NC2, red), favoring a one-state model. For dynamic traces (after addition of NC2, blue) both values decrease with increasing model complexity up to four hidden states. Afterwards the χ^2-value and χ^2-BIC stagnates indicating that a hidden Markov model with four hidden states is sufficient to describe the main feature of the dynamic traces.

Transition-Density Plots

The transition density plots (TDPs) are 2D-density plots and are usually obtained by performing a 2D-binning of the data. The larger the binning size, the more counts there are per bin but with a reduction in the resolution of the maxima. We developed an alternative method to create the 2D-histograms without a loss of resolution. Starting with a picture of an arbitrary size, e.g. 300×300 pixels, the usual binning procedure is performed. This results in a 2D-histogram, where only a few pixels have more than one count. This 2D-histogram is convoluted once with a 2D-Gaussian. Its standard deviation is a free parameter analogue to the binning size of the standard procedure. A higher standard deviation creates maxima with higher counts without loss of the position accuracy. A sharper 2D-Gaussian yields more maxima with lower count rates analogue to a higher binning in the standard procedure.

The transition density plots (Figure 5b and Figure S7, right panel) where created with a resolution of 300×300 pixels and a standard deviation of 4 pixels.

Trace-wise versus global HMM

For a comparison of the four promoters, the transition rates between the two main states, state 2 and state 4, were expected to unravel potential differences between the promoters. However, the intermediate states, only poorly represented by a global HMM, were involved in the determination of that transition rate. Therefore, one new HMM analysis with 10 states for each molecule was independently performed and optimized. The subsequent calculation of the individual Viterbi paths was performed and used to assign different steps. The number of 10 states was sufficient to allocate all rarely appearing intermediate states in each molecule. The inordinate number of states used in the analysis did not disturb the results as states that were not needed were not occupied (5). The results are summarized as 2D-histograms according to their mean FRET-efficiencies and their dwell time in Figure 4.

For the H2B promoter constructs, the steps derived of the molecule-wise HMM around the global HMM states 1 and 3 did not cluster into a clear peak in the histogram shown in Figure 4 in comparison to the two main states. This could support the existence of many different intermediate states rather than a single broad one. In the latter case, a global HMM would describe this region by allocating a single hidden state with a broadened emission function. In contrast, in our case, the HMM tried to sample this area by putting more and more intermediate states with slightly different FRET values. The same was true for the HMM state at $E_{FRET} = 0.2$.

In order to derive global transition rates from these individual states from each molecule (Figure 6), we re-assigned these states back into the four global HMM states of interest, namely states 1, 2 and 4. For this purpose, the most likely global state i_k was determined out of the subset $\tilde{q}$ of those global states for every molecule-wise obtained step k individually:

$$i_k = \arg\max\left(f_{\tilde{q}}(m_k \mid \mu_{\tilde{q}}, \sigma_{\tilde{q}}^2 + \sigma_k^2)\right) \tag{S6}$$

Afterwards, adjacent steps of identical states were merged. This procedure led to local correctly assigned steps, whereas the short-living intermediate states were in this manner transferred to their best fitting neighboring main state. This enabled the determination of the transition rate $r_{i \to j}$ between all states by counting the number of corresponding transitions N_{ij} related to the summed duration of all steps k with assigned state i:

$$r_{i \to j} = N_{ij} / \sum_{k=1}^{N_i} T_{i,k} \tag{S7}$$

$T_{i,k}$ denotes the duration of the k^{th} step assigned to the main state i, N_i counts their number, N_{ij} is the number of all transitions from step i to step j.

Distribution of the FRET efficiency from normal distributed fluorescence intensities

The probability density functions of both fluorescence channels can be, with the assumption of sufficient count rates, approximated by a normal distribution.

The joint density function of both channels is then the product of both normal distributions $f_{I_D} = \mathcal{N}(\mu_{I_D}, \sigma_{i_D}^2)$ and $f_{I_A} = \mathcal{N}(\mu_{I_A}, \sigma_{i_A}^2)$ and is again a normal distribution:

$$f_{I_D} \otimes f_{I_A} = \frac{1}{2\pi\sigma_{I_D}\sigma_{I_A}} \exp\left(-\frac{(I_D - \mu_{I_D})^2}{2\sigma_{I_D}^2} - \frac{(I_A - \mu_{I_A})^2}{2\sigma_{I_A}^2} \right) \tag{S8}$$

I_D and *I_A* are the time dependent photon count rates of the donor- respectively acceptor channel with means μ_{I_D} and μ_{I_A} and variances $\sigma_{I_D}^2$ and $\sigma_{I_A}^2$. The variances are directly related to their corresponding means due to the intrinsic Poissonian characteristics of fluorescence count rates:

$$\sigma_{I_D}^2 = k_D\, \mu_{I_D} \text{ and } \sigma_{I_A}^2 = k_A\, \mu_{I_A} \tag{S9}$$

where k_D and k_A are one in the case of an ideal Poisson distribution. Due to the induced noise by the amplification from the EMCCD-camera used here, these parameters have a value of two (1) and increase further with an increasing level of subtracted background photons I_D^{noise} respectively I_A^{noise} and crosstalk I_D^{cross}:

$$\begin{aligned} \sigma_{I_D}^2 &= 2\,(\mu_{I_D} + \mu_{I_D}^{noise}) && \Rightarrow && k_D = 2\,(1 + \mu_{I_D}^{noise} / \mu_{I_D}) \\ \sigma_{I_A}^2 &= 2\,(\mu_{I_A} + \mu_{I_A}^{noise} + \mu_{I_D}^{cross}) && \Rightarrow && k_A = 2\,(1 + \mu_{I_A}^{noise} / \mu_{I_A} + \mu_{I_D}^{cross} / \mu_{I_A}) \end{aligned} \tag{S10}$$

The following coordination transformation leads to the distribution functions for the measured FRET efficiency *E* and total intensity *I*:

$$\begin{aligned} I_D &= h_1(E,I) = (1-E)\,I & & & E &= g_1(I_D, I_A) = \frac{I_A}{I_D + I_A} \\ & & \Leftrightarrow & & & \\ I_A &= h_2(E,I) = E\,I & & & I &= g_2(I_D, I_A) = I_D + I_A \end{aligned} \tag{S11}$$

The Jacobi determinant for the coordinate transformation is given by:

$$|\mathbf{J}| = \begin{vmatrix} \partial h_1/\partial E & \partial h_1/\partial I \\ \partial h_2/\partial E & \partial h_2/\partial I \end{vmatrix} = \begin{vmatrix} -I & 1-E \\ I & E \end{vmatrix} = |I| \qquad \text{(S12)}$$

The dependence of the new joint density functions of the new variables, *E* and *I*, have the following form:

$$f_{E,I}(E,I) = |\mathbf{J}|\, f_{I_D,I_A}(I_D,I_A)$$

$$f_{E,I}(E,I) = \frac{|I|}{2\pi\sigma_{I_D}\sigma_{I_A}}\exp\left(-\frac{((1-E)I-\mu_{I_D})^2}{2\sigma^2_{I_D}} - \frac{(EI-\mu_{I_A})^2}{2\sigma^2_{I_A}}\right) \qquad \text{(S13)}$$

The final distributions in FRET efficiency and intensity can be determined by summing over the intensity or FRET efficiency respectively. For generality, we integrate from -∞ to +∞, although the density functions should only have amplitudes where the intensity is positive and for FRET efficiencies are between 0 and 1.

$$f_E(E) = \int_{-\infty}^{\infty} f_{E,I}(E,I)\, dI$$

and

$$f_I(I) = \int_{-\infty}^{\infty} f_{E,I}(E,I)\, dE$$

We first consider $f_I(I)$:

$$f_I(I) = \int_{-\infty}^{\infty} \frac{|I|}{2\pi\sigma_{I_D}\sigma_{I_A}}\exp\left(-\frac{((1-E)I-\mu_{I_D})^2}{2\sigma^2_{I_D}} - \frac{(EI-\mu_{I_A})^2}{2\sigma^2_{I_A}}\right) dE = c|I|\int_{-\infty}^{\infty}\exp\left(-O(E^2)\right) dE \qquad \text{(S14)}$$

where *c* represents a constant and $O(E^2)$ the exponent, which is a function of E^2. The Gaussian integral converges. As $I > 0$, the absolute value operation can be ignored and the result is a normal distribution:

$$f_I(I\,|\,\mu_I,\sigma_I^2) = \frac{1}{\sigma_I\sqrt{2\pi}}\exp\left(-\frac{(I-\mu_I)^2}{2\sigma_I^2}\right) \text{ with } \mu_I = \mu_{I_D} + \mu_{I_A} \text{ and } \sigma_I^2 = \sigma^2_{I_D} + \sigma^2_{I_A} \qquad \text{(S15)}$$

The mean values and variances of the fluorescence intensities just sum up to the sum intensity, as one would expect.

Solving the integral for $f_E(E)$ is more involved:

$$f_E(E) = \int_{-\infty}^{\infty} \frac{|I|}{2\pi\sigma_{I_D}\sigma_{I_A}} exp\left(-\frac{((1-E)I-\mu_{I_D})^2}{2\sigma^2_{I_D}} - \frac{(EI-\mu_{I_A})^2}{2\sigma^2_{I_A}} \right) dI$$
$$= \int_{-\infty}^{\infty} |I| exp\left(-O(I^2)\right) dI \qquad \text{(S16)}$$

where $O(I^2)$ represents the exponent, which is a function of I^2. The absolute value of the sum intensity $|I|$ can be replaced by I, because both Gaussians correspond to approximations of Poissonian distributions with μ_{I_D} or μ_{I_A}. The impact of negative I-values is therefore negligible.

Assuming $\mu_{I_D} > 0$ and $\mu_{I_A} > 0$, the integral is soluble and given by:

$$f_E\left(E\middle|\mu_E,\sigma_E^2\right) = \frac{v(E)}{u(E)} \frac{1}{\sqrt{2\pi\sigma_E^2\left(E\right)}} exp\left(-\frac{\left(E-\mu_E\right)^2}{2\sigma_E^2\left(E\right)} \right) \qquad \text{(S17)}$$

where we have used Eqn (S9) to convert the variances $\sigma^2_{I_D}$ and $\sigma^2_{I_A}$ into μ_{I_D} and μ_{I_A} and introduced additional functions $u(E)$, $v(E)$ and $\sigma_E^2\left(E\right)$ to simplify the representation of the formula;

$$\mu_E = \frac{\mu_{I_A}}{\mu_{I_D}+\mu_{I_A}}, \; \mu_I = \mu_{I_D}+\mu_{I_A}$$

$$\sigma_E^2\left(E\right) = \frac{u\left(E\right)}{\mu_I}$$

and

$$u(E) = k_A\ \mu_E\ (1-E)^2 + k_D\ E^2(1-\mu_E)$$

$$v(E) = \mu_E(1-\mu_E)(k_A(1-E)+k_D\ E)$$

To obtain the integral, we rewrote Eqn S16 as the intensity multiplied by an exponential function with a quadratic function as an exponent:

$$f_E\left(E\right) = \int_{-\infty}^{\infty} I\, exp\left(-\left\{ \left[\frac{\left(1-E\right)^2}{2\sigma^2_{I_D}} + \frac{E^2}{2\sigma^2_{I_A}} \right] I^2 - \left[\frac{\mu_{I_D}\left(1-E\right)}{\sigma^2_{I_D}} + \frac{\mu_{I_A}E}{\sigma^2_{I_A}} \right] I + \left[\frac{\mu^2_{I_D}}{2\sigma^2_{I_D}} + \frac{\mu^2_{I_A}}{2\sigma^2_{I_A}} + ln\left(2\pi\sigma_{I_D}\sigma_{I_A}\right) \right] \right\} \right) dI$$

(S18)

This can be equated to the expectation value of a variable x that is distributed with a normal distribution multiplied by a constant c

$$c\langle x\rangle = \int_{-\infty}^{\infty} x \frac{k}{\sqrt{2\pi\sigma_x^2}} exp\left(-\frac{\left(x-\mu_x\right)^2}{2\sigma_x^2}\right) dx$$
$$= \int_{-\infty}^{\infty} x\, exp\left(-\left\{\frac{1}{2\sigma_x^2}x^2 - \frac{\mu_x}{\sigma_x^2}x + \left[\frac{\mu_x^2}{2\sigma_x^2} + \frac{ln\left(2\pi\sigma_x^2\right)}{2} - ln\left(c\right)\right]\right\}\right) dx = c\mu_x \tag{S19}$$

By setting the coefficients of the quadratic exponent equal and solving for μ_x, σ_x and c yields Eqn S17 from the product of $c\mu_x$.

Eqn S17 approximates a normal distribution multiplied by a prefactor, $\frac{v(E)}{u(E)}$ and a variance, $\sigma_E^2\left(E\right)$, that depends on E. We can approximate the prefactor and variance using a Taylor series expansion about $E = \mu_E$, which is the crucial region of the formula. For the prefactor:

$$Taylor\left(\frac{v(E)}{u(E)}, E = \mu_E\right) = 1 - \frac{k_A - k_D}{k_A(1-\mu_E) + k_D\mu_E}(E-\mu_E) + O((E-\mu_E)^2) \tag{S20}$$

and for the variance:

$$Taylor\left(\frac{1}{\mu_I}\left(k_A\mu_E(1-E)^2 + k_D(1-\mu_E)E^2\right), E = \mu_E\right) =$$
$$\frac{\mu_E(1-\mu_E)}{\mu_I}\left(k_A(1-\mu_E) + k_D\mu_E\right) + \ldots$$
$$\frac{2\mu_E(1-\mu_E)}{\mu_I}\left(k_D - k_A\right)(E-\mu_E) + O((E-\mu_E)^2) \tag{S21}$$

Replacing the arguments that depend on E with the zeroth terms of their Taylor-approximations, the pre-factor becomes one and variance is replaced by the zeroth term:

$$\sigma_E^2 = \left(k_A(1-\mu_E) + k_D\mu_E\right)\frac{\mu_E(1-\mu_E)}{\mu_I} \tag{S22}$$

Substituting this into Eqn (S17) yields:

$$f_E(E\,|\,\mu_E,\sigma_E^2) = \frac{1}{\sqrt{2\pi}\sigma_E}\exp\left(-\frac{\left(E-\mu_E\right)^2}{2\sigma_E^2}\right) \tag{S23}$$

Thus, the distribution in FRET efficiency can be well approximated by a Gaussian distribution. In practice, k_A and k_D are equal when there is sufficient signal. In this case, the linear terms in the Taylor series expansions (Eqns (S20) and (S21)) become zero, which further increases the quality of this approximation.

Supplementary Figures

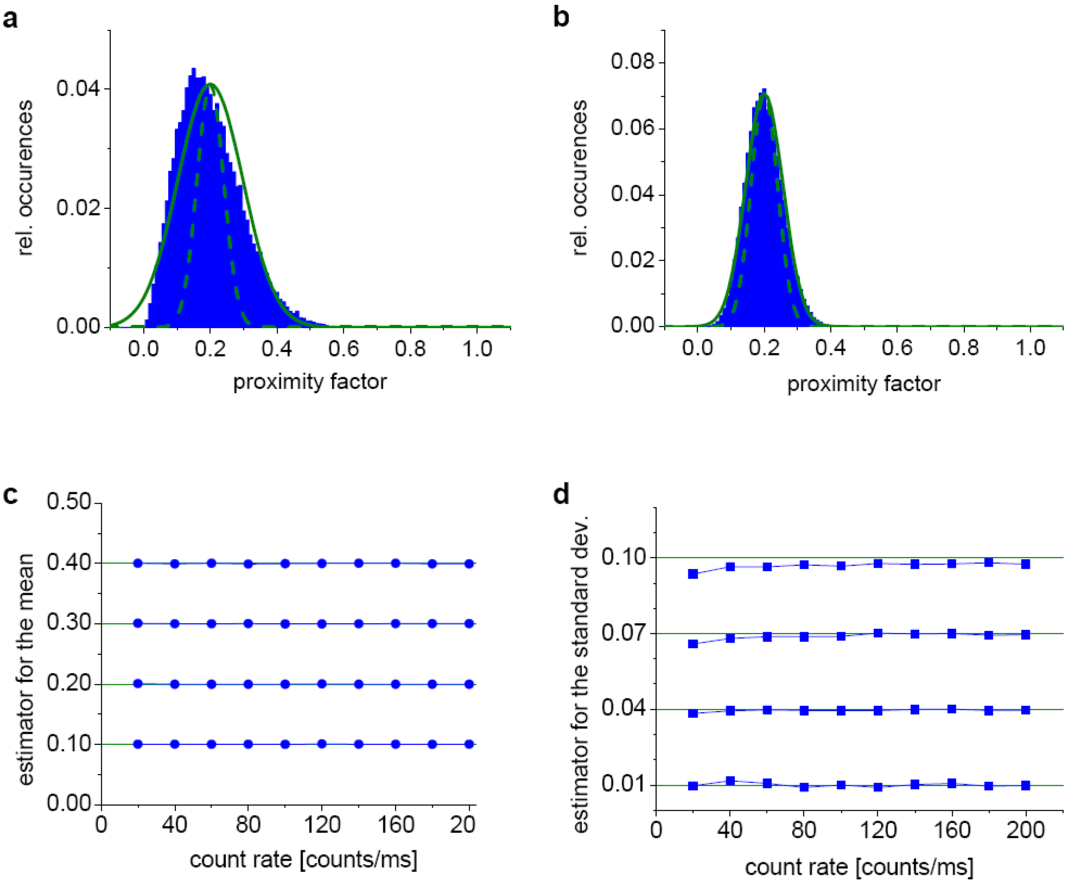

Figure S1: Monte Carlo Simulations and Extended HMM analysis of spFRET histograms. (**a**,**b**) Histogram of the spFRET proximity ratio (blue) with 50,000 simulated data points with an average FRET efficiency of 0.20 and a standard deviation of 0.04 for (**a**) 20 counts/ms and (**b**) 100 counts/ms. Dotted green lines: the underlying hidden FRET distribution. Solid green lines: broadening due to the limited number of measured photons. (**c**) The estimator tested for mean FRET efficiencies of 0.1, 0.2, 0.3 and 0.4 together with an inherent standard deviation of 0.05. The HMM works reliably over a wide range of values even when the underlying Gaussian distribution falls outside of the boarders of 0 and 1 and the appearing asymmetry of the data is not accounted by the estimator. (**d**) The estimation of the variances at standard deviations of 0.01, 0.04, 0.07 and 0.10 together for a mean FRET efficiency of 0.2. A slight systematic deviation at lower count rates and higher standard deviations is observed.

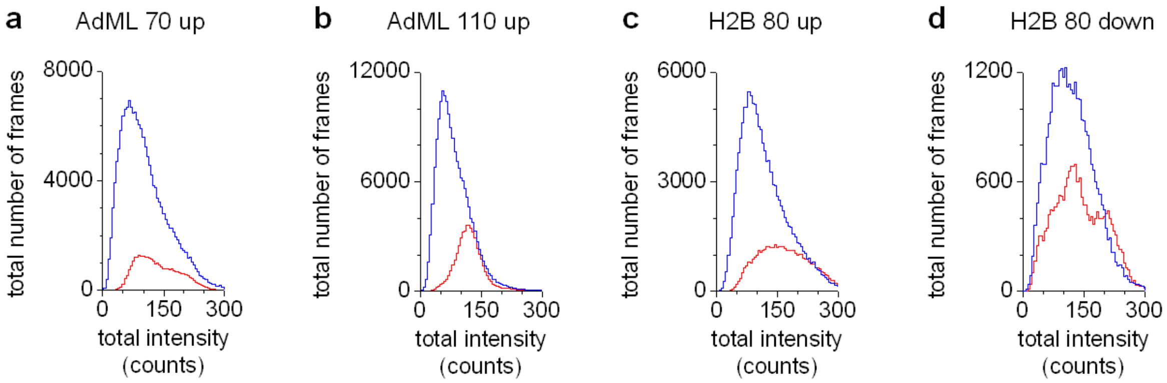

Figure S2: Experimental distribution of the total intensity per frame. Histograms of the total photons detected in the donor and acceptor channels per 5 ms frame are shown for TBP-DNA complexes in the absence (red) and presence (blue) of NC2. The different DNA sequences investigate are (**a**) a 70 bp upstream-labeled DNA containing the AdML TATA box, (**b**) a 110 bp upstream-labeled DNA containing the AdML TATA box, (**c**) an 80 bp upstream labeled DNA containing the H2B TATA box and (**d**) an 80 bp downstream labeled DNA containing the H2B TATA box.

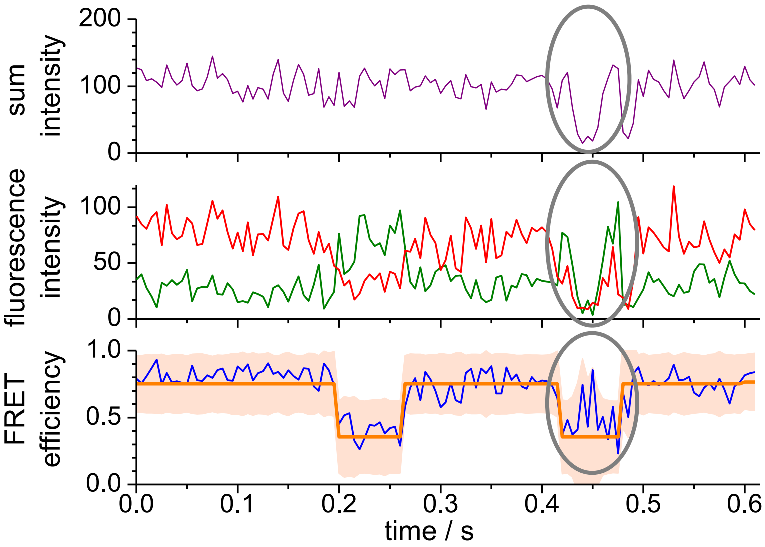

Figure S3: SpFRET trace with donor-quenching. An exemplary spFRET trace of the promoter AdML 110 bp upstream-labeled construct after addition of NC2 where the donor is transiently quenched. Purple: Total intensity per 5 ms, green: donor fluorescence counts per 5 ms, red: acceptor fluorescence counts per 5 ms, blue: FRET efficiency, orange: Viterbi path of the two-state model HMM analysis with the standard deviation due to shot noise shown as the envelope about the Viterbi path. The sudden drop in fluorescence intensity around 0.45 s due to transient quenching of the donor molecule leads to high fluctuations in the FRET efficiency trace. The new estimators are able to account for this effect: the density function (orange) is locally broadened when the total intensity drops, which ensures a correct assignment of the FRET conformation at this time to the low-FRET state.

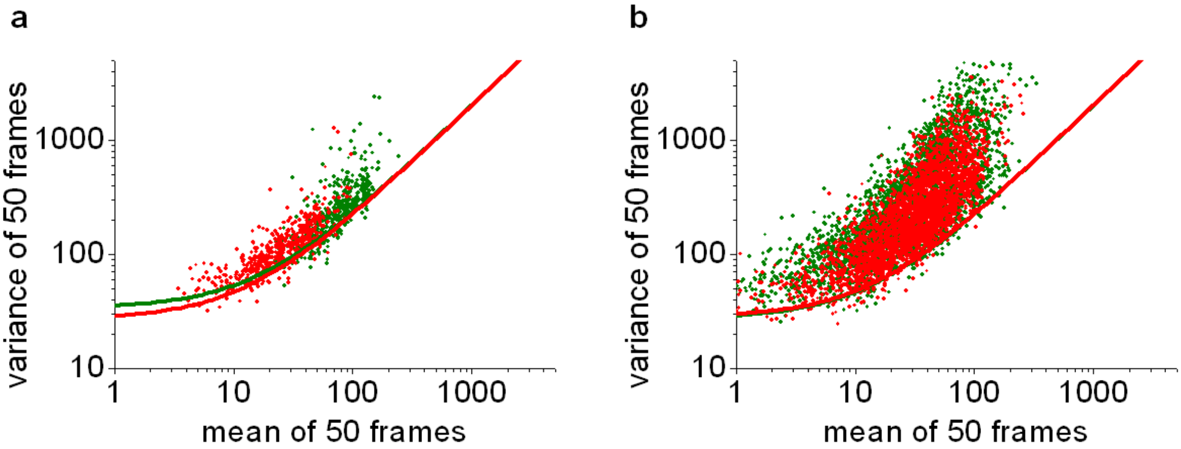

Figure S4: Plots of the experimentally determined variances versus mean intensity. The variance determined from 50 consecutive frames is plotted as a function of the corresponding background-corrected mean value for the donor (green) and acceptor (red) channels in the (**a**) absence (525 points) and (**b**) presence of NC2 (1519 points). The theoretical dependence expected from shot noise:

$$\sigma^2_{D/A,theo}(\mu_{D/A}) = 2\left(\mu_{D/A} + \left\langle I^{background}_{D/A0} \right\rangle\right),$$

are shown as a solid lines for the respective channels. For a Poissonian distribution, the variance is equal to the mean of the detected photons (signal and background). The factor of two accounts for the additional noise generated by the on-chip gain of the EMCCD camera. Before addition of NC2 (a), the data points lie close to the theoretical curve whereas, after addition of NC2 (b), the variances from the data points are much higher than the theoretical curve revealing the presence of additional conformational dynamics.

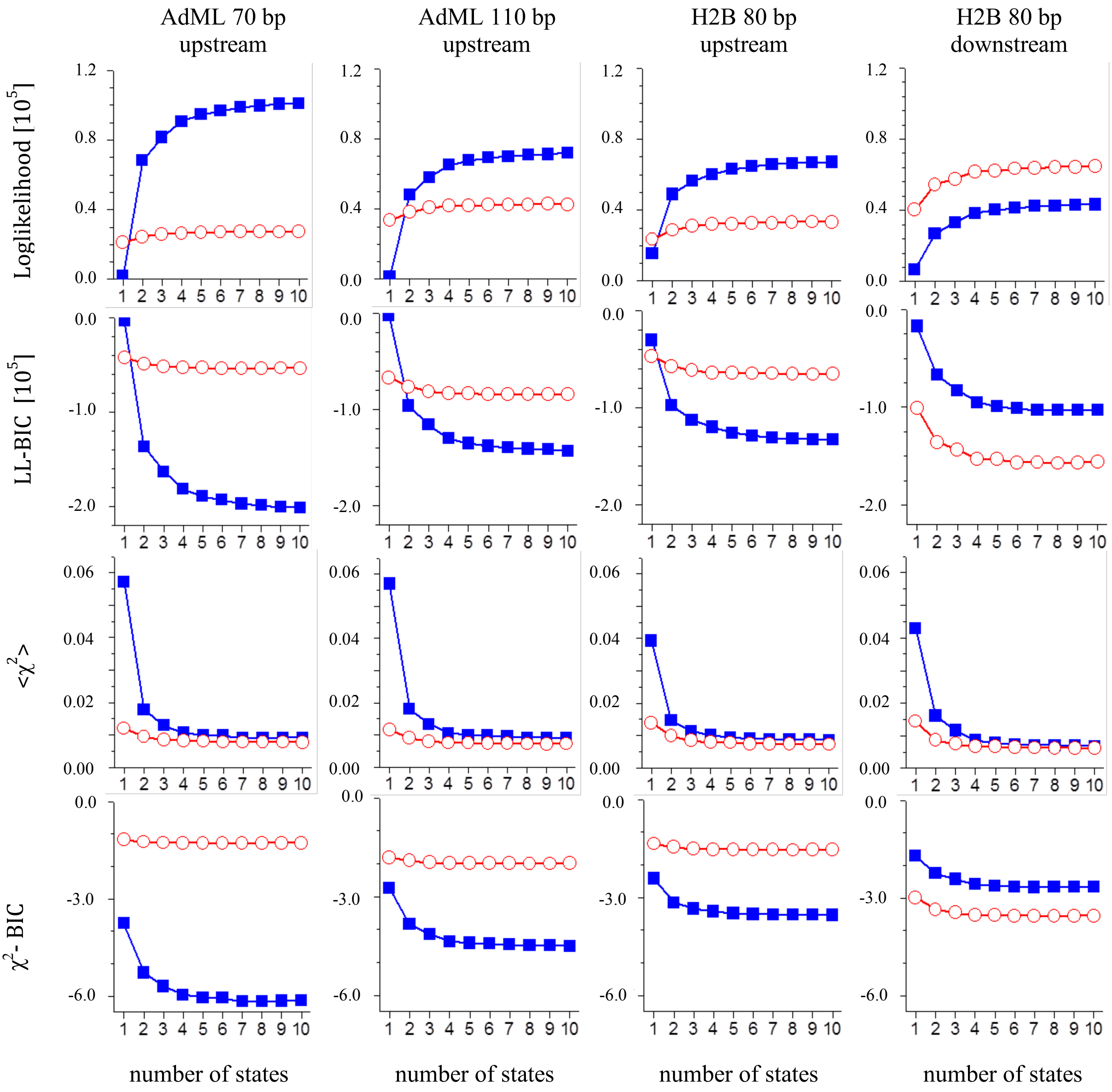

Figure S5: Determination of the number of relevant FRET states. Various criteria were tried to determine the number of distinguishable FRET states in the spFRET measurements. The Loglikelihood, χ^2-value and their corresponding Bayesian Information Criterion (BIC) are shown for the four different samples in the absence (red) and presence (blue) of NC2. First row: The Loglikelihood is plotted as a function of the number of states. Second row: The BIC calculated according to the obtained Loglikelihood-value. Third row: A plot of the χ^2 for the comparison of the Viterbi path to the spFRET data is plotted as a function of the number of states in the HMM. Fourth row: The BIC calculated according to the modified χ^2-value.

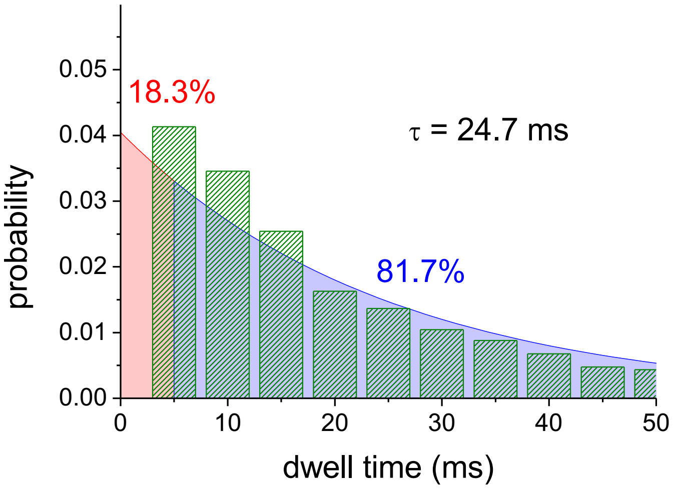

Figure S6: Fraction of Missed Transitions. The survival probability of the 0.64 FRET efficiency state for TBP bound to the 70 bp upstream-labeled DNA containing the AdML TATA box in the presence of NC2 determined from the HMM analysis is shown. An exponential function with a lifetime of 24.7 ms is shown for comparison. Assuming a minimum dwell time of 5 ms for a transition to be detected with the HMM analysis, ~ 18 % of the transitions from $E = 0.40$ to $E = 0.64$ on to $E = 0.83$ would be detected as a direct transition between the $E = 0.40$ and $E = 0.83$ FRET efficiency states.

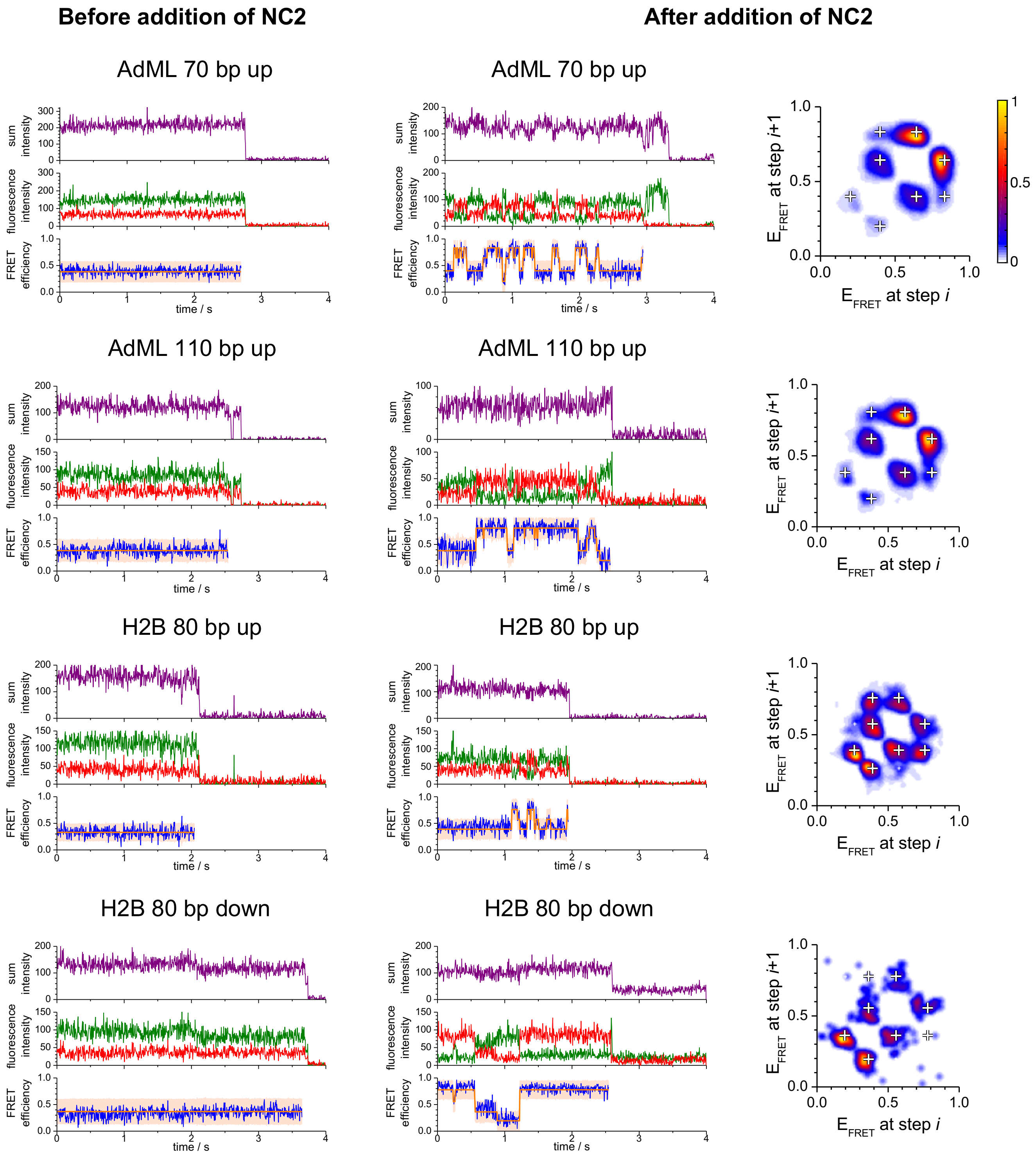

Figure S7: SpFRET traces and TDPs of all four sample preparations. Representative spFRET traces are shown for the four constructs investigated in this work in the absence (*left*) and presence (*middle*) of NC2. The total intensity is shown in purple, the intensity of the donor fluorophore is shown in green, the intensity of the acceptor fluorophore is shown in red, the frame-wise FRET efficiency is shown in blue and the Viterbi path and uncertainty due to shot-noise are shown in orange. (*right*) The TDPs are shown for the different complexes. The optimized Viterbi path from the global four-well HMM analysis was calculated for the individual traces and average FRET efficiency plotted as a Gaussian with a width of 2% for each level. The plots are normalized to the maximum number of transitions and indicates how often the transitions were observed with rare transitions given in blue and more frequent transitions highlighted in yellow. The corresponding color bar is shown to the right. The white pluses represent the values returned from the global four-well HMM analysis.

Supplementary Tables

State	FRET Efficiency μ			Width σ			Dwell Time (ms)		
	Model	Standard HMM	Extended HMM	Model	Standard HMM	Extended HMM	Model	Standard HMM	Extended HMM
s_1	0.250	0.246	0.251	0.002	0.011	0.002	100.0	80.3	106.1
s_2	0.450	0.449	0.450	0.002	0.015	0.002	100.0	90.3	108.8
s_3	0.650	0.651	0.648	0.002	0.014	0.002	100.0	73.9	92.5
s_4	0.850	0.850	0.849	0.002	0.008	0.002	100.0	87.5	110.8

Table S1: Comparison of the Standard HMM and the Extended HMM. Results from a simulation of 20 molecules using either the standard HMM or the HMM where the camera noise is incorporated into the analysis.

	AdML 70 bp up stream	AdML 110 bp up stream	H2B 80 bp up stream	H2B 80 bp down stream
without NC2	103	141	132	62
with NC2	432	315	279	55

Table S2: Number of molecules used by the hidden Markov analysis. Data were collected at 5 ms/frame or 200 Hz. From each molecule, the donor and acceptor intensities where extracted and a FRET trajectory was calculated.

	AdML 70 bp up stream		AdML 110 bp up stream		H2B 80 bp up stream		H2B 80 bp down stream	
	–NC2	+NC2	–NC2	+NC2	–NC2	+NC2	–NC2	+NC2
1 hidden state	103	432	141	315	132	279	62	55
2 hidden states	147	3643	217	2610	252	1113	82	128
3 hidden states	173	8146	241	4128	330	1829	123	184
4 hidden states	160	9546	252	6416	304	1966	112	291
5 hidden states	176	9786	256	6381	323	1984	124	340
6 hidden states	181	9299	234	6008	354	2031	129	366
7 hidden states	183	11797	239	6126	378	1885	140	411
8 hidden states	150	11746	262	7002	315	1853	137	358
9 hidden states	200	12688	263	6489	273	1807	128	367
10 hidden states	180	12582	296	8235	313	1930	132	371

Table S3: Transitions found in Various HMM Analyses. The number of detected transitions for the different hidden Markov models for each sample before (steady) and after (dynamic) the addition of NC2.

Number of hidden states	**AdML 70 bp up stream**			**AdML 110 bp up stream**			**H2B 80 bp up stream**			**H2B 80 bp down stream**		
	μ	σ	*f*	μ	σ	*f*	μ	σ	*f*	μ	σ	*f*
1	0.61	0.22	100%	0.57	0.22	100%	0.46	0.18	100%	0.45	0.19	100%
2	0.39	0.09	43%	0.36	0.12	45%	0.36	0.09	72%	0.33	0.10	66%
	0.78	0.12	57%	0.75	0.09	55%	0.71	0.09	28%	0.68	0.11	34%
3	0.34	0.10	32%	0.26	0.09	21%	0.30	0.08	39%	0.31	0.09	56%
	0.60	0.08	23%	0.48	0.08	30%	0.44	0.06	36%	0.52	0.07	25%
	0.82	0.05	44%	0.77	0.07	49%	0.73	0.08	25%	0.76	0.07	19%
4	0.20	0.08	7%	0.20	0.07	10%	0.26	0.07	19%	0.19	0.06	15%
	0.40	0.07	29%	0.38	0.06	29%	0.39	0.06	47%	0.36	0.06	47%
	0.64	0.07	23%	0.62	0.06	24%	0.58	0.06	14%	0.56	0.06	22%
	0.83	0.04	41%	0.81	0.04	37%	0.76	0.07	19%	0.78	0.06	17%
5	0.18	0.08	6%	0.18	0.07	9%	0.23	0.07	10%	0.19	0.06	13%
	0.36	0.06	21%	0.35	0.05	23%	0.36	0.06	43%	0.35	0.05	42%
	0.50	0.06	15%	0.50	0.05	15%	0.47	0.06	22%	0.49	0.05	18%
	0.70	0.07	24%	0.68	0.06	26%	0.69	0.05	18%	0.63	0.03	12%
	0.85	0.03	34%	0.83	0.03	28%	0.82	0.06	7%	0.79	0.05	14%
6	0.15	0.07	4%	0.16	0.06	6%	0.20	0.07	7%	0.17	0.06	10%
	0.32	0.06	12%	0.30	0.05	13%	0.33	0.06	30%	0.30	0.05	22%
	0.42	0.05	19%	0.41	0.05	20%	0.41	0.05	30%	0.40	0.04	29%
	0.57	0.06	11%	0.57	0.06	13%	0.55	0.06	11%	0.54	0.05	18%
	0.73	0.06	27%	0.71	0.05	26%	0.72	0.04	18%	0.71	0.04	11%
	0.86	0.03	27%	0.84	0.02	23%	0.86	0.05	4%	0.83±	0.04	9%
7	0.14	0.06	4%	0.13	0.06	4%	0.17	0.06	3%	0.15	0.05	8%
	0.28	0.06	7%	0.26	0.05	9%	0.29	0.06	16%	0.27	0.05	16%
	0.39	0.05	20%	0.37	0.05	20%	0.37	0.05	35%	0.37	0.05	34%
	0.51	0.05	10%	0.49	0.05	10%	0.47	0.05	18%	0.49	0.05	16%
	0.65	0.06	16%	0.62	0.06	14%	0.63	0.05	10%	0.62	0.03	11%
	0.80	0.03	31%	0.75	0.04	25%	0.74	0.04	14%	0.75	0.03	11%
	0.89	0.02	13%	0.85	0.02	18%	0.88	0.05	3%	0.86	0.03	5%
8	0.13	0.06	3%	0.11	0.06	3%	0.16	0.06	3%	0.14	0.05	6%
	0.25	0.06	4%	0.22	0.05	7%	0.27	0.06	12%	0.25	0.05	12%
	0.36	0.05	15%	0.33	0.05	14%	0.35	0.05	32%	0.35	0.05	32%
	0.44	0.05	14%	0.42	0.04	15%	0.43	0.05	21%	0.43	0.05	14%
	0.57	0.05	7%	0.55	0.05	9%	0.54	0.05	9%	0.52	0.04	12%
	0.67	0.06	16%	0.66	0.06	16%	0.67	0.04	10%	0.64	0.03	9%
	0.81	0.02	29%	0.78	0.02	22%	0.75	0.03	11%	0.75	0.02	10%
	0.89	0.02	12%	0.86	0.02	14%	0.88	0.04	3%	0.86	0.03	5%

Number of hidden states	**AdML 70 bp up stream**			**AdML 110 bp up stream**			**H2B 80 bp up stream**			**H2B 80 bp down stream**		
	μ	σ	*f*	μ	σ	*f*	μ	σ	*f*	μ	σ	*f*
9	0.12	0.06	3%	0.11	0.05	2%	0.15	0.06	3%	0.14	0.05	5%
	0.24	0.06	4%	0.21	0.05	6%	0.25	0.06	6%	0.22	0.05	9%
	0.34	0.05	11%	0.32	00.05	12%	0.32	0.05	19%	0.31	0.05	19%
	0.42	0.05	14%	0.40	0.04	16%	0.38	0.05	27%	0.38	0.04	26%
	0.52	0.05	9%	0.51	0.05	7%	0.46	0.05	15%	0.49	0.05	16%
	0.66	0.05	11%	0.61	0.06	12%	0.57	0.05	7%	0.62	0.03	10%
	0.68	0.07	9%	0.71	0.04	13%	0.68	0.04	10%	0.73	0.03	8%
	0.82	0.02	28%	0.80	0.02	25%	0.76	0.03	9%	0.80	0.01	6%
	0.89	0.02	11%	0.88	0.01	6%	0.88	0.04	3%	0.87	0.03	3%
10	0.11	0.05	2%	0.08	0.05	1%	0.14	0.06	2%	0.13	0.05	4%
	0.21	0.06	3%	0.19	0.05	5%	0.23	0.05	5%	0.20	0.05	7%
	0.30	0.06	7%	0.29	0.05	10%	0.30	0.06	14%	0.30	0.04	17%
	0.38	0.05	15%	0.38	0.04	15%	0.37	0.05	30%	0.37	0.04	24%
	0.45	0.05	10%	0.47	0.05	10%	0.44	0.05	18%	0.45	0.05	12%
	0.56	0.05	7%	0.59	0.05	9%	0.54	0.05	8%	0.53	0.04	10%
	0.69	0.05	12%	0.63	0.06	6%	0.66	0.04	8%	0.63	0.03	9%
	0.68	0.07	8%	0.71	0.03	13%	0.73	0.03	9%	0.74	0.02	7%
	0.82	0.01	26%	0.81	0.01	25%	0.79	0.03	5%	0.81	0.02	7%
	0.89	0.02	10%	0.88	0.01	5%	0.90	0.03	2%	0.88	0.02	2%

Table S4: Results from various HMM Analyses. The FRET efficiencies *μ* with their residual standard deviations ***σ*** beyond shot-noise broadening and the relative occurrences for a global hidden Markov model analysis with 1 to 10 states for all four samples after addition of NC2.

	Results from Global four-well HMM				**Results from a Global HMM to a linear four-well model**			
Transition Probability Matrix	0.9659	0.0313	0.0022	0.0005	0.9658	0.0342	0.0000	0.0000
	0.0087	0.9459	0.0405	0.0049	0.0093	0.9431	0.0476	0.0000
	0.0007	0.0509	0.8170	0.1314	0.0000	0.0575	0.8050	0.1375
	0.0001	0.0034	0.0789	0.0005	0.0000	0.0000	0.0870	0.9130
Rates (s^{-1}) $k_{\text{row}\rightarrow\text{column}}$	-	6.26	0.44	.01	-	6.84	0.00	0.00
	1.74	-	8.10	0.98	1.86	-	9.52	0.00
	0.14	10.18	-	26.28	0.00	11.50	-	27.5
	.02	0.68	15.78	-	0.00	0.00	17.4	-
Dwell times (ms)	144.3	89.9	24.7	58.1	143.6	85.3	23.1	55.0
Log-Likelihood	9.0771e+004				9.0670e+004			
BIC	1.8140e+005				1.8110e+005			
χ^2	0.008742				0.010625			

Table S5: Transition Rates Matrices. The transition rate matrices and rate matrix for the four state HMM analysis and for a linear four-well model plotted in the presence and absence of direct transitions between the E = 0.40 and E = 0.83 FRET efficiency states.

References

1. Hirsch, M., R. J. Wareham, M. L. Martin-Fernandez, M. P. Hobson, and D. J. Rolfe. 2013. A stochastic model for electron multiplication charge-coupled devices--from theory to practice. PloS one 8:e53671.
2. Borner, R., D. Kowerko, M. Hadzic, S. L. B. Konig, M. Ritter, and R. K. O. Sigel. 2018. Simulations of camera-based single-molecule fluorescence experiments. PloS one 13:e0195277.
3. Schwarz, G. 1978. Estimating the dimension of a model. The annals of Statistic 6:461-464.
4. Irizarry, R. A. 2001. Information and Posterior Probability Criteria for Model Selection in Local Likelihood Estimation. Journal of the American Statistical Association 96:303-315.
5. McKinney, S. A., C. Joo, and T. Ha. 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. Biophysical journal 91:1941-1951.