

# Multiple Lac-mediated loops revealed by Bayesian statistics and tethered particle motion

Stephanie Johnson\*, Jan-Willem van de Meent<sup>†</sup>, Rob Phillips<sup>‡</sup>, Chris H Wiggins<sup>§</sup> and Martin Lindén<sup>¶</sup>  
(February 6, 2014)

The bacterial transcription factor LacI loops DNA by binding to two separate locations on the DNA simultaneously. Despite being one of the best-studied model systems for transcriptional regulation, the number and conformations of loop structures accessible to LacI remain unclear, though the importance of multiple co-existing loops has been implicated in interactions between LacI and other cellular regulators of gene expression. To probe this issue, we have developed a new analysis method for tethered particle motion (TPM), a versatile and commonly-used in vitro single-molecule technique. Our method, vbTPM, is based on a varia-

tional Bayes treatment of hidden Markov models. It learns the number of distinct states (i.e., DNA-protein conformations) directly from TPM data with better resolution than existing methods, while easily correcting for common experimental artifacts. Studying short (roughly 100 bp) LacI-mediated loops, we are able to resolve three distinct loop structures, more than previously reported at the single molecule level. Moreover, our results confirm that changes in LacI conformation and DNA binding topology both contribute to the repertoire of LacI-mediated loops formed in vitro, and provide qualitatively new input for models of looping and transcriptional regulation. We expect vbTPM to be broadly useful for probing complex protein-nucleic acid interactions.

---

\*Dept. of Biochemistry and Molecular Biophysics, California Institute of Technology, Pasadena, CA, USA. Present address: Dept. of Biochemistry and Biophysics, University of California, San Francisco, CA, USA.

<sup>†</sup>Dept. of Statistics, Columbia University, New York, NY, USA.

<sup>‡</sup>Dept. of Applied Physics and Dept. of Biology, California Institute of Technology, Pasadena, CA, USA.

<sup>§</sup>Dept. of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA.

<sup>¶</sup>Center for Biomembrane Research, Dept. of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, and Dept. of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. E-mail: bmelinden@dbb.su.se

## Introduction

Severe DNA deformations are ubiquitous in biology, with a key class of such deformations involving the formation of DNA loops by proteins that bind simultaneously to two distant DNA sites. DNA looping is a common motif in gene regulation in both prokaryotes and eukaryotes [1–3]. A classic example of a gene-regulatory DNA loop-

ing protein is the Lac repressor (LacI), which controls the expression of genes involved in lactose metabolism in *E. coli* [1–3]. LacI has two DNA binding domains, which can bind simultaneously to two specific sites on the DNA, called operators, to form loops. Despite being one of the best-studied model systems of transcriptional regulation, the mechanics of DNA looping by LacI are not well understood. One of the key outstanding issues regarding the mechanics of loop formation by LacI is that theoretical and computational modeling provide evidence for the existence of many conformations of LacI-mediated loops, but it is not clear which conformations are realized for various loop lengths, nor how many of these different conformations are relevant for gene regulation *in vivo* [4, 5]. Quantitative studies of looping and transcriptional regulation would be greatly aided by a better understanding of the conformations of LacI-mediated loops, as many models of looping are sensitive to assumptions about the conformation of the protein and/or the DNA in the loop [5–8]. Moreover, inducer molecules and architectural proteins, which are important influencers of gene regulation *in vivo*, appear to be able to manipulate these parameters [8–12]. In this work we show that at least three distinct loop structures contribute to LacI-mediated looping *in vitro* for a given DNA construct when the loop length is short (on the order of the DNA persistence length), one more than the two structures that are usually reported [13–18].

The naturally-occurring *lac* operon has three operators with different affinities for the LacI protein [1], allowing loop formation between three different pair-wise combinations. Most studies of looping mechanics avoid this complexity by using synthetic constructs with only two operators, but multiple loop conformations are

possible even in these simplified systems. As shown in Fig. 1A, because the DNA-binding domains of LacI are symmetric [19], each operator can bind in one of two orientations, thus enabling four distinct loop topologies. Moreover, loops could form with the LacI protein on the inside or outside of the DNA loop [5, 13]. In addition, it has been shown that LacI has a flexible joint, allowing the V-like shape seen in the crystal structure to adopt extended conformations as well, as in the rightmost schematic in Fig. 1A [13, 14, 20–23]. Finally, the DNA binding domains seem to rotate easily in molecular dynamics simulations [24], which would help LacI to relax strain in the DNA of the loop [5, 7, 8].

Different predicted loop conformations are usually classified as differing in DNA binding topology or in LacI conformation, with a key distinction between the two being that structures differing in DNA topology cannot directly interconvert without LacI dissociating from one or both operators, in contrast to those differing in LacI conformation (e.g., V-shaped versus extended shapes), which should be able to directly interconvert (see, for example, Ref. [13]).

The existence of multiple loop conformations for LacI-mediated loops *in vitro* has been confirmed experimentally, but identifying these experimentally observed loops with particular molecular structures is challenging. One of the most widely-used techniques for studying LacI loop conformations is a non-fluorescent single-molecule technique called tethered particle motion (TPM [25]; see Fig. 1B), which uses the Brownian motion of a microscopic bead tethered to the end of a linear DNA to report on looping [26]. TPM has resolved two looped states with a variety of synthetic and naturally-occurring DNA sequences [13–17, 27]. However, the structural basis of these two states is currently a subject of

debate. Importantly, direct interconversions between the two looped states have been observed in TPM experiments with 138 bp and 285 bp loops, arguing that a conformational change of LacI is involved for these loops, presumably between a V-like and a more extended state since, as noted above, a change of loop topology would require an unlooped intermediate [13, 14].

There is also evidence from both ensemble and single-molecule fluorescence resonance energy transfer (FRET) experiments with synthetic, pre-bent loop sequences, whose conformations can be determined computationally, for at least two [22, 28] and possibly three [23] co-existing loops differing in both DNA topology and LacI conformation. However, it is as yet unclear which of the structures observed by FRET correspond to the states observed by TPM, and whether three loop conformations might also co-exist in the loops formed from generic rather than pre-bent DNA sequences.

One difficulty in determining the number of looping conformations in TPM measurements is that not all loop conformations produce distinct TPM signals [6, 18], raising the possibility that the actual number of conformations might be greater than two. Indeed, elastic modeling consistently predicts the coexistence of more than two conformations for a single looping construct, either through direct arguments (*i.e.*, finding multiple loop structures with comparable free energies [6, 8]), or indirectly, by predicting that the most stable V-shaped loops and the most stable extended loops have different DNA topologies [5, 7]. In the latter case, the lowest energy states of the V-shaped and the extended conformations would be geometrically unable to interconvert directly with each other, since they differ in DNA topology. Thus, previous reports of direct loop-loop interconversions [13, 14] would

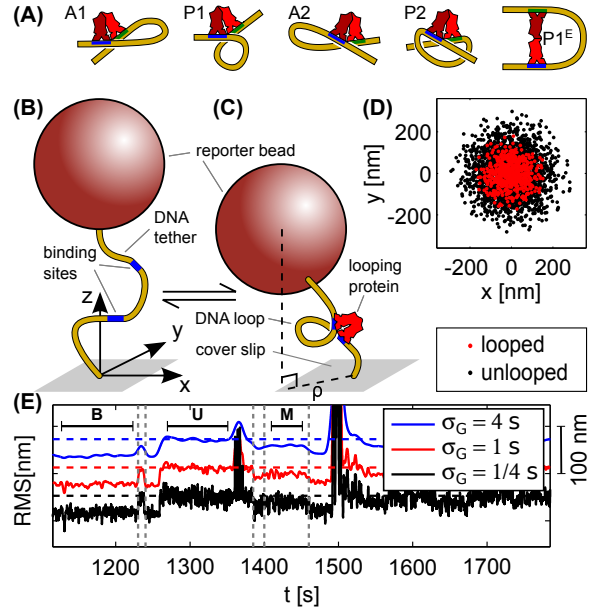


Figure 1: (A) Examples of possible LacI-mediated loops, using the notation of Ref. [7]. (B-C) Tethered particle motion (TPM) setup, in which a reporter bead tethered to a cover slip by a DNA molecule is tracked as it diffuses around the tethering point. The formation of a DNA loop shortens the DNA "leash", which narrows the distribution of bead positions (D). The degree of restriction depends not only on the length of the loop, but also on the relative distance and orientation of the in- and outgoing strands, so that different loop shapes can be distinguished. (E) Root-mean-squared (RMS) signals, time-averaged with Gaussian filters of different kernel width  $\sigma_G$  (see Methods), for an example trace with an unlooped and two looped states (one long stretch of each indicated by U, M, and B respectively). Horizontal dashed lines indicate the unlooped state, offset for clarity, and vertical ones indicate potential loop-loop interconversion events.

have to be explained by the existence of at least one additional loop structure that shares a DNA topology with one of the lowest energy states.

These considerations suggest two questions to address in order to make progress towards identifying the loop structures relevant for looping *in vitro*: (1) is there evidence for more than two loop structures underlying previously reported TPM data, as would be expected from elastic

modeling and from FRET results with pre-bent sequences?, and (2) which of the observed states interconvert directly, identifying them as differing in LacI conformation rather than DNA binding topology?

Shorter loop lengths (*i.e.*, shorter than the persistence length of DNA, roughly 150 bp) tend to enhance the free energy differences between loop structures, and so provide an interesting opportunity to look for detectable signatures of additional loop structures, and to determine which state(s) directly interconvert. We recently reported TPM data of two apparent looped states for loops lengths around 100 bp [16], but the presence or absence of direct interconversions between the two states was not systematically examined. Here, we revisit these data to address the questions of direct interconversions and the number of looped states more rigorously. We demonstrate the presence of a third looped state in addition to the two previously reported, as well as direct interconversions between two of the three states.

These results necessitated the development of an improved analysis tool for TPM, to overcome the technical difficulties of the reduced signal-to-noise in TPM data with short loops compared to longer ones, which impairs the high time resolution required for detecting direct loop-loop interconversions. To meet this challenge, we have developed a powerful set of analysis techniques for TPM data, based on a variational Bayesian treatment of Hidden Markov models that have previously been successfully applied to single molecule FRET [29–33] and single-particle tracking [34]. This toolbox, which we call vbTPM, offers several advantages over existing analysis techniques, including improved resolution, an objective criterion to determine the number of (distinguishable) DNA/protein con-

formational states, robustness against common experimental artifacts, and a systematic way to pool information from many trajectories despite considerable cross-sample heterogeneity.

vbTPM should benefit a broad community of users, as TPM is a versatile and widely-used single molecule technique, with its simplicity, stability, ability to measure DNA-protein interactions at very low applied tension [35, 36], and the potential for high throughput through parallelization [37] making it an attractive tool for *in vitro* studies of protein-nucleic acid interactions that loop or otherwise deform DNA [15–18, 25, 26, 38–49]. Moreover, our results from applying vbTPM to TPM data on short DNA loops provide important new inputs for a comprehensive understanding of LacI-mediated DNA looping *in vitro* and quantitative models of transcriptional regulation *in vivo*.

## Materials and Methods

### TPM data

We present a new analysis of our previously published data [16] on constructs that contain 100 to 109 bp of either a synthetic random sequence called E8 [50, 51] or a synthetic, strong nucleosome positioning sequence called 601TA (abbreviated TA) [50–52] in the loop, flanked by the strongest naturally occurring LacI operator O1 and an even stronger synthetic operator called Oid. We denote these constructs E8x and TA<sub>x</sub>, where x=100–109 and refers to the length of the loop, excluding the operators. The O1 and Oid operators are 21 and 20 bp long, so the distance between operator centers is thus x+20.5 bp. For ease of comparison between our results and others’, we use loop length, not distance between operator centers, when quoting other’s results.

The *in vitro* affinities of LacI for the O1 and Oid operators are roughly 40 and 10 pM respectively [16, 53–56]. The total lengths of the DNA tethers range from 458–467 bp, depending on the length of the loop [16].

For every tethered DNA, we collected 10 minutes of calibration data in the absence of LacI, followed by roughly 20 to 100 minutes of looping data in the presence of 100 pM LacI, purified in-house. Data sets for each loop length typically contain 50–100 TPM trajectories. We used a standard brightfield microscopy-based TPM setup, where 490 nm diameter polystyrene beads are tracked in the xy-plane with video microscopy at 30 Hz, and the resulting trajectories then drift-corrected using a first-order Butterworth filter with a 0.05 Hz cutoff frequency (see Ref. [16] for detailed experimental and analysis procedures). As noted below, this drift-corrected data was used as the input for the HMM analysis (and not the subsequently Gaussian-filtered RMS trajectories that are described in Ref. [16]).

In addition to the pre-existing data, we also obtained calibration trajectories from constructs with total lengths 450 bp (“E894” of Ref. [16]), 735 bp (“wild-type” of Ref. [57]), and 901 bp (“PUC306” of Refs. [15, 57])). Data for these constructs were obtained in the absence of LacI only.

## RMS analysis

The root-mean-square (RMS) trace of a tether is the square root of a running average of the variance of the bead’s position,  $\sqrt{\langle \rho^2 \rangle}$ . We followed the procedures of Ref. [16], except that we varied the standard deviation  $\sigma_G$  of the Gaussian filter kernel for the running average. To count the number of states, we determine the number of peaks in RMS histograms by eye.

## Diffusive HMM for single trajectories

vbTPM is based on a hidden Markov model (HMM) for the bead motion and looping kinetics. For single trajectories, we model the looping kinetics by a discrete Markov process  $s_t$  with  $N$  states (*e.g.*,  $s_t = 1$  when unlooped,  $s_t = 2$  when looped, *etc.*), a transition probability matrix  $\mathbf{A}$ , and initial state distribution  $\boldsymbol{\pi}$ ,

$$p(s_t|s_{t-1}, \mathbf{A}) = A_{s_{t-1}s_t}, \quad p(s_1|\boldsymbol{\pi}) = \pi_{s_1}. \quad (1)$$

This is the hidden part of a hidden Markov model (HMM), and the physics of TPM go into the emission model, which describes the motion of the bead for each hidden state. We use a discrete-time model of over-damped 2D diffusion in a harmonic potential that has been suggested as a simplified model for TPM [58, 59],

$$\mathbf{x}_t = K_{s_t} \mathbf{x}_{t-1} + \mathbf{w}_t / (2B_{s_t})^{1/2}, \quad (2)$$

where the index  $s_t$  indicate parameters that depend on the hidden state, and  $\mathbf{w}_t$  are uncorrelated Gaussian random vectors with unit variance. Hence, the probability distribution of each bead position conditional on the hidden state and previous position is given by

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t, \mathbf{K}, \mathbf{B}) = \frac{B_{s_t}}{\pi} e^{-B_{s_t}(\mathbf{x}_t - K_{s_t} \mathbf{x}_{t-1})^2}. \quad (3)$$

The emission parameters  $K_j$  and  $B_j$  are related to the spring and diffusion constant of the model. More insight into their physical meaning can be gained by noting that with a single hidden state, Eq. 3 describes a Gaussian process with zero mean and

$$\begin{aligned} \text{RMS} &= \sqrt{\langle \rho^2 \rangle} = \sqrt{\langle \mathbf{x}^2 \rangle} = (B(1 - K^2))^{-1/2}, \\ \frac{\langle \mathbf{x}_{t+m} \cdot \mathbf{x}_t \rangle}{\langle \mathbf{x}^2 \rangle} &= K^m \equiv e^{-m\Delta t/\tau}, \end{aligned} \quad (4)$$

where  $\Delta t$  is the sampling time, and  $\tau = -\frac{\Delta t}{\ln K}$  is a bead correlation time. This model captures the diffusive character of the bead motion while retaining enough simplicity to allow efficient statistical analysis.

## Statistical analysis

To analyze TPM trajectories using the above model, we apply a variational Bayesian (VB) [60, 61] technique, that has previously been used in the analysis of other single-molecule data [29–31, 34], but has not been applied to tethered particle motion data so far.

VB methods can determine both the most likely number of hidden states  $N$  and the most likely parameters  $\theta = \{\mathbf{A}, \boldsymbol{\pi}, \mathbf{K}, \mathbf{B}\}$  for the model. Models with more states and parameters can generally model the data more closely, but may ‘overfit’ the data by attributing noise fluctuations to separate states. VB methods perform ‘model selection’ by ranking models according to a lower bound  $F_N$  on the log evidence  $\ln L_N$ . The evidence  $L_N$  is the ‘marginal’ probability of observing the measurement data, obtained by integrating out all model parameters  $\theta$  and hidden state sequences  $\{s_t\}$  from the joint probability  $p(\{\mathbf{x}_t\}, \{s_t\}, \theta | N)$ ,

$$F_N \lesssim \ln L_N = \ln \sum_{s_1, s_2, \dots} \int p(\{\mathbf{x}_t\}, \{s_t\} | \theta, N) p(\theta | N) d\theta. \quad (5)$$

The model with the highest lower bound log evidence  $F_N$  can be interpreted as the model that exhibits the best “average” agreement with the data over a range of parameters, thereby eliminating models that overfit the data and only show good agreement for a narrow parameter range. VB analysis requires us to parameterize our prior knowledge (or ignorance)

about parameter values in terms of prior distributions  $p(\theta | N)$ . We choose ‘uninformative’ priors to minimize statistical bias. VB analysis also yields parameter information in terms of (approximate) posterior distributions on  $\theta$ , which are optimized numerically to maximize  $F_N$  when fitting a model to data. We generally report parameter values as expectation values of these distributions. Further details are given in the Supporting information (SI) and software documentation (see below).

## Downsampling

To decrease the computational cost associated with analysis of large data sets, we downsample by restricting the hidden state changes to occur on multiples of  $n$  data points. By downsampling only the hidden states, and not the TPM data, we avoid discarding valuable information about bead relaxation dynamics [59, 62]. We use  $n = 3$  except where noted otherwise. With an original sampling frequency of 30 Hz and  $K \gtrsim 0.4$  ( $\tau \gtrsim 1/30$  s) in our data (see Results), the shortest possible state lifetime (1/10 s after downsampling) is thus at most three times larger than the bead correlation time.

## Synthetic data

We generate synthetic data by direct simulation of Eqs. (1) and (3), followed by application of a first-order Butterworth filter with 0.05 Hz cut-off frequency to simulate drift-correction [15, 16]. To generate reasonable parameter pairs, we use the empirical fit  $\tau = 0.018 \text{ RMS} - 0.079$ , with  $\tau$  in seconds and RMS in nm, and then compute  $K, B$  from Eq. (4). For analysis, we use the same settings (priors, *etc.*) as for real data.



## Multiple trajectories

To fully utilize the high-throughput capabilities of TPM, it is advantageous to pool information from many trajectories in a systematic way. Indeed, we will see below that this was necessary to unambiguously resolve direct interconversions between looped states. However, pooling information from multiple trajectories raises two challenges. First, TPM data contain artifacts, *e.g.* transient sticking events or tracking errors (described in more detail below). Such spurious events are specific to each trajectory, and should not be pooled. Second, variations in bead size, attachment chemistry, *etc.*, create significant variability between beads in nominally equal conditions (*e.g.* DNA construct length and LacI concentration[16]), hence it would not be feasible to fit a single model to multiple trajectories even without spurious events.

To address the first problem, we extend the single trajectory HMM with a second type of hidden state,  $c_t$  such that  $c_t = 1$  indicates genuine looping dynamics governed by the simple model described above. When  $c_t > 1$ , the apparent bead motion is assumed to arise from some form of measurement artifact, which is modeled by a different set of emission parameters  $\hat{B}_{c_t}, \hat{K}_{c_t}$ . We assume that genuine states,  $s_t$ , to evolve independently of spurious events. Similarly, spurious events  $c_t > 1$  can interconvert independently of the underlying genuine state, but transitions out of  $c_t = 1$  depend on  $s_t$ , to allow for possibilities such as transient sticking events being more frequent in a looped state when the bead is on average closer to the cover slip. The resulting dependency structure of the hidden states resemble a factorial HMM [63], with the joint transition probabilities

$$p(s_{t+1}, c_{t+1} | s_t, c_t) = p(s_{t+1} | s_t) p(c_{t+1} | s_t, c_t), \quad (6)$$

with  $p(s_{t+1} | s_t) = A_{s_t s_{t+1}}$  as earlier, and

$$p(c_{t+1} | s_t, c_t) = \begin{cases} \hat{A}_{s_t c_{t+1}}, & \text{if } c_t = 1, \\ \hat{R}_{c_t c_{t+1}}, & \text{if } c_t > 1, \end{cases} \quad (7)$$

To deal with bead-to-bead variability, we adopt an empirical Bayes (EB) approach that derives from a recently-developed analysis technique for single-molecule FRET data [32, 33]. In EB analysis, the prior is interpreted as the distribution of model parameters across the set of trajectories, and is learned from the data to maximize the total lower bound log evidence. In this manner, similarities between trajectories are exploited to obtain more accurate parameter estimates. We restrict EB analysis to transition probabilities and emission parameters of the genuine states, while priors describing spurious states are held fixed.

Pooled analysis using EB and the extended model is performed in three steps. First, all trajectories in a data set are analyzed individually with the simple HMM, learning the optimal number of states for each trajectory. Second, looped states and artifact states are classified using an automated procedure (see Eq. (8) below), and verified manually using a graphical tool. In practice, very few corrections to the automated classification are needed. Finally, initial factorial models are generated, with additional unoccupied genuine states for trajectories where not all genuine states occur, and these factorial models are analyzed further using an EB procedure to optimize the parameters of the prior distribution, also known as hyperparameters, for the genuine looped states. Information about the whole data set can then be extracted from the optimized hyperparameters. Further details are given in the software documentation.

## Implementation

vbTPM runs on Matlab with inner loops written in C, and includes a graphical tool for manual state classification. The source code and software documentation are available at [vbtpm.sourceforge.net](http://vbtpm.sourceforge.net).

## Results

### Improved resolution on synthetic data

A simple and common way to analyze TPM data is in terms of RMS values, which are the square root of the bead position variance, or the projected distance  $\rho$  between the bead center and tether point (Fig.1E). Transitions can be extracted by thresholding RMS traces, and the number of states by counting peaks in RMS histograms [13, 16, 17, 26, 42, 64, 65]. However, the RMS signal must be smoothed in order for the transitions to appear. This degrades the time resolution [66], and an analysis without smoothing, such as vbTPM, would likely do better [67]. As noted above, good time resolution is of particular interest for one of the issues we wish to address here, namely, determining whether or not apparent loop-loop interconversions are in fact separated by short unlooped intermediates.

We have therefore tested vbTPM on synthetic data, and compared its ability to resolve closely-lying states with that of the RMS histogram method. Two states can be difficult to resolve either due to similar RMS values or short lifetimes. Our state detection tests (see Fig. S1-S3) show that vbTPM offers a great improvement over RMS histograms in the latter case, which is precisely the case that matters most for the question of direct interconversions that we address here. For example, two states separated

by 40 nm are resolved by vbTPM at a mean lifetime of about 0.5 s, while lifetimes of 4-8 s are necessary for states to be resolvable in RMS histograms (Fig. S1). This order of magnitude improvement mainly reflects the detrimental effects of the low-pass filter used in the RMS analysis. The difference diminishes for more long-lived states, and with mean lifetimes of 30 s, the spatial resolution is about 15 nm for both methods (Figs. S2 and S4).

Our tests with synthetic data further show that the parameters, including transition rates, are faithfully recovered by vbTPM, and that all of these results are insensitive to downsampling by the factor of three that we use when analyzing real data (Fig. S4-S6).

### Detection of experimental artifacts

A striking illustration of the improved time resolution of vbTPM is the ability to detect and classify short-lived experimental artifacts in the data. Our normal TPM protocol starts with a short calibration run in the absence of the looping protein for quality control reasons [16]. Here, we expect only one state, that of the fully extended tether. However, analyzing calibration data for three different construct lengths, we find more than one state in most trajectories, although a single state usually accounts for most ( $\sim 99\%$ ) of the trajectory.

Inspection of the coordinate traces (that is, the  $x$  and  $y$  positions of the bead as functions of time) reveals the dominant state to correspond to normal, “genuine” bead motion, while the extra “spurious” states are associated with obvious irregularities in the data. Many of these are too short-lived to show up in the time-averaged RMS traces. Almost all can be interpreted as either transient sticking events (Fig. 2A-B), where the



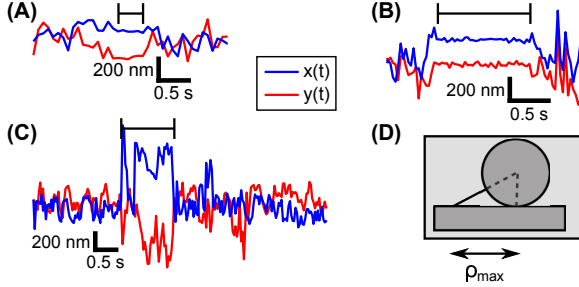


Figure 2: Examples of spurious events in calibration data (*i.e.*, in the absence of repressor). Spurious events are marked by horizontal black lines above the blue and red time traces of the bead’s  $x$  and  $y$  positions. (A,B) show “sticking events” (non-specific, transient attachments of the bead to the surface, the DNA to the bead, *etc*), while (C) contains an excursion larger than the physically possible maximum,  $\rho_{\max}$ , as shown in (D). This could be due to a tracking error, for example when an untethered bead diffuses through the field of view.

motion in  $x$  and  $y$  simultaneously and abruptly goes down dramatically, or brief excursions beyond the limit set by the tether length (Fig. 2C-D), caused by breakdowns of the tracking algorithm when free beads diffuse through the field of view. Some spurious events are described as more than one state in the vbTPM analysis.

A scatter plot of the emission parameters  $K$  and  $B$  for detected states (see the Methods section above) show different patterns for genuine and spurious states (Fig. 3). Genuine states fall along a curve in the  $K, B$  plane, while the spurious states scatter. This makes physical sense, since the genuine dynamics are governed by a single parameter, the effective tether length, while the spurious states are of diverse origins. This pattern persists also in trajectories with looping, with the genuine looped states continuing along the curve indicated by the calibration states (Fig. 4A).

The  $K, B$  values of different trajectories vary significantly, but it turns out that within fitting

uncertainty, most states of individual trajectories satisfy

$$K_{\text{genuine}} \leq K_{\text{cal.}}, \text{ and } B_{\text{genuine}} \geq B_{\text{cal.}}, \quad (8)$$

with  $K_{\text{cal.}}, B_{\text{cal.}}$  parameterizing the genuine state in the calibration trajectory, while most spurious states violate at least one of these inequalities. An intuitive rationale for this rule is that  $K$  ( $B$ ) tends to decrease (increase) with decreasing tether length as seen in Fig. 3. Looping decreases the effective tether length, as does the slight bending of the operator sites upon LacI binding [16, 19].

The upshot of the different behaviors of genuine and spurious states shown in Figs. 3 and 4(A) is that Eq. (8), plus an additional lower threshold on RMS values (see Eq. (4)) to catch sticking events near the tethering point, can be used to computationally label genuine versus spurious states. Very few exceptions remain to be corrected manually.

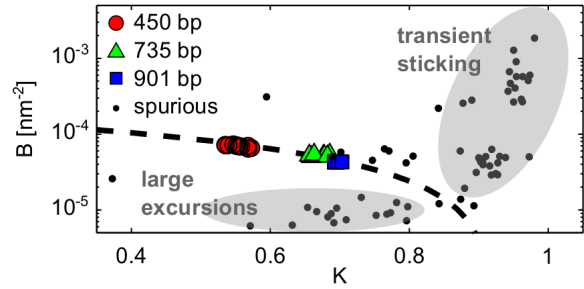


Figure 3: Scatter plot in the  $(K, B)$  plane of genuine and spurious states in trajectories without LacI from three different tether lengths. The genuine states, colored according to tether length, are defined as the most long-lived state in each trajectory, and fall close to the empirical fit  $B = (1.84 - 2K) \times 10^{-4} \text{ nm}^{-2}$  (dashed line, note log-scale on the  $B$ -axis). Spurious states (dots) scatter off of this line. Gray ellipsoids indicate rough parameter trends for sticking and tracking errors (large excursions) respectively.

## More than two looped states

We have used this improved analysis technique to examine looping at 100 pM LacI in E8x and TAx constructs, where “x” indicates the loop length, ranging from x=100 to 109 bp [16], and E8 and TA are two different DNA sequences in the loop (see Materials and Methods). We used Eq. (8) complemented by visual inspection to identify genuine states, and from now on, we will understand all “states” to be genuine unless stated otherwise. Most trajectories exhibit one to three states in the presence of LacI.

We discard trajectories with only one state, as complete lack of looping activity might reflect defective constructs, surface attachment, or LacI molecules [16]. We also discard a small number of trajectories with four states, where inspection reveals either a state split by bursts of spurious events (resulting in artificial differences in state lifetimes), or a genuine-looking state with very low RMS that can be attributed to a sticking event near the tethering point. Thus, our HMM analysis is at first glance consistent with earlier findings of two distinguishable looped states in these constructs [16]. We denote the states from trajectories with three states unlooped (U), “middle” (M), and “bottom” (B), in keeping with the conventions of [16, 17], in which “middle” and “bottom” refer to the tether lengths of the two distinguishable looped states relative to the unlooped state.

The two- and three state-containing trajectories display a striking pattern that we will introduce using the E8106 construct. As shown in Fig. 4A, a scatter plot of the emission parameters for three-state trajectories produces partly overlapping clusters in the  $K, B$ -plane, with these clusters corresponding to the three observed states (U, M, B). Some contributions to

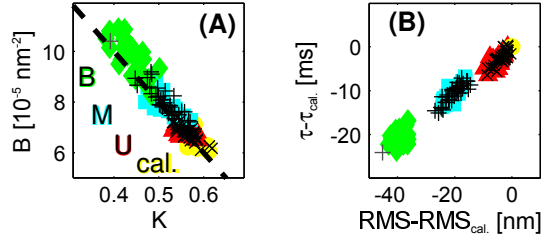


Figure 4: Clustering of LacI-induced looped and unlooped genuine states in the E8106 construct. States U, M, and B in three-state trajectories are represented as filled symbols, while states in two-state trajectories are plotted as +’s (for the looped state) and x’s (for the unlooped state). (A) Raw emission parameters  $K, B$ . The dashed line is the linear fit from Fig. 3. (B) Same states as in A, but plotted as RMS values and relaxation times  $\tau$  (see Eq. (4)) relative to the calibration (that is, no-LacI) states for each tether. From now on, we will plot states in these more intuitive and homogeneous terms.

the parameter noise, such as bead size variations, might be correlated between states, and can thus be reduced by normalization. Indeed, visualizing the states relative to their calibration states (Fig. 4B) produces well-separated state clusters. These clusters allow us to classify the states in the trajectories with only two states, by comparison to the clusters formed by the three-state trajectories. In 37 out of 38 two-state trajectories, the two states coincide with the U and M states. That is, in trajectories that only exhibit one of the two looped states, for the E8106 construct the looped state exhibited is *always* the “middle” state.

This pattern is very unlikely to result from insufficiently equilibrated three-state kinetics ( $P \ll 10^{-6}$ ; see SI, Sec. S3). Similarly, a subpopulation of LacI which is somehow unable to support the B state is unlikely, as different cluster patterns appear with other loops sequences. As shown in Fig. 5, when we subject E8 and TA constructs spanning one helical repeat to the same analysis, we see some constructs (e.g.,

E8103, TA104, E8105, TA106) mimic the 2+3-state pattern of E8106, but others (E8100-101, TA100-101, TA109) seem to have the opposite 2+3-pattern, in that the looped state in two-state trajectories is the B rather than M state. Moreover, while there is also one case for each sequence with almost exclusively 3-state (E8107) or 2-state (TA105) trajectories, the identity of the looped state in two-state trajectories exhibits a clear phasing that correlates with loop length, and therefore with the helical repeat of the DNA. In particular, when the operators are in-phase and looping is maximal, demonstrated in our previous work to occur around 106-bp loops [16], the looped state in two-state trajectories is predominately the M state, whereas when the operators are out-of-phase, around 100 or 110 bp [16], two-state trajectories contain primarily the B state as the looped state.

We propose a structural explanation for these results, namely, that the M-state in trajectories exhibiting only two states corresponds to a different loop structure than the M-state in trajectories with three states. A further line of evidence supporting this explanation concerns the question of whether or not the M and B states in three-state trajectories interconvert: if the M state can interconvert with the B state in three-state trajectories, but the M state in two-state trajectories never interconverts with the B state (because these trajectories show no B state), then it is likely that these two M states (interconverting and not interconverting) are structurally different. Moreover, as noted in the Introduction, the question of direct interconversions can provide insight into what structures might underlie the interconverting and non-interconverting M and B states: if two looped states interconvert without passing through the unlooped state, this would indicate

that the involved states have the same DNA binding topologies, since a change of binding direction would require an unlooped intermediate. To address these questions, we now ask if the looped states in three-state trajectories interconvert directly.

### Direct loop-loop interconversions

Detecting direct interconversions between looped states is difficult. Potential events can be spotted in RMS traces, but as illustrated in Fig. 1E, their interpretation depends on the filter width  $\sigma_G$ , and we cannot exclude the presence of short unlooped intermediates by eye. To test whether the increased temporal resolution of our HMM algorithm could improve upon the detection of short unlooped intermediates, we generated synthetic data using realistic parameters obtained from single-trajectory fits to E8106 and E8107 constructs with three genuine states, with spurious states removed. The transition probabilities  $A_{ij}$  from these fits allow loop-loop interconversions, typically no more than ten per trajectory, but we also generated data without interconversions by setting  $A_{BM} = A_{MB} = 0$ .

Refitting these synthetic data sets with our standard settings, we find that the HMM algorithm over-counts the number of looped state interconversions,  $n_{BM}$ , even when they are absent in the data (Fig. 6A-B). Moreover, models that disallow direct BM-interconversions generally get higher F-values (related to goodness of fit; see Eq. (5)) than models that allow interconversions, when fitted to single trajectories, even when such interconversions are actually present (Fig. 6C-D). Thus, we cannot settle the question of direct loop-loop interconversions by analysis of single trajectories, probably because the

number of such interconversions per trajectory are too few in our data and in the synthetic data we create from it.

To do better, we need to analyze many trajectories at once. The difficulty in this analysis

is that we cannot simply fit a single model to multiple trajectories, because of the large bead-to-bead variations in motion parameters ( $K$ ,  $B$ ) seen in Fig. 4A, and the varying numbers of spurious states in different trajectories (Fig. 3). To solve these problems, we first extend our HMM to split spurious and genuine states into two separate hidden processes (see Methods). Second, we implement an empirical Bayes (EB) approach [32, 33, 61] (see Methods), which optimizes the prior distributions based on the variability of genuine states in different trajectories. This allows information from the whole data set to be used in interpreting each single trajectory, and has been shown to greatly improve the resolution in single molecule FRET data [32].

Analysis of synthetic data, where the true number of interconversion events is known, shows clear improvements when using combined EB analysis in comparison to normal VB methods that analyse each trajectory individually. As shown in Fig. 6A, the tendency to over-estimate the number of BM-interconversions is eliminated when the EB scheme is applied, and almost no such transitions are detected in trajectories where they are absent (Fig. 6B). This shows that the EB scheme can reliably detect the presence of direct B-M interconversions.

When we apply this scheme to real data, we find a substantial number of direct BM-interconversions in the three-state trajectories from the E8106 and E8107 constructs (Fig. 6E-F), which is a strong indication that these constructs do truly display direct loop-loop interconversions. Hence, we conclude that the constructs in Fig. 5 exhibit at least three distinct loop structures, one more than previously reported in a single construct by TPM [13, 14, 16–18]: an M and a B state that can interconvert without an unlooped intermediate, suggesting

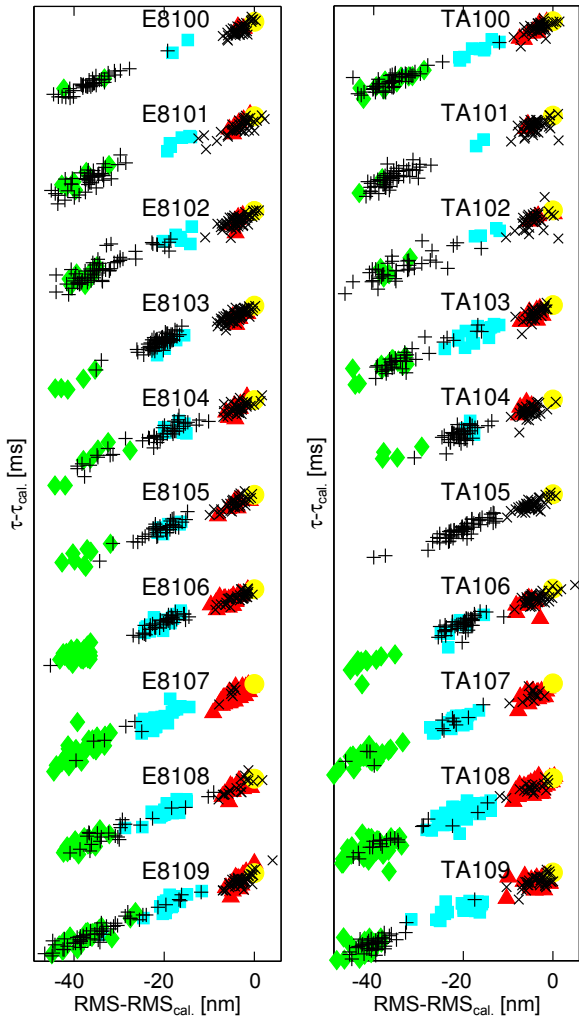


Figure 5: Clustering of looped and unlooped states for E8x and TAx constructs, with loop lengths  $x=100-109$  bp. The states are colored and aligned as in Fig. 4B, and offset in the  $\tau$  direction for clarity.

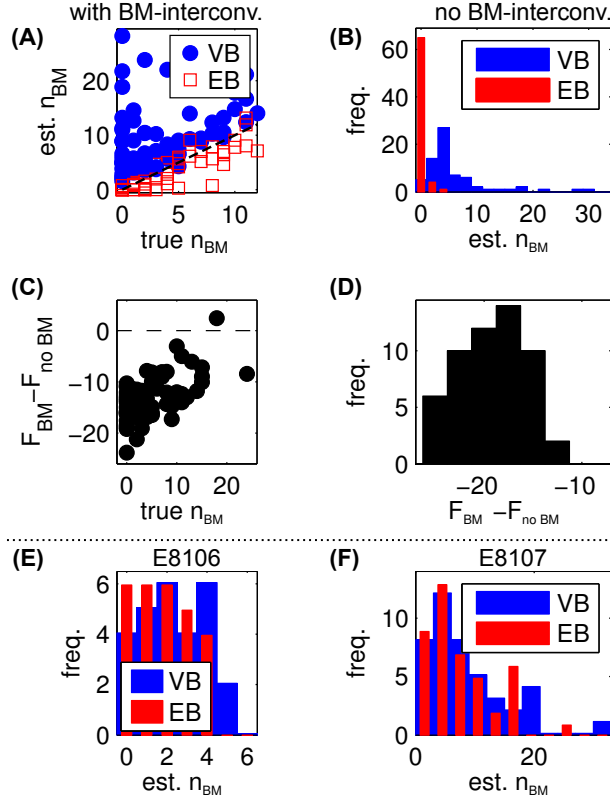


Figure 6: Detecting direct loop-loop interconversions. (A,B) Counting the number of  $B \rightleftharpoons M$  interconversions,  $n_{BM}$ , detected in synthetic data, with (A) and without (B) such transitions present, when trajectories are considered one at a time (“VB”), or using an empirical Bayes approach to analyze all trajectories from the same data set at once (“EB”). The dashed black line in (A) indicates where the estimated number of interconversions equals the true number. (B) shows a histogram rather than a scatter plot, because the true number of interconversions is zero. (C,D) VB analysis of single synthetic trajectories prefers models without B-M interconversions, whether they are present (C) or not (D), probably since they are rare events. Every point and histogram count represent a single trajectory, and  $F_{(\dots)}$  is the approximate log evidence, Eq. (5), for the different models. Higher  $F$ -values indicate better fits, so  $F_{BM} < F_{no BM}$  means that models with no interconversions are preferred by this analysis. (E,F) Analysis of real data yields a substantial number of interconversions even with the EB scheme, a strong indication that they are in fact present.

that they share the same DNA topology but different LacI conformations (*e.g.* a V-shaped and an extended conformation); and an M (for in-phase operators) or B (for out-of-phase operators) state that cannot directly interconvert with another looped state.

## Discussion and conclusion

We have developed a Bayesian analysis method for TPM data based on hidden Markov models called vbTPM. A major advance offered by our method is improved time resolution, which stems from our direct analysis of position data, thus avoiding the time-averaging required to produce readable RMS traces (Fig. 1). We are not the first to exploit this possibility. Beausang and Nelson [62] used manually curated training data to construct detailed models of the diffusive bead motion for the looped and unlooped states, and combined them with a two-state HMM to extract interconversion rates. Manzo and Finzi [67] modeled bead positions as uncorrelated zero-mean random variables, and used change-point and hierarchical clustering methods to segment TPM position traces in order to extract dwell time statistics for further analysis.

Our new analysis tool improves on previous methods in several ways. Compared to the change-point method [67], we use a noise model that accounts for correlations in the bead motion, which removes the need to filter out short dwell times. Compared to the previous HMM treatment [62], which used a more detailed dynamical model, vbTPM does not require curated training data. Instead, it learns the number of states directly from the data along with all other model parameters in a statistically principled way, using a variational Bayes treatment of



Hidden Markov models [29–31, 34]. The number of states, corresponding to, for example, distinct DNA-protein conformations, is often a key quantity of interest, and the possibility to extract it directly from the data will be especially useful for poorly characterized and complex systems (for example, TPM data with three rather than two operators present, as in the wild-type *lac* operon [57]). Also in contrast with previous methods, vbTPM handles common experimental artifacts gracefully, by classifying them in separate states that can easily be filtered out based on their unphysical parameters. Finally, we demonstrate further improved resolution from an ability to pool information from large heterogeneous data sets, using an empirical Bayes approach [32, 33]. Combined, these represent significant improvements over previous analysis methods, which we expect to be useful for a wide range of TPM applications. Our code, implemented in a mixture of Matlab and C, is freely available as open-source software.

We apply our improved method to data from LacI-mediated loop formation in DNA constructs with loop lengths from 100 to 109 bp [16]. Consistent with previous results, we resolve three states that cluster according to the emission parameters of the model,  $K$  and  $B$ , which we denote the unlooped state (U), middle looped state (M), and bottom looped state (B). Using an empirical Bayes approach, we then demonstrate that when the M and B looped states occur in a single trajectory, they can directly interconvert without passing through an unlooped state. This strongly indicates that these M and B states share a DNA binding topology but differ in LacI conformation, because a change of DNA topology would presumably require an unlooped intermediate, as different DNA topologies require the unbinding and re-binding of at least

one LacI DNA binding domain from the DNA. Our finding of direct interconversions between the M and B states are consistent with previous results on longer (138 bp [13] and 285 bp [14]) loops, which were attributed to transitions between a V-shaped and an extended LacI conformation.

Interestingly, at many loop lengths we can distinguish two kinds of trajectories, those that contain both an M and a B state (which can interconvert), and those that exhibit only one of the two looped states (Fig. 5). Which of the looped states (B or M) a two-state trajectory exhibits is the same for all two-state trajectories at a given loop length, but whether this state is the M or B state varies with loop length. The kinetics of loop formation and breakdown are fast enough that such a pattern is unlikely to emerge by chance, and, together with the indication that the single loop state changes with operator phasing (Fig. 5), we propose that this pattern reflects the existence of three different loop structures: two that can interconvert directly via a conformational change in LacI, and one that cannot interconvert directly to any other looped state, but has the same TPM signature as one of the interconverting states. Interconversion between the two- and three-state regimes is slow compared to our typical trajectory lengths, which is the reason we can distinguish them.

We note that a mixture of two- and three-state trajectories was also seen in a 138-bp construct with directly interconverting looped states, flanked by two Oid operators, but the two-state trajectories were not analyzed further [13]. For a 285 bp loop flanked by two O1 operators, only trajectories with two looped states were reported [14]. Closer analysis of these data might be interesting in light of our observations.

Unraveling the structural basis for this behav-

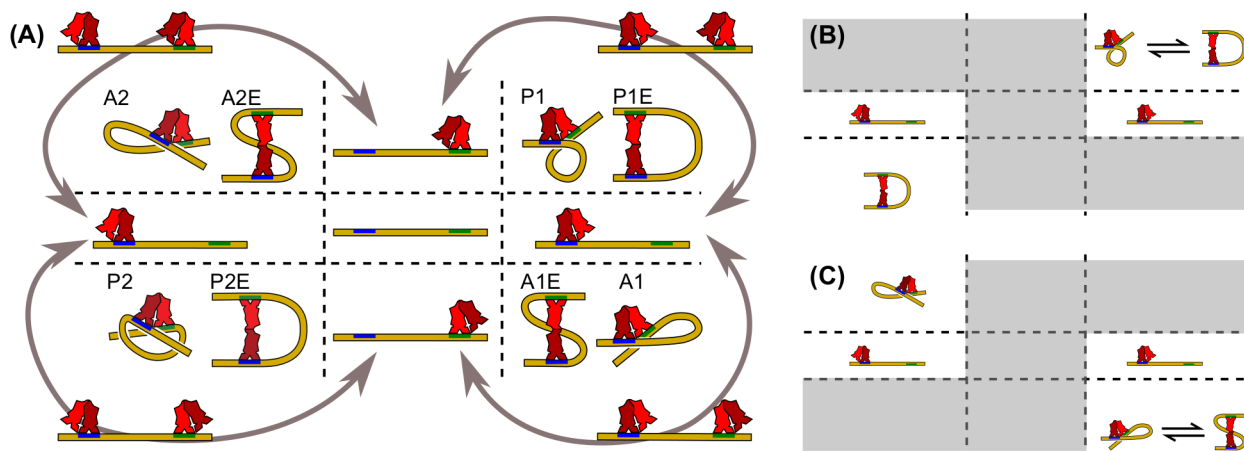


Figure 7: (A) Loop structures arranged by LacI binding directions (loop topology) on the operators Oid (blue) and O1 (green). Transitions between loops of different binding topology (corners) are only possible via unlooped neighbor states. Singly occupied states can also interconvert via unoccupied (center) or doubly occupied states (gray arrows). (B,C) shows two hypothetical ways for the state space in (A) to split up into two slowly-interconverting groups separated by energetically unfavorable states (grayed out). The two different divisions shown in (B) and (C) might represent in-phase versus out-of-phase operators, which differ in which observed looped state (M or B) is present in trajectories with a single looped state (see Fig. 5). In order for these divisions to generate the observed 2+3-state patterns of Fig. 5, one of the two state ‘islands’ must support only one looped state, while the other supports two looped states that can interconvert with one another. Panels (B,C) illustrate two possible ways to realize such behavior, in which the direct  $B \rightleftharpoons M$  interconversions correspond to transitions between V-shaped and extended LacI conformations.

ior will require further experimental, theoretical and computational efforts beyond the scope of this paper, but it is interesting to speculate about possible underlying molecular mechanisms. We propose as a starting point the scheme outlined in Fig. 7. Fig. 7A shows various potential loop structures arranged by binding topology (*i.e.* binding direction on the operators), with loop topology groups separated by unlooped intermediates. Both V-shaped and extended conformations are shown for each group of loop topologies, though it is not clear that all such structures are energetically feasible, nor that all can interconvert. Assuming limited rotations of the DNA binding domains, loop formation and breakdown occur via transitions between vertical and horizontal neighbors in this

plot. Singly occupied unlooped states can also interconvert via doubly occupied intermediates (gray arrows). How could this state-space be split into two slowly interconverting subsets as we observe?

First, we note that for the operators used here, the statistical mechanics analysis from our previous work implies that the no-LacI-bound state (center in Fig. 7A) is essentially unpopulated at 100 pM LacI [16]. Second, we suppose, as shown in Fig. 7B and C, that all energetically feasible loops are found only in two diagonally opposite loop topology groups, which therefore form isolated state “islands” separated by energetically unfavorable states (gray corners). Theoretical and computational work consistently finds some loop topologies to be more stable than others [5–

8], making this supposition tenable. If we further hypothesize that not all extended states can interconvert with their cognate V-shaped topological equivalents (or vice-versa), then we would obtain the mixture of two-state and three-state trajectories that we observe in our data.

A final consideration for this scheme relates to the possibility of passing from one state “island” to the other by way of a doubly-occupied state. That is, it is possible to move from a corner to a singly-bound neighbor state, then to a doubly-occupied state, then to the diagonally opposite corner via unbinding of the original LacI. The relatively low frequency of state transitions in our data combined with the relative dissociation rates of LacI for the Oid and O1 operators we use here make this pathway unlikely on the timescales of our data trajectories. Oid is about four times stronger than O1 [16, 53–56], and off-rates for Oid and O1 under experimental conditions similar to ours have been determined to be about  $0.002\text{ s}^{-1}$  and  $0.005\text{ s}^{-1}$  respectively [13, 68] (similar values have recently been measured *in vivo* as well [69]). Looped and doubly-occupied states are therefore almost three times more likely to decay by O1 unbinding. Therefore, we speculate that the middle column in Fig. 7A acts as a barrier between the two outer columns. Upper and lower loop topology groups in each column can easily interconvert via the singly-occupied Oid-bound state. Moreover, recent work hints at additional types of unlooped states, which might further slow down transitions between different topology groups [12, 18].

The scheme we propose in Fig. 7 illustrates how our results point to new interesting directions for future investigations into LacI-mediated looping. For example, existing theoretical work has focused on looping free energies [5–8], which are not enough to address the question of allowed

interconversions. Another interesting question stems from the putative rotational flexibility of the DNA binding domains [24], which might blur the differences between loop topology groups. Finally, a computational investigation of the RMS signal for different looped states shown in Fig. 7, including the effect of the bead and nearby wall [6], would aid in matching different structural models directly to TPM data.

Regardless of which molecular structures underlie the interconverting and non-interconverting loop states that we observe, it is clear that TPM combined with our novel Bayesian analysis resolves more than two coexisting looped states in a single construct. This is one more than previously observed at the single molecule level [13–18], but in qualitative agreement with theoretical and computational results [5–8] (see Introduction). Our findings are also consistent with recent ensemble FRET studies with loops formed from a library of synthetic pre-bent DNAs, in which at least three loop structures (a mixture of V-shaped and extended) contributed significantly to the observed looping for at least 5 of the 25 constructs examined [23].

The impact of these different loop structures on the ability of LacI to regulate the genes of the *lac* operon *in vivo* remains to be seen. Theoretical work has shown that several classic features of *in vivo* gene repression data with LacI can be best explained by the presence of more than one loop conformation, and that the presence of multiple looped states generally dampens oscillations in gene regulation as a function of loop length [4]. Extending these arguments, the presence of multiple looped states should allow looping under a wider range of conditions, and hence make gene regulation more robust against mechanical perturbations from,

for example, changes in supercoiling state or the presence versus absence of architectural proteins. However, inducer molecules and architectural proteins such as HU have been suggested to also change the relative stability of different loop shapes [4, 8–12] which may add an additional level of regulatory potential to the operon.

Both of the above effects could clearly be present and relevant also in more complex regulatory systems such as in eukaryotic cells. A fuller understanding of the loop structures and interconversion pathways available to the LacI-mediated loops we observe *in vitro*, and how they are influenced by architectural proteins that are known to play a large role in gene regulation *in vivo* [9–11], promises to greatly enhance our understanding of this potential additional layer of gene regulatory information.

**Acknowledgments** We thank members of the Phillips lab for helpful discussions and advice, Jason Kahn for helpful discussions about his lab’s work in relation to ours, and gratefully acknowledge financial support from: the National Science Foundation through a graduate fellowship to S.J.; a Rubicon fellowship (680-50-1016) from the Netherlands Organization for Scientific Research (NWO) to J.W.M.; the National Institutes of Health (NIH) [DP1 OD000217A (Director’s Pioneer Award), R01 GM085286, R01 GM085286-01S1, and 1 U54 CA143869 (Northwestern PSOC Center)], and the Fondation Pierre Gilles de Gennes to R.P.; an NIH National Centers for Biomedical Computing grant (U54CA121852) to C.H.W.; the Wenner-Gren foundation, the foundations of the Royal Swedish Academy of Sciences, and the Foundation for strategic research (SSF) via the Center for Biomembrane research to M.L.

## References

- [1] Oehler, S. and Müller-Hill, B. (2010) High local concentration: A fundamental strategy of life. *J. Mol. Biol.*, **395**, 242–253. doi:10.1016/j.jmb.2009.10.056.
- [2] Schleif, R. (1992) DNA looping. *Annu. Rev. Biochem.*, **61**, 199–223. doi:10.1146/annurev.bi.61.070192.001215.
- [3] Matthews, K.S. (1992) DNA looping. *Microbiol. Rev.*, **56**, 123–136.
- [4] Saiz, L. and Vilar, J.M.G. (2007) Multilevel deconstruction of the *in vivo* behavior of looped DNA-protein complexes. *PLoS ONE*, **2**, e355. doi:10.1371/journal.pone.0000355.
- [5] Zhang, Y., McEwen, A.E., Crothers, D.M. and Levene, S.D. (2006) Analysis of *in-vivo* LacR-mediated gene repression based on the mechanics of DNA looping. *PLoS ONE*, **1**, e136. doi:10.1371/journal.pone.0000136.
- [6] Towles, K.B., Beausang, J.F., Garcia, H.G., Phillips, R. and Nelson, P.C. (2009) First-principles calculation of DNA looping in tethered particle experiments. *Phys. Biol.*, **6**, 025001. doi:10.1088/1478-3975/6/2/025001.
- [7] Swigon, D., Coleman, B.D. and Olson, W.K. (2006) Modeling the Lac repressor-operator assembly: the influence of DNA looping on Lac repressor conformation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 9879–9884. doi:10.1073/pnas.0603557103.
- [8] Czapla, L., Grosner, M.A., Swigon, D. and Olson, W.K. (2013) Interplay of protein and DNA structure revealed in simulations of the *lac* operon. *PLoS ONE*, **8**, e56548. doi:10.1371/journal.pone.0056548.
- [9] Becker, N., Kahn, J. and Maher, L.J. (2005) Bacterial repression loops require enhanced DNA flexibility. *J. Mol. Biol.*, **349**, 716–730. doi:10.1016/j.jmb.2005.04.035.

- [10] Becker, N., Kahn, J. and Maher, L. (2007) Effects of nucleoid proteins on DNA repression in loop formation in *Escherichia coli*. *Nucleic Acids Res.*, **35**, 3988–4000. doi:10.1093/nar/gkm419.
- [11] Becker, N., Kahn, J. and Maher, L.J. (2008) Eukaryotic HMGB proteins as replacements for HU in *e. coli* repression loop formation. *Nucleic Acids Res.*, **36**, 4009–4021. doi:10.1093/nar/gkn353.
- [12] Goodson, K.A., Wang, Z., Haeusler, A.R., Kahn, J.D. and English, D.S. (2013) LacI-DNA-IPTG loops: Equilibria among conformations by single-molecule FRET. *J. Phys. Chem. B*, **117**, 4713–4722. doi:10.1021/jp308930c.
- [13] Wong, O.K., Guthold, M., Erie, D.A., Gelles, J. and Herschlag, D. (2008) Interconvertible Lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol.*, **6**, e232. doi:10.1371/journal.pbio.0060232.
- [14] Rutkauskas, D., Zhan, H., Matthews, K.S., Pavone, F.S. and Vanzi, F. (2009) Tetramer opening in LacI-mediated DNA looping. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 16627–16632. doi:10.1073/pnas.0904617106.
- [15] Han, L., Garcia, H.G., Blumberg, S., Towles, K.B., Beausang, J.F., Nelson, P.C. and Phillips, R. (2009) Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS ONE*, **4**, e5621. doi:10.1371/journal.pone.0005621.
- [16] Johnson, S., Lindén, M. and Phillips, R. (2012) Sequence dependence of transcription factor-mediated DNA looping. *Nucleic Acids Res.*, **40**, 7728–7738. doi:10.1093/nar/gks473.
- [17] Johnson, S., Chen, Y.J. and Phillips, R. (2013) Poly(dA:dT)-rich DNAs are highly flexible in the context of DNA looping. *PLOS ONE*, **8**, e75799. doi:10.1371/journal.pone.0075799.
- [18] Revalee, J.L., Blab, G.A., Wilson, H.D., Kahn, J.D. and Meiners, J.C. (2014) Tethered particle motion reveals that LacI-DNA loops coexist with a competitor-resistant but apparently un-looped conformation. *Biophys. J.* (in press).
- [19] Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G. and Lu, P. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, **271**, 1247–1254. doi:10.1126/science.271.5253.1247.
- [20] Ruben, G.C. and Roos, T.B. (1997) Conformation of Lac repressor tetramer in solution, bound and unbound to operator DNA. *Microsc. Res. Tech.*, **36**, 400–416.
- [21] Taraban, M., Zhan, H., Whitten, A.E., Langley, D.B., Matthews, K.S., Swint-Kruse, L. and Trehwella, J. (2008) Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. *J. Mol. Biol.*, **376**, 466–481. doi:10.1016/j.jmb.2007.11.067.
- [22] Morgan, M.A., Okamoto, K., Kahn, J.D. and English, D.S. (2005) Single-molecule spectroscopic determination of Lac repressor-DNA loop conformation. *Biophys. J.*, **89**, 2588–2596. doi:10.1529/biophysj.105.067728.
- [23] Haeusler, A.R., Goodson, K.A., Lillian, T.D., Wang, X., Goyal, S., Perkins, N.C. and Kahn, J.D. (2012) FRET studies of a landscape of Lac repressor-mediated DNA loops. *Nucl. Acids Res.*, **40**, 4432–4445. doi:10.1093/nar/gks019.
- [24] Villa, E., Balaeff, A. and Schulten, K. (2005) Structural dynamics of the *lac* repressor-DNA complex revealed by a multiscale simulation. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 6783–6788. doi:10.1073/pnas.0409387102.
- [25] Schafer, D.A., Gelles, J., Sheetz, M.P. and Landick, R. (1991) Transcription by single molecules of RNA polymerase observed by



- light microscopy. *Nature*, **352**, 444–448. doi:10.1038/352444a0.
- [26] Finzi, L. and Gelles, J. (1995) Measurement of lactose repressor-mediated loop formation and breakdown in single DNA molecules. *Science*, **267**, 378–380.
- [27] Normanno, D., Vanzi, F. and Pavone, F.S. (2008) Single-molecule manipulation reveals supercoiling-dependent modulation of *lac* repressor-mediated DNA looping. *Nucl. Acids Res.*, **36**, 2505–2513. doi:10.1093/nar/gkn071.
- [28] Edelman, L.M., Cheong, R. and Kahn, J.D. (2003) Fluorescence resonance energy transfer over 130 basepairs in hyperstable Lac repressor-DNA loops. *Biophys. J.*, **84**, 1131–1145. doi:10.1016/S0006-3495(03)74929-7.
- [29] Bronson, J.E., Fei, J., Hofman, J.M., Jr., R.L.G. and Wiggins, C.H. (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. doi:10.1016/j.bpj.2009.09.031.
- [30] Bronson, J.E., Hofman, J.M., Fei, J., Gonzalez, R.L. and Wiggins, C.H. (2010) Graphical models for inferring single molecule dynamics. *BMC Bioinformatics*, **11 Suppl 8**, S2. doi:10.1186/1471-2105-11-S8-S2.
- [31] Okamoto, K. and Sako, Y. (2012) Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.*, **103**, 1315–1324. doi:10.1016/j.bpj.2012.07.047.
- [32] van de Meent, J.W., Bronson, J.E., Wood, F., Gonzalez Jr., R.L. and Wiggins, C.H. (2013) Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *Proceedings of The 30th International Conference on Machine Learning*, **28**, 361–369.
- [33] van de Meent, J.W., Bronson, J.E., Wiggins, C.H. and Gonzalez Jr., R.L. (2014) Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophys. J.* (in press).
- [34] Persson, F., Lindén, M., Unoson, C. and Elf, J. (2013) Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Meth.*, **10**, 265–269. doi:10.1038/nmeth.2367.
- [35] Segall, D.E., Nelson, P.C. and Phillips, R. (2006) Volume-exclusion effects in tethered-particle experiments: Bead size matters. *Phys. Rev. Lett.*, **96**, 088306–4. doi:10.1103/PhysRevLett.96.088306.
- [36] Lindner, M., Nir, G., Medalion, S., Dietrich, H., Rabin, Y. and Garini, Y. (2011) Force-free measurements of the conformations of DNA molecules tethered to a wall. *Phys. Rev. E*, **83**, 011916. doi:10.1103/PhysRevE.83.011916.
- [37] Plénat, T., Tardin, C., Rousseau, P. and Salomé, L. (2012) High-throughput single-molecule analysis of DNA-protein interactions by tethered particle motion. *Nucleic Acids Res.*, **40**, e89. doi:10.1093/nar/gks250.
- [38] Nelson, P.C., Zurla, C., Brogioli, D., Beausang, J.F., Finzi, L. and Dunlap, D. (2006) Tethered particle motion as a diagnostic of DNA tether length. *J. Phys. Chem. B*, **110**, 17260–17267. doi:10.1021/jp0630673.
- [39] Brinkers, S., Dietrich, H.R.C., de Groote, F.H., Young, I.T. and Rieger, B. (2009) The persistence length of double stranded DNA determined using dark field tethered particle motion. *J. Chem. Phys.*, **130**, 215105–9. doi:10.1063/1.3142699.
- [40] Yin, H., Landick, R. and Gelles, J. (1994) Tethered particle motion method for studying transcript elongation by a single RNA polymerase molecule. *Biophys. J.*, **67**, 2468–2478. doi:10.1016/S0006-3495(94)80735-0.

- [41] Vanzi, F., Vladimirov, S., Knudsen, C.R., Goldman, Y.E. and Cooperman, B.S. (2003) Protein synthesis by single ribosomes. *RNA*, **9**, 1174–1179. doi:10.1261/rna.5800303.
- [42] Vanzi, F., Broggio, C., Sacconi, L. and Pavone, F.S. (2006) Lac repressor hinge flexibility and DNA looping: single molecule kinetics by tethered particle motion. *Nucl. Acids Res.*, **34**, 3409–3420. doi:10.1093/nar/gkl393.
- [43] Mumm, J.P., Landy, A. and Gelles, J. (2006) Viewing single  $\lambda$  site-specific recombination events from start to finish. *EMBO J*, **25**, 4586–4595. doi:10.1038/sj.emboj.7601325.
- [44] Broek, B.v.d., Vanzi, F., Normanno, D., Pavone, F.S. and Wuite, G.J.L. (2006) Real-time observation of DNA looping dynamics of type IIE restriction enzymes NaeI and NarI. *Nucl. Acids Res.*, **34**, 167–174. doi:10.1093/nar/gkj432.
- [45] Pouget, N., Turlan, C., Destainville, N., Salome, L. and Chandler, M. (2006) IS911 transpososome assembly as analysed by tethered particle motion. *Nucleic Acids Res.*, **34**, 4313–4323. doi:10.1093/nar/gkl420.
- [46] Chu, J.F., Chang, T.C. and Li, H.W. (2010) Single-molecule TPM studies on the conversion of human telomeric DNA. *Biophys. J.*, **98**, 1608–1616. doi:10.1016/j.bpj.2009.12.4328.
- [47] Rousseau, P., Tardin, C., Tolou, N., Salom, L. and Chandler, M. (2010) A model for the molecular organisation of the IS911 transpososome. *Mobile DNA*, **1**, 16. doi:10.1186/1759-8753-1-16.
- [48] Manzo, C., Zurla, C., Dunlap, D.D. and Finzi, L. (2012) The effect of nonspecific binding of lambda repressor on DNA looping dynamics. *Biophys. J.*, **103**, 1753–1761. doi:10.1016/j.bpj.2012.09.006.
- [49] Fan, H.F. (2012) Real-time single-molecule tethered particle motion experiments reveal the kinetics and mechanisms of Cre-mediated site-specific recombination. *Nucl. Acids Res.*, **40**, 6208–6222. doi:10.1093/nar/gks274.
- [50] Cloutier, T.E. and Widom, J. (2004) Spontaneous sharp bending of double-stranded DNA. *Mol. Cell*, **14**, 355–362. doi:10.1016/S1097-2765(04)00210-2.
- [51] Cloutier, T.E. and Widom, J. (2005) DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 3645–3650. doi:10.1073/pnas.0409059102.
- [52] Lowary, P.T. and Widom, J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol*, **276**, 19–42. doi:10.1006/jmbi.1997.1494.
- [53] Whitson, P.A. and Matthews, K.S. (1986) Dissociation of the lactose repressor-operator DNA complex: effects of size and sequence context of operator-containing DNA. *Biochemistry*, **25**, 3845–3852. doi:10.1021/bi00361a016.
- [54] Whitson, P.A., Olson, J.S. and Matthews, K.S. (1986) Thermodynamic analysis of the lactose repressor-operator DNA interaction. *Biochemistry*, **25**, 3852–3858. doi:10.1021/bi00361a017.
- [55] Hsieh, W.T., Whitson, P.A., Matthews, K.S. and Wells, R.D. (1987) Influence of sequence and distance between two operators on interaction with the lac repressor. *J. Biol. Chem.*, **262**, 14583–14591.
- [56] Frank, D.E., Saecker, R.M., Bond, J.P., Capp, M.W., Tsodikov, O.V., Melcher, S.E., Levandoski, M.M. and Record, M. (1997) Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: effects of converting a consensus site to a non-specific site. *J. Mol. Biol.*, **267**, 1186–1206. doi:10.1006/jmbi.1997.0920.

- [57] Johnson, S. (2012) DNA Mechanics and Transcriptional Regulation in the *E. coli lac* operon. Ph.D. thesis, California Institute of Technology, Pasadena, CA.
- [58] Beausang, J.F., Zurla, C., Finzi, L., Sullivan, L. and Nelson, P.C. (2007) Elementary simulation of tethered brownian motion. *American Journal of Physics*, **75**, 520–523. doi:10.1119/1.2710484.
- [59] Lindner, M., Nir, G., Vivante, A., Young, I.T. and Garini, Y. (2013) Dynamic analysis of a diffusing particle in a trapping potential. *Phys. Rev. E*, **87**, 022716. doi:10.1103/PhysRevE.87.022716.
- [60] MacKay, D. (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge UK ;New York.
- [61] Bishop, C. (2006) Pattern recognition and machine learning. Springer, New York.
- [62] Beausang, J.F. and Nelson, P.C. (2007) Diffusive hidden Markov model characterization of DNA looping dynamics in tethered particle experiments. *Phys. Biol.*, **4**, 205–219. doi:10.1088/1478-3975/4/3/007.
- [63] Ghahramani, Z. and Jordan, M.I. (1997) Factorial hidden Markov models. *Machine Learning*, **29**, 245–273. doi:10.1023/A:1007425814087.
- [64] Laurens, N., Bellamy, S.R.W., Harms, A.F., Kovacheva, Y.S., Halford, S.E. and Wuite, G.J.L. (2009) Dissecting protein-induced DNA looping dynamics in real time. *Nucleic Acids Res.*, **37**, 5454–5464. doi:10.1093/nar/gkp570.
- [65] Laurens, N., Rusling, D.A., Pernstich, C., Brouwer, I., Halford, S.E. and Wuite, G.J.L. (2012) DNA looping by FokI: the impact of twisting and bending rigidity on protein-induced looping dynamics. *Nucleic Acids Res.*, **40**, 4988–4997. doi:10.1093/nar/gks184.
- [66] Manghi, M., Tardin, C., Baglio, J., Rousseau, P., Salom, L. and Destainville, N. (2010) Probing DNA conformational changes with high temporal resolution by tethered particle motion. *Phys. Biol.*, **7**, 046003. doi:10.1088/1478-3975/7/4/046003.
- [67] Manzo, C. and Finzi, L. (2010) Quantitative analysis of DNA-looping kinetics from tethered particle motion experiments. In *Method. Enzymol.*, Elsevier, vol. 475, pp. 199–220. doi:10.1016/S0076-6879(10)75009-6.
- [68] Winter, R., Berg, O. and von Hippel, P. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli lac* repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry*, **20**, 6961–6977. doi:10.1021/bi00527a030.
- [69] Hammar, P., Walldén, M., Fange, D., Persson, F., Baltekin, Ö., Ullman, G., Leroy, P. and Elf, J. (2014) Non-equilibrium contributions to transcription factor mediated gene regulation. *Nat. Genet.* (accepted for publication).

## Supplementary Information

### S1 Choice of priors

We would like to choose uninformative prior distributions in order to minimize statistical bias. This is unproblematic for the emission parameters  $K, B$ , since the amount of data in all states is large enough to overwhelm any prior influence. As derived in the software manual<sup>1</sup>, prior distributions for  $K, B$  are given by

$$p(\mathbf{K}, \mathbf{B} | N) = \prod_{j=1}^N \frac{B_j^{\tilde{n}_j}}{W_j} e^{-B_j(\tilde{v}_j(K_j - \tilde{\mu}_j)^2 + \tilde{c}_j)}, \quad (\text{S1})$$

$$W_j = \frac{\tilde{c}^{-(\tilde{n}_j + \frac{1}{2})} \Gamma(\tilde{n}_j + \frac{1}{2})}{\sqrt{\tilde{v}_j/\pi}}, \quad (\text{S2})$$

with the range  $B_j \geq 0$ ,  $-\infty < K_j < \infty$ . Throughout this work, we use

$$\tilde{\mu}_j = 0.6, \quad \tilde{n}_j = 1, \quad (\text{S3})$$

$$\tilde{v}_j = 5.56 \text{ nm}^2, \quad \tilde{c}_j = 30000 \text{ nm}^2, \quad (\text{S4})$$

which corresponds to

$$\langle K_j \rangle = 0.6, \quad \langle B_j \rangle = 5 \times 10^{-5} \text{ nm}^{-2}, \quad (\text{S5})$$

$$\text{std}(K_j) = 0.3, \quad \text{std}(B_j) = 141.4 \times 10^{-5} \text{ nm}^{-2}. \quad (\text{S6})$$

The prior for the initial state probabilities are Dirichlet distributed,  $p(\boldsymbol{\pi} | N) = \text{Dir}(\boldsymbol{\pi} | \tilde{\mathbf{w}}^{(\boldsymbol{\pi})})$ , and these variables are unproblematic for the opposite reason: the long length of the trajectories makes the initial state relatively unimportant to describe the data. We use a constant prior of strength 5, i.e.,

$$\tilde{w}_j^{(\boldsymbol{\pi})} = 5/N, \quad (\text{S7})$$

<sup>1</sup>See vbtpm.sourceforge.net for the latest version.

where  $N$  is the number of hidden states.

The transition probabilities need more care, because the potentially low number of transitions per trajectory makes the prior relatively more influential. The prior for the transition matrix  $\mathbf{A}$  are independent Dirichlet distributions for each row, parameterized by a pseudo-count matrix  $\tilde{w}_{ij}^{(\mathbf{A})}$ . Following Ref. [34], we parameterize this prior in terms of an expected mean lifetime and an overall number of pseudocounts (prior strength) for each hidden state. In particular, we define a transition *rate* matrix  $\mathbf{Q}$  with mean lifetime  $t_D$ ,

$$Q_{ij} = \frac{1}{t_D} \left( -\delta_{ij} + \frac{1 - \delta_{ij}}{N - 1} \right), \quad (\text{S8})$$

and then construct the prior based on the transition probability propagator per unit time step,

$$\tilde{w}_{ij}^{(\mathbf{A})} = \frac{t_A f_{\text{sample}}}{n_{\text{downsample}}} e^{\Delta t Q}. \quad (\text{S9})$$

Here,  $t_A$  is the prior strength; both  $t_A$  and  $t_D$  are specified in time units to be invariant under a change of sampling frequency. Further, the timestep is given by  $\Delta t = n_{\text{downsample}}/f_{\text{sample}}$ , where  $f_{\text{sample}}$  is the sampling frequency (30 Hz in our case), and  $n_{\text{downsample}}$  is the downsampling factor (we use 3).

Numerical experiments in Ref. [34] show that choosing the strength too low compared to the mean lifetime produces a bias towards sparse transition matrices. This is not desirable in our case, and we therefore use  $t_D = 1$  s, and  $t_A = 5$  s throughout this work.

### S2 Performance on synthetic data

Here, we test the abilities of vbTPM to resolve close-lying states in synthetic data, and compare

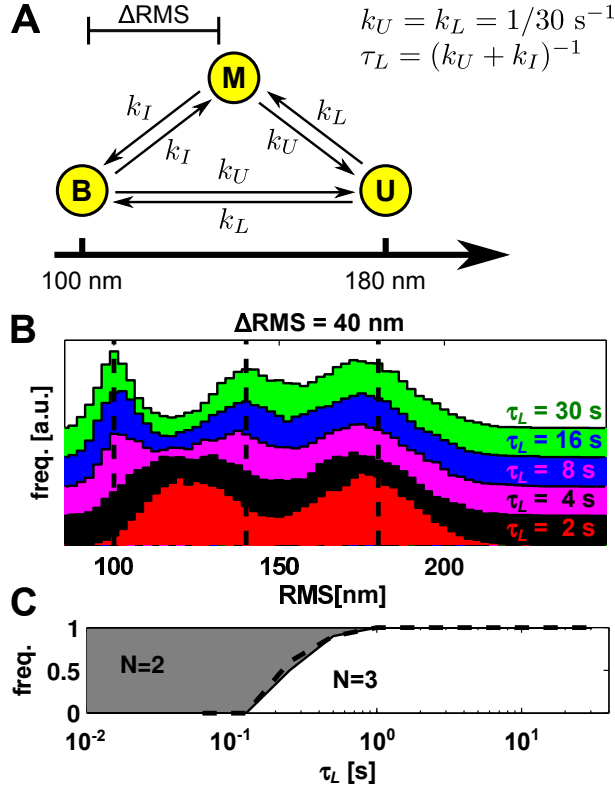


Figure S1: Temporal resolution with vbTPM and RMS histograms. (A) Model for synthetic data, with the difficulty determined by the RMS-separation  $\Delta\text{RMS}$  and mean life time  $\tau_L$  of the two interconverting states M and B. (B) Aggregated RMS histograms from ten 45-min trajectories with  $\Delta\text{RMS}=40 \text{ nm}$  and varying  $\tau_L$ . The M and B states are blurred to a single peak at low  $\tau_L$ , but for  $\tau_L \gtrsim 8 \text{ s}$ , all three states can be resolved. Vertical lines show the true RMS values. (C) Fraction of trajectories in which the HMM algorithm resolved 2 (gray) or 3 (white) states. All three states are resolved already at  $\tau_L \geq 0.5 \text{ s}$ , significantly better than the histogram method. The dashed line shows the result without downsampling, an insignificant improvement. The filter width used in (B) was optimized by eye to  $\sigma_G = 3 \text{ s}$ .

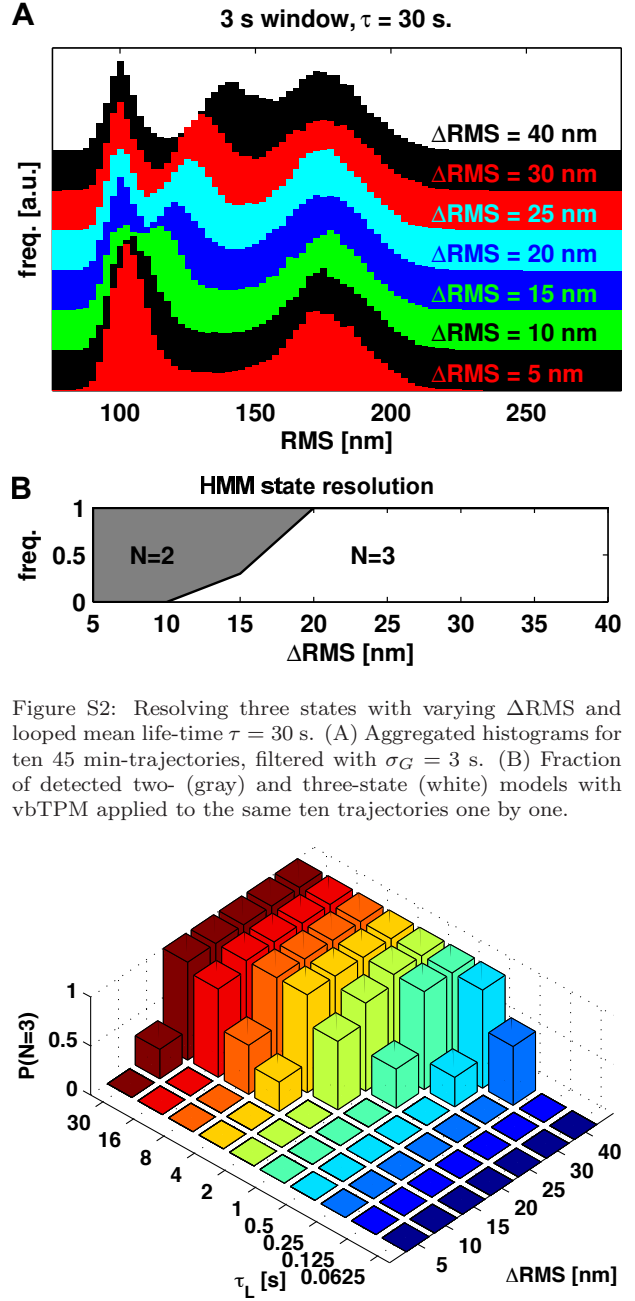


Figure S3: Resolution map of vbTPM shown as the fraction of correctly identified 3-state models at different  $(\Delta\text{RMS}, \tau_L)$ -pairs. Ten 45 min-trajectories were simulated at each parameter set, and 3-fold downsampling was used for the analysis.



it to the RMS histogram method. We also verify that model parameters are recovered correctly, and that these results are insensitive to the factor three downsampling that we use for analysis on real data.

Our test model, depicted in Fig. S1A, has one unlooped (U) and two looped (M and B) states, and the difficulty of resolving states M and B can be tuned by decreasing either their RMS difference  $\Delta\text{RMS}$  or their average life-time  $\tau_L$  (the life-time of the aggregated state B+M is fixed at 30 s, same as the unlooped state). For each parameter setting, we generated and analyzed ten 45 minute trajectories.

Fig. S1B-C shows a comparison of temporal resolution, using  $\Delta\text{RMS}=40$  nm and varying  $\tau_L$ . Resolving states using histograms means resolving peaks, and three distinct peaks emerge at  $\tau_L = 4 - 8$  s. In contrast, vbTPM resolves the correct number of states already at  $\tau_L = 0.5$  s. This order-of-magnitude improvement mainly reflects the detrimental effects of the low-pass filter used in the RMS analysis, and is insensitive to downsampling by a factor of three. The vbTPM limit can instead be compared to the bead correlation time  $\tau$ , which were set to 0.1, 0.17, and 0.25 s for the B,M and U states in this data.

We also compared vbTPM to the histogram method for resolving states that interconvert slowly ( $\tau_L = 30$  s) with varying degrees of separation in RMS. The result is shown in Fig. S2, and indicates that vbTPM does not significantly outperform the histogram method in this case.

To summarize, we mapped out the resolution of vbTPM in the range  $0.0625 \text{ s} \leq \tau_L \leq 30 \text{ s}$ ,  $5 \text{ nm} \leq \Delta\text{RMS} \leq 40 \text{ nm}$ . The results, in Fig. S3, show a nonlinear relation between the spatial and temporal resolution.

Next, we verify that model parameters are also

well reproduced and insensitive to downsampling in this situation. Fig. S4 shows the RMS values for the most likely models fitted to the test data set of Fig. S1. The looped state of the two-state models display an average of the two looped states in the data when those states interconvert too quickly to be resolved. The three-state models generally reproduce the input parameters with a slight downward bias that is more noticeable at high RMS values. We believe that this is an effect of the drift-correction filter we applied to the data. Note that the results with and without downsampling are almost indistinguishable.

The mean lifetimes (Fig. S5) show similar trends of good fit and almost no difference with and without downsampling. Two-state models that do not resolve the two looped states learn their aggregated mean lifetime, which is indeed 30 s in the true model. The tendency to overestimate the short lifetimes can be rationalized by noting that short sojourns are more difficult to resolve, and therefore do not contribute as much to the estimated parameter values.

Individual transition probabilities (elements  $A_{ij}$ ) are presented in Fig. S6. Here there is a clear difference with and without downsampling, since the latter estimates transition probability per timestep, while the former per three timesteps. Low transition probabilities suffer significant fluctuations due to small number statistics, while the higher transition probabilities are well reproduced.

### S3 Estimated P-value for two-state trajectories

How unlikely would it be to observe 38 trajectories with only states U and M together with 27-

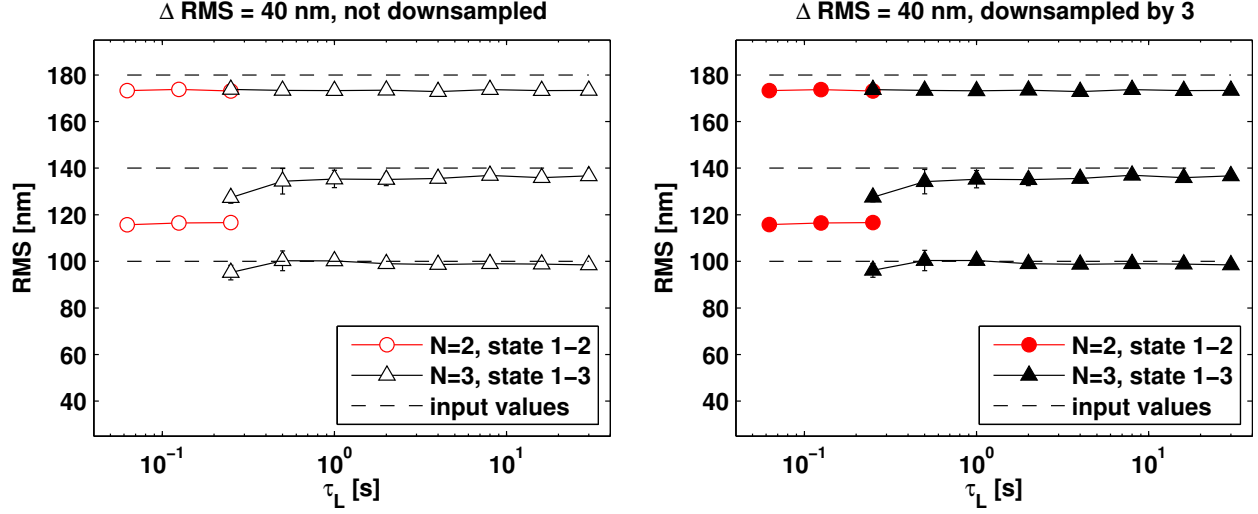


Figure S4: RMS values for the best fit models (symbols) to the data set in Fig. S1, compared to simulated parameters (dashed). Posterior mean value  $\pm$  std. (an estimate of the parameter uncertainty) for two- and three-state models shown separately, according to which model size got the best score for each trajectory. Most error bars are smaller than the symbols. Analysis without (left) and with (right) downsampling give almost identical results.

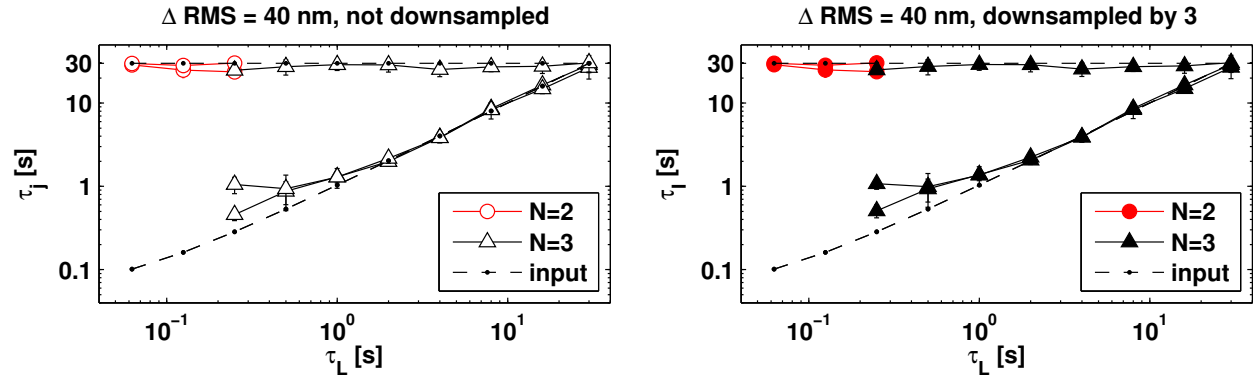


Figure S5: Mean lifetimes, presented as in Fig. S4. The true model (dashed) has one state with mean lifetime 30 s (U), and two states (M and B) with shorter lifetimes. The lower dashed line is not straight because  $\tau_L$  is defined as a rate in a continuous time model, while lifetimes (true and fitted) are defined in a discrete-time setting using the transition probability matrix  $A_{ij}$ , which makes a difference for short lifetimes. The average lifetime of the short-lived states together is always 30 s however, which explains why the two-state models that do not resolve these two states have both lifetimes around 30 s.

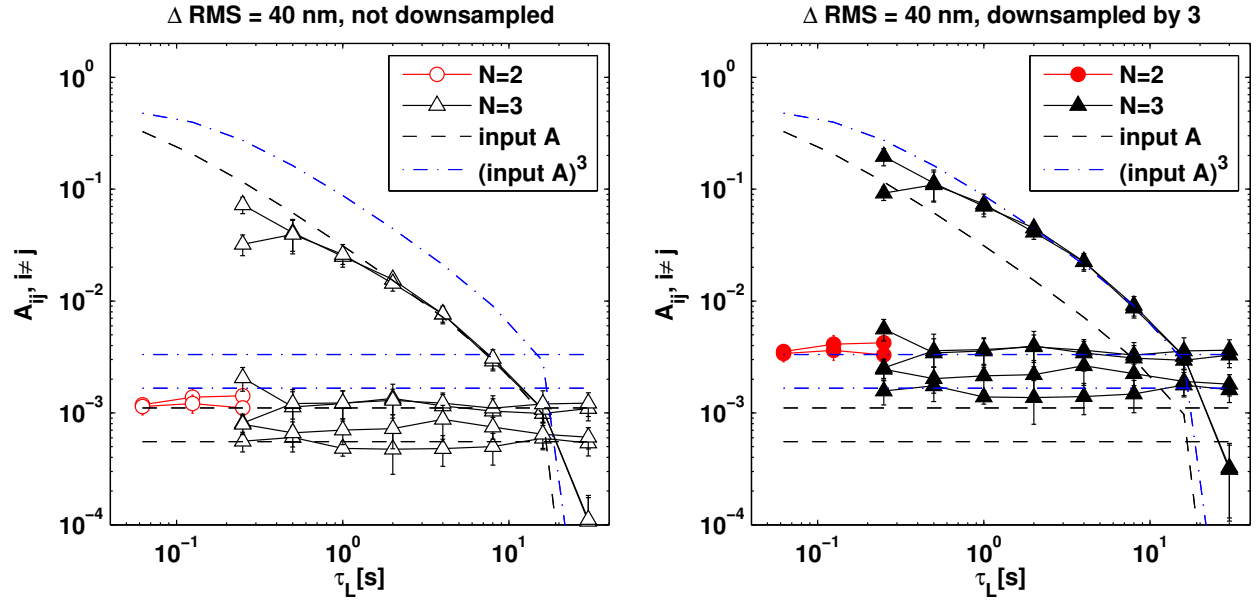


Figure S6: Transition probabilities (non-diagonal elements of  $A_{ij}$ ), presented as in Fig. S4. Due to symmetries of the underlying kinetic model it only contains three distinct transition probabilities. The difference with and without downsampling is due to the fact that the downsampled model effectively estimates transition probabilities per three timesteps, given by  $\mathbf{A}^3$  (blue dash-dotted lines), instead of the single-step probabilities (black dashed lines) used to produce the data. Relative to these different targets, however, the analyses with and without downsampling give very similar results.

three-state trajectories, in the E8106 construct, if all those trajectories were really governed by the “average” three-state kinetics? We estimate this probability here, and find it to be extremely small, indicating that simple 3-state kinetics is not a good description of this construct.

If a Markov chain with three states has initial state probability  $\boldsymbol{\pi}$ , and transition matrix  $\mathbf{A}$ , the probability that it never visits state 3 (state B in this case) during a trajectory with  $n$  steps is given by

$$p(s \neq 3) = \boldsymbol{\pi}_{1:2} \mathbf{A}_{1:2,1:2}^n \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad (\text{S10})$$

where  $\boldsymbol{\pi}_{1:2}$  and  $\mathbf{A}_{1:2,1:2}$  are the sub-arrays that describe only states 1 and 2.

We estimated  $\boldsymbol{\pi}$  and  $\mathbf{A}$  for the average 3-state model from an empirical Bayes analysis of all E8106 trajectories with two or three genuine states (filling out the 2-state trajectories with an empty 3rd state to initialize the EB iterations) and found

$$\langle \boldsymbol{\pi} \rangle = \begin{bmatrix} 0.76 & 0.21 & 0.026 \end{bmatrix}, \quad (\text{S11})$$

$$\langle \mathbf{A} \rangle \approx \begin{bmatrix} 0.9987 & 0.0011 & 0.0002 \\ 0.0010 & 0.9989 & 0.00004 \\ 0.0027 & 0.0006 & 0.9966 \end{bmatrix}, \quad (\text{S12})$$

where  $\langle \cdot \rangle$  here denotes averages over the optimized hyperparameter distributions.

Since the transitions matrix is downsampled to 10 Hz, this means average state lifetimes of 79 s, 90 s, and 30 s respectively. Moreover, the largest non-unit eigenvalue is 0.9979, corresponding to a characteristic time of

$$-\frac{0.1 \text{ s}}{\ln 0.9979} \approx 47 \text{ s} \quad (\text{S13})$$

for approaching the stationary state. With trajectory lengths ranging from 10 to 90 minutes,

one would therefore expect only a few trajectories to not explore all three states.

We then computed  $p(s \neq 3)$  for all trajectories  $i$  with two or three states (it decreases exponentially with trajectory length, which varied between 6100 and 54400 points after downsampling), and used those probabilities to simulate state occupancies. A trajectory was assigned states 1+2 with probability  $p(s \neq 3)$ . No more than 16 such 1+2-state trajectories were generated in  $10^6$  realizations, which means that the p-value for observing at least 38 two-state trajectories with the same two states is  $\ll 10^{-6}$  under the null-hypothesis of uniform three-state kinetics.