1-1-2018

Detection of kinetic change points in piece-wise linear single molecule motion

Flynn R. Hill
University of Wollongong, flynn@uow.edu.au

Antoine M. van Oijen
University of Wollongong, vanoijen@uow.edu.au

Karl E. Duderstadt
Technical University of Munich, Max Planck Institute of Biochemistry

Detection of kinetic change points in piece-wise linear single molecule motion

Abstract

Single-molecule approaches present a powerful way to obtain detailed kinetic information at the molecular level. However, the identification of small rate changes is often hindered by the considerable noise present in such single-molecule kinetic data. We present a general method to detect such kinetic change points in trajectories of motion of processive single molecules having Gaussian noise, with a minimum number of parameters and without the need of an assumed kinetic model beyond piece-wise linearity of motion. Kinetic change points are detected using a likelihood ratio test in which the probability of no change is compared to the probability of a change occurring, given the experimental noise. A predetermined confidence interval minimizes the occurrence of false detections. Applying the method recursively to all sub-regions of a single molecule trajectory ensures that all kinetic change points are located. The algorithm presented allows rigorous and quantitative determination of kinetic change points in noisy single molecule observations without the need for filtering or binning, which reduce temporal resolution and obscure dynamics. The statistical framework for the approach and implementation details are discussed. The detection power of the algorithm is assessed using simulations with both single kinetic changes and multiple kinetic changes that typically arise in observations of single-molecule DNA-replication reactions. Implementations of the algorithm are provided in ImageJ plugin format written in Java and in the Julia language for numeric computing, with accompanying Jupyter Notebooks to allow reproduction of the analysis presented here.

Disciplines

Medicine and Health Sciences

Publication Details

Hill, F. R., van Oijen, A. M. & Duderstadt, K. E. (2018). Detection of kinetic change points in piece-wise linear single molecule motion. Journal of Chemical Physics, 148 (12), 123317-1-123317-9.

Detection of kinetic change points in piece-wise linear single molecule motion

Flynn R. Hill,[1] Antoine M. van Oijen,[1] and Karl E. Duderstadt[2,3,a)]
[1]*Centre for Medical and Molecular Bioscience, Illawarra Health and Medical Research Institute and University of Wollongong, Wollongong, New South Wales 2522, Australia*
[2]*Structure and Dynamics of Molecular Machines, Max Planck Institute of Biochemistry, Martinsried, Germany*
[3]*Physik Department, Technische Universität München, Garching, Germany*

(Received 16 October 2017; accepted 28 November 2017; published online 3 January 2018)

Single-molecule approaches present a powerful way to obtain detailed kinetic information at the molecular level. However, the identification of small rate changes is often hindered by the considerable noise present in such single-molecule kinetic data. We present a general method to detect such kinetic change points in trajectories of motion of processive single molecules having Gaussian noise, with a minimum number of parameters and without the need of an assumed kinetic model beyond piece-wise linearity of motion. Kinetic change points are detected using a likelihood ratio test in which the probability of no change is compared to the probability of a change occurring, given the experimental noise. A predetermined confidence interval minimizes the occurrence of false detections. Applying the method recursively to all sub-regions of a single molecule trajectory ensures that all kinetic change points are located. The algorithm presented allows rigorous and quantitative determination of kinetic change points in noisy single molecule observations without the need for filtering or binning, which reduce temporal resolution and obscure dynamics. The statistical framework for the approach and implementation details are discussed. The detection power of the algorithm is assessed using simulations with both single kinetic changes and multiple kinetic changes that typically arise in observations of single-molecule DNA-replication reactions. Implementations of the algorithm are provided in ImageJ plugin format written in Java and in the Julia language for numeric computing, with accompanying Jupyter Notebooks to allow reproduction of the analysis presented here. *Published by AIP Publishing.* https://doi.org/10.1063/1.5009387

INTRODUCTION

Cellular life depends on a broad array of molecular assemblies that use chemical energy to perform directed processes, ranging from cell division and migration to nucleic acid metabolism and genome maintenance. Many of these systems synthesize new biomolecules or perform complex molecular rearrangements in stages characterized by discrete rate changes.[1] Systems that exhibit this type of behavior have piece-wise linear trajectories of motion that can be quantitatively studied by division into separate linear regimes before and after each rate change. Using classical biochemical methods, these discrete and often stochastic rate changes are obscured by ensemble averaging, preventing their identification and analysis. Single-molecule imaging approaches provide information-rich datasets in which changes in rate by individual members of the population can be followed in real-time, which allows for detailed characterization previously not possible. Nonetheless, events at the molecular scale are dominated by thermal fluctuations, with the resultant noise often obscuring underlying dynamics. While dramatic changes can be visually identified, subtle changes are overlooked. Therefore, automated approaches must be used to avoid bias and ensure the maximum information is retrieved in a rigorous and reproducible manner.

Numerous methods have been developed to detect and analyze transitions in single molecule observations; however, most current approaches depend on non-rigorous user-defined parameters and filtering steps that degrade data quality. While several approaches have been developed to address these issues in detection of intensity changes in single molecule fluorescence observations[2,3] and changes in diffusion in single particle tracking experiments,[4] an approach that is applicable to the piece-wise linear trajectories with Gaussian noise, typically obtained by tracking the motion of processive molecules has been lacking. Kinetic change points, or discrete changes in rate (Fig. 1), commonly occur in observations of biological systems conducted with single molecule force manipulation techniques such as magnetic tweezers, optical tweezers, AFM, or flow stretching.[5] For example, these techniques have been used to visualize the synthesis dynamics of individual RNA polymerases which exhibit changes in speed, stalling, and pausing.[6,7] Kinetic change points also arise in observations made using many other experimental configurations. For example, directly tracking the positions of fluorescently labeled motor proteins (e.g., kinesin or myosin)[8] or replisome components (e.g., polymerases and helicases)[9–11] on surface immobilized

a)Author to whom correspondence should be addressed: duderstadt@biochem.mpg.de

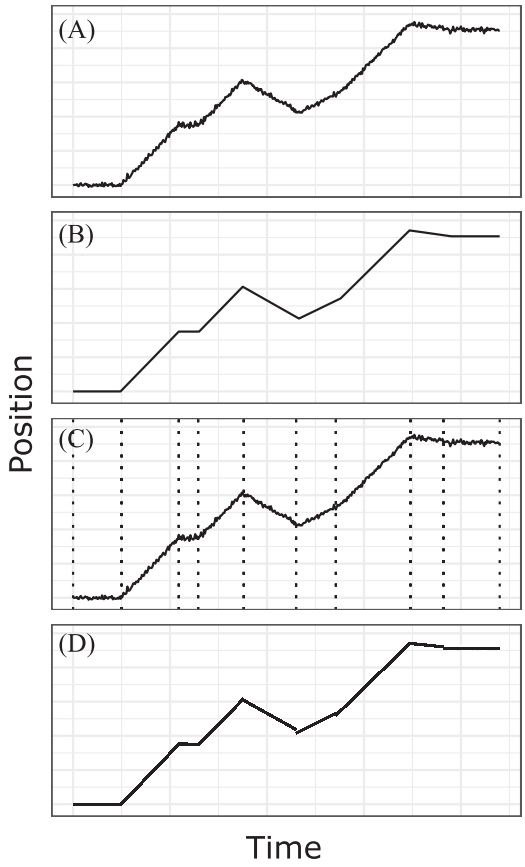

FIG. 1. (a) A simulated piece-wise linear single molecule trajectory, featuring an initial period of zero rate, followed by multiple rate changes. The trace further displays Gaussian noise, masking small rate changes. (b) The same trajectory, with noise removed, representing the target function for kinetic change point detection. (c) Detected change point locations indicated by the vertical dashed lines. The region prior to the start of enzymatic activity can be used to calculate experimental noise σ for the change point detection algorithm. (d) Segments generated by least-squares line fits between the change point locations closely resemble the target function and provide quantitative information about kinetic changes.

substrates likewise yields datasets containing discrete rate changes.

The simplest approach to identify kinetic changes in single molecule observations dominated by noise is to bin or filter (e.g., Savitzky-Golay or Chung-Kennedy[7,12,13]). This method can be used to identify large transitions but is strongly dependent on the choice of filter and bin width. Moreover, thresholds set to distinguish transitions are often arbitrarily chosen, thus introducing bias and complicating reproduction. These issues are well known, and considerable progress has been made to address them. Hidden Markov modeling approaches have been used to extract kinetic parameters in a more rigorous manner.[14,15] These approaches offer substantial fitting power but depend on prior knowledge about the underlying kinetic model, which may not be available or the user would prefer not to predefine one. Further, algorithms have been developed to analyze specific signature behaviors exhibited by certain biological processes. For example, several step fitting algorithms have been developed to analyze the stepping of single motor proteins and translocases.[15,16] Once optimal parameters are determined, these methods may provide the best fit, given the specific behavior of interest. However, these approaches must be hand-tailored to identify specific behaviors in each dataset and lack generality, limiting their applications.

For the more general problem of change point analysis, which can be applied to financial, climatological, and many other types of data, several algorithms have been developed.[17,18] The majority of these methods operate on piece-wise constant data, while some operate on piece-wise linear data—techniques often referred to as segmented regression.[19,20] Typically, these methods involve some means by which the most likely position of a change point is discovered and a method for assessing its statistical significance. If multiple change points are to be discovered, a systematic strategy must be employed to divide and search the trajectory. Only a minority of change point methods can handle an undefined number of change points and those which do typically require extensive user input and assessment on a per-trajectory basis in order to rigorously achieve an optimal solution. Additionally, these methods do not typically offer a means to leverage knowledge of the noise level of the trajectory, information that often is available for single-molecule trajectories. In summary, there is a need for a method of change point detection for piece-wise linear trajectories of known noise levels, with an undefined number of change points and minimal user input.

Here we outline a general approach to detect kinetic change points in the trajectories of processive molecules with uniform Gaussian noise and discrete rate changes, without the need for application specific parameters or data filtering. Given a position versus time trajectory, the algorithm determines all kinetic change point locations and provides linear fits for all regions in between, thus yielding estimates for all distinct rates exhibited by a given system. It should be noted that this method cannot be applied to systems with non-discrete/continuous rate changes or with non-uniform variance. Examples of these are non-processive motor proteins and in force-based experiments any proteins which have a non-linear force-velocity relationship.[21–26] The approach relies on a likelihood ratio test in which the probability of no kinetic change is compared to the probability of a kinetic change occurring, given the experimental noise. The likelihood ratio is calculated for all time points in a given region, and the maximal position is identified as a possible kinetic change point. To ensure experimental noise is handled properly and no overfitting occurs, a confidence interval is set to reduce the frequency of false detections. Applying the method recursively to sub-regions of a single molecule trajectory allows for detection of multiple kinetic change points (Fig. 1). Implementation of the algorithm depends on only two parameters: an acceptable false positive rate (typically below 1%) and the experimental noise. The false positive rate is based on a rigorous statistical foundation, and tabulated values are available based on well-established formulas, given Gaussian noise. The experimental noise can be directly measured for a given experimental setup and then applied to future datasets reducing the possibility of user bias due to arbitrary parameter optimization steps and user-defined parameters.

The method presented has proven tremendously powerful in distinguishing operational modes of the DNA

replication machinery from complex, multistate single molecule observations.[23,24] Here we provide the theoretical foundation for kinetic change point analysis, a detailed description of how the algorithm is implemented, and benchmark the power of detection using simulations. The algorithm provides a very general approach to detect kinetic changes that can be applied to any system exhibiting discrete changes in rate. The method is computationally fast and requires minimal user input and no interactive decision-making or evaluation, allowing for batch processing of massive numbers of trajectories. Therefore, it is appropriate for use with high-throughput single-molecule techniques, where hundreds or thousands of trajectories may need to be processed from a single experiment.[22,23,25] In this way, the identification of change points and their kinetic properties in a large dataset offers a means for fast, unbiased identification of trajectories and features of interest.

THEORY

Identification of kinetic change points in single molecule trajectories is a non-trivial problem due to the complexity of the observations and the significant intrinsic noise level. Two critical issues must be addressed. First, the position and significance level of individual kinetic change points must be determined within a given region. Second, a search method must be developed to ensure all kinetic change points are identified in the entire single molecule trajectory.[18] Here we outline a systematic approach to address these issues leveraging statistical testing theory and information theory in a manner similar to related techniques developed by Yang and colleagues but applied to discrete rate changes in the presence of Gaussian noise.

An ever-increasing number of single-molecule approaches are available to follow individual molecules, leveraging both force manipulation and fluorescence-based methods. In force manipulation experiments, kinetic changes exhibited by biological systems are observed by tracking micron-sized beads physically coupled to the system. For example, length changes during DNA replication can be followed by surface immobilizing the DNA being copied and tracking a bead attached to the end.[27,28] The primary source of uncertainty in these types of experiments is due to thermal fluctuations, giving rise to Gaussian distributed noise. In this example and in many (but not all) other systems of interest, molecular transitions such as changes in stoichiometry and conformation occur at discrete time points, giving rise to discrete changes in rate. In between these rate changes, there is a constant relationship between the rate of change in position and time. Therefore, we assume that individual single molecule trajectories are composed of discrete linear regions convoluted with Gaussian noise. Given this assumption, the likelihood, L_N, of observing a sequence of N positions within a linear region will be the product of the probabilities of observing each position,

$$L_N = p\left(y_1|x_1;a,b\right)\times p\left(y_2|x_2;a,b\right)\times\cdots\times p\left(y_N|x_N;a,b\right), \tag{1}$$

where $p\left(y_i|x_i;a,b\right)$ is the probability density function for obtaining a position y_i at time x_i given a slope a and intercept b ($y_i = a\times x_i + b$). Here we assume that each data point in the series was independently detected and recorded. For Gaussian distributed noise, the probability function will have the following form:

$$p\left(y_i|x_i;a_j,b_j\right) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-a_j\times x_i-b_j)^2}{2\sigma^2}}, \tag{2}$$

where a_j and b_j are the slope and intercept for each linear region j and σ is the standard deviation of the experimental noise. Therefore, the likelihood of a given series of N independently collected observations is

$$\begin{aligned} L\left(Y|X;a_k,b_k\right) &= \prod_1^N p\left(y_i|x_i;a_k,b_k\right) \\ &= \prod_1^N \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-a_k\times x_i-b_k)^2}{2\sigma^2}}. \end{aligned} \tag{3}$$

In practice, since the likelihood functions involve a very large number of products, it is convenient to work with the log-likelihood ratio when performing the likelihood ratio test. Therefore, we take the log of both sides to obtain the log-likelihood function,

$$\begin{aligned} l\left(Y|X;a_k,b_k\right) &= \ln L\left(Y|X;a_k,b_k\right) \\ &= \ln\prod_1^N \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-a_k\times x_i-b_k)^2}{2\sigma^2}} \\ &= N*\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - a_k*x_i - b_k)^2. \end{aligned} \tag{4}$$

This leaves us with the log-likelihood of obtaining a given set of positions within a linear region. To determine the location of a kinetic change point, we must calculate the probability of obtaining a set of positions y_i representing two distinct linear regions A and B with a kinetic change occurring at location k in between. The likelihood of having two regions will be given by the product of the likelihood of each, $L_A * L_B$. We can evaluate this product for all possible time points k in a region to obtain the most likely position for a kinetic change point. However, alone, this will not distinguish between real change points and noise. Therefore, we perform a maximum likelihood ratio test where the numerator is the hypothesis that there are two different linear regions with a change point at k, and the denominator is the null hypothesis that there is only a single linear region,

$$L_N\left(k\right) = \frac{L_N(\text{there is a change at k})}{L_N(\text{there is } \textbf{no} \text{ change at k})} = \frac{L_A * L_B}{L_0}, \tag{5}$$

where L_A is the probability of a linear region from x_1 to x_{k-1}, L_B is the probability of a linear region from x_k to x_N, and L_0 is the probability of a single linear region from x_1 to x_N. The

final expression then becomes

$$\begin{aligned}\mathcal{L}_N &= \ln\left[\frac{L_A * L_B}{L_0}\right] \\ &= \ln[L_A] + \ln[L_B] - \ln[L_0] \\ &= (k-1) * \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{k-1}(y_i - a_A * x_i - b_A)^2 \\ &\quad + (N-k) * \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=k}^{N}(y_i - a_B * x_i - b_B)^2 \\ &\quad - N * \ln\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - a_0 * x_i - b_0)^2. \end{aligned} \quad (6)$$

Given the log-likelihood ratio, the change point position is determined by evaluating the ratio for all possible positions in the region and finding the maximum,[29]

$$\hat{k} = \arg\max_{1\le k\le N}\{\mathcal{L}_N(k)\}. \quad (7)$$

This expression identifies the most likely change point position. However, to determine whether the experimental evidence is consistent with a change point occurring at position k, we must compare the maximum log-likelihood value to a test statistic appropriate for our experimental setup,

$$Z_N = 2\max_{1\le k\le N}\{\mathcal{L}_N(k)\},$$

Algorithm 1. Single change point identification and significance testing.

Input:
1. A trajectory of length N, comprising time points (x_1, x_2, …, x_N) and a kinetic response parameter ($y_1, y_2, \ldots, y_N$).
2. Time points defining the boundaries of the region of interest in the trajectory, x_{start} and x_{end}, with associated length N_{region}.
3. The pre-determined σ value for the trajectory.
4. The user-defined confidence threshold, α.

Process:
1. For every time point x_k in the range $x_{start+2}$:x_{end-2}, calculate the log-likelihood ratio $\mathcal{L}_{N_{region}}(k)$ for a pair of ordinary least squares line fits in the ranges x_{start}:x_k and x_k:x_{end} versus a single line fit in the range x_{start}:x_{end}.
2. Identify the time point $x_{\hat{k}}$ at which the log-likelihood ratio $\mathcal{L}_{N_{region}}(\hat{k})$ is maximized.
3. Look-up in a pre-made table, or otherwise calculate the relevant critical threshold value $C_{N_{region},1-\alpha}$.
4. Perform the significance test: is $2\mathcal{L}_{N_{region}}(\hat{k}) > C_{N_{region},1-\alpha}$?

Output: If the time point passes the significance test, return the change point $x_{\hat{k}}$.

Algorithm 2. The search strategy for multiple change point detection and refinement.

Input:
1. A trajectory of length N, comprising time points (x_1, x_2, …, x_N) and a kinetic response parameter ($y_1, y_2, \ldots, y_N$).
2. The pre-determined σ value for the trajectory.
3. The user-defined confidence threshold $1-\alpha$.

Initialization: An array of change points Q_N is initialised with the values [$Q_1 = x_1$, $Q_2 = x_N$].
An iterator i is initialized with a value of 1.

Initial search: While $i < length(Q_N)$, repeat the following:
1. In the region bounded by the time points at Q_i and Q_{i+1}, test for a change point according to Algorithm 1.
2. If a change point is returned, it is inserted into Q between Q_i and Q_{i+1} and i remains unchanged.
3. If no change point is returned, the value of i is incremented by 1.

Refinement: If $Q_N > 4$, the iterator i is reset to 1 and the following steps are repeated while $i < length(Q_N) - 2$:
1. Remove the change point at Q_{i+1} from Q so that the change point that was at Q_{i+2} is now at Q_{i+1}.
2. In the region bounded by the time points at Q_i and Q_{i+1}, test for a change point according to Algorithm 1.
3. If a change point is returned, it is inserted into Q between Q_i and Q_{i+1}, and the value of i is incremented by 1.
4. If no change point is returned, the value of i remains unchanged.

Output: The array of change point positions Q_N. Line segments can then be fit between neighboring change points to determine their associated kinetic parameters.

with

$$\text{test : if } \begin{cases} \sqrt{Z_N} \geq C_{N,1-\alpha}, \text{there is a change at index } \hat{k} \\ \sqrt{Z_N} < C_{N,1-\alpha}, \text{there is no change in the data set} \end{cases}'$$

where $C_{N,1-\alpha}$ is the critical region for N independent measurements and a confidence interval defined by α. For cases involving Gaussian distributed noise, Gombay and Horváth[30] have derived a closed-form expression to approximate the critical region,

$$\frac{1}{2}C_{N,1-\alpha}^2 \exp\left[-\frac{1}{2}C_{N,1-\alpha}^2\right]\left\{T - \frac{2}{C_{N,1-\alpha}^2}T + \frac{4}{C_{N,1-\alpha}^2}\right\} = 1-\alpha, \tag{8}$$

where $T = \ln\left[\frac{(1-h^2)}{h^2}\right]$ and $h = \ln[N]^{3/2}/N$. Given N data points and a confidence interval α, the critical region $C_{N,1-\alpha}$ can rapidly be calculated using root finding methods.[31]

IMPLEMENTATION

The implementation of the kinetic change point method involves two distinct components: single change point identification and significance testing (Algorithm 1) and the recursive binary search strategy for multiple change point identification (Algorithm 2). The required inputs are the trajectory data, the confidence level, and a measure of the noise. In the identification component, each successive time point in a region of interest is treated as a candidate change point, and line fits to the left and right of this candidate change point are calculated. The time point at which the likelihood for the pair of line fits is maximized is then identified. The likelihood ratio for a two-line fit broken at this candidate change point versus a single, unbroken line fit is then subjected to a critical value significance test, and a change point is returned if it passes this test. Subsequently, the search strategy chooses the regions of interest to which single change point identification and testing will be successively applied. It comprises two steps: an initial sweep for change points, followed by a refinement step. In the first search step, the boundaries of the trajectory are defined as the initial set of change points, and the first change point is sought within these bounds. When a change point is found and passes the significance test, the search region narrows so that the next change point is sought in the region bounded by the first change point and that most recently discovered change point. If no change point is found within the current search region, the search progresses to the next search region to the right. This process continues until the end of the trajectory is reached (Fig. 2).

In the refinement step which follows, each internal change point is successively removed and re-sought within the region bounded by its neighboring change points. It may be returned in its original position, returned in a different position within this region, or removed permanently. In this way, the optimal position of each change point is identified within its local region, within a regime which has already been tentatively marked as kinetically distinct. This constitutes a simple but powerful and necessary means of optimization, and it yields robust results when applied to single molecule data.

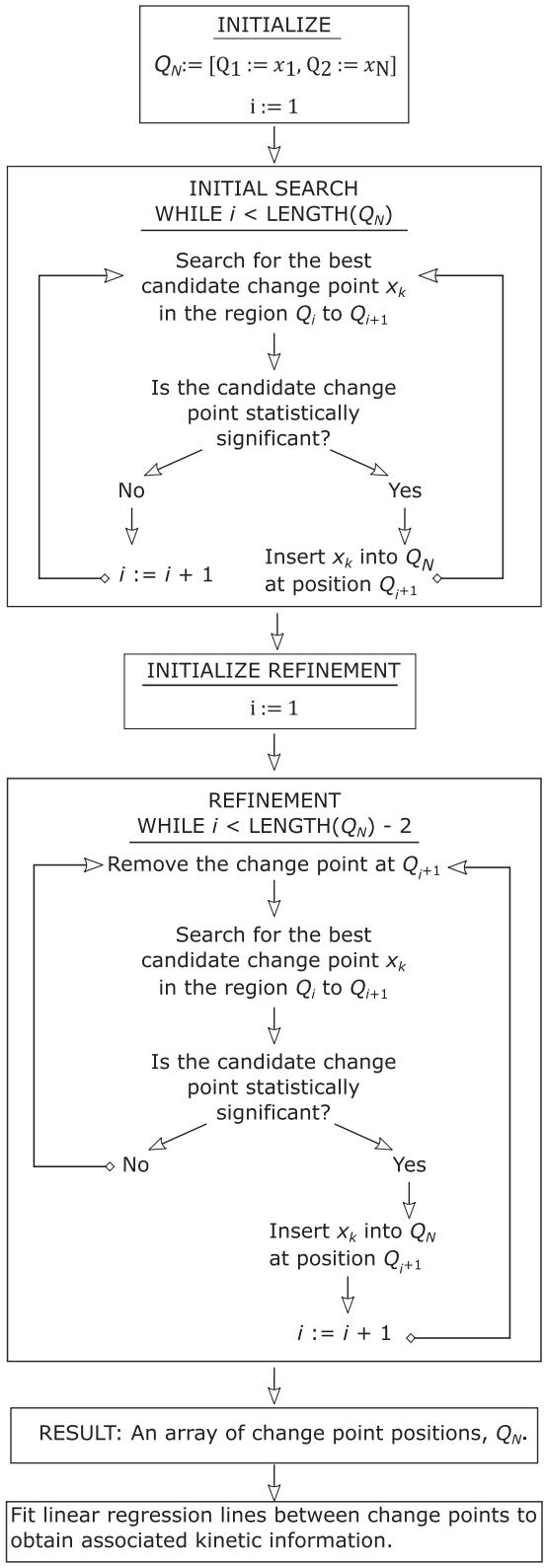

FIG. 2. The scheme of change point detection and refinement involves two phases, an initial search followed by a refinement in which each change point is re-appraised in the context of its neighboring change points, resulting in an array of change point positions. Each change point marks a change in kinetic regime, and by fitting straight lines bounded by these change point positions, the kinetic parameters associated with each regime can be obtained.

RESULTS

To assess the performance of the kinetic change point method, a series of tests were undertaken using simulated trajectory data. These simulations were produced using the Numpy library for numerical computing in Python version 3.6, with time and position being in arbitrary units, and with noise drawn from a random Gaussian distribution with a predefined random seed, allowing the data and analysis presented here to be reproduced from the Github repository (**Software**). All simulated trajectories were analyzed for change points using an implementation of the kinetic change point algorithm written in the Julia language for high-performance numeric computing, version 0.6,[32] with a confidence threshold of 0.99, except where otherwise specified. Subsequent analysis and plotting were performed in the R environment for statistical computing, using the tidyverse set of packages.[33,34]

Likelihood ratio test and false positive rate

Change point analysis inevitably results in false detections, the rate of which is a function of the user-defined confidence threshold (α), as well as how accurately the input σ value represents the true noise. While a confidence threshold of 0.99 putatively gives a false positive discovery rate of 1%, the empirically determined rate is somewhat different, even when allowing for a perfectly accurate σ value. To assess this aspect, 10^5 trajectories each with a length of 500 time points were generated and each exclusively consisting of noise drawn from a Gaussian distribution of width 100. In this and other tests presented here, these values are unitless. Importantly, no true change points were present in these trajectories. The kinetic change point method was applied at confidence threshold values of 0.90, 0.95, and 0.99, with the σ value set to 100 to accurately represent the noise. From this analysis, the empirically determined false positive rates for these confidence thresholds were 15.9%, 7.4%, and 1.4% for the 0.90, 0.95, and 0.99 confidence thresholds, respectively. False positives are predominantly located at the peripheries of the trajectories (Fig. 3).

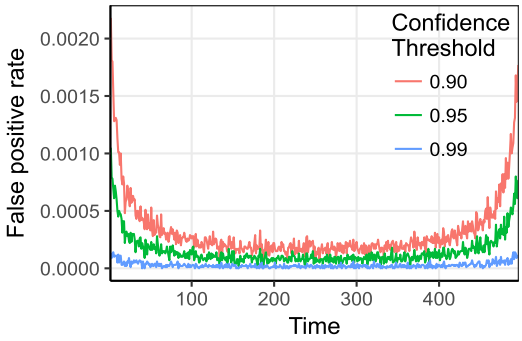

FIG. 3. The rate of false positive detection for various confidence threshold values as a function of position in simulated noise with a duration of 500 time points, averaged over 10^5 trajectories. A greater number of false positives are obtained at the peripheries of the trajectory.

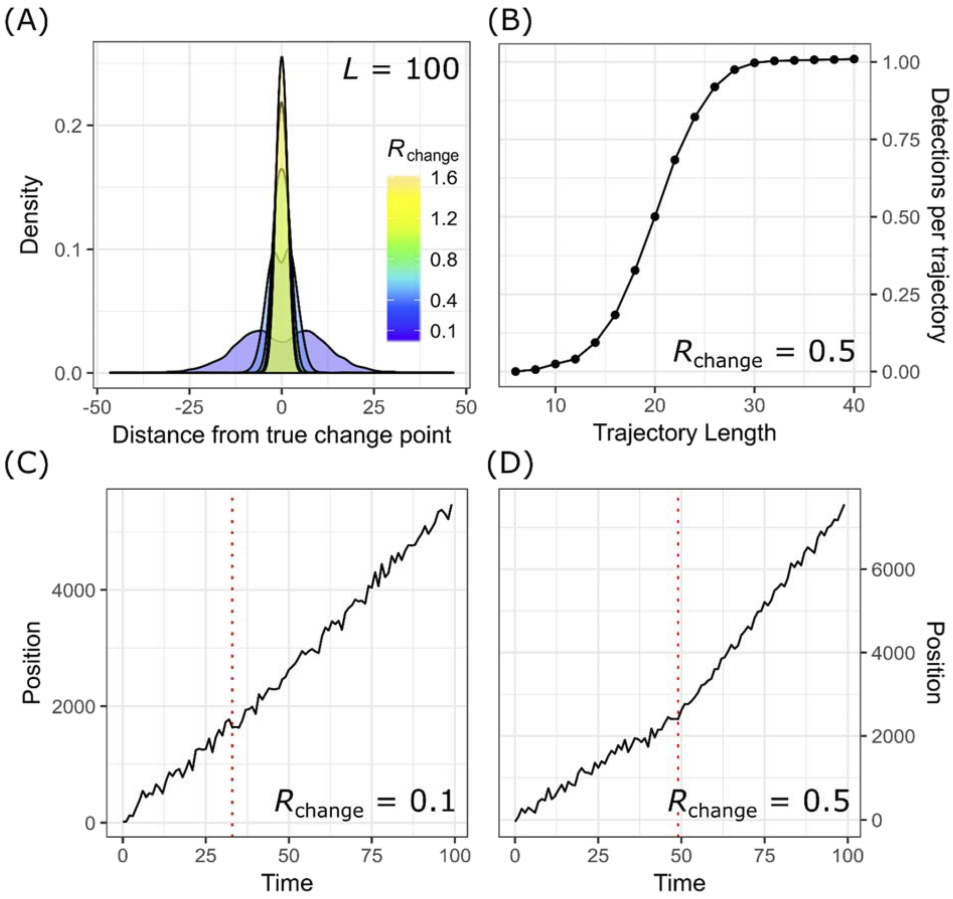

FIG. 4. Accuracy and rate of detection for single change points placed at the halfway points of trajectories. (a) The distribution of discovered change point positions as a function of R_{change}, the ratio of the absolute rate change over the noise, at a fixed trajectory length of 100 time points. (b) The rate of change point discovery as a function of trajectory length, for a fixed R_{change} value of 0.5. For every R_{change} or length condition, 10^4 trajectories were simulated. (c) Example trajectory with a low R_{change} value of 0.1. Although a change point has correctly been detected, it is of low positional accuracy. (d) Example trajectory with an intermediate R_{change} value of 0.5, at which the magnitude of the rate change is equal to that of the noise. The position of the change point has been accurately identified. Time and position are in arbitrary units.

Change point detection accuracy

To assess the accuracy of change point detection, simulated trajectories of enzyme position, P, with length $L = 100$ time points and noise drawn from a Gaussian distribution of width $W = 100$ were produced, in which a single change point occurs between time points 49 and 50. Since arbitrary units were used in the simulations, the ratio of the rate, a, over the noise, i.e., $R = a/W$, is the proper metric to determine the detection power of the algorithm. The initial rate in each trajectory is 50, corresponding to an R_1 value of 0.5, with the second rate taking a range of values between 60 and 200, corresponding to R_2 values ranging between 0.6 and 2. Under these conditions of fixed trajectory length, the challenge to accurate detection is best expressed as the difference in these R values, $R_{\text{change}} = R_2 - R_1$, giving a range of tested R_{change} values of 0.1 to 1.9. At each R_{change} value investigated, 10^4 trajectories were simulated. There were no trajectories that returned type II errors for a trajectory length of 100 points, while supernumerary change points, i.e., type I errors, were discovered for a small fraction of trajectories at a rate consistent with that described in the subsection titled Likelihood ratio test and false positive rate for a confidence threshold of 0.99. The distribution of discovered change point positions narrows as R_{change} increases [Fig. 4(a)]. Another challenge to accurate detection is the number of time points in the search region. At a fixed R_{change} value of 0.5, trajectories of durations ranging between 6 and 40 time points were tested under conditions similar to those above, with the change point located at the halfway position. When the length is short, type II errors are obtained, and the rate of change point detection develops sigmoidally with the trajectory length, with ~50% of change points detected at a length of 19 time points and >99% of change points detected in trajectories of length greater than 27 time points [Fig. 4(b)].

Pause detection

Many biological systems of interest exhibit pausing behavior in their single-molecule trajectories, in which the rate

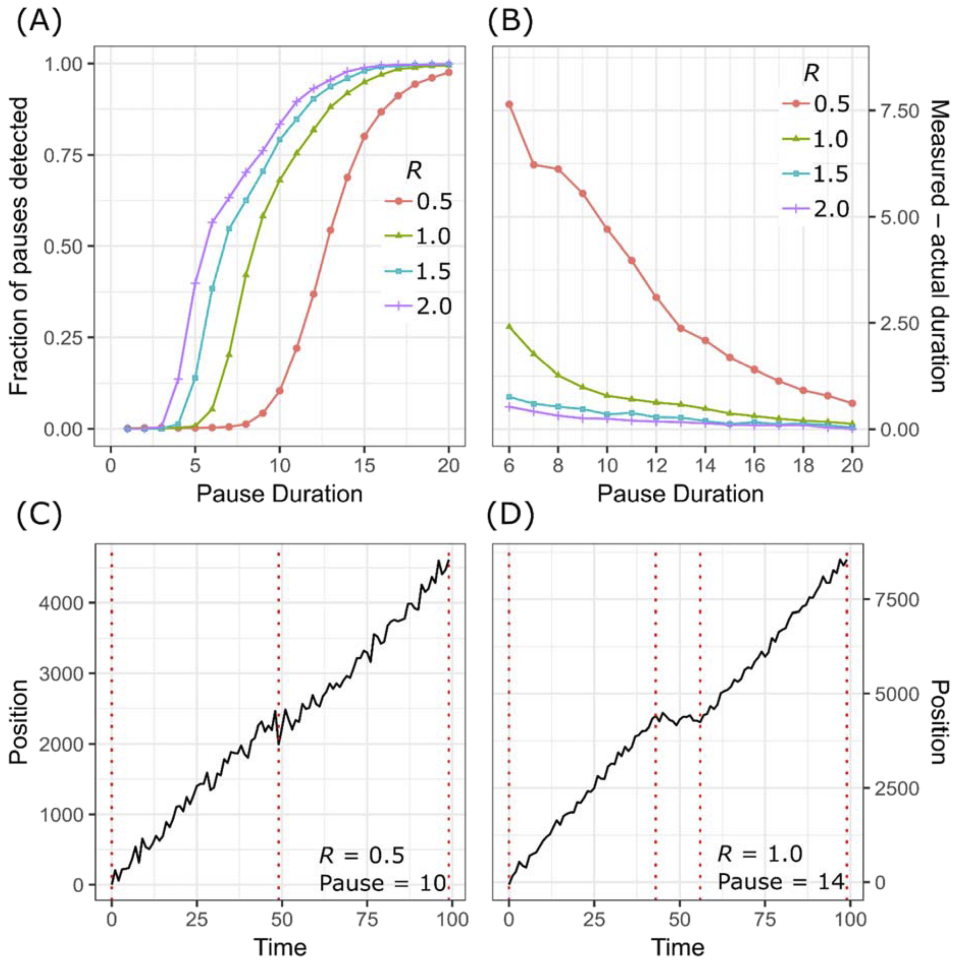

FIG. 5. Pause detection as a function of pause duration and rate. Trajectories of length 100 were simulated, with rate:noise values R ranging between 0.5 and 2.0. Each trajectory contained a pause region of zero rate centered at the midpoint. These pause segments had durations ranging between 1 and 20 time points. For each unique combination of pause duration and R value, 10^4 trajectories were simulated and the change point search was conducted with a confidence threshold of 0.99. (a) The fraction of pauses detected rises as a sigmoidal function with the pause duration, with detection efficiency increasing with rate. (b) The mean discrepancy between the measured pause duration and the actual pause duration rapidly declines to zero for rate values equal to and greater than the noise, i.e., R > 1.0, while at an R value of 0.5, this discrepancy is slower to decline. The discrepancy values are shown for rate and duration combinations for which at least half of the pauses were discovered. (c) An example of a pause that spans only a small number of time points, at a rate that is half the width of the noise, $R = 0.5$. In this example, the pause was not detected; however, a single change point has been detected at the mid-point of the pause. (d) An example of a pause that spans a moderate number of time points, at a rate equal to the width of the noise, $R = 1.0$. In this example, the pause has been accurately identified.

transiently drops to zero before returning to its prior value. It is desirable for any method of single molecule kinetic analysis that these key features of interest are extracted with high confidence and accuracy. However, reliable pause detection can be particularly challenging. The relative prominence of a pause, and therefore its likelihood of detection, depends on the interplay of the rate of the event that has been interrupted by the pause, the pause duration, the noise, the length of the region within which the pause is to be located, and the sampling frequency. To test the pause detection performance of the kinetic change point method, pauses were placed at the center of simulated trajectories of length 100. Rates were varied from 50 to 200, which at a noise of width 100 corresponds to a range of R values of 0.5–2. Here, the challenge parameter is the R value relative to the pause duration. For each rate value, the pause duration was varied from 1 to 20 time points. For completeness, this range includes pauses that are well below the threshold for being visibly discernible; however, the algorithm cannot be expected to dramatically outperform visual perception. At each unique combination of rate and pause duration, 10^4 trajectories were simulated. These were then assessed for change points using a confidence threshold of 0.99.

In each simulated trajectory, successful pause detection was considered the observation of a pair of change points flanking the center of the pause and resulting in a line segment with a slope below 20 [Fig. 5(a)]. The segment lengths of detected pauses were compared with the actual pause durations. For rate values equal to and greater than the width of the noise distribution, and for pause durations longer than six time points, there was a close agreement between the measured and true pause durations, with a slight bias toward overestimation of the duration [Fig. 5(b)]. Of the trajectories where two change points defining the pause were not discovered, typically a single change point was discovered instead, with the mean position of these single change points being located at the center of the pause [Fig. 5(c)]. Since the probability of pause detection decreases with duration, estimates for the mean pause duration will be larger than the true value. This can be corrected, for example, by applying a maximum likelihood estimation-based method to model the measured distribution of pause durations.[35]

Power of detection for complex, multistate data

In real single molecule trajectories, an undefined number of change points representing rate changes and pauses are to be expected. One of the main challenges here is the spacing between change points, which become difficult to detect when they are closer together and when the experimental sampling rate is low. This scenario gives rise to type II errors, and the probability of detection sigmoidally approaches one as the length of the search region increases, even for very small changes (Fig. 4). To evaluate identification of change point pairs, a dataset was produced with defined spacing between change points and a Gaussian noise of width 100. However, the rate change at each change point could take any value drawn from a Gaussian distribution centered at zero, with a width of 200, i.e., double the width of the noise.

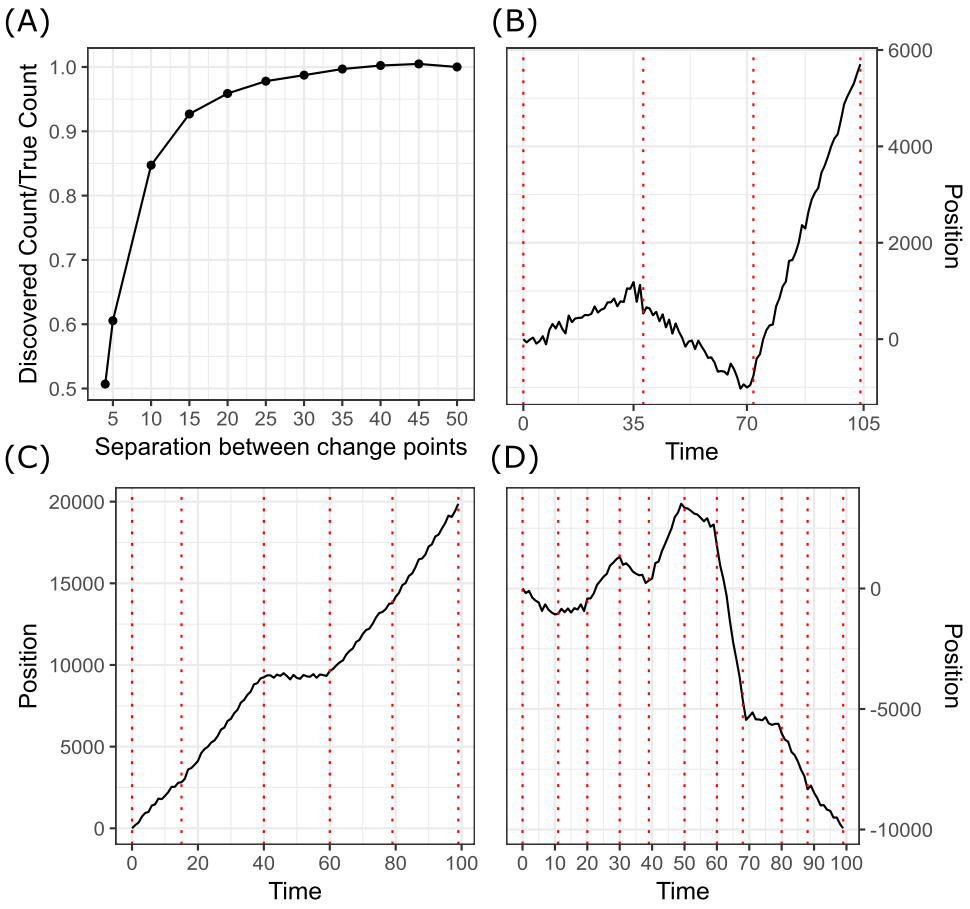

FIG. 6. (a) The ratio of the number of discovered change points over the true number of change points as a function of separation distance for trajectory lengths from 100 to 135 (allowing for integral numbers of change points per trajectory), with a Gaussian noise of width 100, and rate changes at each change point being drawn from a Gaussian distribution centered at zero with a width of 200. 10^4 trajectories were simulated for each separation condition. [(b)–(d)] Example trajectories with change points spaced every 35 [panel (b)], 20 [panel (c)], and 10 time points [panel (d)]. The discovered change point positions are indicated by the vertical dashed lines, while axis labels indicate the first time points that follow each change. Note that when the spacing between change points is adequate, as in (c), there is greater power for subtle changes to be detected.

Separations between change points ranging between 4 and 50 time points were tested, on trajectories of length ~100 (adjusted as necessary to allow an integral number of change points per trajectory), and 10^4 simulated trajectories were generated for each separation value. A substantial proportion of change points represent changes with a very low R_{change} value and are essentially invisible. However, such an inability to separate nearby change points is to be expected in real datasets. Under these conditions, at a separation between true change points of 5 time points, the number of discovered change points was 60% of the number of true change points, i.e., ~40% of the true change points are missed as type II errors. At a separation of 25 time points, the number of discovered change points was 97% of the count of true change points (Fig. 6). These results emphasize the need to record data at a sampling rate which is appropriate for the frequency with which changes occur in the system of interest.

CONCLUSIONS

The rapidly increasing application of single molecule approaches to study a broad range of biological systems has led to numerous techniques for the identification and classification of discrete changes in noisy trajectories. These approaches often provide the best fit for specific data types and hallmark behaviors, but depend on arbitrary user-defined parameters, optimizations, and filtering steps, which complicates reproduction and prevents their broad application. The kinetic change point algorithm outlined here provides a rigorous approach to detect discrete rate changes in noisy single molecule observations with a minimum number of parameters and a sound statistical foundation. The only requirements are that the system exhibits discrete rate changes and the observations are convoluted with Gaussian noise. Optimal performance relies on the data satisfying these conditions and the optimization of a few key parameters. One of these is the noise value provided to the algorithm, which can typically be measured from a subset of data in which it is known that no change points are present. Another key parameter is the sampling rate, which must be sufficient with respect to the minimum interval between change points that is to be resolved. For smaller values of R_{change}, a greater interval between change points is required for them to be accurately identified. This is of extra importance when pauses are present, as the successful and accurate identification of a pause requires that two adjacent change points are accurately identified.

The implementation of this method is computationally fast, allowing rapid batch processing of massive numbers of trajectories. For high-throughput techniques, this approach offers a strategy to rapidly identify trajectories containing features of interest, in a defined manner that is superior to the biased method of manual curation. The ImageJ and Julia packages provided should allow both expert and novice users to broadly apply the approach.

APPENDIX: SOFTWARE AND DATA AVAILABILITY

The Github repository for the Julia implementation of the kinetic change point method, along with Jupyter Notebooks containing all of the code necessary to reproduce the simulations and analysis presented here is available at https://github.com/duderstadt-lab/Julia_KCP_Notebooks. The ImageJ plugin for the kinetic change point method is available at https://github.com/duderstadt-lab/Java_KCP.

[1] R. Phillips, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2013).
[2] J. F. Beausang, Y. E. Goldman, and P. C. Nelson, Methods Enzymol. **487**, 431–463 (2011).
[3] L. P. Watkins and H. Yang, J. Phys. Chem. B **109**(1), 617–628 (2005).
[4] D. Montiel, H. Cang, and H. Yang, J. Phys. Chem. B **110**(40), 19763–19770 (2006).
[5] D. Dulin, J. Lipfert, M. C. Moolman, and N. H. Dekker, Nat. Rev. Genet. **14**(1), 9–22 (2013).
[6] K. M. Herbert, W. J. Greenleaf, and S. M. Block, Annu. Rev. Biochem. **77**, 149–176 (2008).
[7] K. C. Neuman, E. A. Abbondanzieri, R. Landick, J. Gelles, and S. M. Block, Cell **115**(4), 437–447 (2003).
[8] A. Yildiz, M. Tomishige, R. D. Vale, and P. R. Selvin, Science **303**(5658), 676–678 (2004).
[9] K. E. Duderstadt, R. Reyes-Lamothe, A. M. van Oijen, and D. J. Sherratt, Cold Spring Harbor Perspect. Biol. **6**(1), a010157 (2014).
[10] H. J. Geertsema, A. W. Kulczyk, C. C. Richardson, and A. M. van Oijen, Proc. Natl. Acad. Sci. U. S. A. **111**(11), 4073–4078 (2014).
[11] J. J. Loparo, A. W. Kulczyk, C. C. Richardson, and A. M. van Oijen, Proc. Natl. Acad. Sci. U. S. A. **108**(9), 3584–3589 (2011).
[12] M. C. Leake, J. H. Chandler, G. H. Wadhams, F. Bai, R. M. Berry, and J. P. Armitage, Nature **443**(7109), 355–358 (2006).
[13] M. Manosas, J. Camunas-Soler, V. Croquette, and F. Ritort, Nat. Commun. **8**(1), 304 (2017).
[14] S. A. McKinney, C. Joo, and T. Ha, Biophys. J. **91**(5), 1941–1951 (2006).
[15] L. S. Milescu, A. Yildiz, P. R. Selvin, and F. Sachs, Biophys. J. **91**(4), 1156–1168 (2006).
[16] S. Liu, G. Chistol, C. L. Hetherington, S. Tafoya, K. Aathavan, J. Schnitzbauer, S. Grimes, P. J. Jardine, and C. Bustamante, Cell **157**(3), 702–713 (2014).
[17] S. Aminikhanghahi and D. J. Cook, Knowl. Inf. Syst. **51**(2), 339–367 (2017).
[18] R. Killick and I. A. Eckley, J. Stat. Software **58**(3), 1–19 (2014).
[19] C. W. S. Chen, J. S. K. Chan, R. Gerlach, and W. Y. L. Hsieh, Stat. Comput. **21**(3), 395–414 (2011).
[20] R. E. Quandt, J. Am. Stat. Assoc. **53**(284), 873–880 (1958).
[21] K. E. Duderstadt, H. J. Geertsema, S. A. Stratmann, C. M. Punter, A. W. Kulczyk, C. C. Richardson, and A. M. van Oijen, Mol. Cell **64**(6), 1035–1047 (2016).
[22] D. Dulin, B. A. Berghuis, M. Depken, and N. H. Dekker, Curr. Opin. Struct. Biol. **34**, 116–122 (2015).
[23] F. R. Hill, E. Monachino, and A. M. van Oijen, Biochem. Soc. Trans. **45**(3), 759–769 (2017).
[24] J. S. Lewis, L. M. Spenkelink, G. D. Schauer, F. R. Hill, R. E. Georgescu, M. E. O'Donnell, and A. M. van Oijen, Proc. Natl. Acad. Sci. U. S. A. **114**(40), 10630–10635 (2017).
[25] A. D. Robison and I. J. Finkelstein, FEBS Lett. **588**(19), 3539–3546 (2014).
[26] H. Yang, *Change-Point Localization and Wavelet Spectral Analysis of Single-Molecule Time Series* (John Wiley & Sons, Inc., Hoboken, 2011).
[27] H. J. Geertsema, K. E. Duderstadt, and A. M. van Oijen, Methods Mol. Biol. **1300**, 219–238 (2015).
[28] J. B. Lee, R. K. Hite, S. M. Hamdan, X. S. Xie, C. C. Richardson, and A. M. van Oijen, Nature **439**(7076), 621–624 (2006).
[29] L. Horvath, Ann. Stat. **21**(2), 671–680 (1993).
[30] E. Gombay and L. Horváth, J. Stat. Plann. Inference **52**(1), 43–66 (1996).
[31] E. Gombay and L. Horváth, J. Multivar. Anal. **56**(1), 120–152 (1996).
[32] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, SIAM Rev. **59**, 65–98 (2017).
[33] R Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2017).
[34] H. Wickham, Tidyverse, R package version 1.1.1, 2017.
[35] M. S. Woody, J. H. Lewis, M. J. Greenberg, Y. E. Goldman, and E. M. Ostap, Biophys. J. **111**(2), 273–282 (2016).