OPEN ACCESS Freely available online

PLOS ONE

Step Detection in Single-Molecule Real Time Trajectories Embedded in Correlated Noise

Srikesh G. Arunajadai[1]*, Wei Cheng[2]*

1 Department of Biostatistics, Columbia University, New York, New York, United States of America, **2** Department of Pharmaceutical Sciences, College of Pharmacy, University of Michigan, Ann Arbor, Michigan, United States of America

Abstract

Single-molecule real time trajectories are embedded in high noise. To extract kinetic or dynamic information of the molecules from these trajectories often requires idealization of the data in steps and dwells. One major premise behind the existing single-molecule data analysis algorithms is the Gaussian 'white' noise, which displays no correlation in time and whose amplitude is independent on data sampling frequency. This so-called 'white' noise is widely assumed but its validity has not been critically evaluated. We show that correlated noise exists in single-molecule real time trajectories collected from optical tweezers. The assumption of white noise during analysis of these data can lead to serious over- or underestimation of the number of steps depending on the algorithms employed. We present a statistical method that quantitatively evaluates the structure of the underlying noise, takes the noise structure into account, and identifies steps and dwells in a single-molecule trajectory. Unlike existing data analysis algorithms, this method uses Generalized Least Squares (GLS) to detect steps and dwells. Under the GLS framework, the optimal number of steps is chosen using model selection criteria such as Bayesian Information Criterion (BIC). Comparison with existing step detection algorithms showed that this GLS method can detect step locations with highest accuracy in the presence of correlated noise. Because this method is automated, and directly works with high bandwidth data without pre-filtering or assumption of Gaussian noise, it may be broadly useful for analysis of single-molecule real time trajectories.

Citation: Arunajadai SG, Cheng W (2013) Step Detection in Single-Molecule Real Time Trajectories Embedded in Correlated Noise. PLoS ONE 8(3): e59279. doi:10.1371/journal.pone.0059279

Editor: Pratul K. Agarwal, Oak Ridge National Laboratory, United States of America

Received January 3, 2013; **Accepted** February 13, 2013; **Published** March 22, 2013

Funding: This work is supported by National Science Foundation Faculty Early Career Development (CAREER) Award CHE 1149670 (W.C.), NIH grant 1DP2OD008693-01 (W.C.), and also in part by the March of Dimes Foundation Research Grant No. 5-FY10-490 (W.C.). W. C. thanks start-up funding support from the University of Michigan at Ann Arbor and the Ara Paul Professorship fund at the University of Michigan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sarunajadai@columbia.edu (SA); chengwe@umich.edu (WC)

Introduction

The advent of single-molecule techniques [1,2,3,4,5,6,7] in recent years brought many interesting discoveries in chemistry, physics, and life sciences. One unique advantage of single-molecule technique is the ability to measure molecular processes in a heterogeneous environment without the need of synchronizing these molecules, and to unveil the static and dynamic disorders among individual molecules [1]. One broad class of single-molecule measurement is movement of molecular motors in real time. These molecular motors move in steps [8,9,10,11,12]. Statistics on their movement trajectories can reveal rich mechanistic information that is often inaccessible from conventional bulk experiments.

Different types of statistical tools have been developed for analysis of these data to extract characteristics of motor movement. For stepping of molecular motors that can be observed directly from time trajectories, pairwise distance distribution analysis was among the first to be used for this task [13]. A Fourier analysis of the pairwise distance distribution histogram can reveal the periodicity in single-molecule trajectories, which is an objective measure of motor step size. Application of this method to different molecular motors has revealed their apparent step sizes of movement [11,14,15,16,17], although this analysis does not yield information on the dwell time in between motor steps, which is essential in deducing the coupling of fuel molecule to motor movement. To this end, algorithms for detection of both steps and dwells have been developed by investigators [18,19,20,21,22,23,24,25], and the performance of several methods has been quantitatively compared [26]. In particular, the algorithm developed by Kerssemakers et al. [19] (referred as KERS herein) has found increasing use in different motor systems [27,28,29]. In this method, the original data was assumed to be a step function buried in Gaussian noise. The motor steps are found in successive iterations: the plateaus of the steps identified in a previous cycle are further divided to find additional steps. The quality of the fit was assessed using a statistic S, which is the ratio between the Chi-squared of a counter fit and the Chi-squared of the best fit. For molecular motors that can be measured at single-molecule level but whose individual steps are obscured by measurement noise, techniques have also been developed to extract step size information from variance in long trajectories of motor movement [30]. Under these circumstances, even though the individual steps of the motor cannot be identified directly from time traces [31,32,33], estimation of motor step size using this technique has yielded values that are comparable to results from other complementary approaches [34].

Despite the diversity of these different step-detection algorithms, a common practice is the assumption of Gaussian white noise in the experimental data, which is independently distributed and shows no correlation with regard to time. This assumption may be true in certain cases, but has not been thoroughly validated in general. Any noise that has frequency-dependent amplitude will deviate from Gaussian white noise. This so-called 'colored' noise displays autocorrelations and widely exists in nature [35]. For example, colored noise is typically present in lasers that are used to form optical tweezers. Both intensity and pointing stability of the laser display noise whose amplitudes depend on bandwidth [36,37]. As we show, colored noise is present in single-molecule real time trajectories collected from optical tweezers. The assumption of Gaussian white noise for single-molecule data that contains colored noise can result in significant fitting errors. It is thus critical to assess the structure of the noise when analyzing these single-molecule trajectories. We have developed a statistical step detection algorithm based on Generalized Least Squares (referred as GLS herein) that explicitly takes the structure of the noise into account. This algorithm allows one to identify motor steps and dwells directly from time trajectories in the presence of highly autocorrelated noise and provides standard errors and confidence intervals associated with these steps. There is no assumption on a single unique step size in this algorithm. Indeed, variation in size of steps can be fully taken into account [38]. There is no requirement on the motor to be highly processive [30]. The time trajectory can still be analyzed even though the motor can detach from its track prematurely. We present this method in detail and compare it with the KERS method. As we demonstrate, this GLS method can detect steps with highest accuracy in the presence of correlated noise, which can significantly minimize errors in data analysis and interpretation. Because this GLS method can work with high bandwidth data directly without any pre-filtering, it may be broadly applicable to single-molecule data analysis in general.

Results and Discussion

Correlated Noise in Single-molecule Trajectories

The structure of noise in a real time trajectory can be revealed by calculating the autocorrelation function (ACF) of the data. Gaussian white noise will display a delta function while autocorrelated noise will show an exponential decay for its ACF. We have extensively computed the ACF for real time single-molecule trajectories collected with high resolution optical tweezers [38]. A typical result is shown in Fig. 1A, which shows a clear exponential decay. In contrast, a simulated Gaussian white noise shows the expected delta function (Fig. 1B). This result demonstrates that the experimental single-molecule trajectory indeed contains correlated noise, i.e., the noise amplitude at the current moment is a function of past noise and some random error, which induces a correlation structure in the noise. The order of this correlation structure can be further assessed using the plot of partial autocorrelation function (PACF) [39]. Gaussian white noise will display zero everywhere throughout the PACF while for autocorrelated noise of order p, the PACF is zero for lags greater than p and non-zero otherwise [39]. As shown in Fig. 1C, the corresponding PACF of Fig. 1A shows non-zero amplitude before lag 7 and zero thereafter, highlighted by the horizontal dashed lines that indicate the 95% confidence intervals under the null hypothesis of no correlation, thereby suggesting an order 7 for this noise, i.e., the noise is a function of past seven values of noise and some random error. In contrast, the simulated Gaussian noise has zero amplitudes everywhere throughout the PACF (Fig. 1D).

Step Detection in the Presence of Correlated Noise

The above results show that experimental single-molecule trajectories contain noise that is correlated in time. Would it still be fine to assume Gaussian white noise when we analyze these traces? To address this question, we have developed a step detection method using GLS (Materials and Methods). In this method, we have the option of assuming Gaussian white noise for the data to be analyzed, or explicitly take the structure of the noise into account based on PACF analysis. To examine the impact of Gaussian noise assumption on data analysis, we generated simulated single-molecule trajectories embedded in autocorrelated noise, and compared step detection with and without Gaussian assumption for the added noise. In addition, we also analyzed the same set of traces using KERS method in order to compare with the GLS method. Fig. S1A shows a simulated step function that resembles real time RNA unwinding traces based on our recent publication [38]. It consists of a series of upward steps that are occasionally interrupted by downward steps. We represent the time axis by indexing integers for easy identification. We then added noise to the step function to generate mock unwinding traces. One such realization is shown in Fig. S1B. The noise is simulated from an autoregressive process of order 7 (Fig. 1C), with coefficients 0.222, 0.072, 0.035, 0.015, 0.016, 0.003 and 0.013 that are typically found from the published single-molecule trajectories [38]. We independently simulated the noise 100 times to generate 100 mock traces. We then use three different procedures to identify steps and dwells in these traces and compare them: (1) the GLS method; (2) exactly the same procedure as GLS method but ignoring autocorrelation in the noise, i.e., assuming Gaussian white noise even though the added noise is correlated; and (3) the KERS method.

Fig. 2 shows the histograms of the number of steps detected from the above procedures. As listed in Table 1 and shown in Fig. 2A, the GLS method on average detected 34 steps from these traces, ranging between 23 and 40, which compares very well with the total number of 33 steps in the simulated step function. Fig. 3A shows a representative best fit (red line) from this procedure, which shows close resemblance to the original step function (blue dashed line). Repeating the same procedure but ignoring autocorrelation in the noise vastly overestimates the number of steps, with a mean of 66 steps, ranging between 45 and 91 (Fig. 2B). Fig. 3B shows a representative best fit from this second procedure. Comparison between Fig. 3B and Fig. S1 suggests that majorities of the steps in the simulated trace were identified, but a significant fraction of these steps are false positives, because they do not exist in the original trace. *These false positives were identified as a result of the autocorrelated noise, which was not accounted for in this step detection procedure.* To confirm the impact of this correlated noise on step detection, we used the same step function as shown in Fig. S1, but added Gaussian white noise, and repeated the same step detection procedure. Fig. S2A shows a representative best fit from this procedure. Interestingly, it now detects correct number of steps on average (Table 1). These results demonstrate that the structure of the underlying noise in a single-molecule trajectory has a profound impact on the outcome of step detection. The assumption of Gaussian white noise on otherwise correlated noise can lead to a significant overestimation for the number of steps in a trace.

In contrast to the second procedure, the KERS method vastly underestimates the number of steps, with a mean of 5 steps and a range between 4 and 8 (Fig. 2C). Fig. 3C shows a representative best fit from KERS method. Fig. 3D shows the S-statistic obtained throughout the 100 mock traces. For each realization, the S-statistic from the original step function (true S value) was shown as crosses and that from the best attempted fit was shown as red dots.

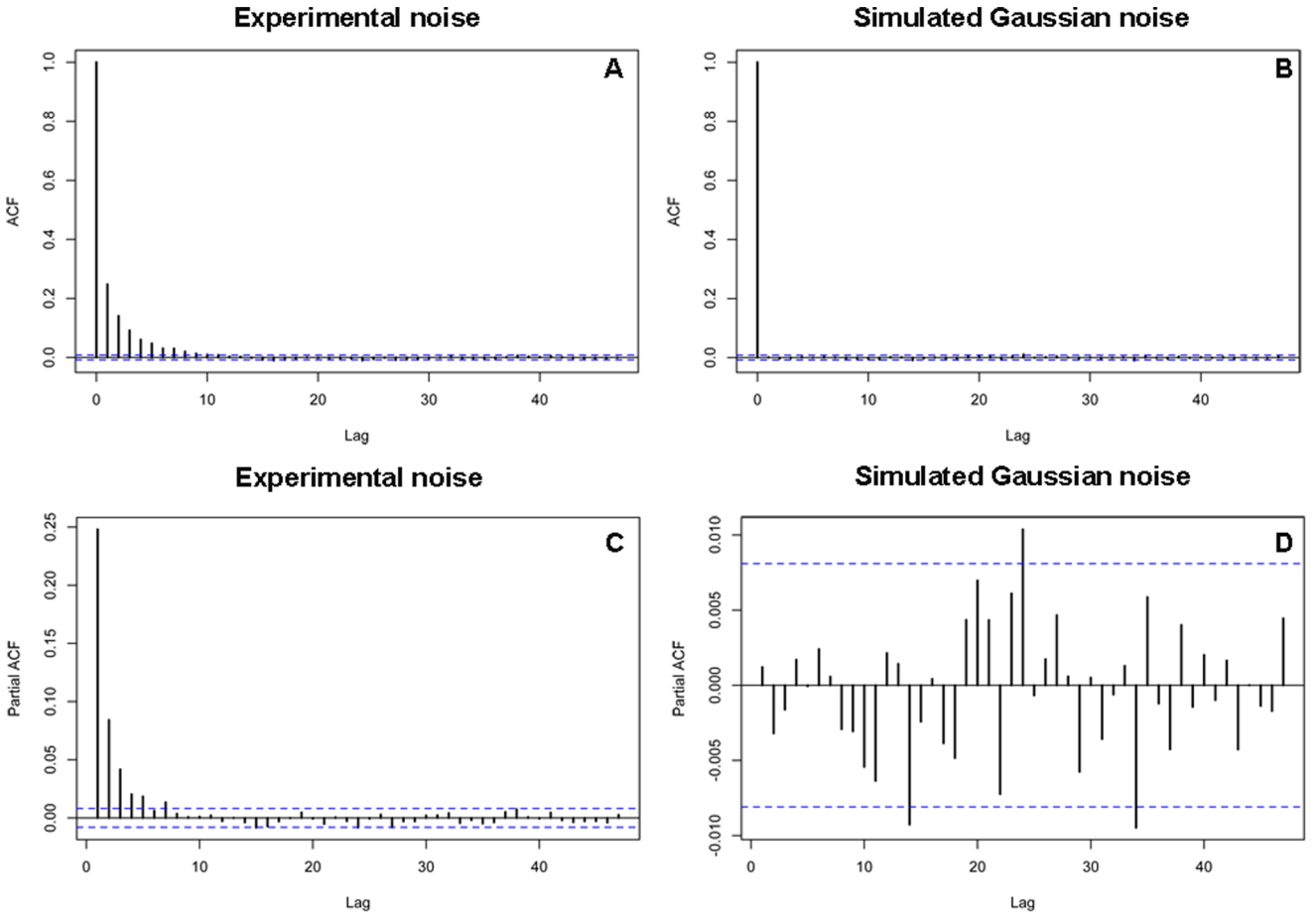

Figure 1. Correlated noise in single-molecule real time trajectories. The autocorrelation function (ACF) and the partial autocorrelation function (PACF) for AR noise of order 7 as observed in a typical RNA unwinding trace (A and C). The plots from simulated Gaussian noise were also shown for comparison (B and D). The horizontal lines indicate the 95% confidence intervals under the null hypothesis of no correlation. For the AR(7) noise, the ACF shows exponentially decay while the PACF gradually cuts off, i.e. goes to zero after lag 7. For Gaussian noise, the ACF is 1 at lag 0 and zero for other lags while the PACF is zero for all lags.
doi:10.1371/journal.pone.0059279.g001

Although the true S values are generally higher than those from the best attempted fits (indicating a better fit), the S-statistic from the attempted fits are within random variable limits of the true S value as indicated by the horizontal 95% confidence interval lines. This result suggests that the KERS method can give results that deviate significantly from reality. One possible reason behind this is the high bandwidth of the data (2.5 kHz). To test this, we used a boxcar filter with a window size of 10 to filter and decimate the trace, and attempted again with KERS method. A representative result is shown in Fig. S2B. It now detects 11 steps instead of 4, closer to reality but still much lower than the true value of 33. This result suggests that the KERS method is highly dependent on the bandwidth of the data, and was not able to correctly identify the steps in the original trace even after filtering. As a result, the clear advantage of GLS method is that it can work with high bandwidth data directly without any filtering. In summary, noise structure should be accounted for in single-molecule data analysis. Assumption of Gaussian white noise can lead to either over- or underestimation of the number of steps depending on the algorithms used. Moreover, the GLS method outperforms the KERS method (which assumes Gaussian white noise) and on average detects the correct number of steps.

Table 1. Summary of number of steps detected from 100 realizations of the simulated traces.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
(a) GLS	23.00	33.00	35.00	33.79	36.00	40.00
(b) Ignoring correlation	45.00	56.00	63.50	65.64	75.00	91.00
(c) True Gaussian	26.00	28.00	30.00	29.75	31.00	36.00
(d) KERS	4.00	4.00	4.00	4.74	5.00	8.00

Four different procedures were used to detect steps, where Procedure (a), (b) and (d) are for traces with correlated noise analyzed with GLS method (a), ignoring noise correlation and assuming Gaussian noise (b) and the KERS method (d); Procedure (c) was done for traces with Gaussian white noise that were analyzed using the GLS method.
doi:10.1371/journal.pone.0059279.t001

Despite being the best among the three procedures, results shown in Fig. 2A and Table 1 indicate that the GLS method can still over- or underestimate the number of steps in a trace. To examine these deviations in more detail, Table 2 shows the statistics of false positives (non-existing steps but detected as a step) and true negatives (true steps that were not detected) from the

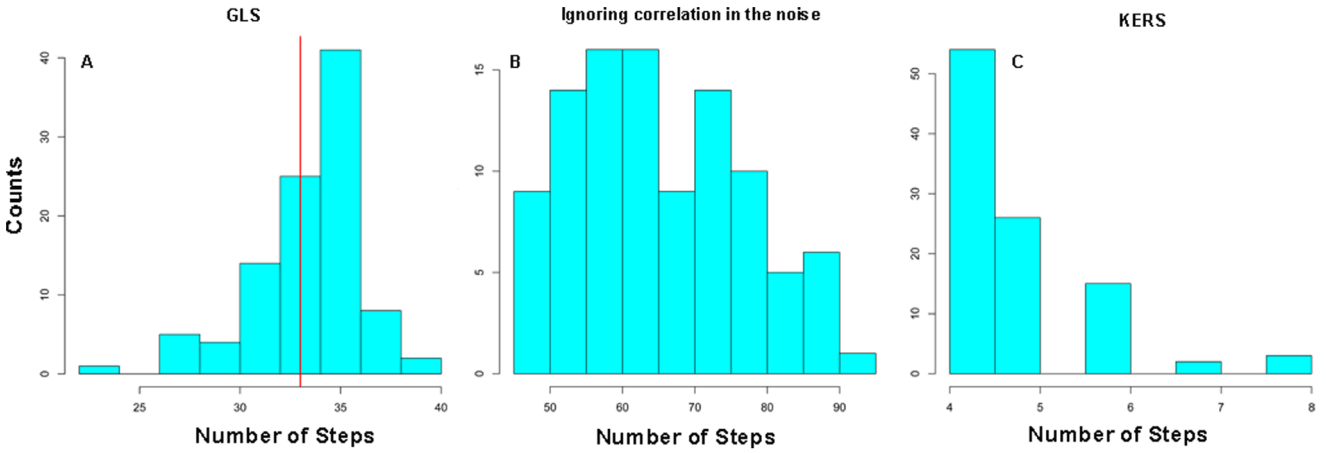

Figure 2. Histogram of the number of steps detected from 100 realizations of the simulated traces. (A), (B) and (C) show the results from GLS method, GLS method but ignoring the correlation in the noise and KERS method, respectively.
doi:10.1371/journal.pone.0059279.g002

mock traces analyzed with the GLS method. On average, the median of false positives was 1. The median of true negatives was zero with 75% of the traces having at most 2 true negatives. This result suggests that GLS does a very good job in identifying almost all the steps that are present, although it can occasionally detect false positives. Fig. 4 further shows the fraction of traces in which a given step was detected. Over the total 33 simulated steps, 70% of them were detected every time by GLS. The steps whose detection efficiency drops below 90% (indicated by the red dashed line) are usually the transient steps, i.e., those steps that have very short dwell times in between, as seen from Fig. S1B. Specifically, these are steps # 1 (31 ms), # 6 (18 ms), # 7(18 ms), # 29(80 ms), # 30(80 ms) and # 31(84 ms), where the dwell times in milliseconds are noted in parentheses. These dwell times are relatively short in comparison to other dwells, which ranged between 0.1 and 2.25 s. Still, the detection efficiency for all these transient events is greater than 65%, which is in contrast to KERS method where detection of transient steps depends on filtering and is below 50% even for filtered data (Fig. S2B). This is a very important feature of GLS method, because one of the distinct advantages of single-molecule real time measurement is to reveal transient events. *If step detection requires data filtering, then these transient events are likely to be masked as a result of filtering.*

The duration of dwells in between steps are of significant interest in single-molecule real time trajectories. These dwells are computed as the time elapsed between two steps. Typically, these dwells represent the waiting time the motor has to take before next motion, which is usually coupled to fuel binding under limiting fuel concentrations. It is therefore important to quantitate the accuracy with which a step location can be identified. To this end, we first detected steps using GLS method from the set of test traces. We then quantified the deviation of the identified step location from its true step location. This deviation is computed as the difference in time (data index) between the two, which is further normalized by the lengths of the true dwell time before and after that step (Fig. 5). For example, imagine the dwell to the left of a true step location be of length 50 and to the right be length 100. If the step is identified 10 points to the left of the true location we indicate its deviation as −20%; if the step is identified at the exact location then it is 0%; and if it is identified 30 points to the right of the true step location, we indicate it as +30%. Fig. 5 shows the percentage of deviation for each step as computed above from GLS method. It can be seen that for most of the steps, the step deviation is close to zero on average. The majorities of the deviations are within ±20% of the true step locations, as indicated by the red solid lines.

In summary, the GLS method can efficiently identify almost all the steps in a single-molecule trajectory, and the step locations were identified with very good precision. Recently, the exact same method has been applied to the single-molecule unzipping trajectories of the hepatitis C virus NS3 RNA helicase [38]. The advantages of this method to work with high bandwidth raw data without any pre-filtering or assumption of Gaussian noise, and its ability to detect transient steps are likely to be useful for single-molecule real time data analysis in general.

Materials and Methods

Partial Autocorrelation Function (PACF)

Compared to ACF, the PACF for a time series is the correlation between time lags after removing the effects of the intermediary points. Please see Supporting Information for further details.

Step Detection Framework using GLS

For a single-molecule real time trajectory measured from time $t = 0$ to $t = T$, we denote the times at which the k steps of the trajectory occur with t_j, $j = 1,\ldots,k$. The steps in a trace can be set up as a regression function given by

$$y_t = \beta_0 + \beta_1 I(t \geq t_1) + \ldots + \beta_k I(t \geq t_k) + \varepsilon_t \qquad (1)$$

where y_t, $t = 0,\ldots, T$ is the observed trace; $I(t \geq t_j; j = 1,\ldots,k)$ are indicator functions such that $I(t \geq t_j) = 1$ and 0 otherwise (similar to the Heaviside step function). ε_t is the underlying noise. Here β_0 is the baseline at which the trace begins at $t = 0$ and β_j, $j = 1,\ldots, k$ are the step sizes of the k steps, with a negative value indicating a downward step.

In general, ε_t is assumed to be independent and identically distributed (i.i.d) as zero-mean Gaussian noise $N(0, \sigma^2)$ with variance σ^2 [18,19,20,21,22,23,26], in which case one can obtain least squares estimates (LSE) for the parameters β_j, $j = 0,\ldots, k$. Let $\Theta = \{\beta_0,\ldots, \beta_k; \sigma^2\}$ denote the vector of parameters to be estimated. Let $f_\Theta(\bullet)$ denote the density function of the error term ε_t dependent on the parameter vector Θ. The parameter Θ, may be estimated by maximizing the likelihood function given by

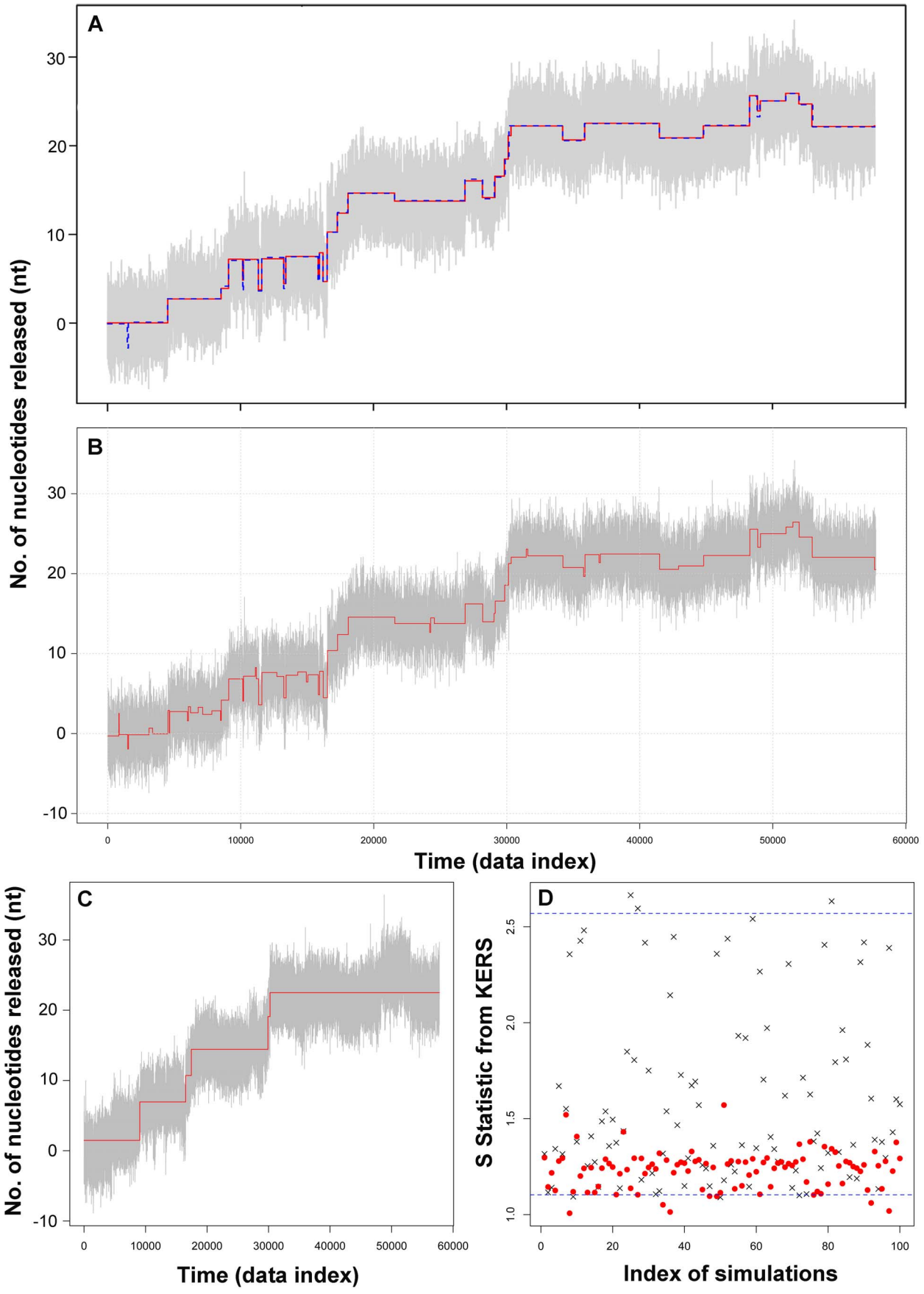
A
B
C
D
No. of nucleotides released (nt)
Time (data index)
S Statistic from KERS
Index of simulations

Figure 3. Over- and underestimates of step numbers in the test simulated traces. (A) A representative best fit by the GLS method for data that contains correlated noise; the 2.5 kHz test trace is shown in grey and the fit is shown in red. The original step function is shown in blue dashed line for comparison. (B) One of the best fits obtained from GLS method by ignoring correlation for data that contains correlated noise; the 2.5 kHz test trace is shown in grey and the fit is shown in red. (C) and (D) Fit and S-Statistic distribution from KERS method. (C) One of the best fits from KERS method; the 2.5 kHz test trace shown in grey and the fit is shown in red. (D) Distribution of S-statistic as a result of fitting using KERS method. The crosses are the S-statistic from the known step function and the red dots from the best fit for each trace.
doi:10.1371/journal.pone.0059279.g003

$$L(\Theta|y_0,...,y_T)=\prod_{t=0}^{T} f_\Theta(\varepsilon_t) \quad (2)$$

or equivalently the log-likelihood function given by

$$l(\Theta|y_0,...,y_T)=\prod_{t=0}^{T} log(f_\Theta(\varepsilon_t)) \quad (3)$$

Estimates obtained using Eq. 2 or 3 is referred to as maximum likelihood estimates (MLE). In the case of i.i.d Gaussian noise, LSE and MLE are identical. As the number of parameters in Θ increases, i.e. the number of steps increases, the likelihood L or l will increase (or equivalently in the Gaussian case, the residual sum of squares will decrease). This can create a tendency to overfit, i.e., more steps can always produce a better fit than less. To avoid over-fitting the trace we present two criteria to choose the optimal number of steps. The Akaike Information Criterion (AIC) given by

$$AIC(p)=-2l(\Theta)+2p \quad (4)$$

and Bayesian Information Criterion (BIC) given by

$$BIC(p)=-2l(\Theta)+p\log(n) \quad (5)$$

where p is the total number of parameters to be estimated in the model, n is the number of observations and l(Θ) is the log-likelihood function given in Eq. 3. As the negative log-likelihood decreases with increasing number of parameters, both AIC and BIC penalize by the number of parameters in the model. For n >7, the BIC offers a higher penalty to the model. The model with the least AIC or BIC value is chosen as the optimal model.

So far we have assumed that the underlying noise ε_t is i.i.d Gaussian. This assumption may be violated, as shown in Fig. 1 from single-molecule traces collected with high-resolution optical tweezers. In such cases, the noise is autocorrelated. A general procedure to model autocorrelated noise is to use an autoregressive (AR) noise of order p [39], given by

$$\varepsilon_t=\phi_1\varepsilon_{t-1}+...+\phi_p\varepsilon_{t-p}+e_t \;\; e_t\sim N(0,\sigma^2) \quad (6)$$

Table 2. Summary of false positive and true negative steps from 100 realizations of the simulated traces.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
False positives	0.00	1.00	1.00	1.52	2.00	6.00
True negatives	0.00	0.00	0.00	1.63	2.00	8.00

The traces contained correlated noise and the steps were identified using GLS method to take the noise structure explicitly into account.
doi:10.1371/journal.pone.0059279.t002

i.e. the noise is a function of p past values of the noise and a random error, which induces correlation in the noise. We assume that AR noise is second order stationary, i.e., the mean is constant (zero) and the correlation between any two time points is dependent only on the lag h between them and not on the absolute time.

In the presence of autocorrelated noise, least squares can be expected to give unbiased estimates of the parameters but will not be efficient, i.e., parameters will have higher variances amongst all unbiased estimators unless ε_t is uncorrelated with constant variance [40,41]. Thus the estimates are not suitable for purposes of inference. To find the optimal solution in the presence of autocorrelated noise, one resorts to GLS [42]. To realize GLS efficiently, we rewrite Eq. 6 as

$$\Phi(B)\varepsilon_t=e_t \quad (7)$$

where

$$\Phi(B)=1-\phi_1B-...-\phi_pB^p \quad (8)$$

and B is the backward shift operator such that $B^p\varepsilon_t=\varepsilon_{t-p}$. Applying the filter given by Eq. 8 to Eq. 1, we get

$$\Phi(B)y_t=\beta_0'+\beta_1\Phi(B)I(t\geq t_1)+...+\beta_k\Phi(B)I(t\geq t_k)+\Phi(B)\varepsilon_t \quad (9)$$

$$y_t^*=\beta_0'+\beta_1x_1^*+...+\beta_kx_k^*+e_t \quad (10)$$

One can see from Eq. 10 that the error term is now i.i.d Gaussian and we can estimate the parameters of the model as before using this transformed equation. This transformation procedure, referred to as the Cochrane-Orcutt scheme [43], is computationally feasible for long time series as in high bandwidth single-molecule data.

Based on the above framework, one needs to know the order of the AR noise p in order to estimate the step size. The order p is determined and the corresponding coefficients are estimated as part of the GLS procedure. First, the steps are fitted assuming i.i.d Gaussian noise. The resulting residuals are examined for any autocorrelation. If the noise is indeed i.i.d Gaussian, then no further steps are required. If the noise is autocorrelated, then the order and coefficients are determined for the noise using standard time series estimation techniques [39]. Having estimated p and the coefficients, the step sizes are re-estimated using the Cochrane-Orcutt scheme described above and the BIC value associated with the fit is computed. After each round of fitting, we used student t-test to compute the p-value for each step and removed the step with the largest p-values from each fitting process. This process is then repeated until no further steps are left, and thus generated a series of fits with different BIC values for the original trajectory. The fit with the lowest BIC among all was chosen as the final model.

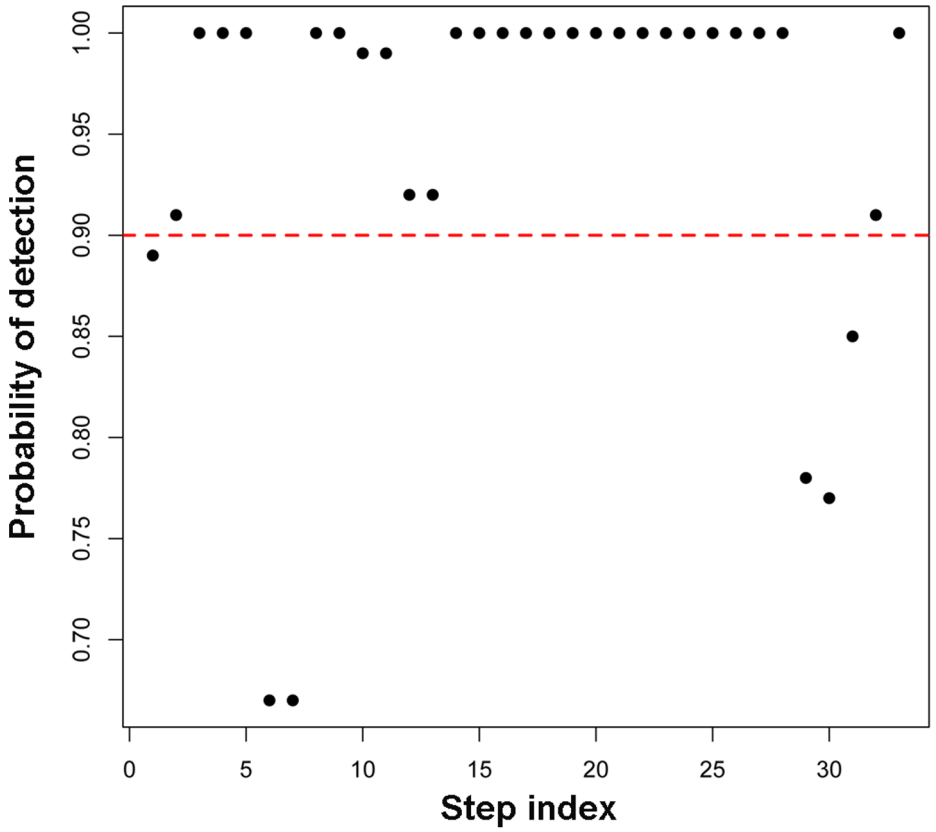

Figure 4. Efficiency of step detection using GLS method. The proportion of the traces in which a given step was detected was plotted as a function of the step index. The red dashed line indicates 90%.
doi:10.1371/journal.pone.0059279.g004

Obtaining a Superset of Plausible Step Locations

To implement the above GLS procedure, one requires a set of plausible step locations to start with, from which an optimal number of steps can be chosen to fit the experimental trajectory. To this end, we developed a statistic η to generate the superset of plausible step locations as follows.

We represent each of the data point in the trace as y_i ($i = 1,\ldots,n$). Consider a window of size 2w+1 centered at the data point, i.e., there are w data points on either side of the given point. We represent this window using the vector $y_{i,w} = \{y_{i-w},\ldots, y_i,\ldots, y_{i+w}\}$. At the ends of the series with indices less than w or greater than n-w, w is set to i-1 and n-i-1 respectively. Let $y_{i,w}^{(min)}$ and $y_{i,w}^{(max)}$ denote the minimum and maximum of the data points in $y_{i,w}$, then $R_{i,w} = y_{i,w}^{(max)} - y_{i,w}^{(min)}$ denotes the range of the points within this window. Now consider the two halves of the window, the left half $y_{i,l} = \{y_{i-w},\ldots, y_i\}$ and the right half $y_{i,r} = \{y_i,\ldots, y_{i+w}\}$. Let $q_{i,l}$ and $q_{i,r}$ denote the vectors comprising the 0.25, 0.5 and 0.75^{th} quantile of the data in $y_{i,l}$ and $y_{i,r}$ respectively. We form the statistic

$$\eta_{i,w} = \frac{(q_{i,l} - q_{i,r})\cdot(q_{i,l} - q_{i,r})^t}{3\cdot R_{i,w}^2} \qquad (11)$$

where t denotes the transpose of the row vector. $\eta_{i,w}$ is thus the mean squared difference of the quartiles on either side of the point i normalized by the square of the range $R_{i,w}$. Normalization provides an upper bound of 1 for $\eta_{i,w}$. If the distribution of the points on either side of i is identical then one would expect $\eta_{i,w}$ to be close to zero. Conversely, if the distribution on either side of i is different, $\eta_{i,w}$ is expected to be greater than zero. At step locations where the difference in the distributions on either side of the step point might be the greatest, one would expect the statistic η to increase to a local maximum right at the step location and decrease thereafter, forming local peaks around the step. The advantage of this statistic as compared to others is its sensitivity to changes in the overall shape and distribution of the data, i.e., the use of quartile that includes both the center and tail regions of the data points *instead of a single mean value used in the popular t-test.* Fig. S3A shows the value of the statistic η for all the points in the simulated trace shown in Fig. S1B using w = 500. The peaks can thus be considered as possible locations of motor steps. The choice of a window size is important. In reality, if the window size is too big, the variation in η will be smoothed out and one may miss the peaks corresponding to motor steps. To avoid this problem, we have used a set of windows of varying size, which range from 10 to 100 in steps of 10 and from 100 to 1000 in steps of 25, thus essentially make this procedure insensitive to data bandwidth and no need to filter data before analysis. Furthermore, a cutoff threshold, either 0.90^{th} or 0.95^{th} quantile of η was adopted. Only data points with η above the cutoff are considered in the superset of plausible steps. This procedure is adopted mainly to reduce computational burden. In practice, this threshold can be changed by the user. The lower the threshold, the greater the number of points chosen and thus greater computational burden. Fig. S3B shows the value of η from various windows (only a subset of the windows plotted for clarity of display) stacked on top of each other. The peaks chosen in each w using a cutoff threshold of 0.9 are highlighted by the red dots, which constitute the superset of

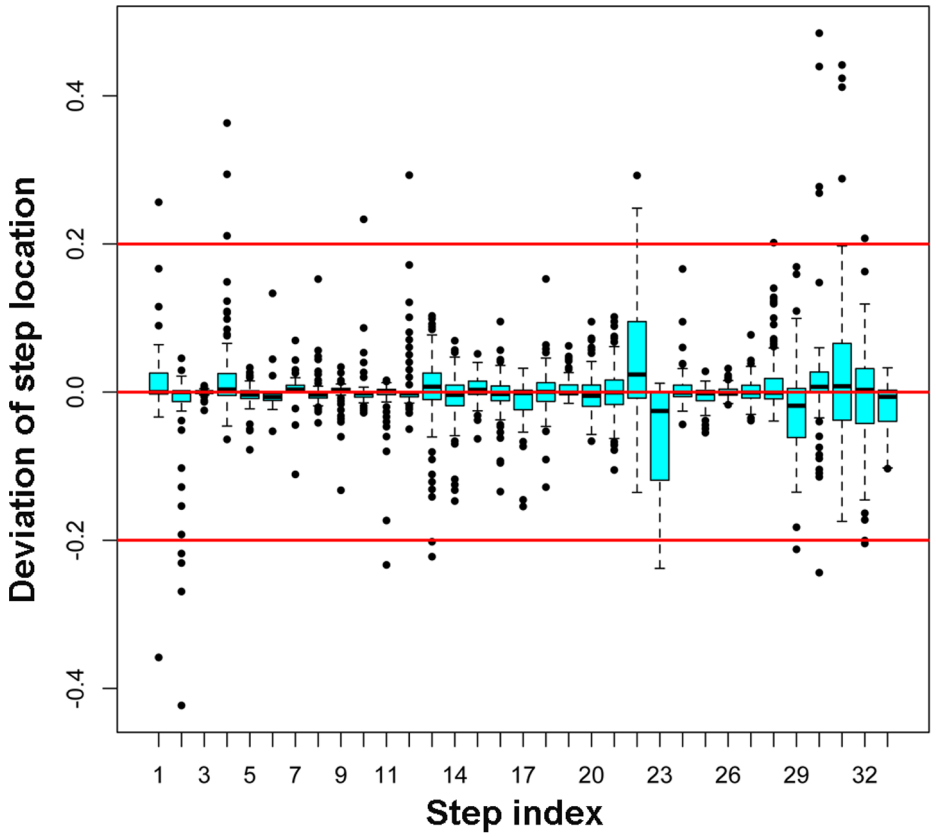

Figure 5. Precision of identified step location using GLS method. Deviation of a step from its true location as a percentage of the plateau length was plotted as a function of the step index in a box plot. The cyan boxes indicate the middle 75% of the data or the interquartile range (IQR). The extended lines or whiskers mark the 1.5×IQR distance. Any point greater or less than this value is an outlier and are shown by the dots.
doi:10.1371/journal.pone.0059279.g005

plausible change points $\mathbf{C_L}$. Because majority of the red dots in Fig. S3B identify the same point, the number of points included in the final $\mathbf{C_L}$ is much less than the total number of red dots in the figure.

The entire GLS algorithm was coded using 'R' [44], the statistical package that is freely available for download (http://www.r-project.org/). The R code together with instructions on how to run the algorithm to detect steps in real time single-molecule trajectory is freely available upon request.

Supporting Information

Figure S1 Simulated single-molecule RNA unwinding trajectory. Panel (A) shows the simulated step function, which indicates the true underlying steps; (B) shows one realization of simulated unwinding trace after addition of AR noise of order 7 on top of the step function shown in (A). The step function is shown in red, and the trajectory with noise is shown in grey.
(TIF)

Figure S2 Representative best fits of simulated RNA unwinding traces from two different procedures. Panel (A) shows one of the best fits obtained from GLS method. The trajectory was simulated from the step function shown in Fig. S1A plus Gaussian white noise. The fit is in red, and the trajectory at 2.5 kHz is in grey. Panel (B) shows one of the best fits for simulated trajectories obtained from KERS method. The trajectory was simulated from the step function shown in Fig. S1A plus correlated noise of AR(7), and further filtered and decimated to 250 Hz using a boxcar filter. The fit is in red, and the trajectory at 250 Hz is in grey.
(TIF)

Figure S3 The statistic η computed for the simulated single-molecule trace in Fig. S1B. (A) from a window size of 500 and (B) shows a stack of η calculated using a set of window size, which includes 10, 30, 40, 50, 60, 70, 80, 90, 100, 125, 200, 275, 350, 425, 500, 575, 650, 725, 800, 875, and 950.
(TIF)

Text S1 Procedures to calculate the partial autocorrelation function for a time series.
(DOCX)

Acknowledgments

We thank Dr. Marileen Dogterom (FOM Institute AMOLF) and Dr. Jeffrey R. Moffitt (Harvard University) for providing the MATLAB code for the KERS algorithm and t-test algorithm respectively. We thank Cheng Lab members, especially Michael DeSantis for a critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: WC SA. Performed the experiments: WC SA. Analyzed the data: SA WC. Contributed reagents/materials/analysis tools: WC SA. Wrote the paper: SA WC.

References

1. Lu HP, Xun L, Xie XS (1998) Single-molecule enzymatic dynamics. Science 282: 1877–1882.
2. Moerner WE, Orrit M (1999) Illuminating single molecules in condensed matter. Science 283: 1670–1676.
3. Weiss S (1999) Fluorescence spectroscopy of single biomolecules. Science 283: 1676–1683.
4. Mehta AD, Rief M, Spudich JA, Smith DA, Simmons RM (1999) Single-molecule biomechanics with optical methods. Science 283: 1689–1695.
5. Gimzewski JK, Joachim C (1999) Nanoscale science of single molecules using local probes. Science 283: 1683–1688.
6. Reisner W, Larsen NB, Silahtaroglu A, Kristensen A, Tommerup N, et al. (2010) Single-molecule denaturation mapping of DNA in nanofluidic channels. Proc Natl Acad Sci U S A 107: 13294–13299.
7. Huguet JM, Bizarro CV, Forns N, Smith SB, Bustamante C, et al. (2010) Single-molecule derivation of salt dependent base-pair free energies in DNA. Proc Natl Acad Sci U S A 107: 15431–15436.
8. Yildiz A, Forkey JN, McKinney SA, Ha T, Goldman YE, et al. (2003) Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. Science 300: 2061–2065.
9. Oster G, Wang H (2003) Rotary protein motors. Trends Cell Biol 13: 114–121.
10. Greenleaf WJ, Woodside MT, Block SM (2007) High-resolution, single-molecule measurements of biomolecular motion. Annu Rev Biophys Biomol Struct 36: 171–190.
11. Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, et al. (2008) Following translation by single ribosomes one codon at a time. Nature 452: 598–603.
12. Bustamante C, Cheng W, Mejia YX (2011) Revisiting the central dogma one molecule at a time. Cell 144: 480–497.
13. Svoboda K, Schmidt CF, Schnapp BJ, Block SM (1993) Direct observation of kinesin stepping by optical trapping interferometry. Nature 365: 721–727.
14. Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM (2005) Direct observation of base-pair stepping by RNA polymerase. Nature 438: 460–465.
15. Dumont S, Cheng W, Serebrov V, Beran RK, Tinoco I, Jr., et al. (2006) RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. Nature 439: 105–108.
16. Mallik R, Carter BC, Lex SA, King SJ, Gross SP (2004) Cytoplasmic dynein functions as a gear in response to load. Nature 427: 649–652.
17. Moffitt JR, Chemla YR, Aathavan K, Grimes S, Jardine PJ, et al. (2009) Intersubunit coordination in a homomeric ring ATPase. Nature 457: 446–450.
18. Carter NJ, Cross RA (2005) Mechanics of the kinesin step. Nature 435: 308–312.
19. Kerssemakers JW, Munteanu EL, Laan L, Noetzel TL, Janson ME, et al. (2006) Assembly dynamics of microtubules at molecular resolution. Nature 442: 709–712.
20. Milescu LS, Yildiz A, Selvin PR, Sachs F (2006) Extracting dwell time sequences from processive molecular motor data. Biophys J 91: 3135–3150.
21. Mullner FE, Syed S, Selvin PR, Sigworth FJ (2010) Improved hidden Markov models for molecular motors, part 1: basic theory. Biophys J 99: 3684–3695.
22. Syed S, Mullner FE, Selvin PR, Sigworth FJ (2010) Improved hidden Markov models for molecular motors, part 2: extensions and application to experimental data. Biophys J 99: 3696–3703.
23. Arunajadai SG (2009) A point process driven multiple change point model: a robust resistant approach. Math Biosci 220: 57–71.
24. Arunajadai SG (2009) RNA unwinding by NS3 helicase: a statistical approach. PLoS One 4: e6937.
25. McKinney SA, Joo C, Ha T (2006) Analysis of single-molecule FRET trajectories using hidden Markov modeling. Biophys J 91: 1941–1951.
26. Carter BC, Vershinin M, Gross SP (2008) A comparison of step-detection methods: how well can you do? Biophys J 94: 306–319.
27. Reck-Peterson SL, Yildiz A, Carter AP, Gennerich A, Zhang N, et al. (2006) Single-molecule analysis of dynein processivity and stepping behavior. Cell 126: 335–348.
28. Myong S, Bruno MM, Pyle AM, Ha T (2007) Spring-loaded mechanism of DNA unwinding by hepatitis C virus NS3 helicase. Science 317: 513–516.
29. Lee G, Bratkowski MA, Ding F, Ke A, Ha T (2012) Elastic coupling between RNA degradation and unwinding by an exoribonuclease. Science 336: 1726–1729.
30. Neuman KC, Saleh OA, Lionnet T, Lia G, Allemand JF, et al. (2005) Statistical determination of the step size of molecular motors. J Phys Condens Matter 17: S3811–3820.
31. Dekker NH, Rybenkov VV, Duguet M, Crisona NJ, Cozzarelli NR, et al. (2002) The mechanism of type IA topoisomerases. Proc Natl Acad Sci U S A 99: 12126–12131.
32. Dessinges MN, Lionnet T, Xi XG, Bensimon D, Croquette V (2004) Single-molecule assay reveals strand switching and enhanced processivity of UvrD. Proc Natl Acad Sci U S A 101: 6439–6444.
33. Saleh OA, Perals C, Barre FX, Allemand JF (2004) Fast, DNA-sequence independent translocation by FtsK in a single-molecule experiment. EMBO J 23: 2430–2439.
34. Ali JA, Lohman TM (1997) Kinetic measurement of the step size of DNA unwinding by Escherichia coli UvrD helicase. Science 275: 377–380.
35. De Los Rios P, Zhang YC (1999) Universal 1/f noise from dissipative self-organized criticality models. Physical Review Letters 82: 472–475.
36. Bustamante C, Chemla YR, Moffitt JR (2009) High-resolution dual-trap optical tweezers with differential detection: an introduction. Cold Spring Harb Protoc 2009: pdb top60.
37. Cheng W, Hou X, Ye F (2010) Use of tapered amplifier diode laser for biological-friendly high-resolution optical trapping. Opt Lett 35: 2988–2990.
38. Cheng W, Arunajadai SG, Moffitt JR, Tinoco I, Jr., Bustamante C (2011) Single-base pair unwinding and asynchronous RNA release by the hepatitis C virus NS3 helicase. Science 333: 1746–1749.
39. Box GEP, Jenkins GM (1994) Time series analysis: forecasting and control. Prentice Hall PTR.
40. Bloomfield P, Watson GS (1975) Inefficiency of Least-Squares. Biometrika 62: 121–128.
41. Lee J, Lund R (2004) Revisiting simple linear regression with autocorrelated errors. Biometrika 91: 240–245.
42. Pinheiro JC, Bates DM (2009) Mixed-eects models in S and S-PLUS. Springer Verlag Chapter 5.
43. Cochrane D, Orcutt GH (1949) Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. Journal of the American Statistical Association 44: 32–61.
44. Team RDC (2012) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.