Steps and Bumps: Precision Extraction of Discrete States of Molecular Machines

Max A. Little,[†‡§] Bradley C. Steel,[‡] Fan Bai,[¶] Yoshiyuki Sowa,[‖] Thomas Bilyard,[‖**] David M. Mueller,[††] Richard M. Berry,[‡] and Nick S. Jones[‡§‡‡*]

[†]Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts; [‡]Department of Physics and [§]Department of Biochemistry, Oxford Centre for Integrative Systems Biology, Oxford, United Kingdom; [¶]Graduate School of Frontier Biosciences, Osaka University, Osaka, Japan; [‖]Department of Frontier Bioscience, Hosei University, Tokyo, Japan; [**]Lawrence Berkeley National Laboratory, Berkeley, California; [††]Department of Biochemistry and Molecular Biology, Rosalind Franklin University of Medicine and Science, The Chicago Medical School, North Chicago, Illinois; and [‡‡]Department of Mathematics, Imperial College London, London, United Kingdom

ABSTRACT We report statistical time-series analysis tools providing improvements in the rapid, precision extraction of discrete state dynamics from time traces of experimental observations of molecular machines. By building physical knowledge and statistical innovations into analysis tools, we provide techniques for estimating discrete state transitions buried in highly correlated molecular noise. We demonstrate the effectiveness of our approach on simulated and real examples of steplike rotation of the bacterial flagellar motor and the F1-ATPase enzyme. We show that our method can clearly identify molecular steps, periodicities and cascaded processes that are too weak for existing algorithms to detect, and can do so much faster than existing algorithms. Our techniques represent a step in the direction toward automated analysis of high-sample-rate, molecular-machine dynamics. Modular, open-source software that implements these techniques is provided.

INTRODUCTION

Nature has evolved many molecular machines such as pumps, copiers, and motors. Biophysical theory proposes that these machines, converting electrochemical energy to linear or rotary motion, do so in a series of thermally driven steplike motions because this maximizes use of available free energy (1). Even in genetically identical cells, each cell shows fundamental variability partly traced to thermal randomness in discrete molecular mechanochemistry (2). Machines such as kinesin moving on microtubules (3), myosin sliding between actin filaments (4), and rotations of protein complexes in the flagellar motor (5), all show hallmarks of discreteness with superimposed thermal fluctuations.

Motion is often highly repetitive and quasiperiodic—composed of several different, superimposed periodicities—as these machines are built of many identical copies of sets of molecular components coupled in interlocking linear, circular, or helical patterns. Of interest are temporal sequences of discrete states observed using advanced experimental techniques such as Förster resonance energy transfer, back focal-plane interferometry, or atomic force microscopy (6). Due to discreteness of steplike transitions, the distribution of states (this is just the distribution of each time point in series, treating the time series as a stationary independent random process) is often multimodal (i.e., bumplike). A common view of molecular machines is that they execute randomly forced motion in a potential energy well around each discrete state (described as generalized Langevin dynamics), leading to temporally correlated noise (1).

The dominant approach to extracting discrete states of molecular machines is to first smooth away (i.e., filter) as much of the random motion as possible, leaving the underlying states (under the Langevin model, these states actually correspond to the minima of each potential well) (4,5,7,8). Classical running-mean filtering is fast and simple, but fundamentally inadequate because: 1), it must smooth away jumps in the data to remove noise; 2), it is the maximum likelihood filter for each window if the noise is Gaussian and uncorrelated (9), but the observed transitions between states is often smooth rather than steplike, due to elastic coupling or friction effects (10), and therefore the observed signal has correlated noise; and 3), it operates on a fixed-length sliding time-window of the series that favors certain dwell times. Adaptations have been proposed (11–13), but none addresses all the above issues simultaneously. Sophisticated alternatives (e.g., Markov chain Monte Carlo, particle filtering, or variational Bayesian techniques) might tackle several of these problems. However, such algorithms are too computationally demanding to rapidly process large numbers of time series, and both in principle and in practice, become intractable as the data size increases (14). After filtering, finding the arrangement of discrete states is a bump-hunting problem—that of finding peaks in a distribution. Histograms of smoothed time-series are easily constructed and popular (5,8), but choosing bin edges and widths is an open-ended problem not solvable without making extra assumptions that may not be appropriate. Furthermore, the resulting histogram distribution estimate is discontinuous (15)—an unrealistic

Submitted January 22, 2011, and accepted for publication May 24, 2011.

*Correspondence: nick.jones@physics.ox.ac.uk

Editor: Hideo Higuchi.

0006-3495/11/07/0477/9 $2.00

doi: 10.1016/j.bpj.2011.05.070

representation of the underlying potential well(s) (1). Many more sophisticated algorithms exist (e.g., mixture modeling (16)), but for the multiple states in typical molecular machines, these algorithms can be computationally demanding. Kernel methods are continuous and more tractable, but, like histogram parameters, selecting kernel bandwidth and shape is an open problem.

In this article, we propose algorithms handling correlated noise and smooth transitions between states by incorporating a simple physical model directly into the filter structure, in such a way that does not require the stipulation of an artificial window size, and that guarantees discovery of the solution without unnecessary computation. We then use contemporary statistical methods permitting us to find quasiperiodic arrangements of bumps in the distribution of molecular states by simple Fourier coefficient selection, circumventing the need to estimate the distribution directly. Finally, we estimate the discrete states by classification of the smoothed time trace to the nearest peak of each bump in the distribution. There are two critical free parameters in our approach and we are able to provide theoretical guidance for how to set them. Fig. 1 gives an overview of the entire process.

Note that, if we know the state transition sizes between discrete states, all our techniques are applicable to linear as well as rotary machines. To ensure that sufficient samples of the dwell states are represented, it is, however, sensible to restrict linear motion to a bounded interval. This is a very common physical situation—many linear molecular machines undergo repetitive sequences of state transitions. If they do not, then the motion can be wrapped onto a bounded interval whose length is a multiple of any fundamental state transition size before application of these techniques, and the process is fundamentally the same. Similarly, these techniques work equally well for back- as for forward-stepping. Finally, if we do not know the state transition sizes, then the step-smoothing methods presented here, which we believe to be novel, are still useful.

We demonstrate by simulation that our algorithms are an improvement in accuracy, precision, and speed over other algorithms in the literature. Moving to real data, we then process a large number of bacterial flagellar motor experimental angle-time traces, unambiguously identifying periodicities in the discrete state locations. We also find a very large number of dwell times to provide clear evidence for non-Poisson stepping in the flagellar motor, and resolve

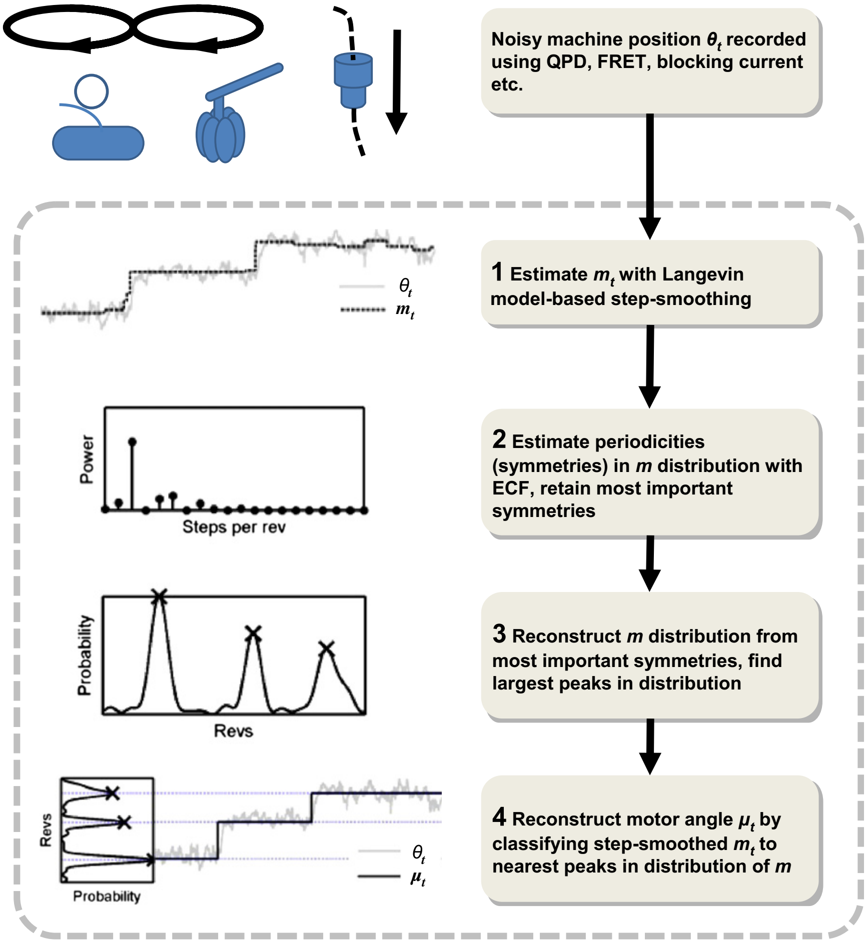
Noisy machine position θ_t recorded using QPD, FRET, blocking current etc.
θ_t
m_t
1 Estimate m_t with Langevin model-based step-smoothing
Power
Steps per rev
2 Estimate periodicities (symmetries) in m distribution with ECF, retain most important symmetries
Probability
Revs
3 Reconstruct m distribution from most important symmetries, find largest peaks in distribution
Revs
Probability
θ_t
μ_t
4 Reconstruct motor angle μ_t by classifying step-smoothed m_t to nearest peaks in distribution of m

FIGURE 1 Overall step finding and bump hunting process. Experimental recording of noisy molecular machine position-time traces θ_t are obtained, by, for example, bacterial flagellar motor angle captured by quadrant photodiode, a charge-coupled device video of flagellar-attached bead, an F_1-ATPase angle using laser dark-field or Forster resonance energy transfer, or a blocking current of DNA translocation through a hemolysin nanopore. (*Dotted box*) Analysis process applied to an example of F_1-ATPase (*1*). Step-driven Langevin model fitted to series (algorithm L1-PWC-AR1, see text), and estimated time traces of machine positions m_t quickly obtained (*2*). Periodicities (symmetries) in distribution of m estimated (using ECF-Bump algorithm, see text), and (*3*) most important symmetries retained and used to reconstruct distribution of m (*4*). Classifying estimated m_t trace to nearest large distribution peaks (algorithm ML-peaks, see text), an estimate of the true machine time-position trace μ_t is recovered, from which secondary properties such as dwell times and dwell-time distributions can be quantified.

rate-limiting reactions in F_1-ATPase stepping. We detect that these reactions are cascaded by revealing unambiguous substeps.

We have released open source code freely available for download, implementing the described algorithms.

MATERIALS AND METHODS

The discussions in the Introduction highlight some of the challenges for current step-smoothing and bump-hunting approaches. Our main observation is that these techniques have been developed to solve different problems than the one with which we are actually concerned. In detecting molecular dynamics, we have a lot of additional information that can serve as constraints on the algorithms to produce results that are more efficient. In particular, the time series can be highly correlated. Thus, general bump-hunting, which throws away time to estimate the distribution of states, will not be as effective as a more specific method that takes time into account. The steps may also have some known geometric relationship to each other. For instance, the steps may have several dominant periodicities that are naturally represented on a Fourier basis.

A plausible model for the typical experimental setup involving damping and potential energy storage is Brownian motion in a potential well $\mathrm{d}\theta = \rho(\mu - \theta)\mathrm{d}t + \sigma \mathrm{d}W$, where μ is the coordinate representing the current center of the well (this might undergo discrete steps), θ is the experimentally observed position, W is a Wiener process, σ is the diffusion coefficient, and ρ is the drift coefficient. This model can be simulated with the discrete-time equivalent $\theta_{t+1} = a\theta_t + (1-a)\mu_t + \varepsilon_t$, for the constant a which can be found from the autocorrelation at one sample lag of the time-angle trace (see Supporting Material). The goal is to find the best approximation m_t of the center of the well where the machine is located, at time t: μ_t (μ_t is thus a piecewise constant signal) observing only θ_t. This problem can be solved by finding the μ_t for the full time range t that minimizes, for example, the sum of square errors $\sum_t \varepsilon_t^2$. In general, this global optimization problem cannot be solved without imposing some additional conditions on μ_t. Therefore, step-smoothness constraints have been introduced, requiring that μ_t is constant in time except at the points where the motor changes state (17). (Note that it is not necessary to consider μ_t as a stochastic process here, although it can be treated as such when finding dwell-time distributions.)

On the surface, this appears to be an intractable optimization problem because there is an unknown number of step-time instants that can occur anywhere in the full range of times t, so that finding these instants requires brute-force testing of every possible combination of time instants for the presence of steps. However, use can be made of recent theoretical innovations. This applies to the particular step-smoothness constraint penalizing the sum of absolute differences between successive instants of μ_t. This theoretical result shows that if there exist only a finite number of steps, solving the resulting optimization problem nearly always (with very high probability) finds the correct positions of all the steps (18). The resulting optimization problem is convex (in this case, the sum of the model fit error with the step constraint has only one minimum with respect to variation in the unknown m'_t):

$$\begin{aligned} m_t = \underset{m'_t}{\arg\min} & \sum_{t=P+1}^{T} \left(\theta_t - \sum_{i=1}^{P} a_i \theta_{t-i} - m'_t \right)^2 \\ & + \gamma \sum_{t=2}^{T} |m'_t - m'_{t-1}|. \end{aligned} \qquad (1)$$

(We call this algorithm L1-PWC-ARP, and with all a_i zero, we call it algorithm L1-PWC; see Supporting Material and Kim et al. (17) for derivations of similar approaches.) Here, T is the length of the time trace, and the a_i values are determined from a biophysical model when this is known, or from analysis of the time trace otherwise. In particular, when $P = 1$, a_1 can be chosen as the first autocorrelation coefficient of the trace. In this article, we demonstrate the case $P = 1$ arising from the integration of a first-order, continuous-time stochastic model with backward Euler integrator, but in general P can be >1 (see Supporting Material for more details). As the constraint constant γ increases, larger weight is placed on the step-smoothing, so that the resulting m_t is increasingly smooth, at the expense of increasing the sum of square errors.

There is a maximum useful value of the parameter γ, which can be calculated based on the length of the data; above this value, m_t is a constant (see Supporting Material for more details). Also, for a step of height h and width w, setting $\gamma > hw/2$ flattens this step. This fact can also be used to argue that if the noise about each known dwell in the trace is Gaussian with standard deviation σ, then the minimum useful value of γ, should be at least 2σ. This follows from the above, when we consider the random fluctuations due to the noise as unit-width steps. Putting these together, a practical bound for γ is $2\sigma < \gamma < hw/2$ (see Supporting Material for details of these calculations, and the resulting values of this parameter used in this study). The quadratic programming problem defined by Eq. 1 has the desirable property that a guaranteed optimal solution can be obtained using standard optimization techniques with computing time and resources that increase very slowly with increasing data size (19).

The variable m'_t is a dummy signal that is varied in minimizing Eq. 1; the optimization algorithm that minimizes Eq. 1 constructs this signal. The principle limitation of this algorithm is that the above bounds on γ could become too restrictive if either the noise spread increases, or the product of step height and width are too small. In addition, we need to take into consideration the effect of uncertainty in the model parameters a_i. For example, for the L1-PWC-AR1 algorithm, a few percent uncertainty in a_1 leads to a small shift in the timing of the jumps detected by the algorithm, and this is tolerable. However, as the uncertainty in a_1 gets larger, small, spurious jumps begin to be introduced near the edges due to residual correlation in the random motion.

As an example of choosing a_1, we explore simulations in the Results and Discussion of the rotary bacterial flagellar motor with typical experimental parameters based on calculated bead load particle of diameter 0.15 μm, $\xi = 0.01\ k_\mathrm{B}T$ s, and measured bacterial flagellar properties (20) $\kappa = 100\ k_\mathrm{B}T$/rad. This gives a smooth step time constant of $\tau = 10^{-4}$ s. At 5τ after an instantaneous transition in the equilibrium angle, the bead will have settled to within 1% of the steady-state μ. With a sampling rate of $\Delta t = 1/104{,}448$ s (chosen to match one of our high-resolution experimental recording setups, although values that match any particular experiment can be used here), the discrete-time first-order AR1 coefficient is $a_1 = 1–0.096 = 0.904$, and the standard deviation of the noise term ε_t is $\sqrt{2\Delta t/0.01} = 0.044$.

Given the smoothed time trace, m_t, we need to identify the most likely locations of the discrete molecular states and their periodicities and this requires an estimate of the distribution of states. Our alternative to peak-finding with histograms or kernel density approaches is to estimate the distribution directly in the Fourier domain, using the Fourier transform of the probability density function $p(m)$, estimated from the finite number of samples m_t:

$$\begin{aligned} P(f_j) &= \int_{-\infty}^{\infty} \exp(if_j m)\ p(m)\mathrm{d}\,m \\ &\approx \int_{-\infty}^{\infty} \exp(if_j m) \frac{1}{T} \sum_{t=1}^{T} \delta(m - m_t)\mathrm{d}\,m \\ &= \frac{1}{T} \sum_{t=1}^{T} \exp(if_j m_t). \end{aligned} \qquad (2)$$

Here, δ is the Dirac delta function. This is also known as the empirical characteristic function (ECF). This is calculated for the K periodicities of interest $f_1, f_2 \ldots f_K$. In this domain, periodicity is naturally represented by only a few dominant nonzero coefficients in the power spectrum (the ECF is a sparse representation for periodic distributions (18)) and we avoid problematic histogram (or other (15)) distribution estimates altogether. Because we only have a finite amount of data and there are experimental confounds, the power spectrum will be the sum of sampling effects and experimental noise and the actual periodicities of the molecular machine. Then the problem is to detect which periodicities are genuine, and which are due to sampling noise and other artifacts. The most important limitation of this approach is the requirement to have sufficient number of dwells at different locations to detect the periodicity.

In this case, the statistical theory of nonlinear threshold estimation provides us with the guarantee that (assuming physically reasonable smoothnesslike constraints on the distribution), on nearly all occasions (with very high probability), this detection problem is solved by simply setting the coefficients for all periodicities with power below a certain threshold to zero, or, which is equivalent, retaining only a fraction φ of the largest power periodicities (18) (see Supporting Material for more in-depth discussion of these claims). Setting coefficients above an upper frequency limit to zero is similar to kernel density estimation of the distribution that smooths away finite sampling effects (15) (see the algorithm ECF-Bump in the Supporting Material,). Thus, we can accomplish both periodicity detection and finite sample effect-smoothing by Fourier coefficient selection. Here, we choose the smoothing frequency parameter to be sufficiently large that we can always capture periodicities of interest. The nonlinear threshold fraction φ is set according to minimax optimality principles.

By choosing $\lambda = \sqrt{2\log K}\varsigma$ where ς is an estimate of the spread of the absolute value of the (nonzero frequency) components (21), the fraction of components retained after thresholding with $|P(f_j)| > \lambda$ varies over the range $\varphi = 0.1–0.2$. The smaller fraction value occurs when the distribution of $|P(f_j)|$ is approximately Gaussian so that choosing the standard deviation for ς is appropriate (in our experiments, the F$_1$-ATPase frequency components are approximately Gaussian in this way). Similarly, the larger fraction occurs when the distribution of $|P(f_j)|$ has prominent outliers, so that a more statistically robust estimate of the spread is appropriate (here this is the case for the bacterial flagellar motor time traces; in this case, we use the robust prescription $\varsigma = 1.482\text{MAD}(|P(f_j)|)$, where *MAD* is the median of absolute deviations from the median (21)).

The noise-reduced distribution of states $p(m)$ is then approximated by applying the inverse Fourier transform to the ECF coefficients:

$$p(m) \approx \sum_{j=-K}^{K} \exp\left(-if_j m\right)\overline{P}\left(f_j\right). \tag{3}$$

Having obtained the distribution, the peaks represent the discrete states, from which, for example, an estimate of the steplike time traces of states μ_t can be recovered. Knowing the dominant periodicity, N, of the distribution, we can expect N peaks. At each time step t, we wish to determine the state of the machine μ_t. This is a statistical classification problem: we wish to find the optimum peak to assign to each time step. A statistically consistent solution to this problem is the maximum likelihood approach obtained by assigning the step-smoothed time series m_t to the nearest peak. This is the solution to the optimization problem,

$$\widehat{\mu}_t = \underset{\mu_n : n=1,2\ldots N}{\arg\min} \left|m_t - \mu_n\right|, \tag{4}$$

where μ_n values are the locations of the N largest peaks in $p(m)$, and $\widehat{\mu}_t$ is our estimate of the state of the machine at each time instant (note that Eq. 4 arises when we assume that m_t is Laplace-distributed, which it will approximately be as a result of minimization of Eq. 1). Finally, the time traces of states can be used to estimate the time spent in each state (the dwell times), and models for the distribution of these dwell times can be found and compared.

Bacterial flagellar motor time-angle traces were obtained by video microscopy of 200-nm fluorescent beads attached to the truncated flagellar filaments of surface-immobilized *Escherichia coli* chimaeras. Please see Sowa et al. (5) for further details. ATPase time-angle traces were obtained by dark-field microscopy of a 60-nm gold bead attached to the γ-subunit of His$_6$-tagged *Saccharomyces cerevisiae* F$_1$-ATPase via a streptavidin-biotin linker. The molecule was immobilized onto a Ni^{2+}NTA surface, at an ATP concentration of 30 μM. Images were captured with a high-speed video camera (model No. PCI1024; Photron USA, San Diego, CA) at a frame and shutter rate of 30 kHz, and the bead position was calculated using the Gaussian mask algorithm described in Thompson et al. (22).

RESULTS AND DISCUSSION

Simulated data performance comparisons

Fig. 2 details an example of the step-smoothing and bump-hunting methods. Validation of our techniques is carried out on simulated bacterial flagellar motor rotation (i.e., Langevin dynamics; see Supporting Material) with known discrete states and state dynamics, over nine test cases that explore the variability of real biological recordings (see Fig. 3 and legend for description of the test case parameters). We compare the step-smoothers in this article (for algorithms L1-PWC and L1-PWC-AR1, see Supporting Material) against the classical median filter (9), the Chung-Kennedy filter (11), and the Kalafut-Visscher step-finder (23) in terms of both the absolute state location recovery error (mean absolute error, MAE) and relative absolute roughness (RAR) of the estimated time series (see Fig. 3). The parameters for these step-smoothers are optimized for the best performance on these simulations (see Supporting Material for further details). The average execution time for the algorithms was, in order of decreasing speed: median filter, 5 s; L1-PWC and L1-PWC-AR1, 7 s; Chung-Kennedy filter, 41 s; and Kalafut-Visscher algorithm, 1020 s (17 min).

The step-smoothers proposed clearly outperform existing methods. By design, median and Chung-Kennedy filters cannot guarantee that the recovered time trace of states will be constant when the machine is actually stationary, so both the recovery error and relative absolute roughness are worse than the L1-PWC filters. Similarly, the Kalafut-Visscher step-finder is confounded by the correlation in the noise, and so finds excessive detail, thus, the relative absolute roughness and the recovery error are large. We note that simulations allowing motor back-stepping lead to similar results.

We tested how well the thresholded empirical characteristic function (algorithm ECF-Bump, see Supporting Material) can recover the known, dominant periodicity of a simulated bacterial flagellar motor, when compared to existing techniques (see Table S1 and Table S2 in the Supporting Material). This proposed algorithm outperforms the alternative methods that are based upon first estimating the distribution of discrete states. This comparison highlights some of the shortcomings of other plausible bump-hunting techniques. For histogram-based methods (5), the histogram bin width sets a fundamental limit on the maximum periodicity that can be identified: increasing the bin-width decreases this frequency. However, reducing

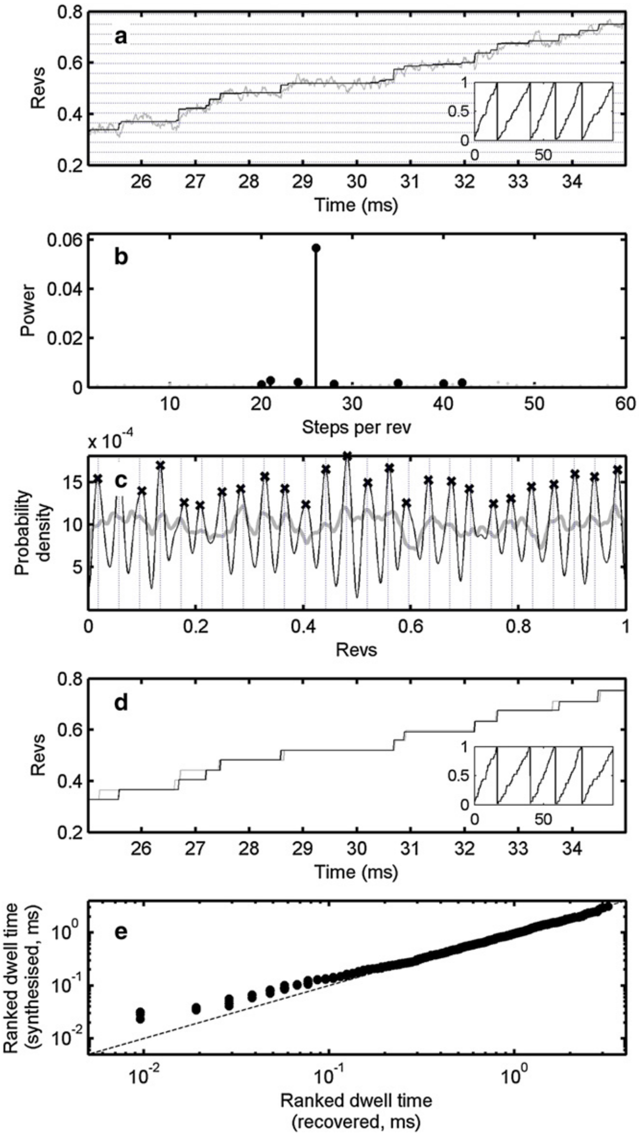

FIGURE 2 Illustration of state location extraction process for synthetic angle-time trace with 26 periodic state locations. (*a*) A small segment of a noisy time trace showing rotation angles θ_t (*light-shaded*), step-driven Langevin model estimates of motor positions m_t (*dark-shaded*), and state locations (*horizontal dotted lines*). (*Inset*) Longer segment of m_t. (*b*) Estimated periodicities (steps per revolution) in distribution of m (*light-shaded*), largest amplitude periodicities kept (*dark-shaded*) and used to reconstruct distribution of m (note that the label "power" is chosen in analogy to power spectrum for clarity of presentation, but formally this is not a power spectrum). (*c*) Reconstructed distribution of m with all periodicities (*light shaded*), and with only the largest magnitude periodicities retained (*dark shaded*). (*d*) The estimated true motor angle time trace μ_t (*dark-shaded*) is obtained by classifying the estimated m_t trace to the largest distribution peaks, the dark crosses in panel *c*. (*Inset*) Longer segment of μ_t. (*e*) Secondary properties such as dwell-time distributions can be reliably quantified, here demonstrating a good fit to the true gamma-dwell-time distribution model (as the points lie close to the *dashed line*). The horizontal axis shows the ranked logarithm of the estimated dwell times, and the vertical axis shows the ranked logarithm of synthesized dwell times drawn from a gamma distribution with the parameters used in the simulation.

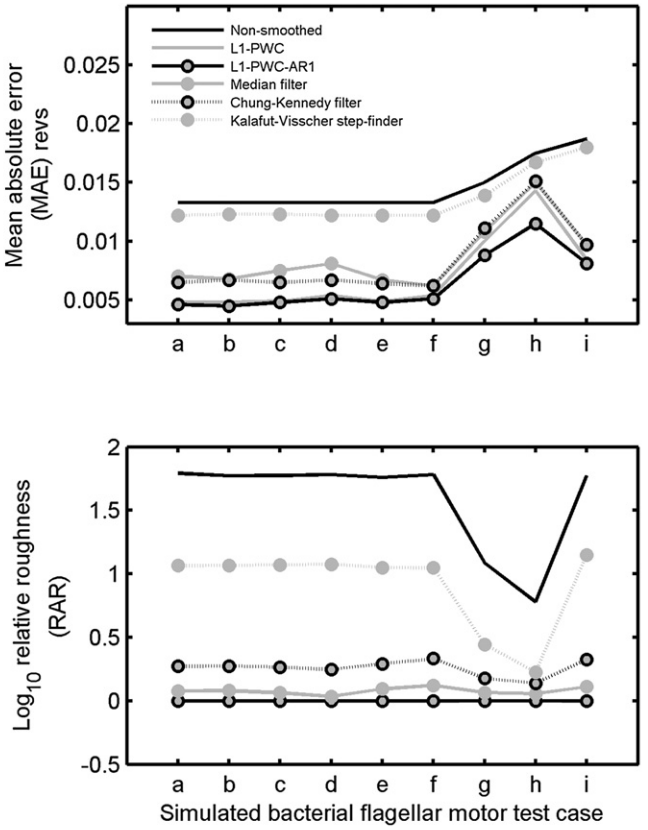

FIGURE 3 (*Top panel*) Mean absolute error, MAE (in revolutions, smaller is better) and (*bottom panel*) logarithm of relative absolute roughness, RAR (closer to zero is better), for five step-smoothing methods and nine test cases of simulated bacterial flagellar motor rotation (see Supporting Material). RAR is the mean absolute difference between adjacent samples in the filter output m_t divided by that in the true (simulated) motor position μ_t. Values are averaged over five replications. Case *a*, default test case, has 26 equally spaced state locations, exponentially distributed dwell times, rotation 10 revolutions/s, a flagellar hook spring stiffness $\kappa = 100\ k_BT$/rad, and coefficient of friction $\xi = 0.01\ k_BT$ values. The other test cases differ from default case in one parameter. Case *b* has 20% dwell state aperiodicity, (*c*) gamma distribution dwell times with shape $k = 2$, (*d*) gamma-dwell-times $k = 10$, (*e*) 30 dwell states, (*f*) 40 dwell states, (*g*) 50 revs/s, (*h*) 100 revs/s, and (*i*) flagellar hook stiffness $\kappa = 50\ k_BT$/rad. Step-smoothing algorithm parameters (see Supporting Material) are L1-PWC, $\gamma = 50$; L1-PWC-AR1, $\gamma = 1$, $P = 1$, $a_1 = 1 - \kappa\Delta t/\xi$, and $\varphi = 0.2$; median filter window size is average dwell time in samples. Chung-Kennedy filter: filter length is half the dwell time in samples, analysis window $M = 16$ samples, weighting parameter $p = 0$. Note that for many test cases, the L1-PWC methods have indistinguishable RAR from the other step-smoothing methods, so that their curves lie almost on top of one another.

the bin sizes increases the error of the bin counts due to finite sample size effects.

This bin-count error is particularly problematic for high numbers of discrete states or significant aperiodicities in state locations where this method returns a wide spread of values. Similarly, for peak-finding in kernel density estimates, the choice of kernel width is a limiting factor: too large, and a small bump in the distribution will be merged into nearby bumps; too small, and spurious bumps will appear. As with the histograms, there is no way to retain small bumps representing real discrete states, and at the same time remove spurious bumps that are due to experimental confounds or finite sample size effects. This is because these methods lack global information about the periodicities in the distribution, making them uncompetitive with the ECF-based algorithm. These sorts of issues with existing techniques confound straightforward bump-hunting in the distribution.

Experimental data

Fig. 4 illustrates the process of applying the step-smoothing and bump-hunting techniques to a 4.2 s long, single experimental time-angle trace of a rotating *E. coli* flagellar motor with attached 200-nm bead; the same data were originally published as Fig. 1 in Sowa et al. (5). We were able to process the entire trace consisting of 10,000 samples at 2.4 kHz sampling rate, in <0.5 s on a standard PC (note that this is faster than real-time: in principle, the analysis could be performed while the experiment is running). Given the flagellar hook stiffness (20) and bead size (200 nm), this sample rate is too slow; this means that all stepping appears instantaneous. Insignificant autocorrelation at positive time-lags confirms this, implying that state transitions are effectively instantaneous and this suggests the L1-PWC algorithm, which does not consider correlated noise (see Supporting Material). The results of the analysis confirm previous findings of 26 discrete states (5). This single trace also shows some evidence for 2-, 11-, and 17-fold periodicity.

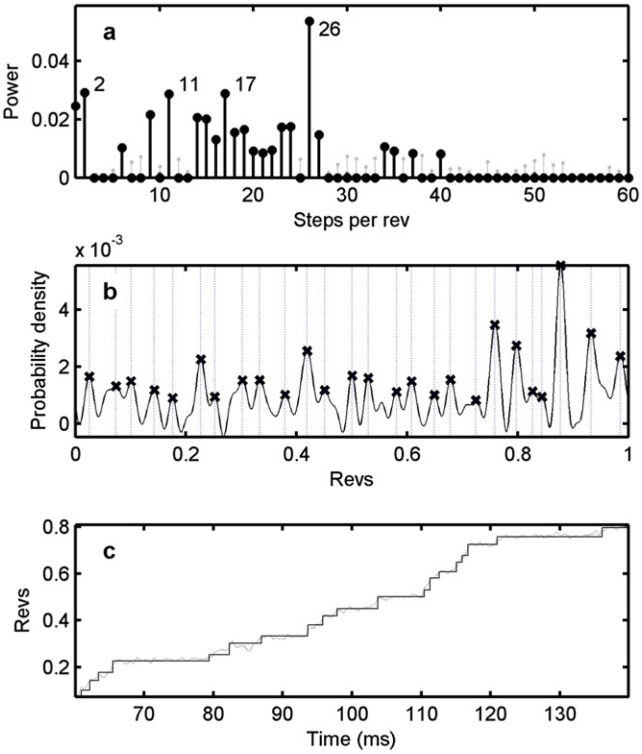

FIGURE 4 Example of processing a single experimental *E. coli* flagellar motor time-angle trace (4.2 s at 2.4 kHz sampling rate) to extract discrete state locations. (*a*) Estimated periodicities in the distribution of the full, smoothed, 4.2 s time series m_t (*light-shaded*), largest magnitude periodicities retained (*dark-shaded*), showing the dominant 26-fold periodicity of this molecular machine. (*b*) The retained periodicities are used to reconstruct the distribution of m (*dark-shaded line*). The 26 largest peaks in this distribution (*dark-shaded crosses*) represent the best estimate of the discrete state locations of the motor, and can be used to estimate the true time series of discrete state transitions. (*c*) A small segment of the recorded time trace showing rotation angles θ_t (*light-shaded*) with classified estimates of motor positions μ_t after L1-PWC smoothing (*dark-shaded*, see text for algorithm descriptions). Algorithm parameters are $\gamma = 1$ (estimated from the data, see Supporting Material for discussion), and $\varphi = 0.2$.

Our methods are robust and fast enough that they can extract discrete states from every time-angle trace in the database, without the need for prior hand-editing. We applied the same process as in Fig. 4 separately to all six traces obtained from the same motor. Fig. 5 *a* shows periodicity analysis averaged over these six traces. This provides clear evidence that the 26- and 11-fold periodicities are properties of the motor, whereas the other periodicities in Fig. 4 *b* are most likely artifacts due to finite sample size effects in this one trace. The increased precision of these techniques has therefore allowed us to show that the weak evidence for 11-fold periodicity in the original study (Fig. 4*b* in Sowa et al. (5)) is, most likely, a real feature of the motor.

Our methods allow us to capture ~6000 dwell times over these six traces, making it possible to characterize, with high statistical power, the distribution of dwell times. In Fig. 5, *b–e*, we fit four different distributions to the dwell times, including an exponential model. There is sufficient data to resolve the extremes of the distribution, clearly indicating that the simple exponential is not a good fit and that the extremes of the distribution have much higher probability than either exponential or gamma distributions would imply. Our model fitting thus reveals new, to our knowledge, features of bacterial flagellar motor stepping.

Fig. 6 shows the process applied to an experimental time-angle trace of a single rotating, surface-immobilized yeast F_1-ATPase molecule with a 60-nm gold bead attached at 30 μM [ATP] (approximately equal to the Michaelis constant ~K_m). The 0.27 s trace, recorded at 30 kHz, has significant autocorrelation and is able to resolve smooth transitions between states. Therefore, we used the L1-PWC-AR1 algorithm with parameter a_1 set equal to the autocorrelation at a time lag of one sample (i.e., 0.8; see Supporting Material for more details). This implies a relaxation to within a fraction e^{-1} of the stationary dwell state of ~0.15 ms. This algorithm was effective at resolving the instantaneous steps buried in the smooth transitions with correlated noise. The pattern of dominant periodicities (Fig. 5, *a* and *b*) is consistent with the existence of six-step

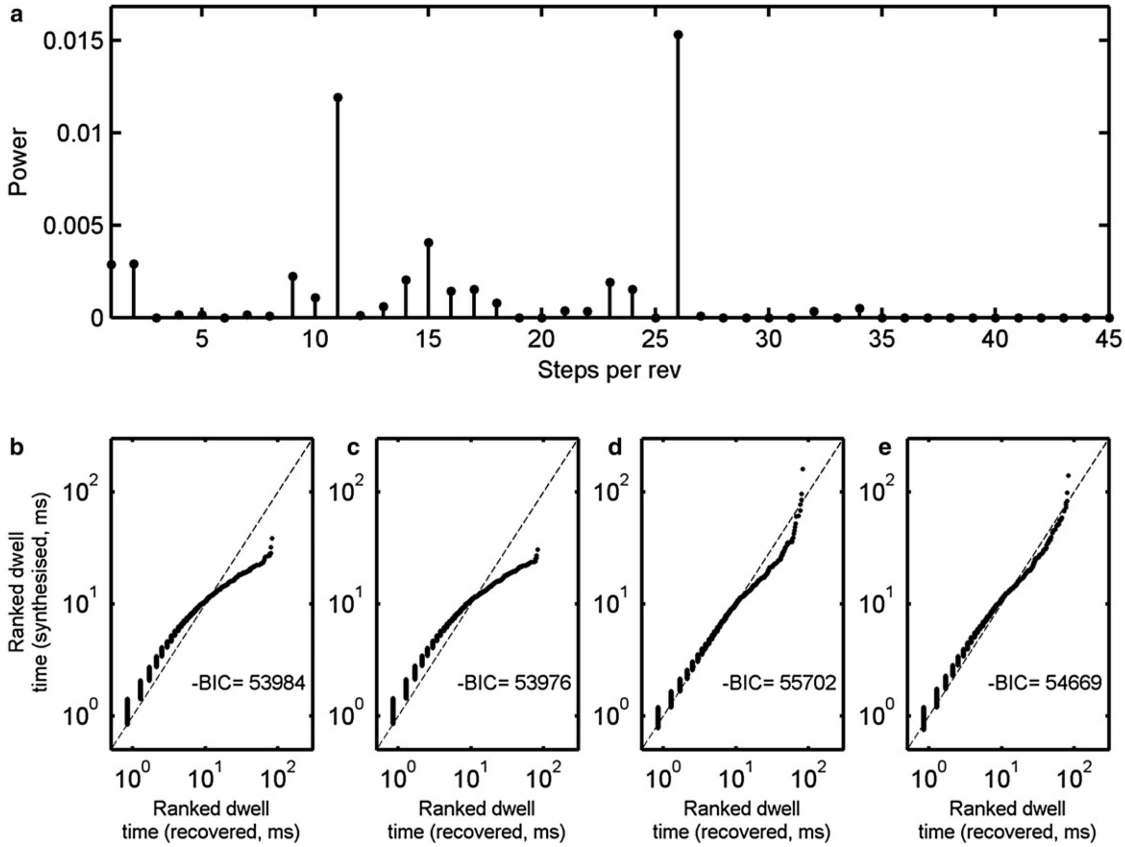

FIGURE 5 Processing of all six traces recorded from the same motor of Fig. 3 (total time 25.2 s). (*a*) Periodicity analysis showing 26 states with superimposed 11-fold periodicity. The power spectrum is the median over all six power spectra obtained, including that shown in Fig. 3 *b*. (*b–e*) Four different distribution models for dwell times. Each plot shows the ranked estimated dwell times (*horizontal*) versus the ranked synthesized dwell times drawn from the particular distribution model with the parameters fitted to the dwell times by a maximum likelihood procedure (*vertical*). Best fitting model is closest to diagonal (*dashed line*). Also shown are the (negative) Bayesian information criterion (BIC) values of the model fit (larger is better). (*b*) Exponential distribution (Poisson stepping), (*c*) gamma distribution, (*d*) Log-normal distribution, and (*e*) generalized Pareto distribution (here with parameters that define a heavy-tailed power-law distribution).

rotation in the pattern $120^\circ \times n$, $120^\circ \times n + 30^\circ$, where $n =$ 1, 2, 3—closely confirming the findings of previous studies on the slower thermophilic F_1 at similar [ATP] (24).

Given that the molecule is expected to have almost perfect threefold periodicity, it is most likely that the aperiodicity of the distribution of in Fig. 6 *b* is due to experimental confounds such as the loose linkage of the bead to the molecule. This aperiodicity manifests as some amplitude in the one-, two-, and fourfold ECF components. This aperiodicity would make it difficult to fit a single model to all of the 120° domains. However, after the discrete state extraction process, a total of 502 dwell times and a median of 83.5 dwell times per state were obtained. This was sufficient data to fit separate distribution models to the dwell times for each of the six states.

Analysis of the quality of fit of various distribution models for each of the six states revealed that different models are appropriate for each of the six states. The best model we could find (in terms of Bayesian information criterion (BIC) over all states, see Supporting Material) was the gamma distribution with scale $k = 2$ for the three states located at $120^\circ \times n$, and the exponential model for each of the three states at $120^\circ \times n + 30^\circ$. This model fitted better than a gamma model for every state (negative BIC difference 3.9), and markedly better than an exponential model for every state (negative BIC difference 49.8). We assumed that the dwell time is a random variable distributed as the sum of two exponentials (see Supporting Material), and when fitting this model to the $120^\circ \times n$ states, we found that the exponentials had equal rate parameters, so that a gamma model with scale $k = 2$ obtained an indistinguishable rate parameter. These findings are consistent with an interpretation that has two cascaded rate-limiting processes at the $120^\circ \times n$ state, and one process at the $120^\circ \times n + 30^\circ$ state.

We find that if the autocorrelation and smooth transitions are ignored by using the L1-PWC algorithm that assumes steplike time-angle traces, at low values of the regularization parameter γ, the algorithm is confounded by the relaxation effect and it returns many small, spurious steps that interpolate the smooth transitions. At the other extreme of

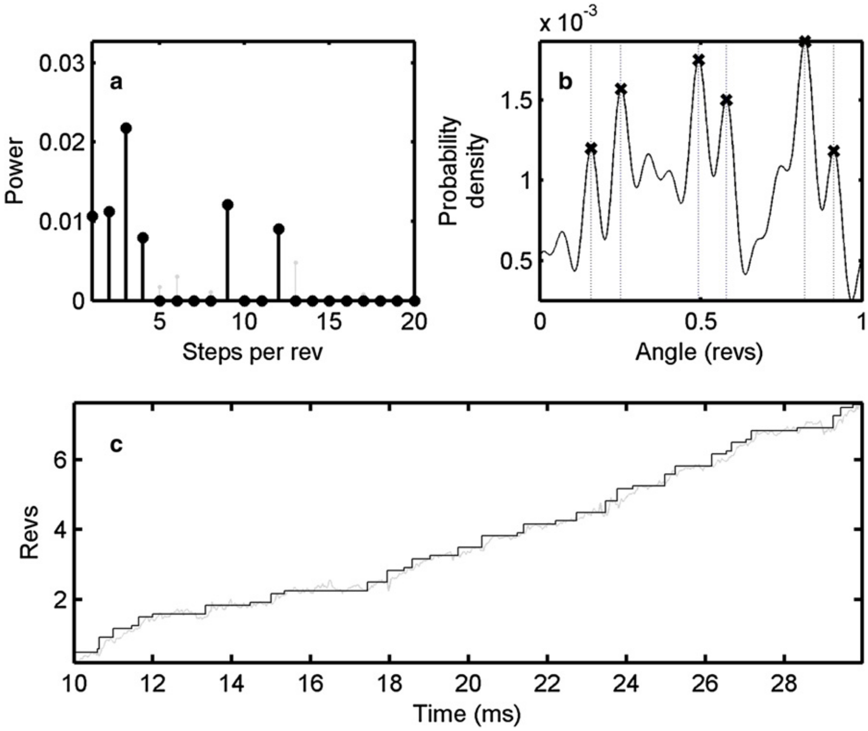

FIGURE 6 Example of processing a single experimental F_1-ATPase enzyme time-angle trace (0.27 s at 30 kHz) to extract discrete state locations. (*a*) Estimated periodicities in the distribution of the full, smoothed, 0.27 s time series m_t (*light-shaded*), largest amplitude periodicities retained (*dark-shaded*), showing a combination of dominant three, nine and twelve-fold periodicities, which is consistent with six dwell locations. (*b*) Reconstructed distribution of states (*dark-shaded line*) estimated using the six periodicities retained (between 1 and 20 steps per revolution) after thresholding in panel *b*. The six largest peaks in this distribution (*dark-shaded crosses*) represent the best estimate of the discrete state locations of the motor, used to estimate the true time series of discrete state transitions (*c*, *dark-shaded*). (*c*) A small segment of the trace showing rotation angles θ_t (*light-shaded*) with classified estimates of motor positions μ_t after L1-PWC-AR1 smoothing (*dark-shaded*, see text for algorithm descriptions). Algorithm parameters are $\gamma = 1.5$ and $a_1 = 0.8$ (estimated from the data), and $\varphi = 0.1$.

large regularization, smaller steps are missed altogether. We therefore cannot find one single, optimal value of the regularization parameter γ. This is because of the fundamental mismatch between the assumption of independence in the model and the temporal dependence in the signal.

CONCLUSION

Our step-smoothing algorithms address the problem of analyzing and recovering discrete state transitions obscured by correlated noise from time series generated by Langevin molecular motion. We show how to process bacterial flagellar motor bead assay time-angle traces faster than real-time. Subsequent application of our distribution bump-hunting algorithm uncovers periodicities in bacterial flagellar motor traces that have hitherto been hinted at, but largely hidden due to the theoretical and practical shortcomings of existing algorithms. By using our discrete state distribution estimation algorithm, we are able to recover the time series of discrete state transitions, from which analysis of the distribution of dwell times shows significant departures from classical Poisson stepping (revealed by the divergence from exponential behavior for extreme dwell times).

A recent cryo-electron microscopy structure of the flagellar rotor indicates that there is a periodicity mismatch between different parts of the rotor, which varies from one motor to the next (25). Non-Poisson stepping might be explained by static heterogeneity: mixed periodicities imply that steps need not be equivalent at all angles, but sufficiently large data-sets may reveal simple Poisson stepping at each angle. Alternatively, heterogeneity may be dynamic, with the state of the motor changing in time due to exchange of stator complexes (26) or other regulatory processes. Analysis using our methods of stepping traces from many motors will be an important tool in the task of understanding heterogeneity in flagellar rotation and finding a model of the flagellar mechanism that explains the periodicities observed in both structural and rotation data.

The F_1-ATPase bead assay shows clear evidence of smooth state transitions, with the relaxation time of the system ~5 times slower than the sampling duration. Therefore, by using Langevin dynamics in the L1-PWC-AR1 algorithm, we could extract periodicities that clearly revealed six states in one revolution of the enzyme in a characteristic angular pattern. We were able to extract a sufficient number of dwell times to detect differences in the number of rate-limiting processes responsible for each dwell. Current models of this rotary enzyme propose that each 120° step comprises an ADP release with ATP binding phase, followed by 80–90° rotation, an ATP cleave with Pi release phase, then a final 30–40° rotation; our findings are consistent with this interpretation (24). To our knowledge, previous studies have not addressed the issue of which distribution best fits these dwell times; therefore, the large amount of high-precision dwell-time data revealed by our techniques provides evidence supporting this model.

Although these analysis tools are quite simple, we have shown that they extend the limits of precision and applicability in the characterization of discreteness and noise in molecular dynamics. The algorithms we describe here require minimal manual intervention, and (assuming that the autocorrelation properties of the trace or the physical characteristics of the Langevin model are known), have only two critical parameters whose choice of values can be guided by associated theory. These tools therefore represent a step toward the automated characterization of the discrete behavior of molecular machines. This is required to exploit fully the promise of high-quality experiments on single and multiple molecules. Furthermore, by speeding up the complete experiment-analysis cycle, we can facilitate the screening of multiple phenotypes or ranges of experimental conditions, so that novel structures and parameter values for mathematical models can be rapidly tested.

SUPPORTING MATERIAL

One figure, two tables, and additional equations are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(11)00705-3.

M.A.L. is funded through the Wellcome Trust-Massachusetts Institute of Technology postdoctoral fellowship WT090651MF, and Biotechnology and Biological Sciences Research Council/Engineering and Physical Sciences Research Council grant BBD0201901. R.M.B. is funded through Biotechnology and Biological Sciences Research Council grant BBE00458X1. N.S.J. thanks the Engineering and Physical Sciences Research Council and Biotechnology and Biological Sciences Research Council including grants EP/H046917/1, EP/I005765/1, and EP/I005986/1. T.B. was funded by the Engineering and Physical Sciences Research Council via the Life Sciences Interface Doctoral Training Centre. B.C.S. is funded through an EMBO Long Term Fellowship and National Institutes of Health grant R01GM066223. D.M.M. is funded through National Institutes of Health grant R01GM066223.

REFERENCES

1. Nelson, P. C., M. Radosavljevic, and S. Bromberg. 2004. Biological Physics: Energy, Information, Life. W.H. Freeman, New York.
2. Raj, A., and A. van Oudenaarden. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell.* 135:216–226.
3. Mori, T., R. D. Vale, and M. Tomishige. 2007. How kinesin waits between steps. *Nature.* 450:750–754.
4. Kitamura, K., M. Tokunaga, …, T. Yanagida. 1999. A single myosin head moves along an actin filament with regular steps of 5.3 nanometers. *Nature.* 397:129–134.
5. Sowa, Y., A. D. Rowe, …, R. M. Berry. 2005. Direct observation of steps in rotation of the bacterial flagellar motor. *Nature.* 437:916–919.
6. Selvin, P. R., and T. Ha. 2008. Single-Molecule Techniques: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
7. Carter, B. C., M. Vershinin, and S. P. Gross. 2008. A comparison of step-detection methods: how well can you do? *Biophys. J.* 94:306–319.
8. Kerssemakers, J. W. J., E. L. Munteanu, …, M. Dogterom. 2006. Assembly dynamics of microtubules at molecular resolution. *Nature.* 442:709–712.
9. Arce, G. R. 2005. Nonlinear Signal Processing: a Statistical Approach. Wiley-Interscience, Hoboken, NJ.
10. Meacci, G., and Y. Tu. 2009. Dynamics of the bacterial flagellar motor with multiple stators. *Proc. Natl. Acad. Sci. USA.* 106:3746–3751.
11. Chung, S. H., and R. A. Kennedy. 1991. Forward-backward non-linear filtering technique for extracting small biological signals from noise. *J. Neurosci. Methods.* 40:71–86.
12. Fried, R. 2007. On the robust detection of edges in time series filtering. *Comput. Stat. Data Anal.* 52:1063–1074.
13. Pawlak, M., E. Rafajlowicz, and A. Steland. 2004. On detecting jumps in time series: nonparametric setting. *J. Nonparametr. Stat.* 16:329–347.
14. Jong-Kae, F., and P. M. Djuric. 1996. Automatic segmentation of piecewise constant signal by hidden Markov models. *In* 8th IEEE Signal Processing Workshop on Statistical Signaling and Array Processing (SSAP '96). 283–286.
15. Silverman, B. W. 1998. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.
16. Hastie, T., R. Tibshirani, and J. H. Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.
17. Kim, S. J., K. Koh, …, D. Gorinevsky. 2009. L1 trend filtering. *SIAM Rev.* 51:339–360.
18. Candes, E. J. 2006. Modern statistical estimation via oracle inequalities. *Acta Num.* 15:257–326.
19. Boyd, S. P., and L. Vandenberghe. 2004. Convex Optimization. Cambridge University Press, Cambridge, UK and New York.
20. Block, S. M., D. F. Blair, and H. C. Berg. 1989. Compliance of bacterial flagella measured with optical tweezers. *Nature.* 338:514–518.
21. Ramírez, P., and B. Vidakovic. 2010. Wavelet density estimation for stratified size-biased sample. *J. Stat. Plann. Inference.* 140:419–432.
22. Thompson, R. E., D. R. Larson, and W. W. Webb. 2002. Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* 82:2775–2783.
23. Kalafut, B., and K. Visscher. 2008. An objective, model-independent method for detection of non-uniform steps in noisy signals. *Comput. Phys. Commun.* 179:716–723.
24. Junge, W., H. Sielaff, and S. Engelbrecht. 2009. Torque generation and elastic power transmission in the rotary F_0F_1-ATPase. *Nature.* 459:364–370.
25. Thomas, D. R., N. R. Francis, …, D. J. DeRosier. 2006. The three-dimensional structure of the flagellar rotor from a clockwise-locked mutant of *Salmonella enterica* serovar Typhimurium. *J. Bacteriol.* 188:7039–7048.
26. Leake, M. C., J. H. Chandler, …, J. P. Armitage. 2006. Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature.* 443:355–358.