Supporting Material

Steps and bumps: precision extraction of discrete states of molecular Machines

Max Andrew Little, Bradley C Steel, Fan Bai, Yoshiyuki Sowa, Thomas Bilyard, David M Mueller, Richard M Berry, and Nick S Jones

Background: smoothing of correlated and step-like data: *Step-smoothing* algorithms remove noise from step-like signals – that is, signals for which the underlying, noise-free time trace to be recovered consists of constant segments with step changes between segments. The literature on step-smoothing is large (1-11) but approaches can be roughly grouped according to how they process the data. *Sliding window* methods (such as the running mean or running median filter) consider a small set of contiguous samples and replace the central sample in that window with the output of an algorithm applied to the samples only in the window. The window is moved along the time series by one sample at a time so that each sample in the series is replaced. *Recursive* methods operate by successive subdivision or merging of constant regions. For example the top-down algorithm described in Kalafut *et al.* (7) starts by finding the location of the best single step for the whole signal. If a stopping criterion is not met, a best step location is found within each constant region either side of this step. This successive subdivision is repeated until the stopping criterion is reached. Conversely, bottom-up algorithms start with every sample as a putative constant region, and merge regions successively until a stopping criterion is reached. Finally, *global* methods consider all the samples simultaneously, either fitting a model to the time series or applying some kind of transformation. The range of algorithms is very diverse. For example, Haar-wavelet de-noising (9) decomposes the signal into a sum of step-like waveforms and the noisy details that fall below a given amplitude are removed from the sum. Hidden Markov models (HMMs) treat the underlying constant regions as unobserved states, obscured by noise, following each other in an unknown sequence, and attempt to find the state transition probabilities, observation noise probabilities, and most likely sequence of states from the signal (6).

All step-smoothing methods have advantages and disadvantages relative to each other, mostly consequences of the mathematical assumptions embodied in the algorithm. Most have algorithm parameters that must be set and the results will depend on the choice of those parameters. Without any guiding theoretical principles, there is always some unavoidable level of subjectivity in choosing parameter values. The goal should always be to minimize the number of parameters, subject to the necessary degree of flexibility determined by the natural variability of the phenomenon under study – too few parameters can be as much as a problem as too many.

Issues due to algorithmic structure: Sliding window methods cannot generally find constant regions larger than the window size; the window size therefore becomes a critical parameter. Recursive methods can find long constant regions, but require specification of the stopping criterion, and this may or may not be appropriate for the underlying trace which is of course unobservable in principle. For example, the top-down method described above (17) produces different results on different-length signals, an unintended side-effect of the mathematical formulation of the stopping criterion. Global methods do not suffer from such problems but, for example, wavelet de-noising requires the selection of noise amplitude which is critical. HMMs are an alternative, but require a-priori choice of the number of states, and the number of free parameters increases with square of the number of states. This can cause considerable uncertainty in the reproducibility of the results for experimental data not used to train the HMM.

Issues due to free parameters: The majority of step-smoothing algorithms make the (explicit or implicit) assumption that the observed noise is independent or specifically, uncorrelated. However, general molecular motion occurs in potential wells centred on each discrete state with correlated, random noise, and this type of motion is often modelled by Langevin dynamics which incorporates general time lag in the motion caused by elastic and friction forces (12) (for example, drag caused by a bead attached to the elastic flagellar hook (13)); the transitions between states are then smooth. Step-smoothing under these circumstances is quite different from the step-smoothing problem as investigated in the existing literature.

Issues due to computational burden: Additionally, many step-smoothing techniques are computationally onerous in practice, either because the number of intensive calculations scales poorly with the length of the signal, or because there is no way of knowing whether a solution constitutes the *optimal* set of steps other than by exhaustive (brute-force) computation. For example, most recursive methods require an exhaustive search for the optimal step location at each iteration (2, 7), and, although there are some incidental computational efficiencies that can be exploited, for this reason such algorithms are intrinsically slow. For the HMM, there are no known algorithms that can evaluate the probabilities in their entirety without infeasible exhaustive computation so the best one can achieve is a solution that may be sub-optimal.

Given these considerations, we did not find any existing step-smoothing methods to be entirely satisfactory, and indeed our experiments on simulated data bear this out (see tables in the main text and Supporting Material).

Background: finding modes (bumps) in distributions: *Mode-finding* (also known as *bump-hunting*) is the activity of finding peak values of a distribution of a random variable. Typically the random variable represents time series of discrete states of a molecular complex or machine. If the states are distinct, then, in theory, the distribution will exhibit a clear set of "spikes" located at each unique state (see Figures 2c, 4b, 6b in main text). These separate states can then be identified from the distribution, and this requires an estimate of the distribution of the time series. Perhaps the simplest and most immediately accessible estimate is the histogram: divide the full range of the random variable into equal-sized bins and count the number of time series values that fall in each bin (14). Assuming each bin is sufficiently small that there are at least two bins per discrete state, the maxima of the histogram can locate estimates of the discrete states of the molecular machine.

However, there are problems with histograms because the separation between states is usually not known in advance, molecular motion is generally obscured by observational noise, and we only have a finite number of samples. As the precision of the peak location estimates is increased by decreasing the bin size, the bins become more sensitive to these confounding factors and spurious peaks start to emerge. Similarly, increasing the bin size to make the peaks less spurious decreases the precision of the peak location estimates and may cause two or more states to become inseparable. Also, location estimates will tend to be sensitive to the choice of bin edges, and non-equal bin sizes appear mainly to introduce complications with no special advantages (14).

Averaging over different bin edges has been proposed (average shifted histograms), but it can be shown that this is a special case of kernel density estimation, where a smooth function (usually with one mode – unimodal) of fixed width is centred on each time series sample, and the distribution at every location is estimated as the equally weighted sum of all these functions (14). Kernel density estimates are an improvement over histograms because smooth distribution estimates stabilise peak-finding in the presence of noise. The choice of kernel width is however crucial, because, if the width is too small, spurious peaks will emerge, and if too large, distinct states may become merged erroneously.

Incorporating additional information about the discrete states may lead to improvements. If the number of states is known in advance, mixture modelling may be used to find the best combination of a weighted sum of component distributions with arbitrary locations and widths (15). With unimodal component distributions, one distribution will be, ideally, located at each individual state. The main difficulty with mixture modelling is that the simultaneous estimation of the component locations and widths is not a *convex problem*, so that we cannot guarantee that any solution we find is the best one, and the computations quickly become onerous as the number of components grows. Similar issues apply to *k*-means clustering which attempts to cluster the time series samples into a given number of states. An algorithm for solving the problem exists (15), and although the problem is simpler than mixture modelling because only the state locations need to be estimated, the results can often depend quite sensitively on the initial choice of assignments required to start the search for a solution.

In estimating discrete states therefore, as with step-smoothing, existing bump-hunting approaches described above are problematic (see Tables S1 and S2).

Physically-based step-smoothing, quasi-periodic bump-hunting and distribution estimation. We seek algorithms for step-smoothing, bump-hunting and distribution estimation that incorporate elementary physical knowledge of molecular conformational dynamics. In the following, we have a time-position trace θ_t, $t = 1, 2$

… T obtained from experimental measurements of molecular dynamics. The series μ_t is the unknown series of positions corresponding to conformations of the molecular machine to be determined (and is assumed to be piecewise constant), and the series m_t is an estimate of μ_t given the time-position trace θ_t. We require that any algorithm results can be obtained with reasonable computational cost, and that they are guaranteed to converge on the globally optimal solution.

Algorithm L1-PWC: *L1-regularized global step-smoothing with independent noise*. The smoothed estimate is constructed by minimizing the *negative log posterior* (NLP) cost function with respect to a possible series of positions m'_t:

$$m_t = \arg\min_{m'_t} NLP(m'_t) = \arg\min_{m'_t} \sum_{t=1}^{T} (\theta_t - m'_t)^2 + \gamma \sum_{t=2}^{T} |m'_t - m'_{t-1}| \tag{S1}$$

The implicit physical model is in the form $\theta_t = \mu_t + \varepsilon_t$, where μ_t is a time series consisting of constant segments with abrupt jumps (steps), and ε_t is a time series of independent Gaussian noise. The problem is to find the series m_t which consists of the piecewise constant steps buried in the noise, but that is simultaneously a good approximation to the recorded time series θ_t. The first term in the NLP represents the error (negative log-likelihood) of the approximation. The second term represents the total absolute difference between consecutive approximation samples (in the Bayesian interpretation this is the negative log prior). When the penalty (*regularization*) term $\gamma = 0$, the solution becomes $m_t = \theta_t$ and no smoothing occurs. Thus, the useful behaviour of this algorithm occurs when $\gamma > 0$, so that increasing weight is placed on minimizing the second term at the expense of the first. There is a maximum useful value for this regularization parameter, γ_1 (see below), and if $\gamma > \gamma_1$ then the solution becomes $m_t = \frac{1}{T}\sum_{t=1}^{T} \theta_t$, i.e. all approximation samples take on the mean of the recorded time series. Assuming that there are only a small number of steps in μ_t amounts to imposing a *sparsity condition* on $\sum_t |m_t - m_{t-1}|$, that is, only a few terms in this expression are non-zero. Under this condition it is (with very high probability) possible to recover an approximation to μ_t that finds the true locations of the steps (16). Increasing γ forces most of the differences between consecutive samples of m_t to zero. This NLP cost function Eq. (S1) is *convex* and *quadratic* so that the optimal approximation can be obtained by minimizing NLP with respect to m_t using standard quadratic programming techniques (17). This algorithm is similar to optimal piecewise linear smoothing (8). In the main text, we demonstrate the use of this algorithm to approximate step-like motion in experimental bacterial flagellar motor time-angle traces where the sampling rate is sufficiently low that the stepping is effectively instantaneous.

Algorithm L1-PWC-ARP: *L1-regularized global step-smoothing with known correlated noise*. This algorithm is an adaptation of the L1-PWC algorithm that incorporates a general, discrete-time Langevin model within the filter structure. It minimizes the following cost function:

$$m_t = \arg\min_{m'_t} NLP(m'_t) = \arg\min_{m'_t} \sum_{t=P+1}^{T} \left(\theta_t - \sum_{i=1}^{P} a_i \theta_{t-i} - m'_t \right)^2 + \gamma \sum_{t=2}^{T} |m'_t - m'_{t-1}| \tag{S2}$$

Here the implicit model is $\theta_t = \sum_{i=1}^{P} a_i \theta_{t-i} + \mu_t + \varepsilon_t$ which captures general Gaussian, linear, discrete-time stochastic dynamics. The real-valued coefficients a_i represent the discrete-time feedback of past values of the recorded time series on the current value. This model incorporates the special case of discrete-time Langevin dynamics with linear drift and diffusion terms, in widespread use as models for general molecular dynamics (12). The forcing term μ_t consists of piecewise constant segments with jumps. Again, minimizing this cost function is convex and quadratic and can be solved for m_t using a standard quadratic programming algorithm. The underlying constant step approximation is then recovered as $m_t \Big/ \left(1 - \sum_{i=1}^{P} a_i\right)$ (note that it is not the unmodified m_t because any constant signal input to an AR model is amplified by the feedback effect of the model). In particular, in the main text we demonstrate use of the special case with P = 1 (L1-PWC-AR1) to capture Langevin motion in experimental F_1-ATPase angle-time trace:

$$NLP(m'_t) = \sum_{t=2}^{T} (\theta_t - a_1\theta_{t-1} - m'_t)^2 + \gamma \sum_{t=2}^{T} |m'_t - m'_{t-1}| \quad \text{(S3)}$$

The useful range of regularization parameter γ can be determined by reference to general principles of convex optimization (8). If $\gamma \geq \gamma_1$ where:

$$\gamma_1 = \left\| \left(DD^T\right)^{-1} D\theta_t \right\|_\infty \quad \text{(S4)}$$

then, upon optimizing Eq. (S1), m_t will be constant; here the notation $\|\cdot\|_\infty$ indicates the elementwise maximum, and the matrix D is the $T \times T$ *first difference matrix* with ones on the main diagonal, and –1 on the next diagonal above the main one and zeros elsewhere. Thus, Eq. (S4) sets the maximum useful value of the regularization parameter. Furthermore, using knowledge that a unit time step of height h is flattened away when $\gamma \geq h/2$ (16), allows us to suggest an estimate for the minimum useful value: if the noise ε_t has standard deviation σ then setting $\gamma \geq 2\sigma$ will flatten away 99% of the noise. Thus, the meaningful range of the regularization parameter is $2\sigma \leq \gamma \leq \gamma_1$, and setting the parameter just above the lower bound retains those steps that are just large enough to be detectable above the noise. In practice, we need to know σ in order to determine this range: this can be estimated from known constant dwells in θ_t if the noise is uncorrelated. Where the noise is correlated, the correlation can first be removed and then the uncorrelated signal would be used to estimate σ. In the case of simulations, we of course know σ a-priori.

Algorithm ECF-Bump: *Nonparametric bump-hunting*. The *characteristic function* is an alternative representation of the distribution $p(m)$ of the step-smoothed time series of molecular states:

$$P(f) = \int_{-\infty}^{\infty} \exp(ifm) p(m) \mathrm{d}\, m \quad \text{(S5)}$$

In this context, $p(m)$ is unknown. However, $P(f)$ can be estimated from the time series m_t, using the *empirical characteristic function* (ECF):

$$P(f_j) \approx \frac{1}{T} \sum_{t=1}^{T} \exp(if_j m_t) \quad \text{(S6)}$$

Here, the ECF is evaluated over a set of chosen frequencies f_j, $j = 1, 2 \ldots K$. The distribution function $p(m)$ can be reconstructed from the coefficients $P(f_j)$, and covers the range $[0, 2\pi]$:

$$p(m) \approx \sum_{j=-K}^{K} \exp(-if_j m) \overline{P}(f_j) \quad \text{(S7)}$$

where the overbar denotes complex conjugation (and $f_{-j} = -f_j$ to ensure that $p(m)$ is a real probability). Each frequency f_j corresponds to a potential symmetry (periodicity) of the molecular machine. In the special case of a rotary machine it corresponds to the number of steps per complete revolution. This characteristic function is closely related to the Fourier transform of the distribution of the time series. Thus, the $P(f_j)$ can be interpreted as *Fourier coefficients* of the distribution function, and are calculated over a range of experimentally relevant symmetries f_j.

The advantage of this representation of the distribution is that it is usually the case for a molecular machine that is has repeating molecular structures and so undergoes motion in a series of repeating steps. This implies that the distribution of states is both bounded and periodic, so the reconstruction converges on the exact distribution extremely rapidly as more frequency components are introduced (K increases). This is because the Fourier transform is a *sparse representation* (18) for smooth, bounded, periodic functions. Sparse representations have the desirable property that only a few of the coefficients are large, and these are the ones that contribute most to the shape of the distribution. The rest of the coefficients will fluctuate due to experimental noise and contribute little, if anything. Thus, considering the coefficients as the sum of true molecular machine symmetries and experimental artefacts and distortions, we can apply *nonlinear thresholding* by ranking coefficients by their

power $|P(f_j)|^2$ and retaining a small fraction φ of the largest coefficients (we retain around 10-20% in this paper, and this represents a good compromise between retaining the main shape of the distribution yet summarising it only by the most important periodicities). It has been more recently shown that this simple procedure for recovering the noisy distribution is statistically optimal (in the *minimax sense*) if the distribution is bounded, smooth and periodic (18), for an appropriate choice of φ related to the amount of experimental noise. This nonlinear thresholding procedure contrasts with *linear thresholding* where only a certain number of the *lowest frequency* components are retained (in this context, kernel density smoothing is effectively a linear thresholding operation, and it therefore gives us no opportunity to retain high frequency components if they are actually important). Because we only have a small number of samples of m_t, the higher frequency coefficients fluctuate due to statistical finite sample effects. Therefore, we can also apply linear thresholding by retaining only those coefficients below a threshold. In practice however, nonlinear thresholding tends to remove most irrelevant high frequency components anyway, and the linear thresholding has negligible or no effect.

Analysis of the characteristic frequency domain reveals important information about the symmetries of the molecular machine, since probability calculations with combinations of random variables become very simple in this domain. If we change the scale of a random variable X by multiplying it by a scaling factor σ and adding a constant μ, the new random variable $Y = \sigma X + \mu$ has the following characteristic function:

$$P_Y(f) = E[\exp(if\{\sigma X + \mu\})] = \exp(if\mu)P(\sigma f) \tag{S8}$$

Hence, shifting the location of the peak of a distribution corresponds simply to multiplying the characteristic function by $\exp(if\mu)$. Similarly, increasing the spread of the peak corresponds to decreasing the width of the characteristic function. In the case when the distribution is composed of superposed bumps of width scaled by σ and with period N, we have that:

$$P(f) = P_{\text{bump}}(\sigma f)\frac{1}{N}\sum_{n=0}^{N-1}\exp\left(if\mu_n\right) = P_{\text{bump}}(\sigma f)\frac{1}{N}\sum_{n=0}^{N-1}\exp\left(\frac{i2\pi nf}{N}\right) = P_{\text{bump}}(\sigma f)\frac{1}{N}\left(\frac{1-\exp(i2\pi f)}{1-\exp(i2\pi f/N)}\right) \tag{S9}$$

For the purposes of analysis, the *power* of the characteristic function is usually more convenient to work with than the characteristic function, and we are interested in whole integer symmetries f only:

$$|P(f)|^2 = |P_{\text{bump}}(\sigma f)|^2 \frac{1}{N^2}\left(\frac{1-\cos(2\pi f)}{1-\cos(2\pi f/N)}\right) = |P_{\text{bump}}(\sigma f)|^2 \times \begin{cases} 1 & f = kN \\ 0 & \text{otherwise} \end{cases} \tag{S10}$$

where $k = 1, 2, \ldots$ is the multiple and σ is the spread due to noise of the bump at each molecular state. This shows that the power of the characteristic function for a periodic distribution consists of a series of non-zero coefficients at integer multiples of the period N, the rest are zero. The absolute square magnitude of these non-zero coefficients is proportional to the absolute square magnitude of the characteristic function of the bump distributions, so that the spikes are attenuated in magnitude as the multiple increases. The characteristic function of many well-known distributions can be found exactly. For example, if the bump distributions are Laplacian, since $P_{\text{bump}}(f) = 1/(1 + \sigma^2 f^2)$, we obtain, at the peaks:

$$|P(f = kN)|^2 = \left(1 + [\sigma kN]^2\right)^{-2} \tag{S11}$$

Similarly, for Gaussian bumps $|P(f = kN)|^2 = \exp(-\sigma^2 [kN]^2)$. Qualitatively, as the noise spread increases, the non-zero coefficients in the characteristic function diminish in magnitude. Therefore, the sharper the bumps in the distribution, the easier it will be to identify the period above the background of finite sample variability and experimental artefacts.

In some cases, there will be an arrangement of molecular states that has more than one superimposed period, in general, a set of Q different periods N_q for $q = 1, 2 \ldots Q$ which are not multiples of each other. In this case, the characteristic function power will be:

$$|P(f = 0)|^2 = 1, \quad |P(f = kN_q)|^2 = |P_{\text{bump}}(\sigma k N_q)|^2 N_q^2\left(\sum_{r=1}^{Q} N_r\right)^{-2} \tag{S12}$$

and $|P(f)|^2 = 0$ otherwise. Thus the power of the characteristic function has a series of non-zero coefficients at every integer multiple of each of the Q constituent periods. The non-zero coefficients for period N_q will have absolute square magnitude proportional to $N_q^2\left(\sum_{r=1}^{Q} N_r\right)^{-2}$, so that larger periods have larger absolute square magnitude. Again, the non-zero coefficients will be attenuated in magnitude by the characteristic function of the bump distribution, and this will typically decrease faster with increasing noise spread. Therefore, superimposed symmetries in the molecular machine can be readily detected from analysis of the largest peaks in the power spectrum.

Algorithm ML-Peaks: *maximum likelihood reconstruction of discrete state time trace from distribution peaks*. Using algorithm ECF-Bump and applying the inverse Fourier transform to the coefficients $P(f_j)$, we can reconstruct the distribution $p(m)$ of molecular states. This distribution may have some small peaks that are due to finite sampling effects or inaccuracies in the reconstruction of the molecular state time trace m. However, the largest peaks are associated with the most dominant, and also most likely, positions of the molecular states. Thus, if the known dominant symmetry is M steps per revolution, this information can be used to select the M largest peaks in the distribution as the dominant discrete molecular state dwell locations. Having located these peaks, the step-smoothed time trace m_t can be used to find an estimate of the true step-like conformational state signal $\hat{\mu}_t$ by classification of each of the m_t to the nearest retained peak in the distribution. This classification is the *maximum likelihood* reconstruction of μ_t (see main text) if the noise around the dwells is Laplace distributed, since we are minimizing the absolute difference between the nearest peak and the state estimate. In fact, this Laplace distribution arises as a consequence of solving Eq. (S1-S2) which has the absolute difference penalty term (17).

Simulations of molecular machines. Here we describe a model of Brownian motion in a potential well for periodic stepping motion of a molecular machine with frictional drag and elastic energy storage. We set up a simple linear stochastic differential equation (SDE) for a typical experiment. We measure the machine conformation through a small load attached to the machine whose observed position is θ. The spring potential of the structure attaching the machine to the load is:

$$U(\theta) = \frac{1}{2}\kappa\theta^2 \tag{S13}$$

where κ is the spring stiffness constant. The load causes drag on the machine, represented using the linear friction model:

$$F(\dot{\theta}) = \xi\dot{\theta} \tag{S14}$$

where ξ is the friction coefficient. Assuming that the machine executes random motion about the equilibrium position $\theta = 0$, a Langevin equation of motion for the experiment can be written as:

$$M\ddot{\theta} + F(\dot{\theta}) + \nabla U(\theta) = \sqrt{2k_B T\xi}\,\varepsilon \tag{S15}$$

where k_B is Boltzmann's constant, T is temperature, M is the machine mass, and ε is an independent, Gaussian random driving force with mean zero and unit standard deviation. Because the ratio M/ξ is very small, the inertial term $\ddot{\theta}$ is negligible and we obtain the equations of motion:

$$\begin{aligned} F(\dot{\theta}) &= -\nabla U(\theta) + \sqrt{2k_B T\xi}\,\varepsilon \\ \xi\dot{\theta} &= -\kappa\theta + \sqrt{2k_B T\xi}\,\varepsilon \end{aligned} \tag{S16}$$

Including the average machine position $\mu(t)$ we obtain the following stochastic differential equation:

$$\mathrm{d}\theta = \frac{\kappa}{\xi}(\mu - \theta)\mathrm{d}t + \sqrt{\frac{2k_B T}{\xi}}\mathrm{d}W \tag{S17}$$

This is an *Ornstein-Uhlenbeck* process with mean μ, drift $\rho = \kappa/\xi$, diffusion $\sigma = \sqrt{2k_BT/\xi}$ and Wiener process $W(t)$. Focusing on the motion of one step to the position μ, we assume that the machine starts at time $t = 0$ at position $\theta = 0$, then the resulting motion is the sum of a deterministic exponential and correlated random fluctuation terms. The load eventually settles into correlated random motion of standard deviation $\sigma/\sqrt{2\rho}$ around the machine dwell conformation μ. The effect of the load drag and stiffness is to delay the transitions by "rounding off" the instantaneous step transition in $\mu(t)$ with step time constant ρ^{-1} s. Therefore, to be effective, a step-smoothing algorithm must take into account this smooth transition.

This is a continuous-time stochastic process, but the experimental angular measurements are available at the sampling interval Δt. Therefore, we need to find a discrete-time version of the model. The simple *Euler method* obtains:

$$\theta_{t+1} = \theta_t + \frac{\kappa}{\xi}(\mu - \theta_t)\Delta t + \sqrt{\frac{2k_BT\Delta t}{\xi}}\varepsilon_t = \left(1 - \frac{\kappa\Delta t}{\xi}\right)\theta_t + \frac{\kappa\Delta t}{\xi}\mu + \sqrt{\frac{2k_BT\Delta t}{\xi}}\varepsilon_t \tag{S18}$$

where $\theta_t = \theta(t\Delta t)$ for the time index $t = 0, 1 \ldots T$ with $\theta(0) = 0$ (we note that although there are a range of generally more accurate methods for discretising such SDEs, most are no more accurate for this particular model and so in this context there is no particular advantage to using a higher order integration scheme). This is also a discrete-time, first order autoregressive (AR) model in the form:

$$\theta_{t+1} = a_1\theta_t + (1 - a_1)\mu_t + \varepsilon_t' \tag{S19}$$

where ε_t' is an independent, zero-mean, constant variance Gaussian process. Therefore, this model is a special case of the implicit model in algorithm L1-PWC-ARP with ($P = 1$) described above, and we can estimate the quantity $a_1 = 1 - \kappa\Delta t/\xi$ directly from experimental time series using the autocorrelation at one time lag Δt of measured bead time traces θ_t.

Step-smoothing and bump-hunting algorithm performance comparisons. Figure 3 (main text) describes nine simulated test cases produced by varying: the symmetry of the discrete state locations (that is, by randomly displacing the state locations from equal spacing), the distribution of dwell times (by changing the gamma shape parameter k), the dominant symmetry (e.g. the number of discrete states), the average speed of rotation (that is, the number of revolutions per second, controlled by scaling the dwell times), and the stiffness parameter κ.

Figure 2 (main text) shows the typical output from the discrete-time model, and Figure 3 (main text) shows the performance of a range of step-smoothing algorithms applied to this test data. We test L1-PWC, L1-PWC-AR1, median filtering (19), the Chung-Kennedy filter (3), and the Kalafut-Visscher step-finding methods (7). We compare the performance of these methods in terms of the accuracy of their ability to extract the simulated, but unobserved motor position $\mu(t)$ using the *mean absolute error*:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\mu_t - m_t| \tag{S20}$$

and smaller is better. Also, the *relative absolute roughness*:

$$RAR = \sum_{t=1}^{T-1}|m_{t+1} - m_t| \Big/ \sum_{t=1}^{T-1}|\mu_{t+1} - \mu_t| \tag{S21}$$

identifies over- and under-smoothing relative to the known, motor position time series, the closer to unity the better. Note that if the $MAE = 0$, then the $RAR = 1$ (although $RAR = 1$ does not necessarily give $MAE = 0$, thus, it is important to interpret the performance with respect to *both* quantities).

Step-smoothing algorithm parameters are optimized on this test data to achieve the best MAE and RAR values. For the L1-PWC algorithm the optimal parameter values were $\gamma = 50$, and for L1-PWC-AR1, $\gamma = 1$, $P = 1$ and $a_1 = 1 - \kappa\Delta t / \xi$ (see above for a description of how we choose these values).

For the median filter, the only parameter is the window size, and as expected the optimum size was found to be the average dwell time. For the Chung-Kennedy filter, the maximum size of all forward/backward moving average predictors plays a similar role to the window size in the median filter. In our implementation, we included predictors of all window sizes up to this maximum window size, and extensive experimentation found that setting this to half the average dwell time optimized performance. Because of the non-instantaneous stepping of the Langevin dynamics, for the nonlinearity $p > 0$, we found that this algorithm introduced numerous spurious steps and non-smoothness that degraded the performance considerably. Therefore, we found that having no nonlinearity (i.e setting $p = 0$) led to the best performance overall, because it was the smoothest possible filter and so was able to perform well for the longer dwell times. The Kalafut-Visscher filter has no explicitly tunable parameters, although we have found that the results depend heavily on the length of the time series.

Bump-hunting algorithm comparisons were made in terms of the median and interquartile range (25% – 75% range) of the recovered number of discrete states (see Supplementary Tables 1 and 2). Algorithm parameters were optimized to achieve the best recovery performance. For the ECF-Bump algorithm, the analysis symmetries (frequencies) ranged from zero to 120 steps per revolution, and the nonlinear threshold was set to retain the top 10-20% largest square magnitude frequencies. The linear threshold was set at 80 steps per revolution. The histogram FFT algorithm used 128 histogram bins and 128-point FFT. The kernel density peak-picking algorithm had a Gaussian kernel with bandwidth parameter of 0.02 rads^2, and peaks smaller than 10% of the maximum peak amplitude were discarded.

Distribution fitting to dwell times. Standard maximum likelihood techniques minimizing the negative log-likelihood have been used to fit each distribution model to the dwell times obtained from bacterial flagellar motor and F_1-ATPase time-angle traces (see below for explicit details of the double exponential model). For distribution model comparison, the Bayesian Information Criterion is calculated as (15):

$$BIC = -2L + p \log N \tag{S22}$$

where p is the number of free parameters in each distribution model, N is the number of dwell times, and L is the log-likelihood of the distribution model. For the exponential, $p = 1$, gamma, lognormal and double exponential, $p = 2$, and for the generalized Pareto, $p = 3$. For the gamma with fixed k, $p = 1$.

When there are multiple subsets of dwell times that require separate distribution models per dwell state, the total *BIC* is obtained by adding the *BIC* for each separate model – this is consistent with assumption that each dwell state is independent of the others.

L1-PWC-AR1 autoregressive parameter estimation. To use the L1-PWC-AR1 algorithm, we use the standard covariance method for autocorrelation analysis to estimate the parameter a_1, which is the autocorrelation at a time lag of one sample. This requires manual identification of a sufficiently long section of the signal where the molecular machine is stationary. In the real F_1 data we studied, unambiguous, long dwells are frequent so that this approach is straightforward.

Double exponential distribution. For more than one reaction cascaded together, a more complex process than the simple Poisson process is usually a better model for the observed discrete state dwell times. Assuming that one reaction has to wait for the other to finish, the total dwell time will be a random variable that is the sum of two exponentially-distributed dwell times $T = T_1 + T_2$ with rate parameters k_1, k_2. Then the distribution of T is the convolution of the distribution of T_1 and T_2. This becomes the product of the moment generating functions of the individual distributions:

$$M_T(s) = \frac{k_1}{s + k_1} \frac{k_2}{s + k_2} \tag{S23}$$

Inverting the moment generating function gives the distribution:

$$p(T) = \frac{k_1 k_2}{k_2 - k_1}\left[\exp(-k_1 T) - \exp(-k_2 T)\right] \tag{S24}$$

with mean $(k_1 + k_2)/(k_1 k_2)$ and variance $(k_1^2 + k_2^2)/(k_1^2 k_2^2)$.

To fit the rate parameters given a set of dwell times T_i, $i = 1, 2 \ldots N$, we can maximize the likelihood, which is equivalent to minimizing the negative log-likelihood:

$$\hat{k}_1, \hat{k}_2 = \arg\min_{k_1, k_2}\left[-N \log \frac{k_1 k_2}{k_2 - k_1} - \sum_{i=1}^{N} \log\left(\exp(-k_1 T_i) - \exp(-k_2 T_i)\right)\right] \tag{S25}$$

This can be solved using a variety of generic nonlinear optimization techniques, with the constraint $k_1, k_2 > 0$. Confidence intervals for the rate parameters are obtained by 1000 bootstrap resampling operations (15).

The degenerate case where $k_1 = k_2$ is the gamma distribution with scale parameter $k = 2$.

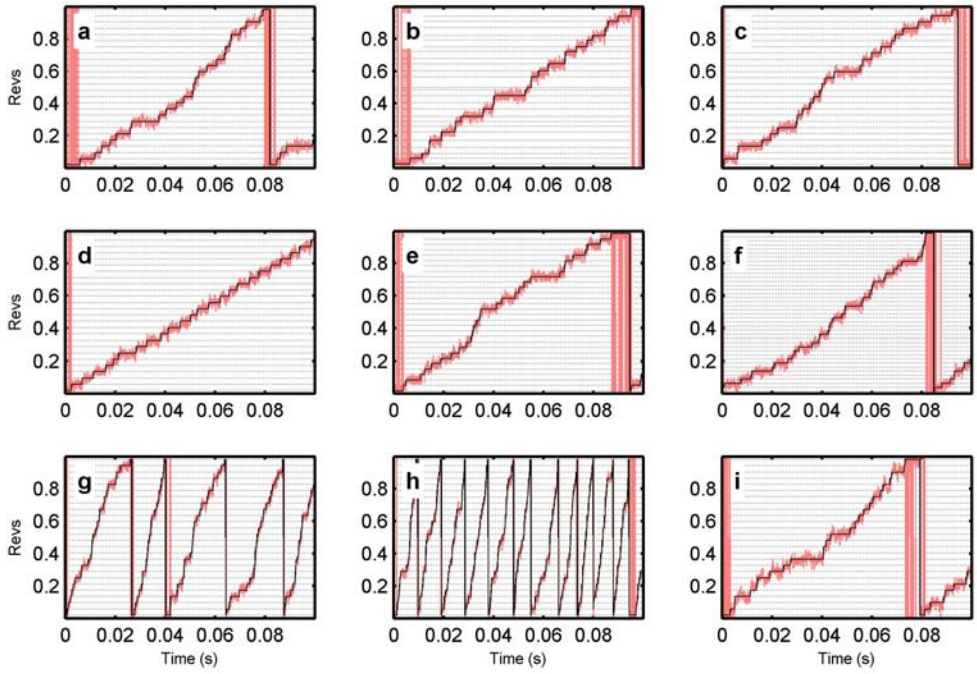
	ECF-Bump	*Kernel density with peak finding*
Default (26)	26 (0.0)	27 (1.5)
20% *dwell aperiodicity* (26)	26 (0.5)	27 (1.5)
Gamma dwell times, k = 2 (26)	26 (0.0)	29 (1.3)
Gamma dwell times, k = 10 (26)	26 (0.0)	29 (2.5)
30 *dwell locations* (30)	30 (0.0)	32 (3.3)
40 *dwell locations* (40)	40 (0.0)	46 (5.0)
50 *revs/sec* (26)	26 (0.0)	39 (5.3)
100 *revs/sec* (26)	26 (0.0)	45 (3.8)
Flagellar stiffness, κ = 50 (26)	26 (0.0)	47 (8.8)

Table S1: Performance of two different bump-hunting methods at recovering the dominant symmetry in the distribution of states for the nine test cases of simulated bacterial flagellar motor rotation time series described in the main text. The figures are the median dominant symmetry over five replications, and the associated interquartile range (difference between 25th – 75th percentile) in brackets. The bracketed number in the first column is the true symmetry. Algorithm ECF-Bump is described above. Kernel density with peak finding estimates the distribution of discrete states using the kernel density method, then counts the number of peaks in the estimated distribution.

	Average exponential dwell time				
Number of dwell locations	2.5ms	1.5ms	1.0ms	0.75ms	0.50ms
ECF-Bump					
20	20.0 (0.0)	20.0 (0.0)	20.0 (0.0)	20.0 (0.0)	20.0 (0.0)
30	30.0 (0.0)	30.0 (0.0)	30.0 (0.0)	30.0 (0.0)	30.0 (0.0)
40	40.0 (0.0)	40.0 (0.0)	40.0 (0.0)	40.0 (0.0)	40.0 (0.0)
50	50.0 (0.0)	50.0 (0.0)	50.0 (2.5)	50.0 (0.0)	54.0 (8.0)
60	60.0 (0.0)	60.0 (0.0)	60.0 (4.0)	60.0 (0.0)	(no result)
Histogram with FFT (13) (see caption)					
20	19.0 (0.0)	19.0 (1.0)	20.0 (1.0)	20.0 (0.0)	20.0 (0.0)
30	29.0 (0.0)	29.0 (0.0)	29.0 (0.0)	29.0 (0.0)	30.0 (1.0)
40	39.0 (0.0)	39.0 (0.0)	39.0 (34.0)	8.0 (16.0)	14.5 (13.0)
50	5.5 (45.0)	4.5 (9.0)	4.0 (14.0)	15.5 (15.0)	15.5 (24.0)
60	20.5 (47.0)	13.0 (16.0)	7.5 (9.0)	14.0 (17.0)	14.0 (18.0)
Kernel density with peak finding (see caption)					
20	21.0 (0.0)	21.0 (1.0)	21.0 (1.0)	21.0 (1.0)	21.0 (1.0)
30	30.5 (1.0)	30.0 (1.0)	31.0 (2.0)	32.5 (3.0)	34.0 (2.0)
40	37.5 (2.0)	38.0 (3.0)	38.0 (2.0)	37.0 (1.0)	39.0 (2.0)
50	40.5 (3.0)	37.0 (2.0)	37.0 (2.0)	36.0 (4.0)	38.0 (2.0)
60	40.0 (3.0)	37.0 (4.0)	37.0 (4.0)	37.5 (1.0)	36.0 (4.0)

Table S2: Performance of different bump-hunting methods at finding the dominant symmetry in the distribution of states of simulated bacterial flagellar motor rotation time series with exponential dwell times, over a wide range of dominant symmetries. Each entry shows the median estimated state periodicity, with the interquartile range (25th – 75th percentile) in brackets. No result indicates that no dominant peak in the ECF could be found. The histogram with FFT method first estimates the distribution of states using a histogram, then finds the fast Fourier transform of that histogram; the largest peak in the spectrum is the estimated periodicity (method used in Sowa *et al.* 2005). Kernel density with peak finding estimates the distribution of states using the kernel density method, then counts the number of peaks in the estimated distribution.

Figure S1: Nine test cases of simulated bacterial flagellar motor time traces; pink line is measured bead angular position θ_t, black line the (unobservable) motor position μ_t. Dotted horizontal lines are the discrete state locations. (a) Default case: 26 regularly spaced states, exponential dwell times, flagellar hook stiffness κ = 100k_BT/ rad, 10 revs/sec. (b) As default, but with 20% dwell location asymmetry (see Supplementary Methods). (c) With gamma-distributed dwell times, k = 2. (d) Gamma dwell times, k = 10. (e) 30 states. (f) 40 states. (g) 50 revs/sec. (h) 100 revs/sec. (i) Flagellar hook stiffness κ = 50 k_BT/rad.



Supporting References

1. Carter, B. C., M. Vershinin, and S. P. Gross. 2008. A comparison of step-detection methods: How well can you do? Biophysical Journal 94:306-319.
2. Kerssemakers, J. W. J., E. L. Munteanu, L. Laan, T. L. Noetzel, M. E. Janson, and M. Dogterom. 2006. Assembly dynamics of microtubules at molecular resolution. Nature 442:709-712.
3. Chung, S. H., and R. A. Kennedy. 1991. Forward-Backward Nonlinear Filtering Technique for Extracting Small Biological Signals from Noise. Journal of Neuroscience Methods 40:71-86.
4. Fried, R. 2007. On the robust detection of edges in time series filtering. Computational Statistics & Data Analysis 52:1063-1074.
5. Pawlak, M., E. Rafajlowicz, and A. Steland. 2004. On detecting jumps in time series: Nonparametric setting. Journal of Nonparametric Statistics 16:329-347.
6. Jong-Kae, F., and P. M. Djuric. 1996. Automatic segmentation of piecewise constant signal by hidden Markov models. In Statistical Signal and Array Processing, 1996. Proceedings., 8th IEEE Signal Processing Workshop on (Cat. No.96TB10004. 283-286.
7. Kalafut, B., and K. Visscher. 2008. An objective, model-independent method for detection of non-uniform steps in noisy signals. Computer Physics Communications 179:716-723.
8. Kim, S. J., K. Koh, S. Boyd, and D. Gorinevsky. 2009. L1 Trend Filtering. SIAM Review 51:339-360.
9. Cattani, C. 2004. Haar wavelet-based technique for sharp jumps classification. Mathematical and Computer Modelling 39:255-278.
10. Hou, Z. J., and T. S. Koh. 2003. Robust edge detection. Pattern Recognition 36:2083-2091.
11. Smith, D. A. 1998. A quantitative method for the detection of edges in noisy time-series. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 353:1969-1981.
12. Becker, O. M. 2001. Computational biochemistry and biophysics. M. Dekker, New York.
13. Sowa, Y., A. D. Rowe, M. C. Leake, T. Yakushi, M. Homma, A. Ishijima, and R. M. Berry. 2005. Direct observation of steps in rotation of the bacterial flagellar motor. Nature 437:916-919.
14. Silverman, B. W. 1998. Density estimation for statistics and data analysis. Chapman & Hall/CRC, Boca Raton.
15. Hastie, T., R. Tibshirani, and J. H. Friedman. 2001. The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. Springer, New York.
16. Strong, D., and T. Chan. 2003. Edge-preserving and scale-dependent properties of total variation regularization. Inverse Problems 19:S165-S187.
17. Boyd, S. P., and L. Vandenberghe. 2004. Convex optimization. Cambridge University Press, Cambridge, UK ; New York.
18. Candes, E. J. 2006. Modern statistical estimation via oracle inequalities. Acta Numerica 15:257-326.
19. Arce, G. R. 2005. Nonlinear signal processing: a statistical approach. Wiley-Interscience, Hoboken, N.J.