J Membrane Biol (2013) 246:57–66
DOI 10.1007/s00232-012-9504-9

Test of Normality for Integrated Change Point Detection and Mixture Modeling

S. P. Parsons · J. D. Huizinga

Received: 31 May 2012 / Accepted: 19 September 2012 / Published online: 16 October 2012

Abstract Single-molecule data often show step-like changes in the quantity measured between constant levels. Analysis of this data consists of detecting the steps, i.e., change point detection (CPD), and determining the levels, i.e., clustering. We describe a novel algorithm which integrates these two analyses, based on a statistical test of a normal distribution. The test of normality (TON) algorithm integrates statistical CPD with gaussian mixture model clustering. We used TON with both simulated data and ion channel patch-clamp recordings. It performed well with simulated data except at a high signal-to-noise ratio and when the frequency of steps was high compared to the sampling frequency. TON has advantages over separate CPD and mixture modeling algorithms, especially for complex single-molecule data. This was illustrated by its application to the maxichannel, an ion channel with multiple subconductance states.

Keywords Mixture model · Cluster analysis · Change point detection · Ion channel · Single molecule

Introduction

A large number of techniques exist for recording single molecules. Examples include force measurements, such as optical tweezers and atomic force microscopy (Carter and Cross 2005; Mejia et al. 2008); electrical recordings of ion channels; electrical recording of non-ion channels (Choi et al. 2012); single-molecule fluorescence resonance energy transfer; and natural radiative emission (Frantsuzov et al. 2008; Kruger et al. 2011). Such recordings often show step-like changes in the measured quantity, intervening between periods where the quantity is near constant except for noise. The latter stationary periods (SPs) correspond to stable conformational states of the molecule with a step or change point (CP) corresponding to the rapid transition from one state to another. Often, the molecule appears to have a limited number of states as the SPs occur at only a few, consistent amplitudes or levels. The object of analysis is to understand and model the kinetics of the molecule, and to this end measurements must be made of the timing of CPs, the length and amplitude of SPs and the amplitude of levels. Algorithms for this are divided into those that detect CPs, change point detectors (CPDs), and those that determine levels, clustering algorithms (CAs).

Cluster analysis is the division of a set of objects into subsets or clusters. In the case of a single-molecule recording, the objects are the amplitude values (samples) of the recording and each cluster is a level. Cluster analysis is a huge field with wide application, so there is a correspondingly diverse menagerie of CAs; but most can be assigned to one of two clusters, hierarchical and optimization CAs (following Everitt 1980; Theodoridis and Koutroumbas 2009). In hierarchical CAs, clustering is performed iteratively to produce a tree-like hierarchy of clusters, with each cluster at a branching point. In *divisive* clustering all objects are first assigned to a single cluster, then this cluster is divided into two, these two into four and so on until each cluster contains one object. In *agglomerative* clustering the reverse occurs. The choice of how to divide or agglomerate is based on interobject distances. The tree hierarchy makes hierarchical clustering a natural approach in phylogenetics. Applied to single molecules it

S. P. Parsons (✉) · J. D. Huizinga
Farncombe Family Digestive Health Research Institute,
McMaster University, 1280 Main Street West, Room 3N5C-H,
Hamilton, Ontario L8S 4K1, Canada
e-mail: parsons_llabh@yahoo.co.uk

can model a molecule with a hierarchy of conformational states, but as yet this idea has had only theoretical application and computer simulation (Frauenfelder 2010; Frauenfelder et al. 1988; Torda and Vangunsteren 1994). In *optimization* CAs the number of clusters is chosen a priori and the task of the algorithm is to partition the objects so that some statistical measure of the resulting clusters is either maximized or minimized. In *mixture modeling* this measure is the fit of each cluster to a particular distribution, the data as a whole being modeled as a mixture of distributions. For example, in a gaussian mixture model each cluster should approximate a gaussian distribution. This is the most common cluster analysis for single-molecule recordings as each level (cluster) is modeled as a single amplitude contaminated with gaussian-distributed noise.

CPDs detect sudden changes in a signal. In the case of a single-molecule recording, this is the step change in amplitude. Of course, by detecting CPs, all CPDs also detect SPs (what is not a CP is an SP). Most CPD algorithms are of four classes: cumulative sum, window statistics, amplitude threshold and derivative threshold. In the first, level amplitudes must be known already and the algorithm uses this information to detect departures from these levels (Basseville and Benveniste 1980; Basseville 1988; Draber and Schultze 1994; Schultze and Draber 1993). The best known is the Page-Hinkley detector. Window statistic algorithms scan a window of samples across the recording, calculating a statistical property of that window. A sudden change in that statistic will indicate a CP. In some cases the statistic can be expressed as a probability (p value) of the null hypothesis that there has been no change, for example, Cochrane's test or Welch's t test (Carter et al. 2008; Carter and Cross 2005; Cochrane 1954; Moghaddamjoo 1988; Pastushenko and Schindler 1997; Riessner et al. 2002; Welch 1947). For others a somewhat more heuristic threshold must be used (Carter et al. 2008; Patlak 1988, 1993; Thompson et al. 2002). This is the case with amplitude and derivative threshold methods based on the amplitude or time derivative of the recording crossing some threshold value (Tyerman et al. 1992; VanDongen 1996).

Clustering and CPD algorithms can be coupled in series to produce a complete analysis of a single-molecule recording. For instance, the levels found by clustering can be used to direct a cumulative sum CPD. Or the mean amplitudes of SPs found by CPD can be clustered instead of individual sample amplitudes. However, there are no algorithms which combine CPD and clustering in one. We describe a novel algorithm that integrates a statistical CPD with gaussian mixture modeling. The algorithm is based on the Jarque-Bera statistic for a gaussian (normal) distribution, and as such, we call it the test of normality (TON) algorithm. The TON algorithm performed quite well with both simulated data and real ion channel recordings and has some advantages over separate mixture modeling and CPD algorithms.

Methods

TON Algorithm: General Principle

The general principle of the TON algorithm is to scan through a record, agglomerating into a level samples that form a single normal distribution. A level is initiated by searching for a contiguous set of samples with (1) number of samples = $N_{\mathrm{T(initiating)}}$, (2) SD ($\sigma$) < σ_{T} and (3) probability of non-normality (ρ) < ρ_{T} (see below for definition of ρ).

Such a set constitutes an initiating SP (Fig. 1). The level is then extended by scanning through the rest of the record for further sets of contiguous samples ("extending SPs," Fig. 1) with (1) a minimum length in ms, $L_{\mathrm{T(extending)}}$; (2) σ for the level (all SPs, up to and including the present) < σ_{T}; and (3) ρ for the level (all SPs, up to and including the present) < ρ_{T}. σ and ρ are calculated with the addition of each sample to an extending SP. When either rises above threshold, the extending SP is terminated. This is the CPD aspect of the algorithm. If the extending SP's length is above $L_{\mathrm{T(extending)}}$, it is included in the level. If its length is below $L_{\mathrm{T(extending)}}$, it is not included in the level (σ and ρ, being returned to their prior values).

Extending SPs are added to the level until the algorithm reaches the end of the record. The properties of the level (mean, SD) are then recorded. The processes of initiation and extension are then repeated from the start of the record to find further levels. As each level is initiated and extended its samples are marked and subsequent level searches ignore these marked samples. This prevents levels from overlapping and provides a mechanism for the algorithm to stop, when a level search cannot find an initiating SP.

To distinguish between levels found by the algorithm and levels of a simulated molecule (the algorithm not being perfect), we use the terms "algorithm level" and "molecule level" from here on.

TON Algorithm: Input Parameter Selection

The values for the four input parameters were chosen as follows.

1. $N_{\mathrm{T(initiating)}}$: this needed to be large enough so that the initiating SP, if it was normally distributed, was asymptotic to a normal distribution. For default we used what we thought was a comfortable value of 200. However, in some cases, where a figure of 200 samples was larger than the lifetime of an SP, we used 50.

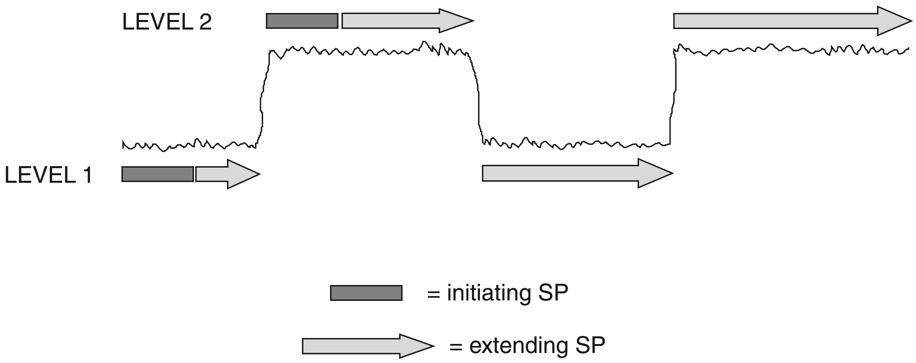

Fig. 1 The TON algorithm. Each level is constructed first by finding an *initiating SP*, a contiguous set of normal samples of length $N_{\mathrm{T(initiating)}}$. The level is then added to with *extending SPs*, further sets of samples which, when added to the level, do not produce a non-normal distribution

2. $L_{\mathrm{T(extending)}}$: by default this was set to the equivalent of $N_{\mathrm{T(initiating)}}$ ($L_{\mathrm{T(extending)}} = N_{\mathrm{T(initiating)}}$/sampling rate); i.e., 20 ms at a sampling rate of 10 kHz. As with $N_{\mathrm{T(initiating)}}$, smaller values were used for recordings with faster kinetics.
3. σ_{T}: for records with high noise or fast kinetics (timescales $<N_{\mathrm{T(initiating)}}$) σ_{T} was important as an initiating SP with an apparently normal distribution (passing ρ_{T}) might actually cover more than one molecule level. The algorithm level will then initiate midway between molecule levels. σ_{T}, set to a value just above the σ of the record's background noise, prevents this. For simulated data σ_{T} was set according to the value of σ_{noise} (see below). For patch-clamp ion channel data σ_{T} was somewhat above the amplifier meter reading of root-mean-square noise (0.5–0.6 pA).
4. ρ_{T}: this was set to 0.95 for all data, a value which seemed appropriate according to the results in Fig. 2.

The Jarque–Bera Test for Normality

The central crux of the TON algorithm is a test of normality. We used one of the simplest tests for a normal distribution, the Jarque–Bera statistic (Jarque and Bera 1987),

$$J = n\left[\frac{\mu_3^2}{6} + \frac{(\mu_4 - 3)^2}{24}\right]$$

$$\mu_p = \frac{M_p}{M_2^{p/2}} \quad M_p = \sum_i^n (x_i - m_1)^p/n \quad m_p = \sum_i^n x_i^p/n$$

where J is the Jarque–Bera statistic, n is the number of samples, x_i is the ith sample, m_p is the pth order moment, M_p is the pth order central moment and μ_p is the pth order standardized moment. J has an approximately χ^2 distribution with two degrees of freedom, so the probability of a non-normal distribution is calculated as

$$\rho = 1 - \Gamma(1, 0.5J)$$

where ρ is the probability of non-normality and Γ is the upper incomplete gamma function.

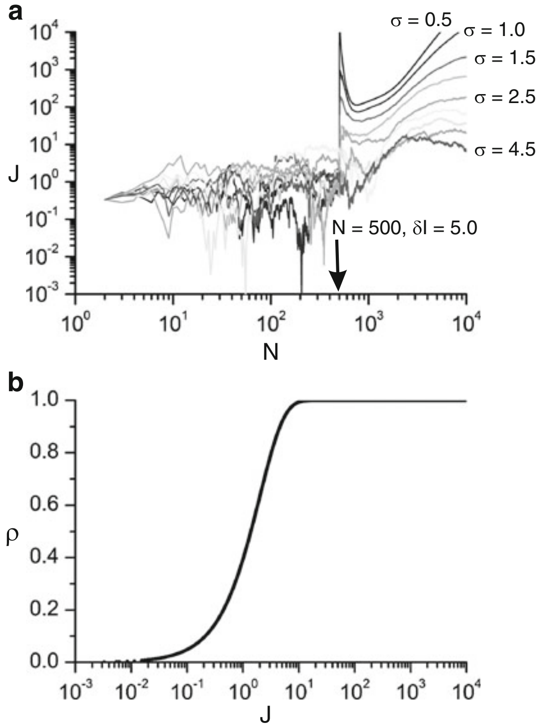

Fig. 2 Illustration of the Jarque–Bera TON. A set of pseudorandom numbers was generated with a gaussian distribution of zero mean and set SD (σ). The Jarque–Bera statistic (J) and probability of non-normality (ρ) were calculated after the addition of each number. After 500 numbers had been added, each subsequent number was increased by 5. **a** Plot of J as a function of set size (N), with different values of σ. **b** Plot of ρ as a function of J for all of the data in **a**

To assess the ability of the Jarque–Bera statistic to detect a CP, a simulation was made. Random numbers were generated with a gaussian distribution of zero mean and constant SD (σ). With the addition of each number to the set, J and ρ were calculated for the set. Once the set reached 500 numbers in size, to each subsequent number a value (δI) was added, simulating a change in level at $n = 500$. The change in level gave a large increase in J, followed by a U-shaped relaxation (Fig. 2a). This transient in J decreased as the ratio of δI to σ decreased. Below a δI/

σ of 2, a transient in J was not apparent. ρ was a steep function of J (Fig. 2b), increasing greatest at values between 1 and 10, which appeared to be the range of the transient increase in J.

Single-Pass, Online Moment Calculation

Traditional algorithms to calculate central moments can be described as "offline" and "double pass." *Double pass* means that the data have to be passed through twice, first to calculate the mean (m_1) and second to calculate the central moment. *Offline* means that if we want to add a sample to the data and then recalculate the moments for this larger data set (as in extending a level), we need to repeat this double pass though all the samples (the moments are calculated from scratch). The latter is very computationally expensive and would give an impossibly slow algorithm. Luckily there is a method to calculate central moments that is online—to recalculate moments with addition of a new sample requires only the old moments and the value of the new sample. This was first derived for the general case (for any order of moment) by Pebay (2008)

$$M_p = \underset{p}{\overset{\prime}{M}}/n$$

$$\underset{p,n}{\overset{\prime}{M}} = \underset{p,n-1}{\overset{\prime}{M}} + \left\{ \sum_{k=1}^{p-2} \left[\binom{p}{k} \underset{p-k,n-1}{\overset{\prime}{M}} \left(\frac{-\delta}{n} \right)^k \right] \right\} + \left\{ \left[\frac{n-1}{n}\delta \right]^p \left[1 - \left(\frac{-1}{n-1} \right)^{p-1} \right] \right\}$$

$$\delta = x_n - m_{1,n-1}$$

$$m_{1,n} = m_{1,n-1} + \frac{x_n - m_{1,n-1}}{n}$$

where M_p is the pth-order unnormalized central moment, $M_{p,n}$ is the pth-order unnormalized moment of the set of n samples, x_n is the value of the nth sample and $m_{1,n}$ is the mean of the set of n samples.

Histogram Data Display

The TON algorithm outputs a set of probability densities (pds) for each record analyzed. The first pd is a histogram of all the samples in the record, normalized to an area of one

$$p_{\mathrm{R}}(a) = \sum_{i}^{n_{\mathrm{R}}} \begin{cases} 1/(n_{\mathrm{R}}\Delta_{\mathrm{bin}}) & a + 0.5\Delta_{\mathrm{bin}} > x_i > a - 0.5\Delta_{\mathrm{bin}} \\ 0 & \text{else} \end{cases}$$

where $p_{\mathrm{R}}(a)$ is the pd of recording sample amplitudes (a), n_{R} is the total number of samples and Δ_{bin} is the bin size (0.01 unless otherwise stated). Other pds are "reconstructed" from the algorithm levels. From the TON we know that the samples of each algorithm level are normally distributed. Therefore, from the mean (μ_{A}), SD (σ_{A}) and number of samples (n_{A}) of each algorithm level we can reconstruct their pd for comparison with $p_{\mathrm{R}}(a)$

$$p_{\mathrm{A},\Sigma}(a) = \sum_{i}^{N_{\mathrm{A}}} p_{\mathrm{A},i}(a)$$

$$p_{\mathrm{A},i}(a) = \frac{f_{\mathrm{A},i}}{\left(2\pi\sigma_{\mathrm{A},i}^2\right)^{0.5}} \exp\left[-\frac{\left(a - \mu_{\mathrm{A},i}\right)^2}{2\sigma_{\mathrm{A},i}^2} \right]$$

$$f_{\mathrm{A},i} = \frac{n_{\mathrm{A},i}}{n_{\mathrm{R}}}$$

where N_{A} is the number of algorithm levels and f_{A} is the fraction of time spent in a level.

Data Simulation

Single-molecule recordings were simulated with N_{C} number of levels. With $N_{\mathrm{C}} > 1$ uniformly distributed random integers between 1 and N_{C} were used for the number of levels stepped during a CP. Exponentially distributed random numbers were used for SP lengths

$$p(t) = \frac{1}{\tau_{\mathrm{SP}}} \exp(-t/\tau_{\mathrm{SP}})$$

where $p(t)$ is the probability of an SP of length t and τ_{SP} is the mean SP length. Gaussian noise with a SD σ_{noise} was added, and then the record was filtered with an eighth-order Bessel low-pass filter with a cutoff of 0.5 kHz (with a sampling rate of 10 kHz).

Computation

All data simulations and analyses were carried out with programs written in C++, using the Dev-C++ IDE. All code is freely available from the author upon request.

Results

Simulated Data: Signal to Noise Ratio

We first tested the TON algorithm with a simulated two-level record. The signal-to-noise ratio (SNR) was varied by changing the amplitude difference between levels (γ), while keeping the SD of the added gaussian noise (σ_{noise}) constant. With an SNR of 1.0 ($\gamma = 1$, $\sigma_{\mathrm{noise}} = 1$) the algorithm was accurate in determining the levels. When the noiseless record was overlapped with the algorithm levels, there was a close agreement in both amplitude and timing (Fig. 3b). Similarly, there was a close agreement between the record's all-points pd and the pd reconstructed from the algorithm levels (Fig. 3c, see "Methods"). From the reconstructed pd it was clear that the algorithm calculated more than two levels. When the amplitude of the algorithm levels was plotted against their lengths (Fig. 3d), it became apparent that each molecule level was approximated by four algorithm levels, one of fairly accurate length and amplitude and three of much shorter length but still of accurate amplitude ("ghost" levels).

When the SNR was reduced to 0.5 ($\gamma = 0.5$, $\sigma_{\text{noise}} = 1$) the TON algorithm was still competent at assigning level amplitudes, though it was a little less competent at the timing of the levels (Fig. 3b). This was despite the fact that the molecule levels were not distinguishable in the all-points pd (Fig. 3c). With SNR reduced further to 0.3 ($\gamma = 0.3$, $\sigma_{\text{noise}} = 1$) the algorithm failed to assign accurate levels. A single algorithm level was calculated with an amplitude midway between the molecule levels, along with several ghost levels across a spread of amplitudes (Fig. 3c, d).

Simulated Data: Fast Kinetics

In the last example the molecule's kinetics were slower than the temporal thresholds of the algorithm ($\tau_{\text{SP}} = 100$ ms, $N_{\text{T(initiating)}}$ and $L_{\text{T(extending)}} = 20$ ms). These temporal thresholds are minimal time periods over which the algorithm can reliably establish normality (see "Methods"). However, many molecules have kinetics which are much faster than this. We looked at two possibilities: (1) what happens if the temporal parameters are reduced and (2) what happens as the timescale of the molecule's kinetics approaches the timescale of the temporal thresholds. To do this we simulated a two-state molecule as before ($\gamma = 1$, $\sigma_{\text{noise}} = 1$)

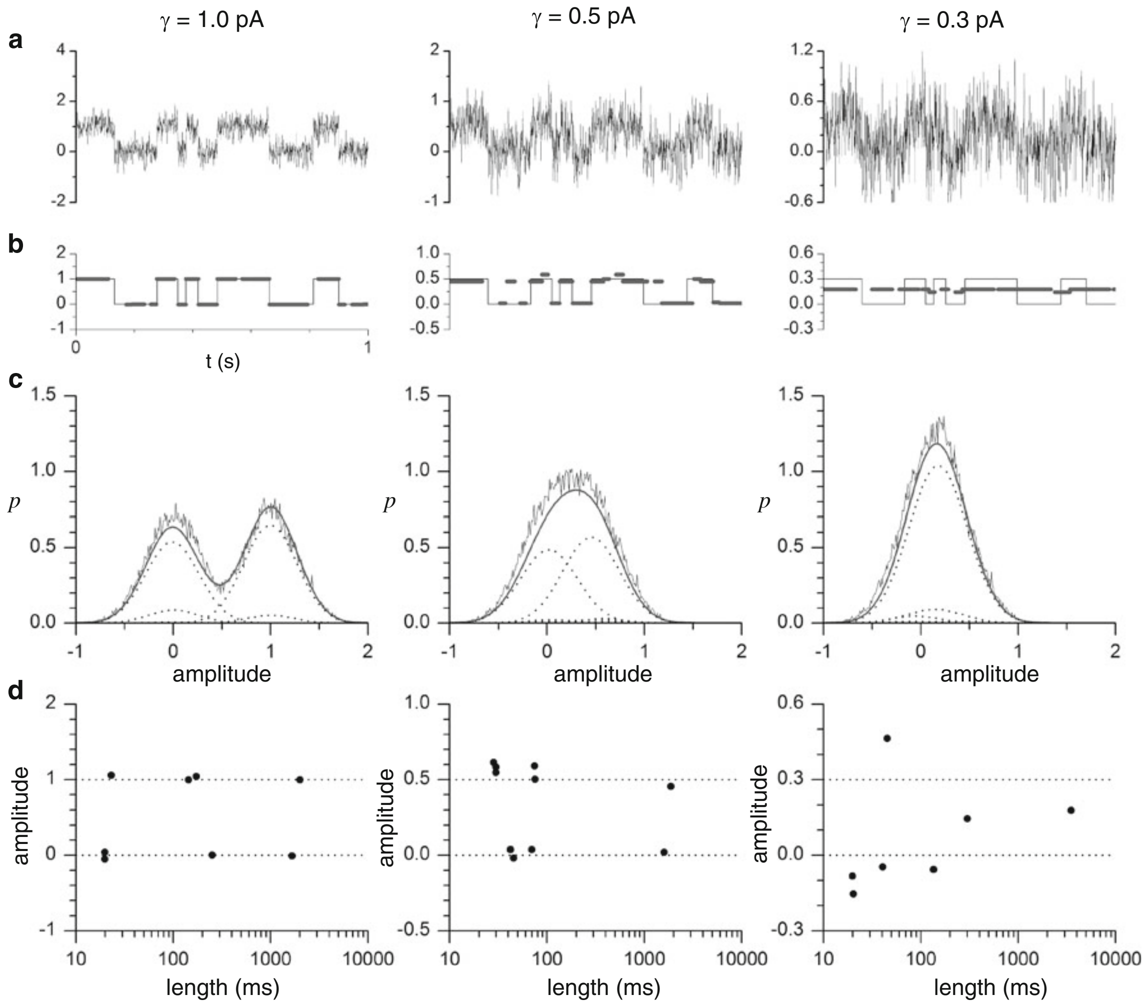

Fig. 3 Effect of varying signal-to-noise ratio on the performance of the TON algorithm. Algorithm parameters were $N_{\text{T(initiating)}} = 200$ samples, $L_{\text{T(extending)}} = 20$ ms, $\sigma_{\text{T}} = 0.3$ and $\rho_{\text{T}} = 0.95$. Simulation parameters were $\tau_{\text{SP}} = 100$ ms, $\sigma_{\text{noise}} = 1.0$, $\gamma = 1.0$, 0.5 or 0.3 (*left to right*). **a** Low-pass-filtered (0.5 kHz cutoff) record. **b** Record without added noise, with algorithm levels superimposed (*thick lines*). **c** Probability densities (pds). *Noisy thin line*, all-points pd ($\Delta_{\text{bin}} = 0.01$); *dotted lines*, reconstructed pds for each algorithm level; *thick smooth line*, sum of reconstructed pds. **d** Plots of algorithm level amplitude against length. *Dotted lines* indicate the two molecule levels

but speeded the molecule's kinetics ($\tau_{SP} = 1$–20 ms) and reduced the temporal thresholds ($N_{T(\mathrm{initiating})} = 50$ samples at 10 kHz sampling frequency = 5 ms, $L_{T(\mathrm{extending})} = 2$ ms).

With $\tau_{SP} = 15$ or 5 ms, the TON algorithm was fairly accurate in assigning both level amplitude and timing (Fig. 4). However, with $\tau_{SP} = 3$ ms (just shorter than $N_{T(\mathrm{initiating})}$ and just longer than $L_{T(\mathrm{extending})}$), there were four algorithm levels of substantial length, two near the amplitude of the molecule levels and two midway in amplitude between the molecule levels.

Simulated Data: Multiple Levels

Many single-molecule recordings show multiple levels. We simulated a recording consisting of seven levels, with CPs between any two levels. As with the first simulation (Fig. 3), the amplitude difference between levels (γ) was varied while keeping the gaussian noise (σ_{noise}) constant. With an SNR of 1.0 ($\gamma = 1$, $\sigma_{\mathrm{noise}} = 1$) the algorithm accurately assigned both level amplitude and timing (Fig. 5b). In fact, there was only one ghost level (Fig. 5d). With an SNR of 0.5 ($\gamma = 0.5$, $\sigma_{\mathrm{noise}} = 1$) most algorithm levels were still accurate, despite the fact that the molecule levels could no longer be distinguished in the all-points pd (Fig. 5c). There were still relatively few ghost levels. With an SNR of 0.3 ($\gamma = 0.3$, $\sigma_{\mathrm{noise}} = 1$) the accuracy of the algorithm began to fail. Though at least three molecule levels (0, 0.6 and 0.9) were fairly approximated by the algorithm, the other molecule levels had no equivalent algorithm level. Instead, there was a long algorithm level at 1.67 (midway between the molecule levels of 1.5 and 1.8) and multiple ghost levels (Fig. 5d).

Real Data

We tested the TON algorithm with patch-clamp records of two types of ion channel expressed by interstitial cells of Cajal, the pacemaker cells in the gastrointestinal tract. Transient-outward currents (also known as A-type currents) reflect the expression of voltage-dependent potassium

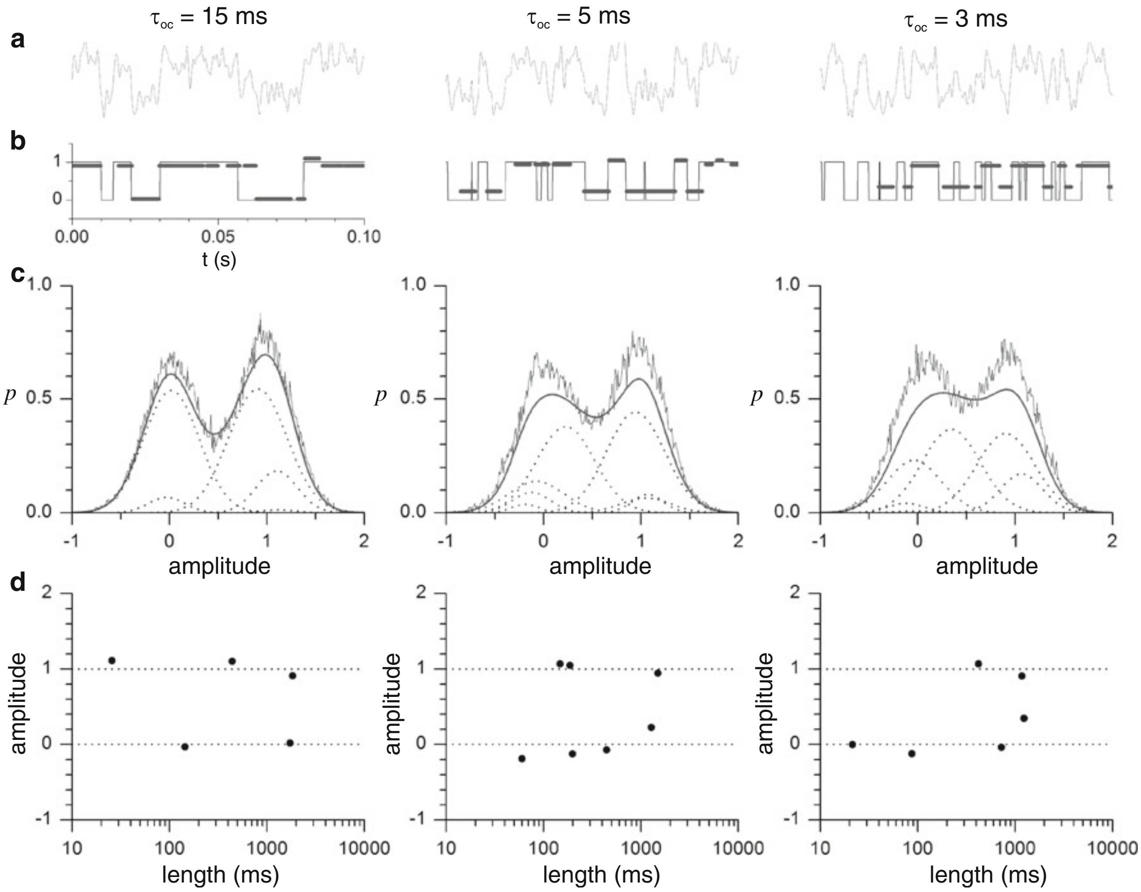

Fig. 4 Effect of fast molecule kinetics on the performance of the TON algorithm. Algorithm parameters were $N_{T(\mathrm{initiating})} = 50$ samples, $L_{T(\mathrm{extending})} = 2$ ms, $\sigma_T = 0.3$ and $\rho_T = 0.95$. Simulation parameters were $\tau_{SP} = 15$, 5 or 3 ms (*left to right*); $\sigma_{\mathrm{noise}} = 1.0$; and $\gamma = 1.0$. **a** Low-pass-filtered (0.5 kHz cutoff) record. **b** Record without added noise, with algorithm levels superimposed (*thick lines*). **c** Probability densities (pds). *Noisy thin line*, all-points pd ($\Delta_{\mathrm{bin}} = 0.01$); *dotted lines*, reconstructed pds for each algorithm level; *thick smooth line*, sum of reconstructed pds. **d** Plots of algorithm level amplitude against length. *Dotted lines* indicate the two molecule levels

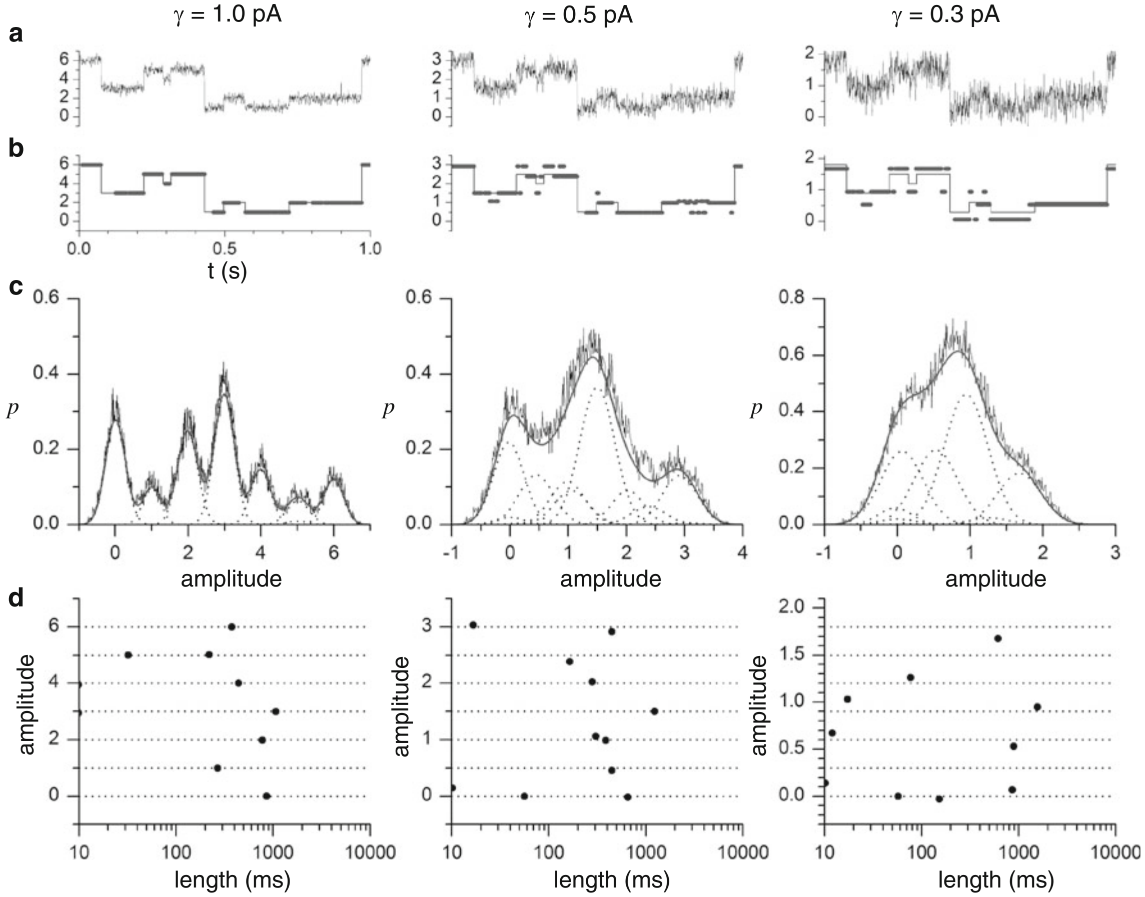

Fig. 5 Effect of varying signal-to-noise ratio on the performance of the TON algorithm with multiple levels. Algorithm parameters were $N_{\mathrm{T(initiating)}} = 100$ samples, $L_{\mathrm{T(extending)}} = 5$ ms, $\sigma_{\mathrm{T}} = 0.3$ and $\rho_{\mathrm{T}} = 0.95$. Simulation parameters were $\tau_{\mathrm{oc}} = 100$ ms, $\sigma_{\mathrm{noise}} = 1.0$ and $\gamma = 1.0$, 0.5 or 0.3 (*left to right*). **a** Low-pass-filtered (0.5 kHz cutoff) record. **b** Record without added noise, with algorithm levels superimposed (*thick lines*). **c** Probability densities (pds). *Noisy thin line*, all-points pd ($\Delta_{\mathrm{bin}} = 0.01$); *dotted lines*, reconstructed pds for each algorithm level; *thick smooth line*, sum of reconstructed pds. **d** Plots of algorithm level amplitude against length. *Dotted lines* indicate the molecule levels

channels of the Kv1–Kv5 family (Parsons and Huizinga 2010). Some of the recordings of these channels showed quite fast kinetics, so we used the same TON parameters as for the fast simulated molecules (Fig. 4) ($N_{\mathrm{T(initiating)}} = 50$ samples at 10 kHz sampling frequency, $L_{\mathrm{T(extending)}} = 2$ ms, $\rho_{\mathrm{T}} = 0.95, \sigma_{\mathrm{T}} = 0.3$ pA). The algorithm performed well for channels that displayed both fast and slow kinetics (upper and lower panels of Fig. 6a). The algorithm levels appeared appropriate in comparison to the channel records (Fig. 6a), and the reconstructed pds closely approximated the all-points pds (Fig. 6b). In plots of algorithm level amplitude against length, the longer levels approximated a level interval of 1.3 pA for both slow and fast channels (Fig. 6c). There were also a number of ghost levels spread across a range of amplitudes.

Maxichannels have a large conductance (over 200 pS) with numerous subconductance states (levels) and permeability to both anions and cations (Parsons et al. 2012; Parsons and Sanders 2008). A recording of a single maxichannel showed multiple subconductance states (Fig. 7a). Several peaks could be distinguished in the all-points pd (Fig. 7e), but many of these appeared too broad or asymmetric to be attributable to a single level. The TON algorithm broke these peaks up into several levels (α–θ) and the reconstructed pd approximated well the all-points pd, except for the γ level which was underestimated by the algorithm (Fig. 7e). When the channel record was overlapped with the algorithm levels, the algorithm levels appeared to be fair approximations of the channel levels (Fig. 7b–d).

Discussion

We have described an algorithm that integrates statistical CPD and gaussian mixture modeling. Some advantages spring from just this integration. The number of input parameters in TON is not negligible but would undoubtedly

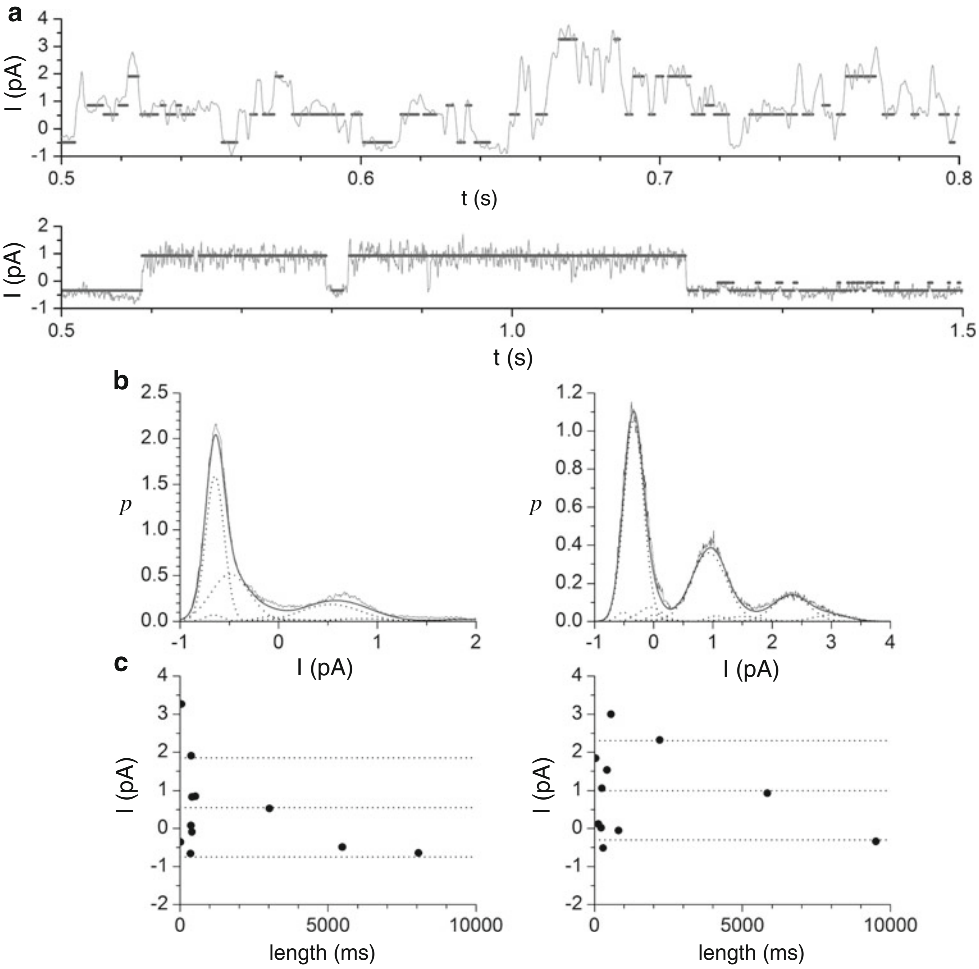

Fig. 6 Performance of the TON algorithm with two recordings of transient-outward potassium channels. Both records were 21.6 s long, with a 5 kHz sampling rate. Algorithm parameters were $N_{\mathrm{T(initiating)}} = 50$ samples, $L_{\mathrm{T(extending)}} = 2$ ms, $\sigma_{\mathrm{T}} = 0.3$ pA and $\rho_{\mathrm{T}} = 0.95$. **a** Low-pass-filtered (0.5 kHz cutoff) channel records with algorithm levels superimposed (*thick lines*). *Upper record* had four channels with relatively fast kinetics. *Lower record* also had four channels but with relatively slow kinetics. **b** Probability densities (pds). *Left panel* = *upper record*, *right panel* = *lower record*. *Noisy thin line*, all-points pd ($\Delta_{\mathrm{bin}} = 0.01$ pA); *dotted lines*, reconstructed pds for each algorithm level; *thick smooth line*, sum of reconstructed pds. **c** Plots of algorithm level amplitude against length. *Left panel* = *upper record*, *right panel* = *lower record*. *Dotted lines* are spaced 1.3 pA apart

be larger with separate clustering and CPD algorithms. Also the output of TON is integrated, unlike clustering and CPD in series. For instance, if we started with a CPD and then used the SP amplitudes for gaussian mixture modeling, would the resulting distributions be the same as if we performed mixture modeling directly on the samples of those SPs or of all the samples? Certainly not if some SPs have non-normal distributions because the CPD did not detect a CP (Schroder et al. 2004). With TON the match between level and SP distributions is inherent in the algorithm.

Purely as a CA, TON also has advantages. With optimization clustering the number of clusters has to be chosen a priori. If we have a kinetic model in mind (e.g., regularly spaced subconductances) or the recording appears very simple (e.g., just two levels), then this is not a problem. However, if this is not so, we have to assess the number of clusters either subjectively (by eyeballing the data) or iteratively through a number of choices, comparing the statistical plausibility of each outcome (Djuric et al. 1996; McManus et al. 1988; Sansom et al. 1989), which introduces further parameters. Also, with many mixture model algorithms, initial guesses have to be made for the parameters of each cluster's distribution if the algorithm is to settle on the optimal fit, rather than some local minima. Again, with a model or simple data this is not a big problem, but otherwise we have to resort to subjective methods or other algorithms which introduce further parameters. TON is unique for a mixture model algorithm in that it itself determines the number of clusters. This is a tremendous advantage with complex recordings such as for the maxichannel.

The disadvantage of TON as a gaussian mixture model algorithm is its production of ghost and other small

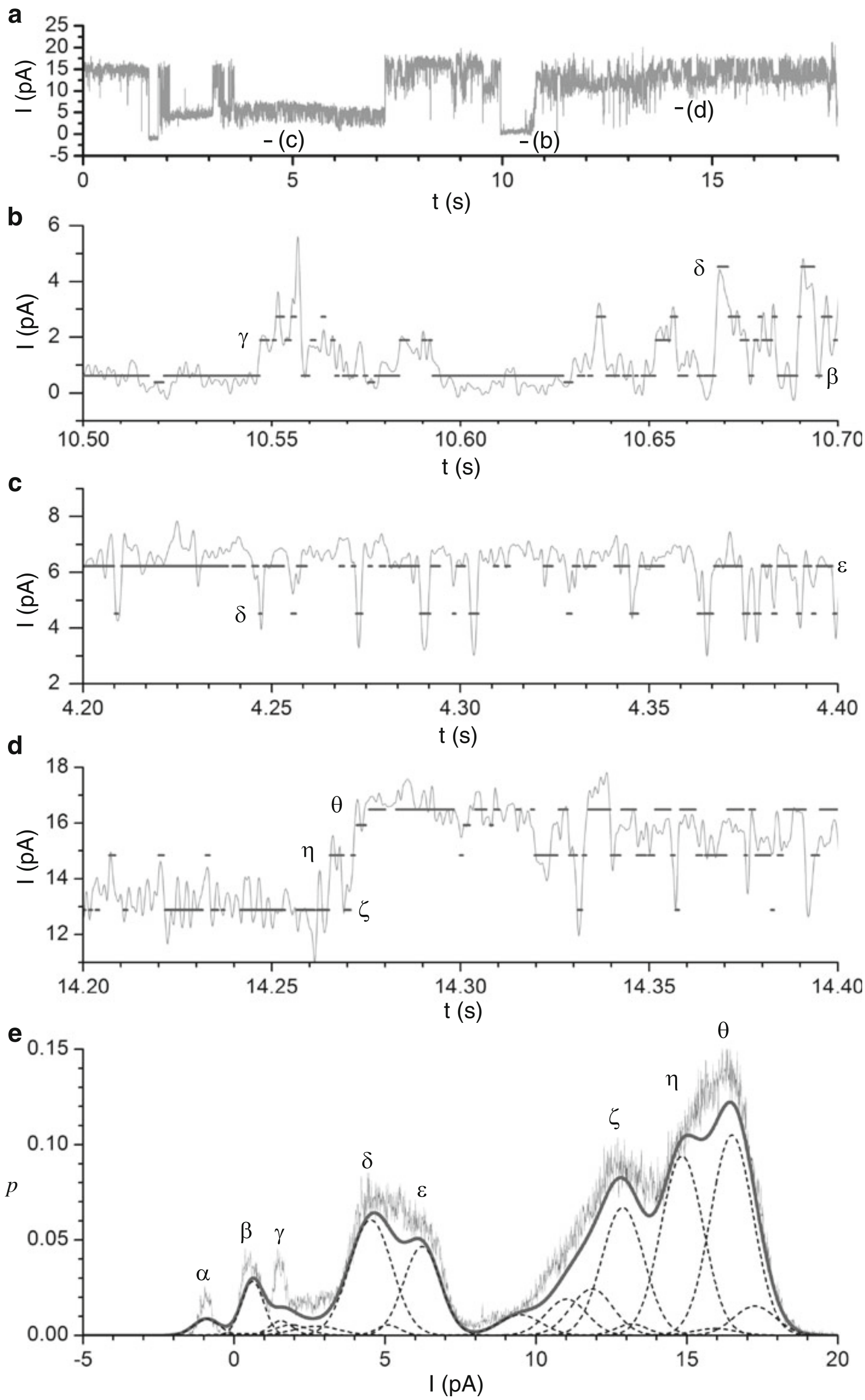

Fig. 7 Performance of the TON algorithm with a recording of a maxichannel. The record was 21.6 s, long with a 10 kHz sampling rate. Algorithm parameters were $N_{\mathrm{T(initiating)}} = 100$ samples, $L_{\mathrm{T(extending)}} = 1$ ms, $\sigma_{\mathrm{T}} = 0.7$ pA and $\rho_{\mathrm{T}} = 0.95$. **a** Low-pass-filtered (0.5 kHz cutoff) channel record. **b–d** Short sections of this record with algorithm levels superimposed (*thick lines*). **e** Probability densities (pds). *Noisy thin line*, all-points pd ($\Delta_{\mathrm{bin}} = 0.01$ pA); *dotted lines*, reconstructed pds for each algorithm level; *thick smooth line*, sum of reconstructed pds. Main levels are labeled α to θ

clusters. That is, TON has a tendency to overfit the data. Purely as a CPD, TON is not the best. The requirement for a large initiating SP to accurately establish a normal distribution (larger than a window required to accurately establish variance in other statistical CPDs) makes TON inadequate for data where transition rates between levels are quick compared to the sampling rate. TON can also be confused by fast transition rates when the distribution of two levels approximates normality (Fig. 4). A remedy for this may be possible by combining TON with another CPD—extension is stopped not just by non-normality but also by some other CP threshold.

The main objective of TON was to fuse analysis by CPD and mixture modeling into one simple algorithm that would work with complex single-molecule data. The algorithm does this admirably in the basic implementation presented. Undoubtedly, it could be improved with further modifications, for example, by introducing other CPD methods. However, here we wanted to demonstrate the basic concept of integration of CPD with mixture modeling.

Acknowledgement This work was supported by a CIHR operating grant (MOP12874) and an NSERC Discovery grant (386877).

References

Basseville M (1988) Detecting changes in signals and systems—a survey. Automatica 24:309–326

Basseville M, Benveniste A (1980) Detection of abrupt changes in signals and dynamical systems. Springer, New York

Carter NJ, Cross RA (2005) Mechanics of the kinesin step. Nature 435:308–312

Carter BC, Vershinin M, Gross SP (2008) A comparison of step-detection methods: how well can you do? Biophys J 94:306–319

Choi Y, Moody IS, Sims PC, Hunt SR, Corso BL, Perez I, Weiss GA, Collins PG (2012) Single-molecule lysozyme dynamics monitored by an electronic circuit. Science 335:319–324

Cochrane WG (1954) Methods for strengthening the common χ^2 tests. Biometrics 10:417–451

Djuric PM, Fwu JK, Jovanovic S, Lynn K (1996) On the processing of piecewise-constant signals by hierarchical models with application to single ion channel currents. In: 1996 IEEE international conference on acoustics, speech, and signal processing (ICASSP 96), vol 5, IEEE Service Center, Piscataway, NJ, pp 2762–2765

Draber S, Schultze R (1994) Detection of jumps in single-channel data containing subconductance levels. Biophys J 67:1404–1413

Everitt B (1980) Cluster analysis. Halsted, New York

Frantsuzov P, Kuno M, Janko B, Marcus RA (2008) Universal emission intermittency in quantum dots, nanorods and nanowires. Nat Phys 4:519–522

Frauenfelder H (2010) Conformational substates. In: The physics of proteins. Springer, New York, pp 97–112

Frauenfelder H, Parak F, Young RD (1988) Conformational substates in proteins. Annu Rev Biophys Biophys Chem 17:451–479

Jarque CM, Bera AK (1987) A test for normality of observations and regression residuals. Int Stat Rev 55:163–172

Kruger TP, Ilioaia C, van Grondelle R (2011) Fluorescence intermittency from the main plant light-harvesting complex: resolving shifts between intensity levels. J Phys Chem B 115:5071–5082

McManus OB, Weiss DS, Spivak CE, Blatz AL, Magleby KL (1988) Fractal models are inadequate for the kinetics of four different ion channels. Biophys J 54:859–870

Mejia YX, Mao H, Forde NR, Bustamante C (2008) Thermal probing of *E. coli* RNA polymerase off-pathway mechanisms. J Mol Biol 382:628–637

Moghaddamjoo A (1988) Step-like signal processing with distinct finite number of levels. IEEE Trans Ind Electron 35:489–493

Parsons SP, Huizinga JD (2010) Transient outward potassium current in ICC. Am J Physiol Gastrointest Liver Physiol 298:456–466

Parsons SP, Sanders KM (2008) An outwardly rectifying and deactivating chloride channel expressed by interstitial cells of cajal from the murine small intestine. J Membr Biol 221:123–132

Parsons SP, Kunze WA, Huizinga JD (2012) Maxi-channels recorded in situ from ICC and pericytes associated with the mouse myenteric plexus. Am J Physiol Cell Physiol 302:C1055–C1069

Pastushenko VP, Schindler H (1997) Level detection in ion channel records via idealization by statistical filtering and likelihood optimization. Philos Trans R Soc Lond B Biol Sci 352:39–51

Patlak JB (1988) Sodium channel subconductance levels measured with a new variance-mean analysis. J Gen Physiol 92:413–430

Patlak JB (1993) Measuring kinetics of complex single ion channel data using mean-variance histograms. Biophys J 65:29–42

Pebay P (2008) Formulas for robust, one-pass parallel computation of covariances and arbitrary-order statistical moments. Sandia National Laboratories, Albuquerque

Riessner T, Woelk F, Abshagen-Keunecke M, Caliebe A, Hansen UP (2002) A new level detector for ion channel analysis. J Membr Biol 189:105–118

Sansom MS, Ball FG, Kerry CJ, McGee R, Ramsey RL, Usherwood PN (1989) Markov, fractal, diffusion, and related models of ion channel gating. A comparison with experimental data from two ion channels. Biophys J 56:1229–1243

Schroder I, Huth T, Suitchmezian V, Jarosik J, Schnell S, Hansen UP (2004) Distributions-per-level: a means of testing level detectors and models of patch-clamp data. J Membr Biol 197:49–58

Schultze R, Draber S (1993) A nonlinear filter algorithm for the detection of jumps in patch-clamp data. J Membr Biol 132:41–52

Theodoridis S, Koutroumbas K (2009) Pattern recognition. Academic Press, San Diego

Thompson RJ, Nordeen MH, Howell KE, Caldwell JH (2002) A large-conductance anion channel of the Golgi complex. Biophys J 83:278–289

Torda AE, Vangunsteren WF (1994) Algorithms for clustering molecular-dynamics configurations. J Comput Chem 15:1331–1340

Tyerman SD, Terry BR, Findlay GP (1992) Multiple conductances in the large K^+ channel from *Chara corallina* shown by a transient analysis method. Biophys J 61:736–749

VanDongen AM (1996) A new algorithm for idealizing single ion channel data containing multiple unknown conductance levels. Biophys J 70:1303–1315

Welch BL (1947) The generalisation of "student's" problem when several different population variances are involved. Biometrika 34:28–35