analytical chemistry

Article

pubs.acs.org/ac

MOSAIC: A Modular Single-Molecule Analysis Interface for Decoding Multistate Nanopore Data

Jacob H. Forstater,[†,‡] Kyle Briggs,[§] Joseph W. F. Robertson,[†] Jessica Ettedgui,[†,‡] Olivier Marie-Rose,[‖] Canute Vaz,[†] John J. Kasianowicz,[†] Vincent Tabard-Cossa,[§] and Arvind Balijepalli*[,†]

[†]Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States

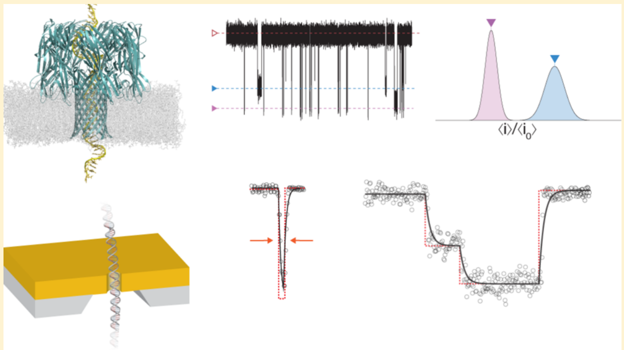
[‡]Department of Chemical Engineering, Columbia University, New York, New York 10027, United States

[§]Department of Physics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

[‖]Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States

S *Supporting Information*

ABSTRACT: Biological and solid-state nanometer-scale pores are the basis for numerous emerging analytical technologies for use in precision medicine. We developed Modular Single-Molecule Analysis Interface (MOSAIC), an open source analysis software that improves the accuracy and throughput of nanopore-based measurements. Two key algorithms are implemented: ADEPT, which uses a physical model of the nanopore system to characterize short-lived events that do not reach their steady-state current, and CUSUM+, a version of the cumulative sum statistical method optimized for longer events that do. We show that ADEPT detects previously unreported conductance states that occur as double-stranded DNA translocates through a 2.4 nm solid-state nanopore and reveals new interactions between short single-stranded DNA and the vestibule of a biological pore. These findings demonstrate the utility of MOSAIC and the ADEPT algorithm, and offer a new tool that can improve the analysis of nanopore-based measurements.



Protein and solid-state nanopores (Figure 1A) are the basis for single-molecule measurements of a variety of analytes including ions,[1,2] single-stranded RNA and DNA,[3−10] double-stranded DNA,[11,12] proteins,[13−19] synthetic polymers,[20−23] and metallic nanoparticles.[24,25] The method is conceptually simple. An electric potential applied across a nanopore that spans two electrically isolated chambers (filled with electrolyte solutions) results in an ionic current with a mean value $\langle i_0 \rangle$ (Figure 1B). Single molecules that reversibly partition into the pore cause a series of pulses or current blockades (Figure 1B). The change in pore conductance is caused by the volume exclusion of mobile ions from the pore[20,21] and interactions between the ions and the analyte.[5,21,26] The change in conductance[4,20,21,26,27] and the residence time of analytes in the pore[8,20,21] are used to estimate the analyte size,[20,21] effective charge,[21] and dipole moment.[28]

Analyte-induced events appear as single or multiple conductance state levels, arising from changes in the analyte conformation or interactions in the pore[9,11,20,21,27,29,30] (Figure 1C). Multiple conductance level events differ from the gating of ion channels, where the channel fluctuates between two states, open and closed.[31] These fluctuations are well characterized using hidden Markov models[32−34] and kinetic simulations.[34,35] On the other hand, several analysis techniques have been applied to analyze nanopore-based single-molecule data including threshold detection,[4,21] slope- or area-based techniques,[36,37] the cumulative sum (CUSUM) algorithm,[38] charge conservation,[39] and probabilistic machine-learning techniques.[20,40] While these approaches are effective when the residence times of analytes in the nanopore are long (compared to the characteristic time constant of the system), they are not useful for characterizing short-lived events. To more accurately characterize short events, we developed a technique that models the ionic current response with an equivalent electrical circuit.[26,41] This algorithm, when applied to the interaction of a polydisperse mixture of a synthetic polymer with the *Staphylococcus aureus* α-hemolysin (αHL) nanopore, recovered 18-fold more events per unit time at high measurement bandwidth (B = 100 kHz), reduced the constraints on data acquisition by permitting polymers to be separated at lower bandwidth (B = 10 kHz), and improved the resolving power in the low mass regime (to polymers with molecular weight ≈ 370 g/mol).

Received: September 20, 2016
Accepted: October 31, 2016
Published: October 31, 2016

DOI: 10.1021/acs.analchem.6b03725

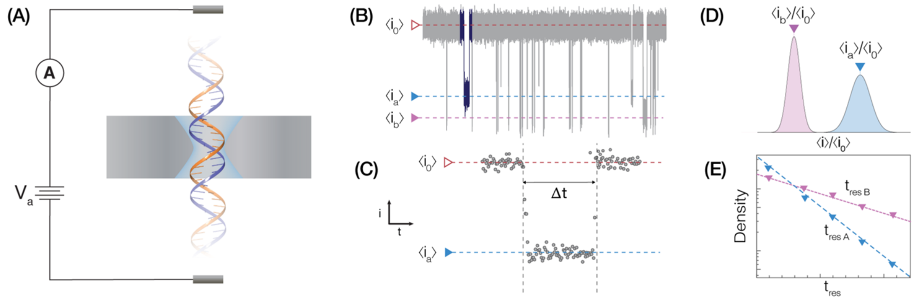


Figure 1. (A) Schematic illustration of DNA translocation through a solid-state nanopore. An electric potential applied across the pore produces an ionic current. (B) The partitioning of DNA into the pore causes well-defined current reductions with different mean current blockade amplitudes (e.g., $\langle i_a \rangle$ and $\langle i_b \rangle$). (C) A single level event is characterized by the ratio of the mean currents for the occupied and fully open pore ($\langle i_a \rangle / \langle i_0 \rangle$) and the level residence time (Δt). (D) A histogram illustrating the relative current blockade depth for two species obtained by analyzing the events. (E) Residence time distributions for the two blockade depth populations. The mean residence times are estimated from fits to the distributions.

Here, we describe Modular Single-Molecule Analysis Interface (MOSAIC), an improved data analysis tool for analyzing nanopore data. We implemented two algorithms in MOSAIC: ADEPT, the equivalent electrical circuit model described above,[26,41] and an improved version of CUSUM that is suitable for analyzing events with relatively long residence times in the pore. The software is extensible, and allows many commonly used data formats, signal conditioning, and data processing algorithms to be seamlessly integrated. Below, we demonstrate the features and utility of MOSAIC when applied to data measured with both biological and solid-state nanopores.

■ MATERIALS AND METHODS

Solid-State Nanopore Measurements. Solid-state nanopore data were reanalyzed from Briggs et al.[42] Briefly, nanopores were fabricated in 50 μm × 50 μm, 10 nm thick low-stress silicon nitride (SiN_x) TEM windows (Norcada, Canada) via the controlled breakdown (CBD) method[47] in a 1 M NaCl buffer (pH 10, 10 mM $NaHCO_3$). Nanopore measurements of double-stranded DNA (dsDNA) were performed in 3.6 M LiCl, 10 mM HEPES (pH 8) using highly purified 50 base pair (bp) dsDNA fragments (NoLimits no. SM1421, Life Technologies). Data were low-pass filtered at 100 kHz with a hardware 4-pole Bessel filter (Axopatch 200B) and digitized using a National Instruments USB-6351 DAQ card (Austin, TX) at a sampling rate, F_s = 500 kHz.

Biological Nanopore Measurements. Nanopore measurements were performed using quartz capillaries with a (≈ 1 μm diameter) aperture on one end,[26,43,44] within a custom polycarbonate test cell with ≈ 200 μL volume (Electronic Biosciences, San Diego, CA). Analytes were dissolved in the working buffer and added directly into the capillary or to the external test cell.

For single-stranded DNA measurements, the quartz capillary was filled with a 10−20 μM solution of different length homopolymeric adenosine dA_{20}, dA_{40}, or dA_{100} (Integrated DNA Technologies, Coralville, IA) dissolved in 1 M NaCl, 1× TE buffer (10 mM Tris, 1 mM EDTA in DNase-free water, titrated to pH 7.2 with 3 M HCl).

For poly(ethylene glycol) measurements (PEG), data from two previous studies that span a wide range of polymer sizes were combined.[22,26] In both cases, the capillary was filled with a solution containing a combination of polydisperse PEG (Fluka, Switzerland) and a highly purified calibration standard (Polypure, Oslo, Norway), dissolved in 4 M KCl (Sigma-Aldrich, St. Louis, MO), buffered with 10 mM Tris (Schwarz/Mann Biotech, Cleveland, OH) and titrated to pH 7.2 with 3 M citric acid. The two different solutions were as follows: (a) 20 μM PEG-600 (MW_{avg} = 600 g/mol), 40 μM PEG-400 (MW_{avg} = 400 g/mol) and 2 μM purified PEG-502 (M_w = 502 g/mol) or (b) 30 μM PEG-1000 (MW_{avg}= 1000 g/mol), 30 μM PEG-1500 (MW_{avg} = 1500 g/mol), and 1 μM purified PEG-1251 (M_w = 1251 g/mol).

Planar lipid bilayers were formed across the quartz capillary aperture using a 10 mg/mL solution of 1,2 diphytanolyl-*sn*-glycero-3-phosphatidylcholine (DPhyPC; Avanti Polar Lipids, Alabaster, AL) in *n*-decane (Sigma-Aldrich).[26] Subsequently, wild-type *S. aureus* α-Hemolysin (αHL) was introduced to the test cell by adding a solution containing either ≈250 ng of monomeric αHL (List Biological Laboratories, Campbell, CA) or ≈2.5 ng of purified preformed heptamers. To facilitate channel incorporation, the bilayer was thinned and enlarged by applying a transmembrane potential of ≈300 mV and a static back pressure within the capillary. Following the insertion of a single channel, the static pressure was reduced and the voltage decreased to the value used for the measurement to prevent further channel incorporation.

The potential was applied across the membrane by a pair of Ag/AgCl electrodes. Immediately prior to use, the electrode placed within in the capillary was prepared by abrading an Ag wire (Alfa Aesar) with 600 grit sandpaper and soaking it in bleach for ≈10 min. The external electrode in the test cell bath was a 2 mm Ag/AgCl disk electrode (E202, In Vivo Metric). Data were acquired with a custom high-impedance amplifier system (Electronic BioSciences, San Diego, CA) and conditioned with a low-pass antialiasing filter. The analog signal was digitized by a National Instruments PCI-6120 DAQ card with a sampling rate (F_s) of 1 MHz, further conditioned using a software-based 8-pole low pass Bessel filter with a cutoff frequency of 100 kHz and resampled at 500 kHz.

Data Processing and Analysis. Nanopore data were processed using a Python based program (MOSAIC) developed in-house. The software implements the ADEPT and CUSUM+ algorithms, which are described below. The compiled program and source code are freely available at https://pages.nist.gov/mosaic/. MOSAIC consists of a modular data processing pipeline which allows users to analyze ionic current data from single-molecule nanopore experiments. The software is designed using object-oriented principles, which ensures that modules remain interoperable. This also makes it straightforward to implement new features into the software, such as alternative analysis algorithms or custom data formats.

In many cases, users can interact with MOSAIC using a front end graphical user interface.

Algorithms. MOSAIC consists of a pipeline with five modules: (i) load data, (ii) optional signal conditioning and filtering, (iii) event detection, (iv) event analysis, and (v) results storage. A detailed description of the software architecture is presented in the Supporting Information.

In this section, we discuss two algorithms implemented in MOSAIC: (i) CUSUM+, an improved version of the CUSUM algorithm[38] that provides robust statistical analysis of events which converge to a steady-state, and (ii) ADEPT, an implementation of a previously developed theory[26,41] that uses a physical model of the nanopore system to accurately characterize very short events that do not approach a steady-state ionic current.

Cumulative Sum Analysis (CUSUM+). Cumulative Sum (CUSUM) is a commonly used method to detect step-like changes in time-series data[45] but was only recently implemented in nanopore analysis.[38] It assumes that the interaction of an analyte with the pore causes a series of instantaneous changes in the ionic current from its baseline value (defined as states and well-approximated by step functions[45]), and the ionic current noise follows a known distribution (e.g., Gaussian). A statistical test identifies when the current level changes. The instantaneous log-likelihood ratios of sequential data points for both positive and negative step changes are calculated. The positive values of these ratios are independently summed and a negative log-likelihood resets the sum to zero. These form a two-sided decision function, which detects level changes that correspond to either an increase or decrease in the current level. A new state is identified when one of the decision functions exceeds a threshold determined automatically by the software. The locations of state changes are determined from the minima of related functions,[45] and the mean ionic current between sequential states is calculated and used to determine the local blockade depth, defined as ratio of the ionic current when the pore is occupied to that of the open pore ($\langle i\rangle/\langle i_0\rangle$, Figure 1C).

We implemented an improved version of the CUSUM algorithm in MOSAIC (CUSUM+), which is less sensitive to artifacts that can be falsely identified as a state change. This is achieved by specifying a minimum time between successive triggers (to exclude transients from the state change detection and blockade depth calculations) and by requiring that identified state levels differ by a minimum value (corresponding to a physically significant change). The efficiency is improved by eliminating or reducing redundant computations (e.g., maintaining running calculations of the mean and variance).

Adaptive Time-Series Analysis (ADEPT). For very long events ($>5\tau$, Figure 2 left), the blockade depth is easily estimated (e.g., with CUSUM+). However, that process fails for short-lived events ($<5\tau$, Figure 2 right) that do not reach a steady-state mean value. In this case, the blockade depths are estimated by fitting the data to an electrical circuit model of the nanopore (implemented as ADEPT[26,41] in MOSAIC). The algorithm assumes a molecule partitioning into the nanopore instantaneously increases the nanopore resistance (R_p) by ΔR. However, the system capacitance causes the ionic current change to occur over a finite time. For a constant applied voltage, V_a, the predicted ionic current is $i(t) = i_0 - \beta(1 - e^{-t/\tau})$, where $i_0 = \frac{V_a}{R_s + R_p}$, $\beta = \frac{V_a \Delta R}{(R_s + R_p + \Delta R)(R_s + R_p)}$, and $\tau = \frac{C_m R_s (R_p + \Delta R)}{R_s + R_p + \Delta R}$. The difference between the time constants (τ) leading up to and following an event are much shorter than typical sampling rates (see the Supporting Information for the calculation). We therefore use a single fit parameter for τ, which reduces the degrees of freedom. There is an option to override this constraint.

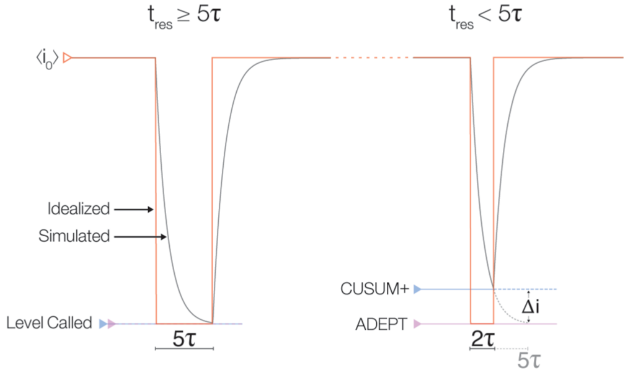


Figure 2. ADEPT and CUSUM+ analysis applied to a simulated nanopore measurement (gray). Two events with the same current blockade (red) but different residence times (t_{res}), with respect to the system characteristic relaxation time (τ) are shown. (Left) For a long event ($t_{res} \geq 5\tau$), the ionic current converges close to its steady-state value, and the current levels estimated by ADEPT and CUSUM+ are equivalent. (Right) For short events (e.g., $t_{res} \approx 2\tau$), the current does not reach the steady-state value of the idealized pulse (red). In this case, CUSUM+ and other algorithms used in nanopore analysis systematically underestimate the steady-state current (blue) by an amount Δi (gray; dashed). In contrast, the physical model underlying ADEPT allows the algorithm to accurately estimate an event's steady-state current.

RESULTS AND DISCUSSION

Analysis of Short dsDNA Fragments Measured with SiN$_x$ Nanopores. We compared the results of CUSUM+ and ADEPT on measurements of 50 base pair (bp) double-stranded DNA (dsDNA) translocating through a ≈2.4 nm diameter SiN$_x$ nanopore (≈2800 events).[42] At an applied potential of 400 mV, the mean residence time of dsDNA in the pore is ≈440 μs, more than an order of magnitude longer than the characteristic time constant of the system (τ = 10 μs; B = 100 kHz). Both algorithms produce two distinct peaks in the blockade depth histogram ($\langle i\rangle/\langle i_0\rangle$ = 0.070 ± 0.001 and 0.488 ± 0.004). Peak positions were obtained using an error-weighted Gaussian fit and are reported with an expanded uncertainty, k = 2 (see Supporting Information for a full listing of the analysis and fit parameters). The leftmost peak corresponds to DNA translocation, whereas the rightmost peak is likely due to the helical structure of dsDNA unwinding to transition from the B-form to the S-form dsDNA,[46,47] where the chain elongates by 1.7 fold because of the strong electric field gradient across the pore.[42]

At 800 mV, the blockade depth histogram produced by ADEPT has two overlapping peaks (Figure 3B) consisting of a narrow peak ($\langle i\rangle/\langle i_0\rangle$ = 0.080 ± 0.002), the expected location for B-form dsDNA,[46] and a broader peak ($\langle i\rangle/\langle i_0\rangle$= 0.138 ± 0.002). The latter is comprised of short, single-level events, which likely result from transient interactions between the dsDNA and the access region outside the pore.[48] This was not accurately identified in our previous analysis[42] (*see below*). A third, low amplitude, broad peak is visible at $\langle i\rangle/\langle i_0\rangle$ = 0.226 ±

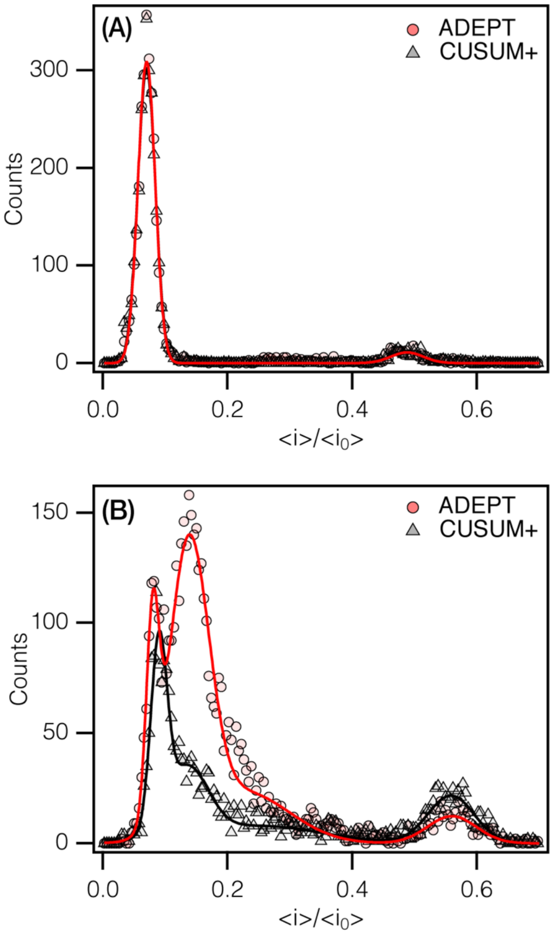


Figure 3. Blockade depth histograms for a 50 base pair double-stranded DNA measured with a 2.4 nm diameter SiN_x nanopore.[42] (A) At 400 mV, both CUSUM+ (gray), and ADEPT (red) are in excellent agreement. (B) At 800 mV, the mean residence time decreases to ≈36 μs (estimated from ADEPT analysis). The analyses by ADEPT and CUSUM+ are markedly different.

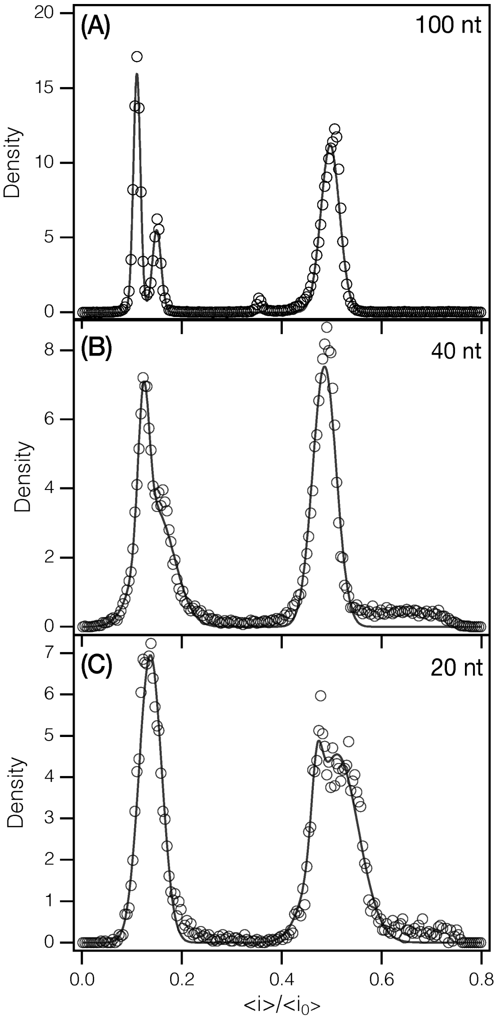


Figure 4. Normalized blockade depth histograms for (A) dA_{100}, (B) dA_{40}, and (C) dA_{20} single-stranded DNA interacting with the αHL nanopore estimated using the ADEPT algorithm. The applied potential is V = 140 mV and the polynucleotides are added to the cis side of the pore.[49]

0.018. It is probable that this peak is associated with these access-region interactions. Alternatively, it is possible that these events could represent partial unwinding of the dsDNA secondary structure at forces below the B−S stretching transition threshold.[47] As seen in Figure 3B, CUSUM+ also detects the first peak, $\langle i\rangle/\langle i_0\rangle$ = 0.090 ± 0.002 (albeit slightly shifted compared to the ADEPT value). However, it only characterizes ≈20% of the events in the second peak ($\langle i\rangle/\langle i_0\rangle$ = 0.134 ± 0.010) where the mean residence time of the events ($\langle t_{res}\rangle$ = 47 ± 4 μs) is less than 5τ. In addition, CUSUM+ misses most of the events from the third peak detected with ADEPT. While CUSUM+ could be allowed to characterize events with lifetimes less than 5τ, it will underestimate the blockade depth ratios (see the Supporting Information, Figure S3). Both algorithms identify the fourth peak ($\langle i\rangle/\langle i_0\rangle \approx 0.56$), which arises from the stretching transition of dsDNA noted above (S-form of dsDNA[46,47]). CUSUM+ recovers more events here than ADEPT, which utilizes a fitting routine that may not converge for some very long events (t_{res} > 25 ms; 50 000 points; F_s = 500 kHz). Clearly, the results would be improved if ADEPT is used for relatively short-lived events and CUSUM+ is used on events with residence times >5τ (Supporting Information, Figure S4). This functionality will be implemented in a future version of MOSAIC.

While both CUSUM+ and ADEPT produce comparable results for events with residence times >5τ, CUSUM+'s statistical approach is on average ≈10× faster than the Levenberg−Marquardt least-squares fitting used in ADEPT.[48] Therefore, CUSUM+ is preferred for events with mean residence times considerably longer than the recovery time of the system (≫5τ). Furthermore, the processing time per event for each algorithm scales linearly with the residence time, and therefore the number of data points in an event (see Supporting Information, Figure S5).

Analysis of Single-Stranded DNA Oligonucleotides with ADEPT. We use ADEPT to determine the blockade depth ratio histograms for three different length single-stranded DNA (ssDNA) homopolymers (dA_{100}, dA_{40}, and dA_{20}) entering an αHL nanopore from the cis side.[49] We consider events with up to 6 discrete states, with each state containing at least 5 data points (t_{res} > 10 μs, F_s = 500 kHz). Events are partitioned from the time series data with a thresholding algorithm that identifies when the current deviates by more

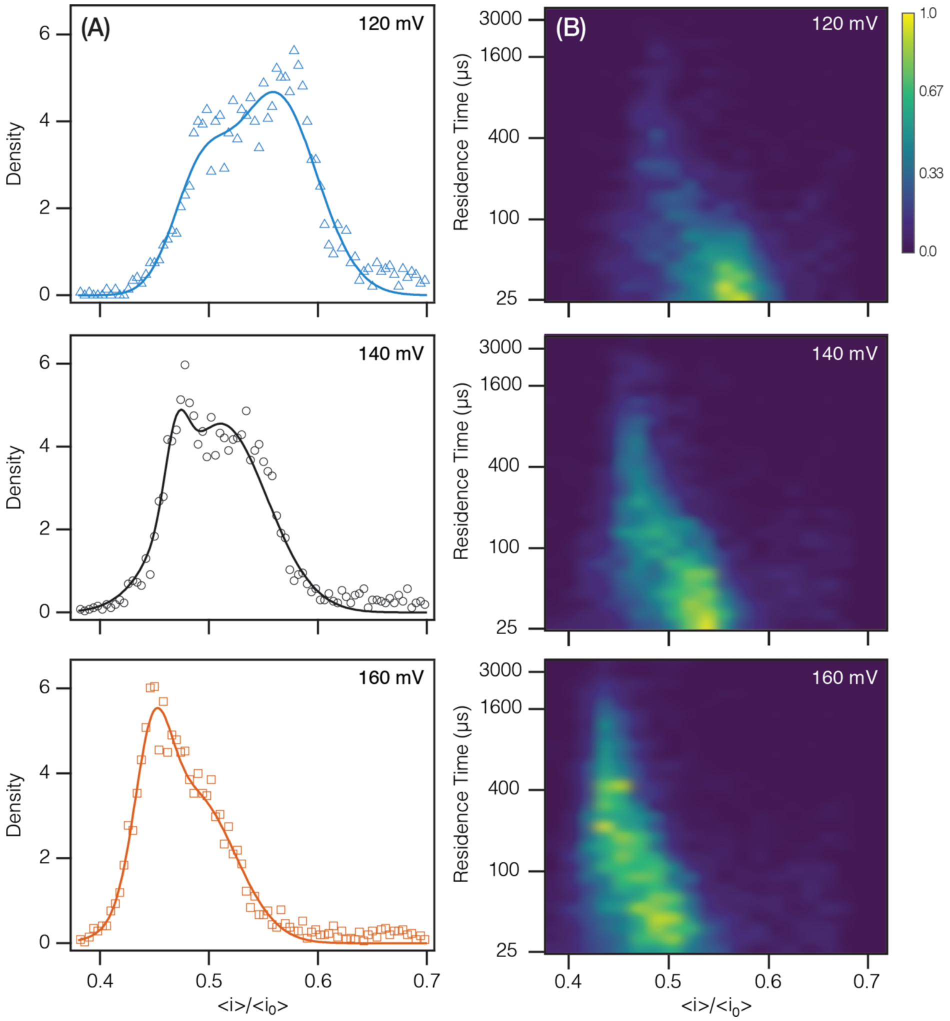


Figure 5. Voltage dependence of dA_{20} shallow blockade depth and residence time. (a) The normalized blockade depth histogram as a function of voltage yields peaks with changing morphology. (B) Joint residence time-blockade depth distribution (log−linear) as a function of voltage. Z-scale (color) was normalized and smoothed using a Gaussian interpolation.

than 5 standard deviations from the mean open channel current. A complete listing of the analysis parameters is shown in Supporting Information, Table S3.

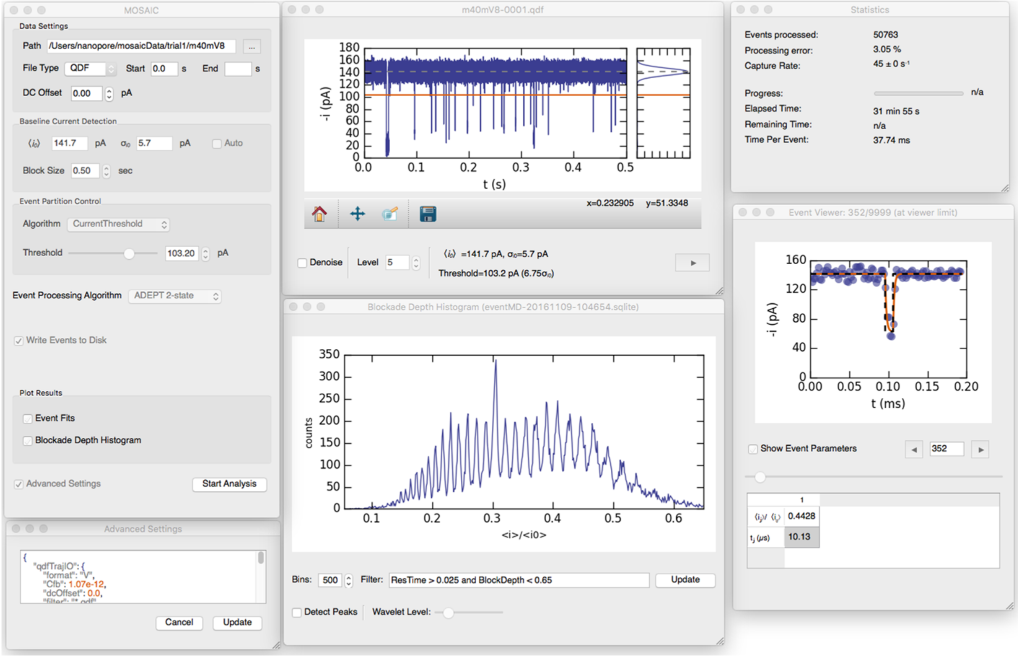
The current blockades observed with poly(dA) appear as either one level (a shallow or deep blockade) or two levels (a shallow blockade followed by a deeper one[8,50,51]). Approximately 60% of the events have two levels. As shown in the blockade depth ratio histograms for dA_{100}, dA_{40}, and dA_{20} (Figure 4), these two observed levels are comprised of several different states. Figure 4A shows that the blockade depth histogram for dA_{100} has three peaks, $\langle i\rangle/\langle i_0\rangle = (0.11 \pm 0.03)$, (0.15 ± 0.05), and (0.50 ± 0.05). The first two peaks (denoted $\langle i\rangle/\langle i_0\rangle_{3'}$ and $\langle i\rangle/\langle i_0\rangle_{5'}$) are consistent with the dependence of the blockade depth on the orientation of the leading end of the DNA (3′ vs 5′) entering the pore.[4,52,53] The location of these two peaks agrees with previous measurements of dA_{100} where two highly overlapping peaks were observed at these locations.[6] In contrast to earlier measurements we resolve the 3′ and 5′ events with a separation better than 3σ.

The differences between the 3′ and 5′ blockade depth peaks (Figure 4A, two leftmost peaks) are progressively more difficult to discern for the shorter polynucleotides (Figure 4B,C). The amplitude of the 5′ peak decreases substantially for dA_{40} (Figure 4B) and is not resolved for dA_{20} (Figure 4C). These results are likely due to the lower probability of the 5′-end entering the pore and the decreasing residence time of shorter ssDNA molecules.[4,53]

Interestingly, the shallow blockade level ($\langle i\rangle/\langle i_0\rangle \approx 0.5$) is characterized by a single peak for dA_{100} and dA_{40} and two peaks for dA_{20}. Previous studies have either not reported this peak[6] or only noted it for molecules as short as dA_{50}.[50] Furthermore, the sharp decrease in residence time associated with shorter polymers complicates their analysis and has thus far limited the analysis of polynucleotides as short as dA_{20}.

The algorithms within MOSAIC improve the characterization of short polynucleotides. Figure 5 shows the voltage-dependent behavior of the dA_{20} shallow blockade peaks and their residence time distributions. The shallow blockade depth distributions are qualitatively different than those measured for longer polymers (Figure 4), as noted previously. Furthermore, we observe a change in the morphology of the peaks with increasing voltage as seen in Figure 5A. In particular, increasing the magnitude of the applied potential: (i) shifts the peaks to smaller $\langle i\rangle/\langle i_0\rangle$ values, i.e., the polynucleotide blocks more

DOI: 10.1021/acs.analchem.6b03725
Anal. Chem. 2016, 88, 11900−11907



MOSAIC

Data Settings

Path /Users/nanopore/mosaicData/trial1/m40mV8 ...

File Type QDF Start 0.0 s End s

DC Offset 0.00 pA

Baseline Current Detection

⟨i₀⟩ 141.7 pA σ₀ 5.7 pA □ Auto

Block Size 0.50 sec

Event Partition Control

Algorithm CurrentThreshold

Threshold 103.20 pA

Event Processing Algorithm ADEPT 2-state

☑ Write Events to Disk

Plot Results

□ Event Fits

□ Blockade Depth Histogram

☑ Advanced Settings

Start Analysis

Advanced Settings

```
{
  "qdfTrajIO": {
    "format": "V",
    "Cfb": 1.07e-12,
    "dcOffset": 0.0,
    "filter": "*.qdf"
```

Cancel Update

m40mV8-0001.qdf

x=0.232905 y=51.3348

□ Denoise Level 5

⟨i₀⟩ =141.7 pA, σ₀=5.7 pA

Threshold=103.2 pA (6.75σ₀)

Blockade Depth Histogram (eventMD-20161109-104654.sqlite)

Bins: 500 Filter: ResTime > 0.025 and BlockDepth < 0.65 Update

□ Detect Peaks Wavelet Level:

Statistics

Events processed:	50763
Processing error:	3.05 %
Capture Rate:	45 ± 0 s-1
Progress:	n/a
Elapsed Time:	31 min 55 s
Remaining Time:	n/a
Time Per Event:	37.74 ms

Event Viewer: 352/9999 (at viewer limit)

□ Show Event Parameters ◀ 352 ▶

	1
⟨i₁⟩/ ⟨i₀⟩	0.4428
t₁ (μs)	10.13

Figure 6. Graphical user interface (GUI) used to setup and run an analysis in MOSAIC. A main settings window is used to configure the analysis. The results of an analysis are displayed in real-time in the adjacent panels. Completed analysis can also be reexamined within the GUI by opening the saved database files.

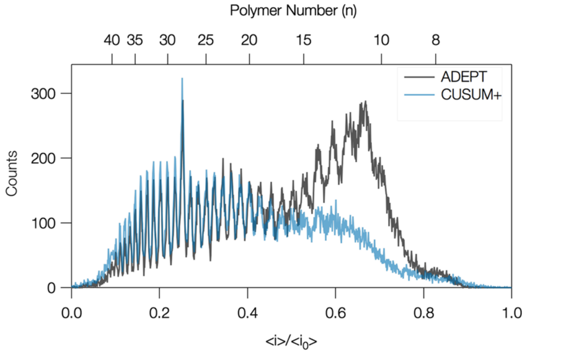


Figure 7. Comparison of ADEPT (gray) and CUSUM+ (blue) analyses of a polydisperse polyethylene glycol (PEG) solution measured with an αHL nanopore. Blockade depth histogram of events recovered by each algorithm show that the number of events recovered by CUSUM+ decreases sharply for small polymers (n < 20) that exhibit fast residence times (<5τ) in the pore.

current (Figure 5A), (ii) increases the residence times (Figure 5B) (in contrast to the mean residence time of the deep blockades in Figure 4 that are associated with translocation),[4,5] and (iii) increases the number of events observed per unit time (capture rate) (see the Supporting Information, Figure S7). Moreover, with increasing voltage, the shallow blockades were more likely to exhibit two states with the shallow blockade preceding a deep blockade.

The above results strongly suggest that the shallow blockade peaks correspond to dA_{20} interacting with the vestibule but not translocating through the pore. Furthermore, the observed increase in the residence time of the shallow blockade (Figure 5B) with voltage suggests that the change in the leftmost peak amplitude in Figure 5A is likely due to interactions between the analyte and different regions of the vestibule, rather than a loss of signal. Interestingly, of the three measured analytes (dA_{100}, dA_{40}, dA_{20}), this voltage-dependent change in peak structure was only observed with dA_{20}, indicating that the phenomenon may be length-dependent.[49,54]

Example of Using MOSAIC: Single Molecule Mass Spectrometry with a Biological Nanopore. We show a typical analysis using MOSAIC's graphical user interface. Specifically, we use both CUSUM+ and ADEPT to separate monomers in polydisperse PEG samples with an αHL nanopore.[20−22,26,55,56] Previous studies showed that the blockade depth ratio (⟨i⟩/⟨i_0⟩) and mean residence time of these events scale monotonically with polymer size.

Analysis of the PEG data is set up using the GUI shown in Figure 6 and is configured using drop down menus. After selecting a data source (MOSAIC accepts most common electrophysiology data formats: Axon ABF, QUB QDF, as well as raw binary and comma separated value, CSV, data), a segment of the time series is displayed, which assists in determining the mean open channel current (<i_0>), noise (σ_{i0}) and threshold values for preliminary event identification.

A key feature of MOSAIC is the ability to integrate custom algorithms into the processing pipeline. Within the GUI, the user can select the analysis algorithm. The PEG data were analyzed independently using both the ADEPT and CUSUM+ algorithms. The blockade depth histogram of the events and the processing statistics are presented in real time. Fits of the physical model (ADEPT) or detected states (CUSUM+) of individual analyzed events are also displayed to monitor the progress and quality of the analysis. The results are stored in a SQLite database (or can be exported as a CSV file from within the GUI) for further analysis.

This example further illustrates the differences between the CUSUM+ and ADEPT algorithms. Only events that deviate from the open channel current baseline by at least 2.7σ were analyzed. Events shorter than 5τ were excluded from the CUSUM+ analysis (the default value when running CUSUM+

from the GUI). On the other hand, when using ADEPT, we excluded events shorter than 2.5τ (25 μs) to minimize fitting errors (set with the *Advanced Settings* dialogue in the GUI).

Both algorithms produce well-resolved peaks for PEGs larger than 17-mers (Figure 7 and Supporting Information, Figure S8). For smaller PEGs that have shorter mean residence times, CUSUM+ shows a single broad peak whereas ADEPT easily identifies additional individual species. Here, CUSUM+ recovers significantly fewer events than ADEPT because only a fraction of the total events (those with $t_{\text{res}} > 5\tau$) are considered. This effect is particularly significant because it amounts to examining the tail of the exponentially distributed lifetimes.[20,21]

The histograms in Figure 7 are fit to a sum of Lorentzian functions using Igor Pro 6.3 (Wavemetrics Inc., Portland, OR). To directly compare the blockade depth histograms, we use the peak positions from the ADEPT data set as initial guesses for a fit of the CUSUM+-derived blockade depth distribution (Figure 7, gray). As expected, for both algorithms, where the peaks are resolved, the peak positions are in good agreement. For PEGs ($n < 17$), the signal-to-noise ratio of the CUSUM+ peaks is lower than those recovered by ADEPT (Supporting Information, Figure S8), consistent with the lower number of events recovered by CUSUM+ in this region.

■ CONCLUSIONS

We developed a new open source platform for the analysis of single-molecule data (MOSAIC) and implemented two robust optimized algorithms (ADEPT and CUSUM+) for biological or solid-state nanopore measurements. When applied to dsDNA measurements with a 2.4 nm SiN_x nanopore, MOSAIC found previously undetected states most likely arising from the transient interactions between dsDNA and the access region of the solid-state pore. Additionally, when measuring short oligonucleotides poly(dA)$_n$, MOSAIC accurately analyzed events with residence times $<5\tau$, thereby characterizing previously unreported interactions of dA_{20} with the αHL nanopore. Such analysis can be used to provide greater insight into the underlying physics of analyte-nanopore interactions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b03725.

> Additional analysis of ssDNA and dsDNA blockade depth histograms and extended discussion of MOSAIC software contents and structure (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: arvind.balijepalli@nist.gov. Phone: (301) 975-3526.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by Grant R01HG007415 from the National Human Genome Research Initiative (J.J.K) and by the Natural Sciences and Engineering Research Council of Canada (NSERC). K.B. acknowledges the financial support provided by NSERC and the Vanier CGS program for postgraduate fellowships. We gratefully acknowledge the Ju Laboratory (Columbia University) for providing the preheptamerized alpha-hemolysin used in some of these studies. We thank the numerous research groups who provided feedback on the software during its development.

■ REFERENCES

(1) Bezrukov, S. M.; Kasianowicz, J. J. *Phys. Rev. Lett.* **1993**, *70* (15), 2352−2355.

(2) Kasianowicz, J. J.; Bezrukov, S. M. *Biophys. J.* **1995**, *69* (1), 94−105.

(3) Zahid, O. K.; Wang, F.; Ruzicka, J. A.; Taylor, E. W.; Hall, A. R. *Nano Lett.* **2016**, *16*, 2033.

(4) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (24), 13770−13773.

(5) Henrickson, S. E.; Misakian, M.; Robertson, B.; Kasianowicz, J. J. *Phys. Rev. Lett.* **2000**, *85* (14), 3057−3060.

(6) Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (3), 1079−1084.

(7) Clarke, J.; Wu, H.-C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H. *Nat. Nanotechnol.* **2009**, *4*, 265−270.

(8) Henrickson, S. E.; DiMarzio, E. A.; Wang, Q.; Stanford, V. M.; Kasianowicz, J. J. *J. Chem. Phys.* **2010**, *132* (13), 135101.

(9) Manrao, E. A.; Derrington, I. M.; Laszlo, A. H.; Langford, K. W.; Hopper, M. K.; Gillgren, N.; Pavlenok, M.; Niederweis, M.; Gundlach, J. H. *Nat. Biotechnol.* **2012**, *30* (4), 349−353.

(10) Cherf, G. M.; Lieberman, K. R.; Rashid, H.; Lam, C. E.; Karplus, K.; Akeson, M. *Nat. Biotechnol.* **2012**, *30* (4), 344−348.

(11) Merchant, C. A.; Healy, K.; Wanunu, M.; Ray, V.; Peterman, N.; Bartel, J.; Fischbein, M. D.; Venta, K.; Luo, Z.; Johnson, A. T. C.; Drndic, M. *Nano Lett.* **2010**, *10* (8), 2915−2921.

(12) Rodríguez-Manzo, J. A.; Puster, M.; Nicolaï, A.; Meunier, V.; Drndic, M. *ACS Nano* **2015**, *9*, 6555−6564.

(13) Marshall, M. M.; Ruzicka, J.; Zahid, O. K.; Henrich, V. C.; Taylor, E. W.; Hall, A. R. *Langmuir* **2015**, *31* (15), 4582−4588.

(14) Kasianowicz, J. J.; Henrickson, S. E.; Weetall, H. H.; Robertson, B. *Anal. Chem.* **2001**, *73* (10), 2268−2272.

(15) Oukhaled, G.; Mathé, J.; Biance, A. L.; Bacri, L.; Betton, J. M.; Lairez, D.; Pelta, J.; Auvray, L. *Phys. Rev. Lett.* **2007**, *98* (15), 158101.

(16) Oukhaled, A.; Cressiot, B.; Bacri, L.; Pastoriza-Gallego, M.; Betton, J.-M.; Bourhis, E.; Jede, R.; Gierak, J.; Auvray, L.; Pelta, J. *ACS Nano* **2011**, *5* (5), 3628−3638.

(17) Pastoriza-Gallego, M.; Rabah, L.; Gibrat, G.; Thiebot, B.; van der Goot, F. G.; Auvray, L.; Betton, J.-M.; Pelta, J. *J. Am. Chem. Soc.* **2011**, *133* (9), 2923−2931.

(18) Rotem, D.; Jayasinghe, L.; Salichou, M.; Bayley, H. *J. Am. Chem. Soc.* **2012**, *134* (5), 2781−2787.

(19) Larkin, J.; Henley, R. Y.; Muthukumar, M.; Rosenstein, J. K.; Wanunu, M. *Biophys. J.* **2014**, *106* (3), 696−704.

(20) Robertson, J. W. F.; Rodrigues, C. G.; Stanford, V. M.; Rubinson, K. A.; Krasilnikov, O. V.; Kasianowicz, J. J. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (20), 8207−8211.

(21) Reiner, J. E.; Kasianowicz, J. J.; Nablo, B. J.; Robertson, J. W. F. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (27), 12080−12085.

(22) Balijepalli, A.; Robertson, J. W. F.; Reiner, J. E.; Kasianowicz, J. J.; Pastor, R. W. *J. Am. Chem. Soc.* **2013**, *135* (18), 7064−7072.

(23) Baaken, G.; Halimeh, I.; Bacri, L.; Pelta, J.; Oukhaled, A.; Behrends, J. C. *ACS Nano* **2015**, *9* (6), 6443−6449.
(24) Angevine, C. E.; Chavis, A. E.; Kothalawala, N.; Dass, A.; Reiner, J. E. *Anal. Chem.* **2014**, *86* (22), 11077−11085.
(25) Ettedgui, J.; Kasianowicz, J. J.; Balijepalli, A. *J. Am. Chem. Soc.* **2016**, *138* (23), 7228−7231.
(26) Balijepalli, A.; Ettedgui, J.; Cornio, A. T.; Robertson, J. W. F.; Cheung, K. P.; Kasianowicz, J. J.; Vaz, C. *ACS Nano* **2014**, *8* (2), 1547−1553.
(27) Movileanu, L.; Cheley, S.; Bayley, H. *Biophys. J.* **2003**, *85* (2), 897−910.
(28) Yusko, E. C.; Bruhn, B. R.; Eggenberger, O. *arXiv.org.* **2015**, arXiv:1510.01935.
(29) Schneider, G. F.; Kowalczyk, S. W.; Calado, V. E.; Pandraud, G.; Zandbergen, H. W.; Vandersypen, L. M. K.; Dekker, C. *Nano Lett.* **2010**, *10* (8), 3163−3167.
(30) Barati Farimani, A.; Min, K.; Aluru, N. R. *ACS Nano* **2014**, *8*, 7914−7922.
(31) Bezanilla, F. *Physiol Rev.* **2000**, *80* (2), 555−592.
(32) Rabiner, L. R.; Juang, B. *ASSP Magazine, IEEE* **1986**, *3* (1), 4−16.
(33) Qin, F.; Auerbach, A.; Sachs, F. *Biophys. J.* **2000**, *79* (4), 1915−1927.
(34) Magleby, K. L.; Weiss, D. S. *Biophys. J.* **1990**, *58* (6), 1411−1426.
(35) Colquhoun, D. *J. Physiol.* **2003**, *547* (3), 699−728.
(36) Pedone, D.; Firnkes, M.; Rant, U. *Anal. Chem.* **2009**, *81* (23), 9689−9694.
(37) Gu, Z.; Ying, Y.-L.; Cao, C.; He, P.; Long, Y.-T. *Anal. Chem.* **2015**, *87* (2), 907−913.
(38) Raillon, C.; Granjon, P.; Graf, M.; Steinbock, L. J.; Radenovic, A. *Nanoscale* **2012**, *4* (16), 4916.
(39) Gu, Z.; Ying, Y.-L.; Cao, C.; He, P.; Long, Y.-T. *Anal. Chem.* **2015**, *87*, 10653−10656.
(40) Schreiber, J.; Karplus, K. *Bioinformatics* **2015**, *31* (12), 1897−1903.
(41) Balijepalli, A.; Ettedgui, J.; Cornio, A. T.; Robertson, J.; Cheung, K. P.; Kasianowicz, J. J.; Vaz, C. *ACS Nano* **2015**, *9* (12), 12583−12583.
(42) Briggs, K.; Kwok, H.; Tabard-Cossa, V. *Small* **2014**, *10* (10), 2077−2086.
(43) White, R. J.; Ervin, E. N.; Yang, T.; Chen, X.; Daniel, S.; Cremer, P. S.; White, H. S. *J. Am. Chem. Soc.* **2007**, *129* (38), 11766−11775.
(44) Kumar, S.; Tao, C.; Chien, M.; Hellner, B.; Balijepalli, A.; Robertson, J. W. F.; Li, Z.; Russo, J. J.; Reiner, J. E.; Kasianowicz, J. J.; Ju, J. *Sci. Rep.* **2012**, *2*, 684.
(45) Page, E. S. *Biometrika* **1954**, *41* (1/2), 100.
(46) Cluzel, P.; Lebrun, A.; Heller, C.; Lavery, R.; Viovy, J. L.; Chatenay, D.; Caron, F. *Science* **1996**, *271* (5250), 792−794.
(47) Strick, T. R.; Allemand, J. F.; Bensimon, D.; Bensimon, A.; Croquette, V. *Science* **1996**, *271* (5257), 1835−1837.
(48) Newville, M.; Stensitzki, T.; Allen, D. B.; Ingargiola, A. *LMFIT: non-linear least-square minimization and curve-fitting for Python; Zenodo* **2014**, DOI: 10.5281/zenodo.11813.
(49) Song, L. Z.; Hobaugh, M. R.; Shustak, C.; Cheley, S.; Bayley, H.; Gouaux, J. E. *Science* **1996**, *274* (5294), 1859−1866.
(50) Butler, T. Z.; Gundlach, J. H.; Troll, M. *Biophys. J.* **2007**, *93* (9), 3229−3240.
(51) Butler, T. Z.; Gundlach, J. H.; Troll, M. A. *Biophys. J.* **2006**, *90* (1), 190−199.
(52) Mathé, J.; Aksimentiev, A.; Nelson, D. R.; Schulten, K.; Meller, A. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (35), 12377−12382.
(53) Muzard, J.; Martinho, M.; Mathé, J.; Bockelmann, U.; Viasnoff, V. *Biophys. J.* **2010**, *98* (10), 2170−2178.
(54) Mathé, J.; Visram, H.; Viasnoff, V.; Rabin, Y.; Meller, A. *Biophys. J.* **2004**, *87* (5), 3205−3212.
(55) Krasilnikov, O. V.; Rodrigues, C. G.; Bezrukov, S. M. *Phys. Rev. Lett.* **2006**, *97* (1), 018301.
(56) Bezrukov, S. M.; Vodyanoy, I.; Brutyan, R. A.; Kasianowicz, J. J. *Macromolecules* **1996**, *29* (26), 8517−8522.