Detecting translocation of DNA nanostructures through nanopores: first steps towards structural barcode readout

Pratima Upretee*, Sybren Santermans†, Koen Martens†, Juliette Gevers†,
Sanjin Marion†, Wouter Van Den Bosch†, Jan Fostier*, Nilesh Madhu*
*IDLab, Ghent University - imec, Belgium, †imec, Belgium
{sybren.santermans, koen.martens, juliette.gevers, sanjin.marion, wouter.vandenbosch}@imec.be,
{pratima.upretee, jan.fostier, nilesh.madhu}@ugent.be

***Abstract*—Nanopore sequencing works on the principle of detecting the patterns in the current as a biomolecule translocates through an electrically charged nanopore. Robustly detecting such translocations is the first, key step in nanopore signal analysis. As the current changes in a step-wise manner when a molecule translocates, state-of-the-art approaches rely on straightforward thresholding of the signal to identify the start and end of translocation events. For robustness, the threshold is adaptively set, based on a local signal analysis. User expertise is, however, essential in selecting the appropriate parameters for this purpose. To reduce user dependency, and implicitly incorporate all information in the signal for event detection, we propose a Bayesian framework. This allows us to robustly infer the underlying nanopore state at each instant, based on a stochastic model of the nanopore signal. Experiments on real nanopore data, obtained from two separate devices, conclusively demonstrate the advantages of such modelling compared to the state-of-the-art. Further, traditional detection metrics only quantify one aspect of detection performance – giving an incomplete picture. Therefore, we also propose a new, complementary metric: complete event detection rate (CEDR) which measures the quality of the detected events, providing a more holistic view of event detection.**

***Index Terms*—Nanopore signal analysis, translocation event detection, Gaussian mixture models, Bayesian statistics**

I. Introduction

A biomolecule translocating through a nanopore modulates the electrical current running through the nanopore. By analysing the *pattern* of current changes as the molecule translocates, information relevant to the molecule can be inferred – this is the working principle of nanopores sequencing. Building upon the success of nanopore-based DNA sequencing, research is now (also) headed towards protein sequencing and protein identification [1], [2]. As a first step, the idea is to do this using barcoded DNA nanostructures. A barcode is formed by introducing molecular structures (e.g. dumbbells, also referred to as *labels*) in a specific pattern along a DNA backbone to encode information. The backbone of each unique barcode has antigens binding to antibodies of specific proteins to be detected. Using this in an assay, followed by nanopore-based read-out of the barcoded DNA, proteins present can be identified. In the work by the Keyser group [3], [4] the DNA origami method is used to create the barcode molecules. A *high-throughput* and *accurate* read-out of barcoded DNA can not only boost developing omics techniques, such as high-throughput protein identification [2], [5], [6], but can contribute to materializing DNA memory as well. We focus on solid-state nanopores, produced by etching a nanoscale hole in a thin membrane, typically made of silicon nitride. These pores are highly parallelisable, and can be engineered to control their size and shape – making them versatile in application. Such a solid-state nanpore set-up is schematically shown in Fig. 1.

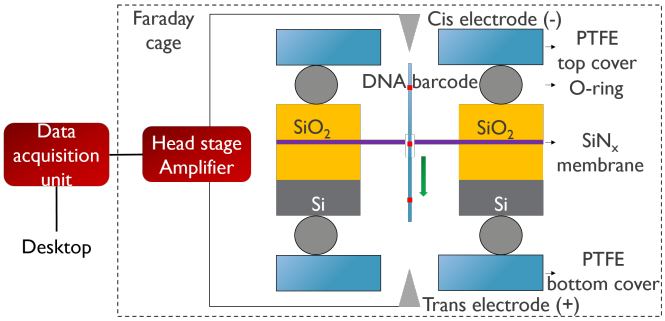


Fig. 1: Schematic of solid-state nanopore and data acquisition setup. The DNA molecule translocates through the pore from the cis- to the trans-layer, generating a current signal. Depending on the thickness of the molecule along its length, different current levels are observed during the translocation.

Biomolecule identification [7] from the recorded nanopore current signals is a two-stage approach. The first stage is the *detection* of each single molecule translocation (which we term an *event*) through the nanopore. Extracted events are then further analysed in a second stage to infer the barcode embedded in the molecule. Accurate event detection is, therefore, essential for subsequent analysis and bioanalyte identification. Fig. 2 (left) depicts the nanopore current signal in a 10 s period in our set-up (Sec. III-A). The dips in the signal correspond to translocation events. Fig. 2 (right) shows the current pattern for a *single* event, where the three 'spikes' indicate the passage of a thicker section of the DNA - corresponding to the three labels (dumbbells) added to the backbone.

Because of the step-wise nature of the current pattern, state-of-the-art (SOTA) for event detection is based on a simple thresholding of the current signal. The threshold is typically based on local, short-term analysis of data [8], as implemented in the various nanopore signal analysis toolboxes: OpenNanopore [9], Transalyzer [10], MOSAIC [11], EventPro [12], EasyNanoPore [13], and Autonanopore [14]. However user expertise is essential in choosing appropriate parameters in

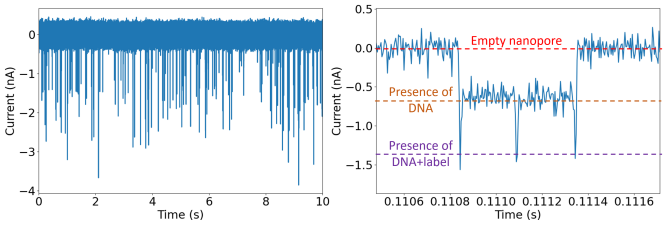


Fig. 2: Left: nanopore current signal. Significant dips in the current correspond to events. Right: single event with 3 spikes due to presence of 3 labels (dumbbells) on the DNA backbone.

all cases. In [15] the threshold is calculated using the open channel current, the diameter of the nanopore and that of the translocating molecule. While this can yield a straightforward threshold estimate, it has the following limitations: first, it depends on an unstable parameter, the diameter of pore, which can change over time, impacting the performance. Second, it cannot account for change in molecule thickness due to e.g., folding. Thus, robust event detection remains an ongoing, important topic of research.

We propose a Bayesian framework for event detection, that addresses the drawbacks in the SOTA. By fitting a Gaussian mixture model (GMM) [16] to the distribution of the current signals, we realise a 3-state description for the nanopore at every sampling instant. By analysing the state-probabilities, a robust decision on event start and end is obtained – leading to a more accurate event detection. The parameters of the GMM are estimated directly from the data, eliminating the need for user-defined parameters. We demonstrate the efficacy of this method on data captured using two commercially available nanopore systems. Further, we show that simply quantifying event detection performance in terms of standard detection metrics is insufficient. Therefore, we introduce a new metric: Complete Event Detection Rate (CEDR) to qualify the detected events with respect to the ground truth. Together with the detection metrics, this yields a more holistic comparison of event detection approaches.

II. Event detection

Let $x(n)$ be the sampled nanopore current signal as in Fig. 2. The zoomed-in plot clearly shows the three possible states of the nanopore, namely: **0** - empty (no molecule); **1** - DNA backbone region within nanopore; and **2** - backbone+label region in nanopore. Event detection may then be summarised as detecting change *from* state 0 (event begin), followed by a *return* to state 0 (event end). The event length is the time between these changes. We first discuss baseline event-detection approaches, based on signal-adaptive thresholds.

A. MOSAIC-based event detection

Since the change in the current is step-wise when it transitions from one state to another, it can be well-approximated by a piecewise constant function in each state. From this representation the *occurrence* and *time-span* of translocation events can be straightforwardly inferred. In the SotA ([9]–[13]), state-change is detected by thresholding the signal, where the threshold Γ is estimated from MOSAIC [11] as:

$$\Gamma = \mu - C\sigma\,, \tag{1}$$

where μ and σ are, respectively, the *mean* and *standard deviation* of $x(n)$, computed over the observation interval. C is a tuning parameter: larger C leads to more conservative event detection whereas lowering C would lead to more false detections. The two-state indicator function $y(n)$ is then obtained as:

$$y(n) = \begin{cases} 0 & x(n) \geq \Gamma \\ 1 & \text{otherwise}\,, \end{cases} \tag{2}$$

where $y(n) = 0$ indicates an open nanopore and $y(n) = 1$ indicates an event occurring. Additional *heuristic* considerations can be applied to fine-tune Γ, e.g., by using moving averages for μ and σ or estimating them based on *a priori* knowledge on the range of $x(n)$ in state 0. Similarly, additional smoothing can be applied to $y(n)$ to remove spurious detections. These require user expertise.

B. Autonanopore-based event detection

Autonanopore, implicitly considers a two-state (0 or 1) characterisation of the nanopore. It first partitions $x(n)$ into a specified number K of non-overlapping segments $\mathbf{x}_k = \{x(n + kN_0)\}$, $(k = 0, 1, \ldots K-1)$, of N_0 samples. Next, for each segment, the sample with the maximum negative excursion in $\mathbf{x}_k$ is selected. Denote this by $x_{p,k}$. The open-channel current level $x_{b,k}$ is estimated by averaging the signal for a suitable range of samples around $x_{p,k}$. $x_{p,k}$ is then corrected as $x_{p,k} \leftarrow x_{p,k} - x_{b,k}$, and their 1st quartile (Q1) and interquartile range (IQR) are calculated. Only the $x_{p,k}$ satisfying:

$$x_{p,k} < Q1 - C * IQR\,, \tag{3}$$

are retained. In (3), C is a user-defined tuning parameter. Under the assumption that each segment contains *at most* one event, the retained $x_{p,k}$ are considered as samples representative of the current level during an event. The start and end of the events are then determined by a heuristic search across the respective segments, as a function of $x_{p,k}$ and $x_{b,k}$. To further reduce false positives due to low amplitude excursions, an additional threshold criterion is applied to the retained $x_{p,k}$.

Assuming at most one event per slice limits the performance of the algorithm. In addition, the open-channel current might not be correctly computed in case of longer events or where the two-state characterisation of the nanopore is incorrect - as in our case. Lastly, the choice of the thresholds and the parameters for the heuristic search of the open-channel current level and the event duration again depends on the user.

C. Proposed probabilistic modelling of nanopore state

As seen previously, the nanopore can be in one of 3 possible (hidden) states. The hidden state can be inferred, given a probabilistic model of the observed current signal $x(n)$. Histogram analysis of $x(n)$ indicates that, depending on the state, $x(n)$ can be characterised by a Gaussian distribution, centered around the current level *typical* for that state (Fig. 3). Subsequently, $x(n)$ is treated as a stochastic signal whose

distribution corresponds to a Gaussian mixture model (GMM) with 3 components:

$$x(n) \sim \sum_{\ell=0}^{2} w_\ell \mathcal{N}(\mu_\ell, \sigma_\ell^2)\,, \tag{4}$$

where μ_ℓ, σ_ℓ^2 and w_ℓ indicate, respectively, the mean, variance and *prior* of state ℓ. This model may now be used in two ways.

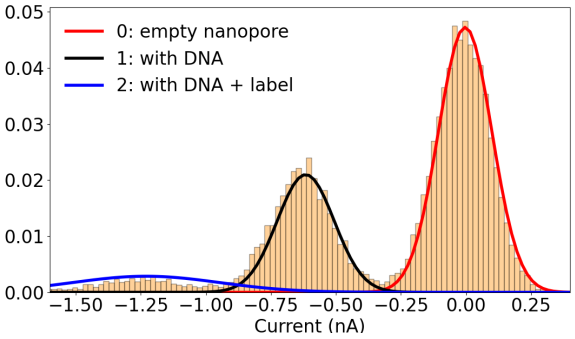


Fig. 3: Histogram analysis of $x(n)$, along with its GMM fit. Three distinct components are evident, corresponding to the 3 states of the nanopore.

A first idea is to use μ_0 and σ_0 to calculate the threshold as in (1), from which events can be detected as in Sec. II-A. This still requires a user-specified parameter C. Further, it does not optimally use the information in the full distribution.

A more intuitive and robust way would be to compute the probability of the hidden states at *every sample instant* n by using the Bayes theorem. Denoting by $P(\ell|x(n))$ the *posterior* probability of state ℓ at time n, we have from (4):

$$P(\ell|x(n)) = \frac{w_\ell\, p(x(n)|\ell)}{p(x(n))}\,, \tag{5}$$

where

$$p(x(n)) = \sum_{\ell=0}^{2} w_\ell\, p(x(n)|\ell) \tag{6}$$

and $p(x(n)|\ell) \sim \mathcal{N}(\mu_\ell, \sigma_\ell^2)$. The posterior probability for state 0 is shown in Fig. 4, with the current signal for that event. It may be seen that applying (5) yields near-binary state-decisions, *without* additional tuning parameters. Spurious peaks in $P(0|x(n))$ are eliminated by a simple 3-point median filter, yielding $\overline{P(0|x(n))}$. The data-dependent threshold $\Gamma = P(1|x(n)) + P(2|x(n))$ yields the indicator function:

$$y(n) = \begin{cases} 0 & \overline{P(0|x(n))} \geq \Gamma \\ 1 & \text{otherwise}\,. \end{cases} \tag{7}$$

The parameters of the GMM are estimated by the expectation maximization (EM) algorithm [17]. While σ_ℓ and x_ℓ are randomly initialized, μ_ℓ is initialized by peak-finding on the kernel density estimate (KDE). A Gaussian kernel with bandwidth estimated by Silverman's rule [18] is used for KDE.

III. Experiments & Results

A. Data

1) Data acquisition: DNA translocation data are obtained using two different measurement systems: the Elements srl 10 MHz nanopore reader [19] and the Axon Axopatch 200B [20]. For the Axopatch 200B and 10 MHz nanopore reader, the sampling rate is 250 kHz and 40 MHz, respectively, and the system bandwidths 100 kHz and 10 MHz, respectively.

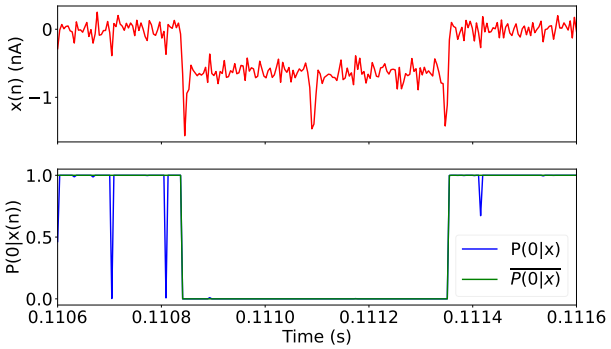


Fig. 4: Observed signal ($x(n)$) for one event (top plot), with the corresponding *posterior* probability $P(0|x(n))$ (bottom plot). Spurious peaks in $P(0|x(n))$ are eliminated by a 3-point median filter, yielding $\overline{P(0|x(n))}$ (green line), for a robust event detection.

2) DNA translocation experiment setup: Nanopore measurements are performed in a solution containing 4M LiCl, 10 mM Tris buffer and 1 mM EDTA (Ethylenediaminetetraacetic acid) with a pH of 8. The solution is degassed prior to use. 15 μl of this measurement solution is placed on both sides of the nanopore chip and is contained using two o-rings and two Teflon flow cells (top and bottom). Afterward, the cis and trans Ag/AgCl pseudo reference electrodes are inserted into the top and bottom flow cells respectively. These pseudo-reference electrodes contact the measurement solution at both sides of the chip and are hooked up to the head stage of the measurement system. To trigger the translocation of DNA barcode molecules through the nanopore, 2 μl of DNA barcode solution (87.1 nM) is pipetted into the flow cell at the cis side of the nanopore. The negative voltage applied to the cis electrode then drives the negatively charged DNA barcode through the pore. The trans electrode is connected to the ground. We apply respectively -200 mV and -400 mV to the cis electrode for the Axopatch 200B and 10 MHz nanopore reader experiments. The commercially available nanopore chips (Northern Nanopore [21]) used consist of a SiNx membrane with a thickness of 30 nm. Pore diameter is 12 nm for the Axopatch 200B measurements and 20 nm for the 10 MHz nanopore reader experiments.

3) DNA barcode: The DNA barcode molecules are obtained from tilibit [22] and consist of a single-stranded scaffold DNA molecule (M13mp18 - 7228 bases) with three DNA labels attached to it. The three labels are identical and comprise 17 dumbbell units. Each dumbbell unit is 20 base pairs long resulting in a 340 base pair (∼115 nm) length of one label. The outermost labels are placed 104 bases from the end of the scaffold DNA. The distance between the labels equals 3000 bases resulting in a 1 μm spacing.

4) Ground truth generation: The ground truth set is created by a manual search of the events in the signals.

B. Evaluation metrics

An event is said to be detected if the the samples of the detected region overlap with the ground truth. Standard detection metrics: the true positive rate (TPR) [23] and false discovery rate (FDR) [24] are therefore used to benchmark the detection performance of the proposed approach (false alarm rate is inapplicable here because of the disproportionately large number of true negatives). However, these metrics cannot describe whether the detected event is *complete*, where a complete event comprises at least all the data points from the corresponding ground truth. To assess the quality of the predicted event in terms of its entirety, a new metric – Complete Event Detection Rate (CEDR), is proposed. CEDR measures the ratio of the energy of a detected event to the energy of the corresponding ground truth:

$$\text{CEDR} = \frac{\sum_{n\in\mathcal{I}_{\text{DET}}} x^2(n)}{\sum_{n\in\mathcal{I}_{\text{GT}}} x^2(n)}, \tag{8}$$

where $\mathcal{I}_{\text{DET}}$ is the set of sample indices for the detected event and $\mathcal{I}_{\text{GT}}$ is the corresponding set for that event in the *ground truth*. In determining $\mathcal{I}_{\text{DET}}$, samples other than those present in the ground truth are removed. Thus $|\mathcal{I}_{\text{DET}}| < |\mathcal{I}_{\text{GT}}$, where $|\cdot|$ represents the cardinality of the set.

C. Experimental setup

Event detection is performed on two datasets: dataset A recorded using the Elements srl; and dataset B recorded using the Axon Axopatch 200B (Table I). All data were *downsampled* to 250kHz before evaluation.

TABLE I: Dataset details

Dataset	*duration (s)*	*# events*
A	60	829
B	1556	281

The two variants of the proposed GMM-based approach are considered: (i) using μ_0 and σ_0 to compute Γ as in (1), followed by thresholding to obtain (2) (denoted as PF_{GMM}), and (ii) determining the indicator function as in (7) (denoted as PF_{PROB}), where 'PF' indicates the piecewise constant indicator function. They are compared to the MOSAIC-based method of Sec. II-A ($\text{PF}_{\text{MOSAIC}}$) and Autonanopore (AN) baselines. The parameters for all approaches are estimated as described, on the same span of data. For AN, the default settings from the open-source implementation were already a good fit for our data (segments of 30ms, and $C = 1.5$)

D. Results

Detection rates: We first plot the TPR versus FDR curve for dataset A (largest number of events) in Fig. 5. The curves are obtained by varying C in the threshold-based approaches $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM}. They show the missed detection vs false discovery, trade-off typical to threshold-based approaches. PF_{PROB} only has a single point as the threshold is *implicit*.

Fig. 5 further shows that both $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM} have a good balance of TPR and FDR at $C = 3$. Thus, this value is taken for the evaluation of dataset B (Table. II). Note that AN could not be applied on dataset A, as the data was not available in the '.abf' format required for this approach.

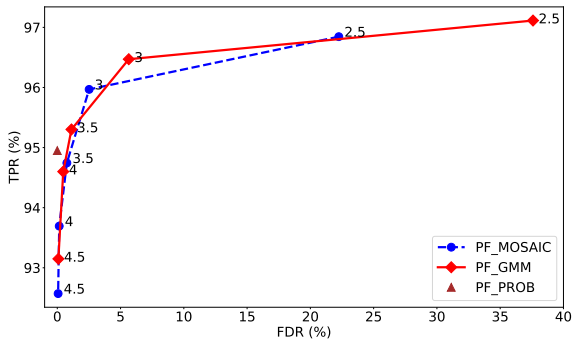
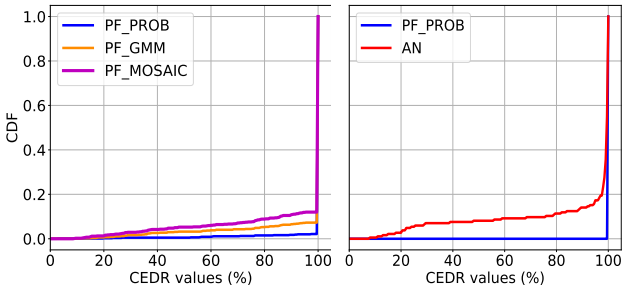


Fig. 5: ROC for evaluation of different algorithms.The performance of $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM} with different value of C is shown. Results for PF_{PROB} is represented by a triangle (in brown) .

TABLE II: TPR and FDR of different algorithms for dataset B.

Dataset B	AN	$\text{PF}_{\text{MOSAIC}}$	PF_{GMM}	PF_{PROB}
TPR	73.54%	98.21%	**100%**	**100%**
FDR	6.71%	**0%**	0.34%	**0%**

***Completeness* of detected events:** Fig. 6 shows the cumulative distribution of the CEDR of the detected events. For dataset A (Fig. 6a), we observe that the number of complete events is higher for PF_{PROB} compared to $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM}. Indeed, only few incomplete events are identified by PF_{PROB}, whereas both $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM} suffer more from incomplete event detection. Similarly, for dataset B, we observe 100% complete event detection for PF_{PROB} whereas the AN method suffers from incomplete event detection. The cdf plot for $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM} are not shown, because they also have 100% complete event detection for dataset B, and overlap with the curve of PF_{PROB}.



(a) Dataset A (b) Dataset B

Fig. 6: Cumulative distribution function (CDF) of the CEDR values for the different algorithms.

IV. DISCUSSION

When comparing the different algorithms in terms of detection metrics, AN performs the worst whereas $\text{PF}_{\text{MOSAIC}}$, PF_{GMM} and PF_{PROB} show a comparable performance. Whereas $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM} can be further fine-tuned using the parameter C in (1), it results in the typical TPR vs FDR trade-off: an increasing TPR is obtained at a higher FDR. The sensitivity of these approaches to parameter settings is also seen in Fig. 5 - FDR drastically drops from $C = 2.5$ to $C = 3$.

This highlights the importance of user experience during parameter selection for threshold-based approaches. PF_{PROB} takes into account the distribution of all states when computing $P(0|x(n))$. Thus, it makes better use of the information, allowing it to generalise better across datasets and *without* user-defined tuning parameters.

Whereas the evaluation based on detection metrics indicate a very similar performance for PF_{PROB}, $\text{PF}_{\text{MOSAIC}}$ and PF_{GMM} (for properly chosen parameters in PF_{GMM} and $\text{PF}_{\text{MOSAIC}}$), a more *nuanced* image is obtained when considering the cumulative density of the CEDR metric in Fig. 6 – which evaluates completeness of the detected events. Fig. 6a indicates that $\text{PF}_{\text{MOSAIC}}$ has fewer complete event detection compared to its counterparts. This happens when the indicator function falsely returns to state 0 for outliers in the current signal *during* an event. The cause lies in the use of the *complete* signal statistics for computing Γ, which is biased by the signal behaviour in the other states. As the GMM better models the signal distribution across the states, the threshold determined in PF_{GMM} is more robust – indicated by the improved cdf for this method. The best performance is obtained for PF_{PROB}, which explicitly incorporates information of other states in computing the state probabilities. The poor performance of AN lies in its inherent, limiting assumption of a single event in each segment – making it impossible to detect multiple, closely spaced events.

The higher performance for dataset B can be explained by the speed of the translocation. Due to the lower voltage applied, molecules translocate more slowly in dataset B, which allows for a better (more robust) detection of state-change.

V. Conclusions

We proposed a Bayesian framework for detecting translocation events in solid-state nanopores. The nanopore is modelled as being in one of 3 (mutually-exclusive) hidden states (0: empty, 1: DNA backbone present, 2: backbone+label present), and the state is inferred from the observed nanopore current signal. Histogram analysis of the signal amplitudes show that a 3-component Gaussian mixture model well characterises the probability density function of the observed signal, where each component is associated with one state. This allows a straightforward computation of *posterior* probability of each state, leading to a more robust detection. Experiments on real data conclusively demonstrate that such statistical models, which implicitly incorporate information of the other states, outperform the state-of-the-art. Further, they are not dependent on fine-tuning parameters – reducing the need for user expertise in the analysis. We also showed that existing metrics are insufficient for fully quantifying the performance of event detection approaches. In combination with our newly proposed CEDR metric, which measures the *completeness* of the detected events, a more holistic view is obtained. Associated code will be made available to the community.

References

[1] B. D. Reed, M. J. Meyer, V. Abramzon, O. Ad, P. Adcock, and F. R. Ahmad et al, "Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device," *Science*, vol. 378, no. 6616, pp. 186–192, Oct. 2022.

[2] M. Chee, K. Gunderson, and M. P. Weiner, "Macromolecule analysis employing nucleic acid encoding," WIPO Patent WO2 017 192 633A1, 2017. [Online]. Available: https://patents.google.com/patent/WO2017192633A1

[3] N. A. W. Bell and U. F. Keyser, "Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores," *Nature Nanotechnology*, vol. 11, no. 7, pp. 645–651, Jul. 2016.

[4] K. Chen, J. Kong, J. Zhu, N. Ermann, P. Predki, and U. F. Keyser, "Digital data storage using DNA nanostructures and solid-state nanopores." *Nano Letters*, vol. 19, no. 2, pp. 1210–1215, 2018.

[5] K. Martens, D. Barge, L. Liu, S. Santermans, C. Stoquart, J. Delport, K. Willems, and et al., "The nanopore-FET as a high-throughput barcode molecule reader for single-molecule omics and read-out of DNA digital data storage." in *In 2022 International Electron Devices Meeting (IEDM)*. IEEE, 2022, pp. 17–4.

[6] L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. Brody, J. Carter, and et al., "Aptamer-based multiplexed proteomic technology for biomarker discovery." *Nature Precedings*, pp. 1–1, 2010.

[7] C. Wen, D. Dematties, and S.-L. Zhang, "A guide to signal processing algorithms for nanopore sensors," *ACS Sensors*, vol. 6, no. 10, pp. 3536–3555, Oct. 2021.

[8] N. Das, N. Mandal, P. K. Sekhar, and C. RoyChaudhuri, "Signal processing for single biomolecule identification using nanopores: a review," *IEEE Sensors*, vol. 21, no. 11, pp. 12 808–12 820, 2020.

[9] C. Raillon, P. Granjon, M. Graf, L. J. Steinbock, and A. Radenovic, "Fast and automatic processing of multi-level events in nanopore translocation experiments," *Nanoscale*, vol. 4, no. 16, pp. 4916–4924, Jul. 2012.

[10] C. Plesa and C. Dekker, "Data analysis methods for solid-state nanopores," *Nanotechnology*, vol. 26, no. 8, p. 084003, Feb. 2015.

[11] J. H. Forstater, K. Briggs, J. W. F. Robertson, J. Ettedgui, O. Marie-Rose, C. Vaz, J. J. Kasianowicz, V. Tabard-Cossa, and A. Balijepalli, "MOSAIC: A modular single-solecule analysis interface for decoding multistate nanopore Data," *Analytical Chemistry*, vol. 88, no. 23, pp. 11 900–11 907, Dec. 2016.

[12] Y. M. Bandara, N. D. Y., J. Saharia, B. I. Karawdeniya, P. Kluth, and M. J. Kim, "Nanopore data analysis: baseline construction and abrupt change-based multilevel fitting," *Analytical Chemistry*, vol. 93, no. 34, pp. 11 710–11 718, Aug. 2021.

[13] J. Tu, H. Meng, L. Wu, G. Xi, J. Fu, and Z. Lu, "EasyNanopore: A ready-to-use processing software for translocation events in nanopore translocation experiments," *Langmuir*, vol. 37, no. 33, pp. 10 177–10 182, Aug. 2021.

[14] Z. Sun, X. Liu, W. Liu, J. Li, J. Yang, F. Qiao, J. Ma, J. Sha, J. Li, and L.-Q. Xu, "AutoNanopore: an automated adaptive and robust method to locate translocation events in solid-state nanopore Current Traces," *ACS Omega*, vol. 7, no. 42, pp. 37 103–37 111, Oct. 2022.

[15] X. Liu, Z. Sun, W. Liu, F. Qiao, L. Cui, J. Yang, J. Sha, J. Li, and L.-Q. Xu, "Multi-level translocation events analysis in solid-state nanopore current traces," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 1648–1653.

[16] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *CTIT technical reports series*, 1998.

[17] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[18] S. Weglarczyk, "Kernel density estimation and its application," *ITM Web of Conferences*, vol. 23, p. 00037, 2018.

[19] "Elements srl," accessed: 03/2023. [Online]. Available: https://elements-ic.com/nanopore-reader-10-mhz

[20] "Axon instruments," accessed: 03/2023. [Online]. Available: https://www.moleculardevices.com/products/axon-patch-clamp-system

[21] "Norcada," accessed on 02/2023. [Online]. Available: https://www.norcada.com/products/nanopore-products

[22] "tilibit nanosystems," accessed on 03/2023. [Online]. Available: https://www.tilibit.com

[23] J. Muschelli, J. Betz, and R. Varadhan, "Binomial Regression in R," in *Handbook of Statistics*. Walthm, MA, USA: Elsevier, Jan. 2014, vol. 32, pp. 257–308.

[24] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995.