2D MoS$_2$ nanopores: ionic current blockade height for clustering DNA events

To cite this article: Angel Diaz Carral et al 2019 2D Mater. 6 045011

View the article online for updates and enhancements.

Recent citations
- Deep learning for nanopore ionic current blockades
  Angel Diaz Carral et al
- Nanopore sensing of single-biomolecules: a new procedure to identify protein sequence motifs from molecular dynamics
  Adrien Nicolaï et al
- Pavel Sen et al
2D Materials

PAPER

2D MoS2 nanopores: ionic current blockade height for clustering DNA events

Angel Diaz Carral, Chandra Shekar Sarap, Ke Liu, Aleksandra Radenovic and Maria Fyta

1 Institute for Computational Physics, Universität Stuttgart, Allmandring 3, 70569 Stuttgart, Germany
2 Laboratory of Nanoscale Biology, Institute of Bioengineering, School of Engineering, EPFL, 1015 Lausanne, Switzerland

E-mail: mfyta@icp.uni-stuttgart.de

Keywords: 2D nanopores, molybdenum disulfide, DNA, clustering, machine learning, DNA translocation

Abstract

2D nanopores can be used to electrophoretically drive DNA molecules, which can in turn be identified through measurable electronic current blockades. In this work, we use experimental data from molybdenum disulfide nanopores threading DNA nucleotides and propose a methodological approach to interpret DNA events. Specifically, the experimental ionic traces are used to train an unsupervised machine learning model for identifying distinct molecular events through the 2D nanopore. For the first time, we propose a clustering of experimental 2D nanopore data based on the ionic current blockade height and unrelated to the traditional dwell time for each DNA event. Within this approach, the blockade level information is implicitly included in the feature space analysis and does not need to be treated explicitly. We could show the higher efficiency of the blockade height over the traditional dwell time also in coping with sparse nanopore data sets. Our approach allows for a deep insight into characteristic molecular features in 2D nanopores and provides a feedback mechanism to tune these materials and interpret the measured signals. It has, thus, a high impact on the efficiency of 2D nanopore-based DNA sequencers.

1. Introduction

Nanometer-sized pores opened in materials are efficient single-molecule detectors. These can electrophoretically thread biomolecules, such as DNA, RNA, and proteins to identify molecular events and single molecule properties such as the length of a molecule and its sequence [1–3]. Among many solid-state materials used for nanopores, 2D pores made from graphene [4–7] or molybdenum disulfide (MoS2) [8, 9] can play a valuable role in ultra-fast DNA sequencing [10–13]. The precise information on the length and type of molecule, as well as on the characteristics of molecular events is typically mapped onto ionic current blockades measured across the pore [14]. The duration of a current blockade denotes the translocation or dwell time for the event, i.e. the time required for the DNA translocating through the nanopore [15]. These signals can also be used to discriminate among different homopolymers [16] or different folding events of the molecule through the pore [17]. Typically, the dwell time is used to measure the length of a translocating molecule [18, 19]. This feature, though, cannot efficiently identify the sequence of a molecule nor capture the rich dynamics of each translocation event given through the typical multi-level nature of the ionic current blockades [17, 20]. In that respect, important information on molecular aspects is missing in the traditionally used dwell time.

On top of this, one needs to cope with the often broad ionic current distributions from a nanopore and the additional errors introduced through thermal fluctuations, the surrounding salt solution, the applied electric field, the flexibility of the biomolecules, etc. These factors decrease the detection fidelity [21], which though can be enhanced through a tuning of the experimental setup [22] or by post-processing of the ionic current measurements from the nanopore [23]. The latter is more efficiently achieved through the use of machine learning (ML) algorithms, which deal with strategies for translation of raw sequencing data into base calls [24]. Towards this goal, different ML protocols have been developed [25–28] for a proper feature extraction and classification. Machine learning techniques such as Hidden-Markov-based algorithms [29–32] or neural networks [33–36] have been implemented to process real-time long-read length
sequencing data from the most promising nanopore device in the market, the MinION from Oxford Nanopores [35, 37–39]. In order to deal with overlapping nanopore signals, segmentation techniques such as the cumulative sum (CUSUM) [40] or Bayesian-statistic-based [41] algorithms are applied to dissect the ionic current into subunits or levels. Additionally, the analysis of nanopore data involves a classification based on descriptors/features [42] such as the dwell time or the mean current blockade pointing to the most probable DNA translocation paths [40, 43]. Overall, the ML schemes for nanopores involve supervised learning and labeling of the data. De novo clustering [44] is—to our knowledge—the only unsupervised learning method, which though focuses in accelerating the data processing. In any case, ML techniques in conjunction with nanopores mainly focus on either guiding the learning process for optimizing the feature space, thus the error rates or improving the algorithmic scaling of the data processing.

In order to fill in the gap in using ML approaches to gain insight onto the molecular features inherent in the nanopore data, here we work on two novel aspects: (a) we propose a new very efficient feature to identify molecular events from the nanopore and (b) we use for this an unsupervised learning algorithm in order to classify nanopore data without the use of labels, allowing the algorithm to act on the nanopore information without guidance. The first point is essential in obtaining more information on the underlying Physics of the molecular translocation and is clearly missing from the literature. The second point highlights the suitability also of unsupervised learning for nanopore data. On a time scale lower than the traditionally used dwell time of a whole molecule, the search of an ‘appropriate’ feature space for training a ML algorithm is quite complex. Along these lines, we analyse and classify the various possible molecular configurations or the events for the translocation of single nucleotides through 2D MoS$_2$ nanopores using a clustering algorithm. In order to achieve this, the paper is organized as follows: we begin with an outline of the proposed methodology and present the experimental 2D nanopore ionic current traces from DNA translocation experiments. At a next step, different features of these data are analyzed and clustered in order to identify certain molecular events of the DNA molecules in the nanopore. In the end, we discuss the relevance towards nanopore sequencing.

2. Methodology

2.1. Data collection and concatenation of raw signals

The data used in this work are ionic current traces obtained from experiments of DNA translocation through a two-dimensional molybdenum disulfide (MoS$_2$) nanopore [9]. The size of the nanopore was 2.8 nm, while a voltage difference of 200 mV was applied and the ionic current for the translocation events are measured in a 0.1 M aqueous KCl solution. The experiments were carried out by applying an additional viscosity gradient across the nanopore in order to control the dynamics of the DNA molecules and delay their translocation through the pore to allow detection. We use four different datasets obtained from the translocation of the four single nucleotides dAMP, dTMP, dGMP and dCMP [9]. The corresponding raw datasets of the ionic traces are shown in figure 1.

Figure 1. The raw ionic current signals with time for the DNA single nucleotides (a) dAMP, (b) dTMP, (c) dGMP, (d) dCMP. The insets show the concatenated events of single nucleotides translocating the pore. These single translocations are marked through the red rectangular regions.
In order to examine and interpret the information hidden in the raw ionic current signals, the datasets are being concatenated using the CUSUM algorithm as implemented in the open Nanopore software [40]. CUSUM fits the raw signal from the experiments providing structure data files. These contain all the characteristic features such as the number of events, the start and end point of the translocation events, the number of levels for each event, as well as details on the dwell time, and the current blockade values. As mentioned above, four datasets for the translocation of the single nucleotides were extracted from the raw data. The length (number of events) of the data sets for all the nucleotide translocation experiments was 3887, 672, 127, and 119, for dAMP, dCMP, dTMP, and dGMP, respectively. As representative examples of concatenated data, the insets in figure 1 show a short series of single translocation events inspection of these insets clearly identifies the different types of events for each nucleotide. This difference is implicitly included in the ionic current blockade height, as will be revealed in the next section. Each block in the red rectangle represents a translocation event involving a certain configuration of the DNA nucleotide threaded through the nanopore.

2.2. Clustering and optimization

The main objective of this work is to propose an efficient feature in order to interpret the nanopore data shown in figure 1. For this we will attempt to cluster these data using an unsupervised learning technique [45], which can identify similarities and patterns in the data sets. This classification of translocation events will be done here with the open-source clustering algorithm k-means from the scikit-learn library [46]. We use this algorithm without any modifications. k-means is a randomly-initialized iterative clustering method, which is widely used in data mining, vector quantization, data compression, and pattern recognition [47, 48]. For a given k number of clusters, the algorithm can iteratively find centroid positions according to the data position in the feature space [48]. In the field of nanopore sequencing, it is possible to improve the read accuracy by applying the k-means algorithm, as it reduces the systematic and random noise in the experimental datasets [49].

In order to have a better control of the data clustering, we use the standard k-means algorithm and not any enhanced version that automatically selects the best k number of clusters [48]. This choice is also based on our analysis using two concrete statistical scores and allows us to better tune our unsupervised learning process. Two well-known statistical clustering scores the (a) silhouette (S) score [50] and (b) Calinski-Harabasz (CH) score [51] were used for choosing the optimization parameter k for all data sets. These have been chosen based on their high intrinsic differences. In general, the S-score is a good statistical tool, but fails in the case of sub-clusters and very close or similar clusters [52]. The CH-score is one of the best schemes for identifying sub-clusters. This scheme mainly uses an internal cluster validity measure, which can grade the generated clusters individually.

2.3. Feature selection

In order to interpret the translocation experiments and classify the various feasible and unique events, it is necessary to transform the concatenated information into features. These will be organized in feature spaces that can be clearly clustered. For the data sets in figure 1 mapping the ionic current patterns onto single DNA nanopore events will be done by considering physically intuitive aspects for the feature extraction. Accordingly, for a possible clustering of translocation events, the following four features referring to single DNA translocation events are considered:

(a) the traditionally used dwell or translocation time, which maps the duration of an event,

(b) the height of the ionic current blockade, defined as the difference between the maximum and minimum values of a single current blockade peak,

(c) the ionic blockade mean current (mean), defined through the average current value, and

(d) the levels, which are the number of the presumably different DNA configurations through the nanopore.

The dwell time and mean of the ionic current blockade are typically the features used for statistical analysis of nanopore experiments [23, 40, 53]. The novelty in our study is the proposal of the ionic blockade height as a new feature for clustering. The choice of this feature was physically intuitive, as it is related to the dynamics of a single ionic current blockade. In this way, we attempt to include additional information inherent in a blockade, which is not involved in the dwell time and mean ionic current features. It will be shown in detail in section 3 that the blockade height compared to the other features can much more efficiently cluster nanopore events. In order to understand the DNA topology, i.e. the various molecular configurations in the nanopore, the number of levels within a single blockade, as well as their ionic current values will also be discussed in the end. The information on these levels will allow for a comparison between the number of clustered translocation events and the possible DNA configurations in the pore.

3. Results and discussion

3.1. Feature extraction

In order to classify the various possible configurations of the DNA molecules threading through the nanopore, we have analysed the feature space using a scatter plot for the dAMP translocation. Different combinations of the features of the dwell time, mean, and height have
been examined. Figure 2 represents the scatter plots of the ionic traces with respect to the blockade mean and height versus dwell time and the blockade height versus the mean. As already mentioned, the dwell time is typically used as the main indicator to demonstrate the distinct translocation events [40, 43]. For supervised learning, the scaling of this feature is often necessary in order to accelerate the learning process [54]. However, the variation in the scaling of this feature can be problematic in unsupervised learning, like clustering, as it can bias and modify the data. It can change the distances between the samples leading to different clustering combinations. In figure 2(a), it is evident that the scatter plot of the respective features (mean or height versus time) does not point to clearly distinct clusters that could lead to possible configurations of dAMP through the nanopore. In fact, the combination of the mean current with the dwell time only identifies a large cluster. When using the blockade height in combination with the dwell time, the results are improved. Two clusters seem to form, but are still overlapping without giving a very clear picture. As an alternative, we select the combination of blockade height versus mean. This choice, as seen from figure 2(b) clearly points to two clusters mapping two different types of molecular configurations through the nanopore. Accordingly, the blockade height feature has the quality of clearly being clustered providing better information than the commonly used dwell time. This emphasizes the novel feature of the approach we propose here.

3.2. Feature efficiency in clustering nanopore events

Based on the feature extraction and analysis mentioned above, we move on by applying a k-means clustering algorithm to the translocation data from all nucleotides. The feature space analysis with respect to blockade mean and height for all the single nucleotides is given in figure 3. Indeed, two distinct clusters can be seen in the case of adenine demonstrating two distinct molecular configurations for dAMP. For the other nucleotides, the picture is similar, though not always as clear as for dAMP due to the higher sparsity of the other data. In any case, even for the few data for dGMP two distinct clusters seem to form, while dCMP has the tendency to form four clusters. In order to find the optimum number of clusters for all data sets and confirm the findings above, we calculate the Silhouette (S) and Calinski-Harabasz (CH) scores and scan cluster sizes $k$ in the range 2–10. The results for all data sets are summarized in figure 4. The scores in this figure are obtained by taking two sets of features, the blockade height versus dwell time and the blockade height versus mean. For visibility and comparison with the S-score, the CH-score was scaled using the max–min normalization technique within the [0–1] range [55]. For a given data set, the highest values of these scores represent the most probable number of clusters. Similar S-score values for different $k$ values point to the existence of sub-cluster configurations. In order to further evaluate this, we turn to the CH-score in the same $k$ range. Note, that we do not focus on the best possible way for optimizing classifiers, rather on finding the global density of clusters for the translocation data.

As evident from this figure, the choice of the dwell time in the feature space leads to an almost stable S-score value with increasing $k$ for all nucleotides. As opposed to this, a linear increase is observed in the CH-score. The different behavior of the two scores underlines the fact that the dwell time is not a good feature for classification. In contrast to this, when choosing the blockade height and the mean current as the feature space, both scores show a similar variation with the number of cluster size. For dAMP, the S-score is similar for cluster sizes $k = 2$ and $k = 3$. For the CH-score, a cluster size of $k = 2$ shows the largest score. This $k$ is related to a higher score also for dTMP and dGMP. Accordingly, dAMP, dTMP, and dGMP assume two distinct types of configurations through the pore as denoted by the cluster size of $k = 2$. A deviation can be observed for dCMP, for which the CH-score is higher for larger
Figure 3. Feature space analysis for the (a) dAMP, (b) dTMP, (c) dGMP, and (d) dCMP translocation experiments from figure 1.

Figure 4. Silhouette (S) (solid lines) and normalized Calinski-Harabasz (CH) (dashed lines) scores for all single-nucleotide translocation experiments. The scores are obtained from the analysis of the feature spaces dwell time—blockade height (filled circles) and blockade height—mean current (filled squares), as denoted in the legends.
cluster sizes. In order to visualize this in figure 3 cluster sizes of \( k = 3 \) and \( k = 4 \) are sketched for dCMP. These denote the high probability of finding four different types of dCMP configurations in the pore. Overall, when the number of clusters increases, the CH-score is higher than the S-score due to the small distance of the samples to the cluster. Applying the k-means algorithm in sparse data sets cannot distinguish among optimum cluster sizes when using the feature space of dwell-time and mean current. Nevertheless, including the ‘blockade height’ in the feature space (especially without the dwell time) leads to a more efficient classification even for the more sparse data sets. In this case, the variation of both scores with \( k \) is similar supporting the higher efficiency of ‘height’ over ‘dwell time’.

3.3. Interpretation of clustering analysis—dealing with data sparsity

In order to interpret the cluster sizes in figure 4, we seek additional information through an analysis of the current blockade levels and events. In this way, a physical meaning can be assigned to the clustering obtained through our ML approach. For this, we consider the current blockade levels instead of the mean current blockade and attempt to classify the number of configurations observed. We have used the OpenNanopore software [40] for extracting statistics from the translocating data. In this way, the length of the data is increased by resolving also the number of different current blockade levels within a single current blockade. A scatter plot from this level analysis as a function of the dwell time and the ionic current data for dAMP is shown in figure 5. The data are represented through a histogram revealing two distinct peaks. These two peaks arising from the levels correspond to two most probable conformations of a dAMP threading the nanopore. These results again support the fact that using the current blockade height values of the translocation levels instead of the events themselves leads to similar results. This is also a very efficient approach for extracting information on the molecular conformations in the pore. The representation in this figure clearly denotes that the most dense cluster in a distribution maps a higher occurrence probability of a certain molecular conformation. In the case of longer DNA molecules, such two peaks would be related to unfolding (lower current, slower translocation) and folding (higher current, faster translocation) conformations [17, 56]. In the case of single nucleotides, these peaks can be assigned to either different entrance configurations of the molecules in the pore or possible molecular isoforms. This remains to be shown in a separate study. Here, our aim was to provide a feature analysis without the need of a levels analysis. We could reveal a direct robust relation between the molecular topology and the feature height.

Another very important aspect is the handling of sparse data sets. Intuitively, the larger the data set the better the clustering of the events. We have shown here the significantly higher efficiency of the blockade height in clustering nanopore data (see figure 2). These results, though, were shown for the rich dAMP data sets. We further support our findings, by performing the same analysis for the much more sparse data for dGMP in figure 6. Overall, we can again underline the findings from the rich data sets: the use of the blockade height can identify two clusters. These are very distinct when the second feature is not the dwell time. The exclusion of the height in the feature analysis cannot identify more than one clusters, despite the fact that the better Silhouette score led to an optimum size of two clusters (figure 4).

As a final note, the fact that the height feature is more efficient than the traditional dwell time for clustering

![Figure 5. Blockade level analysis for dAMP showing all experimental data (left) and their distributions (right). Two distinct peaks representing two most probable configurations can clearly be identified. In the inset, the red lines correspond to the separate levels taken for one event. These levels point to the respective circles in the events plot.](image-url)
events relies on the implicitly richer statistics included in this feature. Specifically, to each ionic blockade the difference of the start and end stage of the translocation is assigned when taking the dwell time or the mean blockade. However, the blockade height is defined through the difference between the highest and lowest peak within a blockade. In this way more distinct information is included for each event. As clearly seen from the insets of figure 1, there is a much more rich dynamics in the many levels (i.e. peak heights) in a single blockade and these are very different for each event. These multi-level characteristics typically defined by the blockade extrema are explicitly taken into account in the height feature. In terms of the underlying Physics of the translocation process, one could argue that from the three features we have examined here, the dwell time and mean blockade are essentially coarse-graining the detailed features of each blockade and through this the molecular details in the pore. On the contrary, the blockade height is inherently including the characteristics of the detailed Physics of a single translocation event. The extrema in the single blockade taken for the height definition are the characteristic ones for the process. These carry for example, information on the two extreme configurational changes of the molecule through the pore (relevant to the peaks in figure 5). In this respect, the blockade height includes much more rich dynamics than the dwell time and naturally enhances the differences among different molecules.

4. Summary

We have proposed here a new feature used with an unsupervised learning model for identifying different molecular topologies from translocation events of DNA through ultra-thin MoS₂ nanopores. The novelty of our approach is based on using the ionic blockade height instead of the ‘traditional’ dwell time or the level information. We have shown, that the commonly used dwell time is not the best feature for unraveling molecular events in the pore. Instead we could manifest that the novel feature of the ionic blockade height is highly more efficient in forming well defined clusters of events. In order to support this, we have performed a classification for clustering experimental data for different DNA molecules. Our approach does not use labeling of the data, thus is not biasing the leaning process. We were able to distinguish distinct types of the most probable molecular conformations of the DNA threading the pore. Accordingly, we could clearly identify two nucleotide events. Our novel approach uses solely the feature space made of the mean ionic current and the ionic current blockade height, while avoiding sparse patterns obtained from a blockade level or event analysis of the raw data using the dwell time feature. Our proposal for the relevance and importance of the ionic blockade height was supported by the fact that this feature can also cope with data set sparsity. The blockade height is proven efficient in all cases where the traditional dwell time fails to identify clusters of events.

The novel feature selection we propose here together with unsupervised learning can be valuable in providing a focal point on 2D nanopore experimental data and a valuable insight into the nanopore. We begin here with the identification of molecular events within the pore. Having shown the applicability of our scheme to DNA, we can further extend it to other types of biomolecules, which could involve a more rich conformational variation. These rich dynamics is an aspect implicitly carried within the ionic current blockade height feature. In order to reveal the exact configurational aspects, more data are needed. Our framework, though, can also be used with data from other nanopores in order to gain insight into the differences among nanopores and guide the design of the nanopore material. Overall, the relevance of the proposed scheme, in view of sensing DNA and also detecting its sequences is very high. An efficient 2D nanopore should be able not only to detect the molecule, but also read-out the order of its nucleobases with the lowest possible errors. These errors arise from different sources and are often difficult to control or reduce. Our classification scheme based on using the blockade height as a feature can be further tested, improved, and applied to other types of DNA sequences in order to reduce errors and filter out the
information on the identity of the nucleobase along a translocating DNA. The classification scheme has the potential to deal with different nanopore materials and sizes, as well as varying experimental conditions (applied voltage, salt concentration, pH, etc.). In this respect, the highest gain of our approach would be to assist in generating a database of 2D nanopore events inherently including all the above details. This can in turn be valuable not only for interpreting events, but also predicting DNA sequences in 2D nanopore experiments. Our work provides an important methodological input towards this direction. In the end, the importance of this study is two-fold: it provides additional support from the EXC 310 SimTech Cluster of the University of Stuttgart. CSS and MF thank I Steinwart for an insightful discussion. ADC and MF acknowledge financial support from the Swiss National Science Foundation (Grant No. 200021.153653, BSGC10.157802).

Acknowledgments

ADC, CSS, and MF thank I Steinwart for an insightful discussion. ADC and MF acknowledge financial support from the EXC 310 SimTech Cluster of the University of Stuttgart. CSS and MF thank the Juniorprofessorenprogramm funded by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) Forschungsgemeinschaft (DFG) for financial support. AR and KL would like to acknowledge financial support from the Swiss National Science Foundation.

ORCID iDs

Ke Liu https://orcid.org/0000-0002-1976-2297
Alekksandra Radenovic https://orcid.org/0000-0001-8194-2785
Maria Fyta https://orcid.org/0000-0002-5425-7907

References

[17] Li J, Gershov M, Stein D, Brandin E and Golovchenko J A 2003 Nat. Mater. 2 611
[27] Kretić P, Ashcroft B and Lindsay S 2015 Nanotechnology 26 084001
[29] Landry M and Winters-Hilt S 2007 BMC Bioinform. 8 S12
[31] Schreiber I and Karplus K 2015 Bioinformatics 31 1897–903
[34] Misiunas K, Ermann N and Keyser U F 2018 Nano Lett. 18 4040–5
[39] Li Y, Han R, Bi C, Li M, Wang S and Gao X 2018 Bioinformatics 1 10