Nanoscale

Nanopore Sensing of Single-Biomolecule: A New Procedure to Identify Protein Sequence Motifs from Molecular Dynamics

Journal:	*Nanoscale*
Manuscript ID	NR-ART-07-2020-005185.R2
Article Type:	Paper
Date Submitted by the Author:	14-Oct-2020
Complete List of Authors:	Nicolaï, Adrien; Université Bourgogne Franche-Comté, Laboratoire Interdisciplinaire Carnot de Bourgogne Rath, Aniket; Université Bourgogne Franche-Comté, Laboratoire Interdisciplinaire Carnot de Bourgogne Delarue, Patrice; Université Bourgogne Franche-Comté, Laboratoire Interdisciplinaire Carnot de Bourgogne Senet, Patrick; Université Bourgogne Franche-Comté, Laboratoire Interdisciplinaire Carnot de Bourgogne

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Nanopore Sensing of Single-Biomolecule: a New Procedure to Identify Protein Sequence Motifs from Molecular Dynamics†

Adrien Nicolaï,* Aniket Rath, Patrice Delarue and Patrick Senet

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Solid-state nanopores have emerged as one of the most versatile tools for single-biomolecule detection and characterization. Nanopore sensing is based on measuring the variations in ionic current as charged biomolecules immersed in an electrolyte translocate through nanometer-sized channels, in response to an external voltage applied across the membrane. The passage of the biomolecule through the pore yields information about its structure and chemical properties, as demonstrated experimentally with sub-microsecond temporal resolution. However, extracting the sequence of the biomolecule without the information about its position remains challenging due to the fact there is a large variability of sensing events recorded. In this paper, we performed microsecond time scale all-atom non-equilibrium Molecular Dynamics (MD) simulations of peptide translocation (motifs of alpha-synuclein, associated to Parkinson disease) through single-layer MoS_2 nanopores. First, we present an analysis based on current threshold to extract and characterize meaningful sensing events from ionic current time series computed from MD. Second, a mechanism of translocation is established, for which side chains of each amino acid are oriented parallel to the electric field when they are translocating through the pore and perpendicular otherwise. Third, a new procedure based on permutation entropy (PE) algorithm is detailed to identify protein sequence motifs related to ionic current drop speed. PE is a technique used to quantify the complexity of a given time series and it allows to detect regular patterns. Here, PE patterns were associated to protein sequence motifs composed of 1, 2 or 3 amino acids. Finally, we demonstrate that this very promising procedure allows the detection of biological mutations and could be tested experimentally, despite the fact that reconstructing the sequence information remains unachievable at this time.

1 Introduction

Solid-state nanopore (SSN) technology for the detection and analysis of proteins is an emerging experimental tool with promising applications in medical diagnostics[1,2]. In SSN sensing experiments[3], charged biomolecules, which are suspended in an ionic solution, are driven by a transverse electric field through a nanopore within an ultrathin membrane. During that time, the ionic current $I(t)$ is monitored to detect the passage of biomolecules through the pore at a sub-microsecond temporal resolution[4]. Typically, translocation events are detected as drops in ionic current signal, *i.e.* ΔI, lasting for a certain time, *i.e.* dwell time τ_d, and contain information about the biomolecule structure and chemical properties[5]. However, as shown recently from experiments for DNA sensing[6] and from simulations for protein sensing[7], no consistent levels of ionic current can be visually attributed to DNA segments (nucleotides) or to protein motifs (amino acids), respectively. In fact, the fluctuations and noise observed in ionic current traces due to the fast translocation of biomolecules through the pore are very complicated[8] and visually reading the primary structure of biomolecules from raw signals remains very challenging.

To overcome these limitations, various approaches have been investigated experimentally in the past decade. For instance, different materials have been tested to fabricate ultrathin nanoporous membranes such as silicon nitride[6], graphene[9], hexagonal boron nitride[10] or molybdenum disulfide[11]. 2D materials offer large signal-to-noise ratio (SNR) due to the fact membranes are a few Angström thick. However, graphene shows a lower SNR[12] than MoS_2[13], even though the thickness of graphene is one atom thick (3 atoms thick in MoS_2). Moreover,

Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Université Bourgogne Franche-Comté, 9 Av. A. Savary, BP 47 870, F-21078 Dijon Cedex, France; E-mail: adrien.nicolai@u-bourgogne.fr

† Electronic Supplementary Information (ESI) available. See DOI: 00.0000/00000000.

various effort have been made to either increase the time resolution[14] or slow down the translocation process[15]. Finally, different experiments integrating additional detection methods have been carried out using optical and/or electrical techniques[16–19].

Another strategy is to consider the ionic current time series as they are measured (raw signal) and apply time series analysis tools in order to get insights into the origin of such a large variability in translocation events. In the past few years, sophisticated algorithms have been developed to detect and statistically characterize amplitude and time of translocation events from experimental measurements[20,21]. More recently, several mathematical approaches have been used for pattern recognition from nanopore raw data such as Hidden Markov Models (HMM)[22], Artificial Neural Networks (ANN)[23] or Support Vector Machine (SVM)[24] for the classification/clustering of translocation events. These algorithms from Machine Learning (ML) have been mainly applied to DNA sequencing using biological nanopores. To the best of our knowledge, only a few works report the application of such ML methods for DNA/protein sequencing through solid-state nanopores[25,26]. In addition, these methods require a very large amount of data (thousands of translocation events) from experiments or from extensive simulations[27]. In the later case, biases are usually applied in order to generate sufficient data in a reasonable computational time.

In the present work, these approaches are not of interest since our goal is first, to understand the variability of event signatures by establishing *the non-linear relationship existing between the presence of the biomolecule inside the nanopore and the ionic current variations measured* and, second, to characterize their variability with a meaningful and reliable physical parameter. As already mentioned above, ionic current time series extracted from nanopore sensing experiments are very complex time series since they are characterized by large fluctuations and noise. In physics, the complexity of a time series is associated to disorder degree, *i.e.* randomness and unpredictability. In order to evaluate the complexity of a given time series, entropy is one of the most powerful metrics. For instance, Shannon entropy[28] or Kolmogorov-Sinaï entropy[29] have been widely used. However, in these entropy definitions, there is no information about the dynamics and they can be computationally very expensive. In 2002, Bandt and Pompe combined the concept of entropy and temporal order in a time series, the so-called permutation entropy[30] (PE). PE measures information based on the occurrence of absence of certain permutation patterns of the ranks of values in a time series. In addition, the main advantage of PE is the fact that it can be calculated for arbitrary real-world time series, the method being extremely fast and robust. Last but not least, PE is preferable compared to the methods mentioned above for huge data sets due to the fact there is no need for pre-processing and fine-tuning of parameters.

In this paper, we performed microsecond timescale all-atom Molecular Dynamics (MD) simulations to investigate the translocation of peptides through single-layer MoS_2 nanopores, providing the knowledge of the exact position of the peptide that is translocating through the pore at any time. From these MD runs, we computed ionic current time series, as they are measured experimentally. Thanks to MD, we can analyse the non-linear relationship between the actual peptide position and the ionic current variations. The sequence of the peptide was extracted from α-Synuclein protein, an intrinsically disordered protein which is a major constituent of Lewy bodies[31], the insoluble aggregates that are the hallmark of one of the most prevalent neurodegenerative disorders, Parkinson's disease. This protein is characterized by the presence of repeat motifs in its primary sequence, KTKEGV, which are key mediators for the neurotoxicity[32]. The paper is organized as follows. We present first the methods. Next, MD trajectories are analysed. Particularly, a threshold of ionic current was determined to ensure the detection of true sensing events from raw MD data. Second, a statistical analysis of current drops ΔI along the amino acid sequence of the peptide is also presented and mechanisms of translocation observed in MD simulations and their origins are discussed. Third, permutation entropy algorithm is applied to ionic current time series extracted from MD to quantify their complexity. From this procedure, we extracted patterns of current drops characterized by a new parameter $\Delta I/\Delta t$, named ionic current drop speed. Finally, we explored the effect of biological mutations onto ionic current drop speed characteristics. The paper ends with concluding remarks.

2 Materials and Methods

2.1 Molecular Dynamics

All-atom MD simulations using periodic boundary conditions were performed using the LAMMPS software package (http://lammps.sandia.gov)[33]. Each simulation box of dimension $8.0 \times 8.0 \times 16$ nm^3 is comprised of a MoS_2 nanoporous membrane, a biological peptide (capped with ACE and NME groups at N anc C-terminal) plus a 1M KCL electrolyte and is globally neutral (Fig. 1). Peptide translocation in MD simulations was enforced by imposing a uniform electric field, directed normal to the nanoporous membrane (z-direction), to all atomic partial charges in the system. The corresponding applied voltage simulated is $V = -EL_z$, where L_z is the length of the simulation box in the z-direction, with $V = 1$ V for all MD runs presented. Five independent MD runs of 500 ns each were performed for each peptide presented here, KTKEGV, KTKKGV and KTKEGR, for a total simulation time of 2.5μs for each peptide. Finally, an open pore MD simulation (no peptide) of 500 ns using the same procedure as the one described above for translocation simulations, was performed. Technical details are available in Electronic Supplementary Information[†].

2.2 Ionic Current Time Series

Ionic current time series were computed from MD production runs using z-coordinates of K^+ and Cl^- ions as a function of time, as:

$$I(t) = \frac{1}{\Delta t L_z} \sum_{i=1}^{N} q_i [z_i(t+\Delta t) - z_i(t)] \quad (1)$$

where Δt is the time between MD snapshots chosen for the calculations ($\Delta t = 1$ ns), L_z is the dimension of the simulation box in the z-direction, which is the direction of the applied electric field, N is the total number of ions, q_i is the charge of the ion i and $z_i(t)$

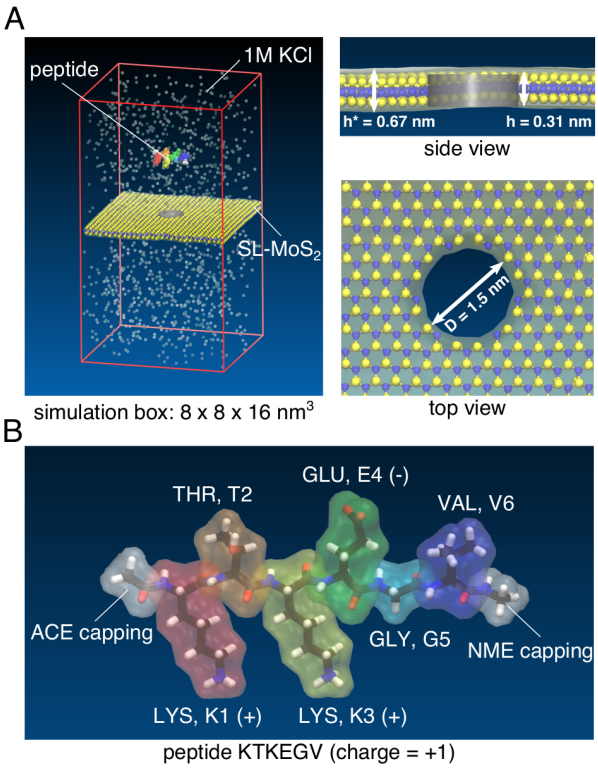

Fig. 1 A) Atomic representation of the nanopore system simulated in the present work. Left panel shows the simulation box used in the present work (red lines). MoS_2 nanoporous membrane is represented in ball and stick (Mo atoms in blue and S atoms in yellow), peptide is drawn in surface representation and KCl electrolyte is shown with transparent spheres (K^+ ions in gray and Cl^- ions in green). Water molecules are not drawn for more clarity. Right panels represent top and side views of the nanopore (in gray). Pore characteristics (diameter D, thickness h and effective thickness h^*) are indicated. B) Atomic representation of KTKEGV peptide translocated through single-layer MoS_2 nanopores. The licorice representation is used here. Positively (+) and negatively (-) charged amino acids are indicated. The figure was created using VMD 1.9.3 software[34].

is the z-coordinate of the ion i at time t.

2.3 Permutation Entropy

A time series is defined by successive measurements of a variable x as a function of time, at discrete time values, regularly spaced or not. Hence, given a time series X of length T, $X = \{x_i\}_{i=1,2....,T}$, the calculation of the permutation entropy involves considering the temporal order of the values x_i in the time series. It gives the rank order of successive x_i in sequences of length n in the time series X. The calculation of the permutation entropy on a given time window T depends on the initialization of 2 parameters: n, which is the order of permutation or the number of elements that need to be compared with each other and τ_{lag}, which is the time separation between the elements that need to be compared. For example, consider the time series $X = \{5,7,8,1,3,2,4\}$ with $T = 7$, the permutation order being set to $n = 3$ and the time separation being set to $\tau_{lag} = 1$. Then, we extract the following triplets ($n = 3$): [5,7,8], [7,8,1], [8,1,3], [1,3,2], [3,2,4], which are shifted by 1 time value ($\tau_{lag} = 1$). In each triplet, the order of the values are labeled as 0,1, and 2 by increasing values. Therefore [5,7,8] is associated to the pattern $\pi_1 = (0,1,2)$, [7,8,1] to $\pi_4 = (1,2,0)$, [8,1,3] to $\pi_5 = (2,0,1)$, [1,3,2] to $\pi_2 = (0,2,1)$ and finally [3,2,4] to $\pi_3 = (1,0,2)$. Overall, each triplet is associated to one of the $j = 1,...,n! = 6$ possible permutation patterns π_j.

Permutation entropy measures the disorder of successive values by using the probability of the different permutation patterns in each time window. Hence, we define normalized PE as:

$$PE = -\frac{1}{log_2(n!)} \sum_{j=1}^{n!} p_j log_2(p_j), \quad (2)$$

where p_j represent the relative frequencies of the possible permutation pattern π_j. For $T \gg n$, the probability of each permutation of a completely random signal would be equal to $1/n$ and the PE maximum. Finally, the smaller PE is (minimum value PE=0), the more regular and more deterministic the time series is. Contrarily, the larger PE is (maximum value PE=1), the more noisy and random the time series is. More details about PE algorithm can be found in reference[35]. In the present work, PE algorithm applied to ionic current time series of translocation data was calculated using a dimension $n = 3$ and a time lag $\tau_{lag} = 1$ for a given time window $T = 1,000$ samples of the signal, then sliding the time window by 500 samples.

3 Results and Discussion

3.1 Sensing Event Detection from Ionic Current Time Series

In nanopore experiments, single-biomolecule sensing events are usually detected from ionic current time series using a threshold value, which is extracted from open pore measurements[6] (defined as no biomolecule present in the electrolyte). A drop of ionic current ΔI is considered to be a sensing event if values of current measured during a certain amount of time, called dwell time τ_d, are below the threshold. The main advantage using MD to perform in silico simulations of nanopore sensing experiments is that the position of the biomolecule as a function of time is known at every single time step of the simulation. Therefore the passage of a biomolecule through the pore can be validated from its coordinates, as already done in a previous work[7]. In order to mimic as close as possible experimental investigation of sensing events from nanopore measurements, we decided, first, to perform a "blind" detection of sensing events from ionic current time series extracted from MD trajectories based on a threshold as in experiments.

Fig. 2A represents ionic current time series computed from translocation of KTKEGV peptides through SL-MoS_2 nanopores. First, fluctuations of the signal extracted from MD are very large and noisy, the corresponding probability distribution $P(I)$ being very broad and unimodal. It means that drops of current associated to the passage of biomolecules through the pore cannot be distinguished from the raw signal. From this observation, we filtered the data in order to remove high frequency fluctuations by computing the moving mean of the ionic current signal over $T = 10,000$ samples. As shown in Fig. 2A, current drops visually appear in the filtered signal, which is confirmed by the bimodal characteristics of the probability distribution. The maximum peak

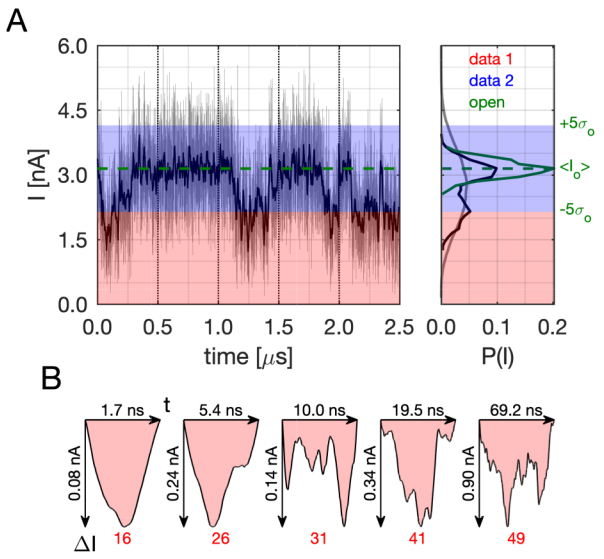

Fig. 2 A) Typical ionic current I [in nA] as a function of time [in μs] computed from five independent concatenated MD simulations (500 ns each) of translocation. The translocation of KTKEGV peptide through SL-MoS_2 nanopore is shown (left panel). Green dashed line represents the mean open pore current value $< I_o >$. Gray and black lines correspond to raw and filtered data, respectively. Right panel represents probability distributions of ionic current time series $P(I)$. The color code is the following: raw data (gray), filtered data (black) and open pore data (green). Red (*data 1*) and blue (*data 2*) shaded areas represent current values below and above the $5\sigma_o$ threshold, respectively. B) Ionic current drop ΔI *vs* time patterns for 5 of the 49 sensing events detected in the present work. Maximum current drop ΔI^{MAX} and dwell time τ_d for each event are indicated. Event index is given in red.

of $P(I)$ is centered around the mean value of open pore current $< I_o >= 3.1$ nA, with the same width as the one computed from open pore current simulation. This clearly demonstrates that this part of the translocation signal is associated to an open pore situation. In addition, the second peak of the distribution, which is centered around 2.0 nA (1.1 nA smaller than $< I_o >$), corresponds to a decrease associated to the passage of the biomolecule through the nanopore. Finally, the SNR calculated from MD simulations is around 5.5 ($I_{RMS} = 0.2$ nA and $\Delta I = 1.1$ nA). This value is very close to experimental SNR of single-layer MoS_2 nanopore of comparable dimension (D =1.4 nm)[8].

From the standard deviation of the open pore current extracted from MD ($\sigma_o = 0.2$ nA), we defined the threshold for the detection of sensing event as $5\sigma_o$. In other words, with this threshold, each sensing event in our simulations corresponds to an event for which at least one amino acid is present inside the pore. Using this value, we avoid the detection of false sensing events for which the peptide is not inside the pore during the recording of ionic current values. In fact, these particular events correspond to a shadow effect of the peptide on the top (or bottom) of the pore which reduces the current values beyond the fluctuations of the open pore case (data not shown). As shown in Fig. 2A, the $5\sigma_o$-threshold corresponds to the part of the signal which belongs to the tail of the ionic current probability distribution. A total of 49 sensing events were detected, representing cumulatively $\sim$20% of the total 2.5μs ionic current time series. Five of them are shown in Fig. 2B, the others being shown in Fig. S1[†]. As observed experimentally for DNA[6], there is a large variability of current versus time signatures within sensing events. For instance, some events maintain fairly constant current drop and others show switching levels and bumps. The origin of such a variability and the relationship between levels and bumps observed in time series and the sequence of the peptide being inside the pore during these events is discussed next.

3.2 Statistical Characterization of Amino-Acid Patterns from Sensing Event

For each of the 49 sensing events detected from ionic current time series (see Fig. 2), we associated ionic current drop values ΔI to each amino acid of KTKEGV peptide. Namely, an ionic current drop value ΔI at time t is associated to a specific amino acid if the amino acid is inside the pore at the same time t in the simulation. Similarly, a motif is associated to a drop ΔI at time t if all amino acids of the motif are present in the pore at the same time. This analysis is described in detail for a specific event in ESI[†]. Knowing which amino acids are in the nanopore as a function of time, we were able to statistically differentiate amino-acid patterns from ionic current fluctuations in our in silico simulations of the experimental device. As shown in Fig. 3A, glutamic acid E4 is the most consistently sensed amino acid from ionic current data (Fig. 2), this specific negatively charged amino acid being present in the pore for 48 over 49 of sensing events. Glycine G5 is the second most sensed residue followed by threonine T2 and lysine K3. The two terminal parts, N-terminal lysine K1 and C-terminal valine V6 are the least sensed residues, with almost a zero probability for K1. This statistical analysis demonstrates that the position of a residue in the primary structure of the peptide drastically influences the sensing of amino acids, more than the size or other properties of its side-chain.

In addition, we computed average ionic current drops $< \Delta I >$ associated to each amino acid of the peptide (Fig. 3B). Except for valine V6 which presents the largest ionic current drop pattern, all $< \Delta I >$ patterns are similar within the error bars. Therefore, despite a large variability of sensing event dwell times and levels of current drop, $< \Delta I >$ patterns along the primary structure of the peptide may not be an appropriate characteristics to sequence proteins. By computing $< \Delta I >$ associated with the presence of protein sequence motifs in the pore (Fig. 3C), there is a much larger variability in current drop values, which confirms that from ionic current traces recorded at the microsecond time scale, measurement of current drops are associated to protein sequence motifs and not specifically to single amino acids. This behaviour comes from the fact that at a given time t and despite the fact that single-layer MoS_2 is an extremely thin material (Fig. 1), only E or T amino acids can reside alone inside the pore. Most of the time, pairs or triplets even quadruplets of amino acids are simultaneously inside the pore during a sensing event. For instance, G residue is always associated to motifs EG, EGV, TKEG, TEG, TKG. Therefore, characterizing amino acids with an average current drop does not make sense in the context of protein

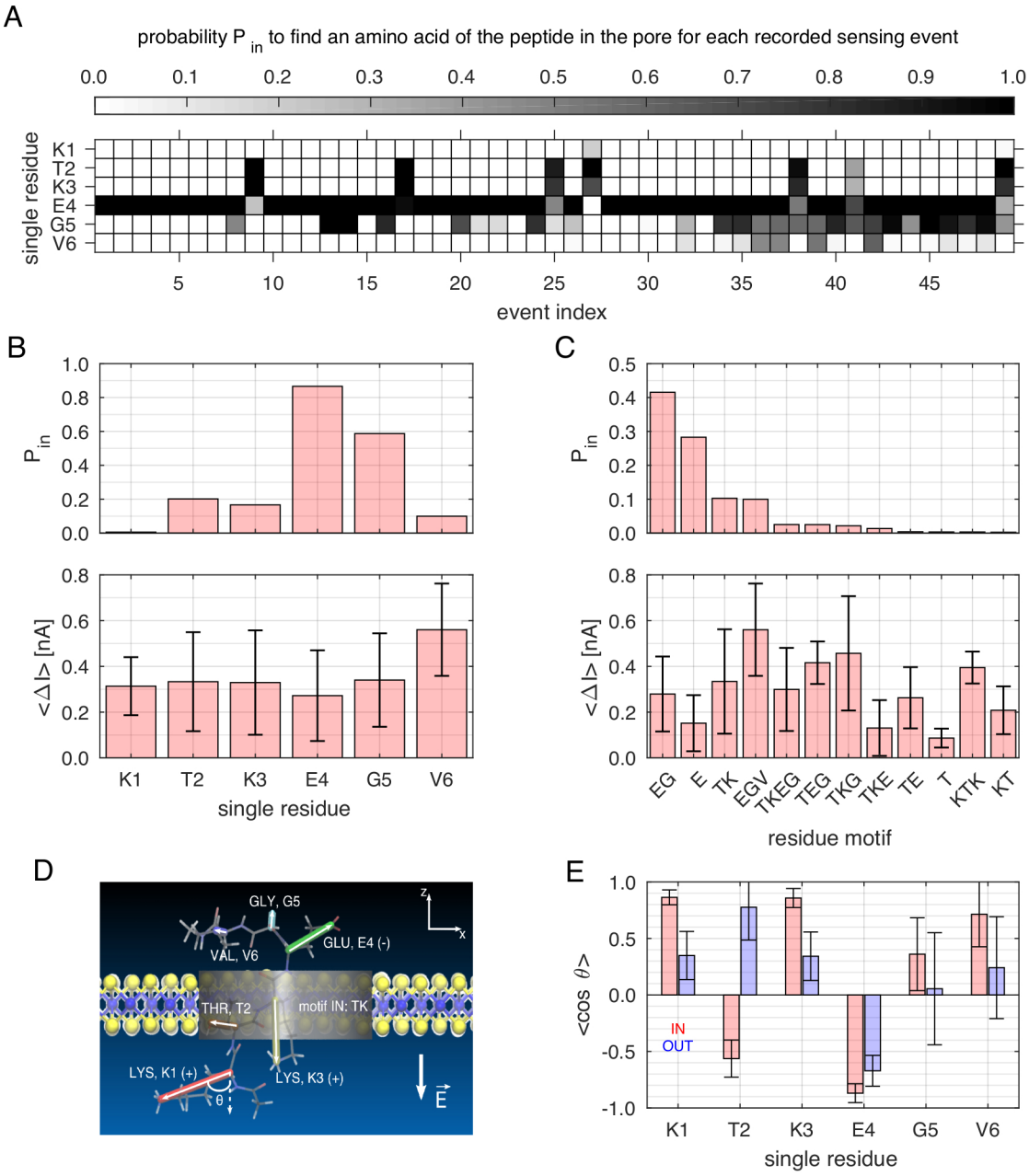

Fig. 3 A) Probability P_{in} to find an amino acid of KTKEGV peptide inside the nanopore *vs* event index computed from MD data shown in Fig. 2. White squares represent an exact zero probability. B) Probability P_{in} to find a single amino acid inside the pore for all 49 sensing events (top panel) and average current drop [in nA] associated to it. Error bars correspond to standard deviation. C) Same as panel B but for protein sequence motifs. D) Atomic representation of KTKEGV peptide inside SL-MoS_2 nanopore during a sensing event. The color code is similar to the one used in Fig. 1. Side chains of each amino acid are shown with tube and their orientation compared to the electric field θ is indicated for K1. E) Average side-chain orientations $< \cos\theta >$ of amino acid along the peptide sequence. Red and blue bars represent average values when the amino acid is inside or outside the pore, respectively. Error bars representing the standard deviation are shown with black lines.

sensing.

Finally, to get insights into what causes drops of ionic current during the passage of the peptide through the nanopore, we studied different mechanisms of peptide translocation. From MD trajectories, we extracted three time series that may cause directly or indirectly ionic current drops: the variation of volume inside the pore due to the presence of the peptide $\Delta V(t)$, the variation of charge $\Delta q(t)$ and, last but not least, the orientation of amino-acid side-chains $\cos\theta(t)$ during the diffusion of the peptide through the pore or at the surface of the nanoporous membrane (Fig. S3[†]). First, we computed temporal correlation between variation of volume and charge and ionic current drops. From this analysis (described in detail elsewhere[†]), it appears that, statistically, all different mechanisms exist, *i.e.* drops are due to increase/decrease of volume or increase/decrease of charge. Second, we computed averaged orientations of amino-

acid side chains when they are inside the pore or diffusing at the surface of the membrane (Fig. 3D). From this analysis, we statistically observed that side-chain orientations change drastically when amino acids are present inside the pore. By averaging side-chain orientations $< \cos\theta >$ over all the 49 sensing events (Fig. 3E), a mechanism is established, *i.e.* side chains of each amino acid are parallel to the electric field when they are inside the pore whereas side chains of amino acid are oriented perpendicular to the electric field when diffusing at the surface of MoS_2 membrane. This mechanism is verified for K, E, G and V residues. In the case of T2, the residue is parallel to the electric field and in the same direction when inside the pore whereas T2 is parallel to the electric field and in the opposite direction when outside the pore. It might come from the fact that threonine amino acid is comprised of an OH group at the end of its side-chain, which means that the orientation is extremely sensitive to the chemistry groups composing amino acids. To conclude, the fact that side-chain orientations are parallel to the electric field when inside the pore means that the volume occupied by the peptide when translocating is relatively small. The consequence is that the nanopore sensor is less sensitive to the side-chain size and volume, at least for single-layer MoS_2 membranes. This may be problematic to design a nanopore protein sequencing device in the future.

3.3 Quantifying the Complexity of Ionic Current Time Series: Permutation Entropy

The statistical analysis of ionic current time series presented in Fig. 3 cannot be directly applied to experimental nanopore measurements due to the fact that this analysis is based on the knowledge of peptide exact position compared to the nanopore and this information is not accessible from experiments. Only sensing event detection using a $5\sigma_o$-threshold can be extracted from experimental time series. Moreover, as described above, the variability of current levels observed in ionic current time series arises from different mechanisms of translocation and most of them cannot be visually associated to the primary structure of the peptide. Nevertheless, information about biomolecule that goes through the pore might be hidden in ionic current time series and the question is where is the information localized. In other words, which levels of current and its variations are relevant for the sequencing technology and how much information do they contain. To answer these questions, we decided to quantify the complexity of ionic current time series extracted from MD using permutation entropy algorithm. As explained in the Materials and Methods section, we applied PE to shifting windows of 1 ns duration along the ionic current time series presented in Fig. 2A. Results are presented in Fig. 4.

We applied PE algorithm to three different sets of data extracted from ionic current time series: *data 1*, which corresponds to all ionic current values below the $5\sigma_o$-threshold recorded from translocation MD simulations (red area in Fig. 2); *data 2*, which corresponds to all ionic current values above the $5\sigma_o$-threshold and recorded from translocation MD simulations (blue area in Fig. 2) and finally, *open*, which corresponds to values of ionic current recorded from an independent open pore MD simulation. As shown in top panel of Fig. 4A, the probability to have PE=0 is much larger from *data 1* than from *data 2* and *open*, which means that much more regular patterns exist in ionic current time series when the peptide is inside the pore. In addition, by looking at the probability distribution for larger PE values (bottom panel of Fig. 4A), distributions look very similar for the three set of data. Since we are interested in regular patterns associated to sensing events, we extracted from *data 1* sub-events for which PE is lower than $< PE > -\sigma_{PE} = 0.11$ (Fig. 4B). We applied this method to all 49 events. As shown in panel B of Fig. 4 for a specific sensing event (index 41, Fig. 3A) of 19.5 ns duration and shown in Fig. 2, 4 sub-events were detected, for which PE is almost null, corresponding to regular linear drops of ionic current, the correlation coefficient of linear fitting R^2 being larger than 0.97 for all the sub-events detected this way (data not shown).

Then, from PE analysis, each sub-event was characterized by the absolute value of its slope, *i.e.* $\Delta I/\Delta t$, which corresponds to an ionic current drop speed. Fig. 4C represents probability distributions of $\Delta I/\Delta t$ extracted from the three sets of data. From the distributions, we can clearly see that *data 2* and *open* time series contain similar information. It confirms that the $5\sigma_o$-threshold is a correct threshold to extract sensing events. Moreover, the distribution from *data 1* is distinct from the two others. Therefore, PE analysis clearly points out that *data 1* contains more regular patterns and that those patterns are characterized by lower values of $\Delta I/\Delta t$, which arises from the fact that drops of current last for longer time. Finally, from *data 1*, we computed average values of ionic current speed $< \Delta I/\Delta t >$ per protein sequence motif, as performed for $< \Delta I >$ in Fig. 3C. As shown in Fig. 4D and E, 4 most probable motifs were extracted: EG, E, TK and EGV. These four motifs are exactly the same as the ones extracted from ionic current characteristics $< \Delta I >$ presented in Fig. 3C and the probabilities P_{in} associated to each motif are also very similar. *It means that applying PE as a filter to extract sub-events within a sensing event is consistent and contains the same quantity of information than the one extracted from ionic current drops using peptide position. This result is crucial since in experiments, the position of the peptide is not available*. Moreover, the variability of $< \Delta I/\Delta t >$ values associated to each motif is not very large, which confirms that extracting the sequence of amino acids translocating through the pore only from PE (Fig. 4) or ionic current drops (Fig. 3) analysis remains challenging. Nonetheless, the detection of biological mutations using PE analysis might be reachable.

3.4 Effect of Biological Mutations on Sensing Event Detection and Characterization

In order to study the impact of biological mutations onto ionic current characteristics shown above for KTKEGV peptide (Fig. 4), we first had a look at amino acids that show singular behaviours. The goal here is to see if similar motifs are sensed by the nanopore and if so, if the same sensed motifs are characterized by the same ionic current drop speed. As already shown in Fig. 3B, E4 residue (glutamic acid, negatively charged) is the best candidate for a mutation since E4 is the amino acid being inside the pore for

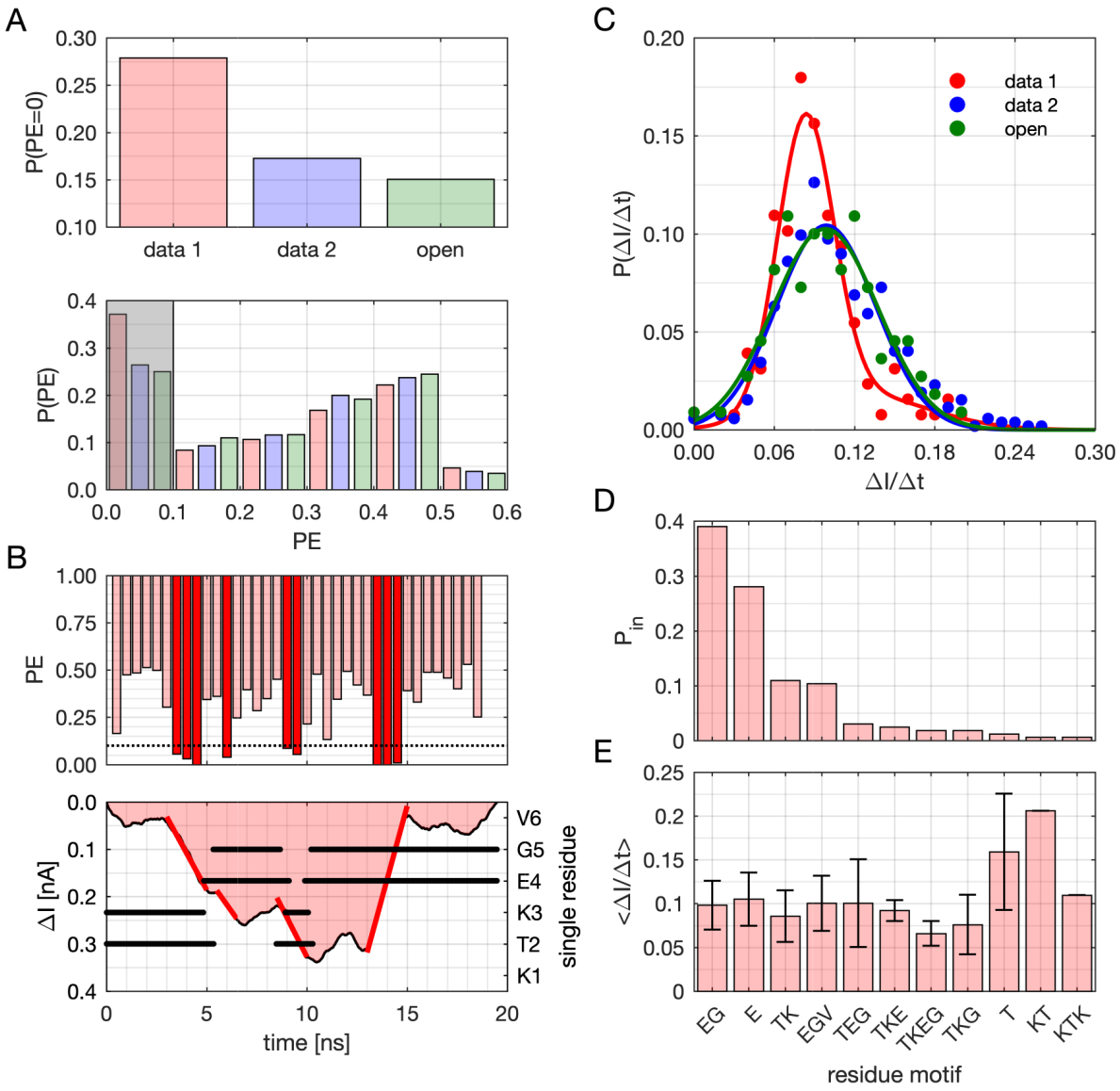

Fig. 4 A) Probability distributions of PE computed from ionic time series presented in Fig. 2A. Time series from translocation simulations were separated in 2 sets: *data 1* and *data 2* (see main text for details). Data named *open* was extracted from open pore simulation. Top panel represents the probability to have exactly PE=0. Bottom panel represents the total probability distribution. B) PE *vs* time (top panel) for a specific sensing event (index 41, Fig. 3A) of 19.5 ns duration and shown in Fig. 2. Ionic current drop *vs* time (bottom panel) is also shown for comparison. Red lines correspond to parts of the signal for which PE<0.11, named hereafter sub-events. C) Probability distribution of ionic current drop speed, which corresponds to the slope of red lines shown in panel B. D) Probability to find a protein sequence motifs inside the pore from sub-events. E) Average current drop speed [in nA/ns] associated to each motif. Error bars correspond to the standard deviation.

the longest period during in silico simultions of nanopore experiments. Therefore, we decided to drastically changed the electrostatic properties of the peptide by performing the mutation E4->K4 (total charge of the peptide from +1 to +3), expecting relatively important repercussions on ionic current time series. Next, from $< \Delta I >$ characteristics shown in Fig. 3B, V6 (Valine, neutral) shows also a different behaviour than the other amino acids with a much larger averaged current drop. We also decided to perform a drastic mutation but less than the previous one by replacing V6 with a positively charged residue, an arginine (V6->R6, total charge of the peptide from +1 to +2). Using the same two procedures as the ones presented above, we characterized sensing events using on one hand, $< \Delta I >$ from the position of the peptide extracted from MD and using, on the other hand, $< \Delta I/\Delta t >$ from PE time series analysis. The goal here is to compare both procedures and to show if they exhibit consistency from independent simulations of mutant peptides.

As shown in Fig. 5, KTKKGV mutant (total charge +3) shows a qualitatively and quantitatively different behaviour than the two others sequence, KTKEGV (+1) and KTKEGR (+2). Drops of current associated with the passage of KTKKGV peptide through the pore are much larger than for the two other sequences, with a maximum peak in the distribution around 0.7 nA (Fig. 5A). KTKEGV and KTKEGR show similar current drop values and dis-

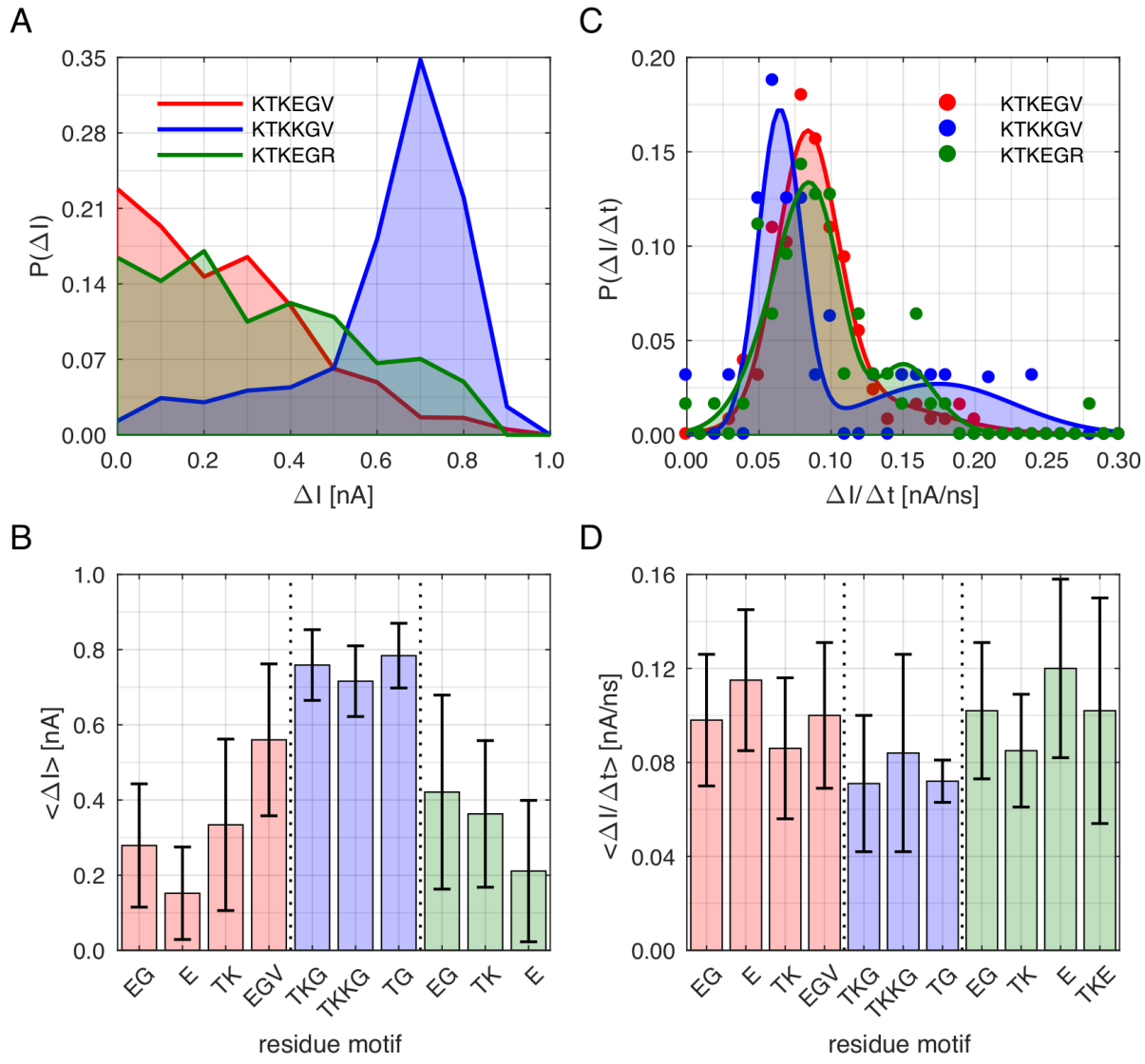

Fig. 5 A) Probability distribution of ionic current drops computed from sensing event detection extracted from MD simulations of the translocation of KTKEGV (red), KTKKGV (blue) and KTKEGR (green) peptides. B) Average current drops [in nA] per protein sequence motif. Error bars represent the standard deviation. C) Probability distribution of ionic current drop speed $\Delta I/\Delta t$ computed from sensing sub-event extracted from MD simulations of the translocation of KTKEGV, KTKKGV and KTKEGR peptides. The color code is the same as in panel A. D) Average current drop speed [in nA/ns] per protein sequence motif. Error bars represent the standard deviation.

tributions look similar. In addition, from each independent sequence, four, three and three protein sequence motifs were extracted from $< \Delta I >$ analysis for KTKEGV, KTKKGV and KTKEGR sequences, respectively (Fig. 5B). Among those ten motifs, three are common to the detection of KTKEGV and KTKEGR sequences, *i.e.* EG, E and TK. Motifs from KTKKGV are all different which means that E4 $\rightarrow$ K4 mutation is drastic and, in this case, MoS_2 nanopore does not sense motifs similar to KTKEGV and KTKEGR and their characteristics cannot be directly compared to those of the other peptides. They are shown in Fig. 5B and D and in Table 1 for information only. Next, we will focus on the three motifs common to KTKEGV and KTKEGR experiments.

First, EG motif is characterized by an average current drop $< \Delta I >= 0.279$ nA from the KTKEGV sequence detection whereas the same motif is characterized by a $< \Delta I >= 0.421$ nA from the KTKEGV sequence detection (Table 1). It means that for measurements of the same motif in different sequences, there is a variation of 50% in the characteristics $< \Delta I >$ of the motif detection. In short, $< \Delta I >$ of a motif is context dependent. Similarly, E motif is showing a difference of 40%. Only protein motif TK shows a relatively small difference of 10%. Therefore, using the procedure based on ionic current drops $< \Delta I >$ *computed from the knowledge of the position of the peptide, does not demonstrate consistency from independent simulations, as similar motifs show different characteristics*.

From PE analysis, probability distributions $P(\Delta I/\Delta t)$ extracted from independent simulations for KTKEGV, KTKKGV and KTKEGR sequences are much more similar than the ones from ΔI (Fig. ref-fig6C). The main differences arise for KTKKGV peptide with a maximum peak shifted to lower values and the presence of a

Table 1 Protein sequence motifs detected from ionic current drop analysis and from permutation entropy analysis applied to independent MD simulations of mutant peptides KTKEGV, KTKKGV and KTKEGR. $<\Delta I>$ and $<\Delta I/\Delta t>$ characteristics are presented.

peptide sequence	protein motifs	$<\Delta I>$ [nA]	$<\Delta I/\Delta t>$ [nA/ns]
KTKEGV	EG	0.279	0.098
	E	0.152	0.115
	TK	0.334	0.086
	EGV	0.560	0.100
KTKKGV	TKG	0.759	0.071
	TKKG	0.716	0.084
	TG	0.784	0.072
KTKEGR	EG	0.421	0.102
	TK	0.363	0.085
	E	0.211	0.120
	TKE	0.375	0.102

shoulder for large values. Although differences are less pronounced than from ΔI analysis, distinguishing sensing sub-events between mutants remains possible and it can be seen as something positive since the characteristics of protein motifs may be similar. From PE, a total of eleven relevant motifs were extracted, four from KTKEGV, three for KTKKGV and four from KTKEGR sequences. Ten of them are the same as the ones extracted from current drops, only TKE motif is detected from KTKEGR. In fact, as shown in Fig. 5D and in Table 1, among those eleven motifs, three are common to the detection of KTKEGV and KTKEGR sequences, *i.e.* EG, E and TK. First, EG motif is characterized by an average current drop velocity $<\Delta I/\Delta t>=0.098$ nA/ns from KTKEGV sequence detection whereas the same motif is characterized by a $<\Delta I>=0.102$ nA from the KTKEGV sequence detection (4% difference) Similarly, motif E shows a difference of 4% and the TK motif a difference of 1%. This result confirms that applying PE algorithm to extract relevant characteristics of protein sequence motifs from ionic current time series is appropriate and particularly consistent.

4 Conclusion

In the present work, we investigated, using MD simulations, nanopore sensing of single-biomolecule through MoS_2 nanopores and the possibility to identify protein sequence motifs. First, we showed from ionic current time series that a threshold of $5\sigma_o$ is necessary to avoid detection of false sensing events. Using this threshold, 49 sensing events were detected for KTKEGV peptide, a motif of alpha-synuclein protein related to Parkinson disease and these events are visually characterized by a large variability of current levels and bumps, as already observed experimentally for DNA sensing. For each of the 49 sensing events, we established the relationship between the presence of the biomolecule inside the pore and current drops measured. This relationship was established using a standard procedure where the average ionic current drop of each residue along the protein sequence was computed from the values of time series for which the residue is inside the pore. To do so, the knowledge of the position of the peptide is needed. Moreover, we showed that single amino acid are mainly characterize by similar average current drops, which makes the sequencing of protein not conceivable. It is much more appropriate to associate sensing events with protein sequence motifs than with single amino acid. Indeed, more than one or two amino acids can be inside the pore at the same time. From this observation, a larger variability of $<\Delta I>$ was observed. This behaviour comes from the fact that mechanisms of translocation of peptides through MoS_2 nanopores are of different kinds, some being related to electrostatic effects, some related to steric effects, some related to both and even some related to none. The major consistent result in terms of mechanisms of translocation observed in the present work is that the orientation of the side chain of amino acids during their passage through the pore is collinear to the applied electric field. Therefore, current drops are less sensitive to the side-chain size and volume of amino acids.

This procedure, based on both current drops and the position of each residue of the peptide cannot be applied to experimental measurements since the later are not accessible from experiments. In the present work, we developed a new procedure that allows to extract similar information by applying a filter to the measured data: the permutation entropy algorithm. PE measures the complexity of a given time series and allows to extract regular patterns. From MD, PE analysis confirms that the $5\sigma_o$-threshold is appropriate to extract relevant sub-events related to the presence of the peptide in the pore. Moreover, PE used as a filter to extract sensing events contains the same information as the procedure which uses the position of the peptide plus the ionic current measurements. The exact same motifs were extracted from both procedure with less inputs using PE. The characteristics extracted from PE is $<\Delta I/\Delta t>$, which corresponds to an ionic current drop speed, *i.e.* how fast (or slow) a drop of current happens when residues of the peptide are located inside the pore. Finally, PE applied to the detection of biological mutations showed consistency from independent measurements, protein sequence motifs detected being identified with similar values of the new parameter defined here, $<\Delta I/\Delta t>$ in different sequences. This result confirms that nanopore sensing of single-biomolecule using MoS_2 nanopores is very promising although reconstructing the sequence of a protein from ionic current time series even using PE procedure remains challenging.

Acknowledgements

The simulations were performed using HPC resources from DSI-CCuB (Université de Bourgogne). The work was supported by a grant from the Air Force Office of Scientific Research (AFOSR), as part of a joint program with the Directorate for Engineering of the National Science Foundation (NSF), Emerging Frontiers and Multidisciplinary Office (grant No. FA9550-17-1-0047). Moreover, this work is part of the project NANOSEQ (2018-2021) and NANO-NEURO-MED (2019-2022) supported by the EIPHI Graduate School (contract ANR-17-EURE-0002), the Conseil Régional de Bourgogne Franche-Comté and the European Union through the PO FEDER-FSE Bourgogne 2014/2020 programs.

Notes and references

1 J. W. F. Robertson and J. E. Reiner, *Proteomics*, 2018, **18**, e1800026.

2 Y. Luo, L. Wu, J. Tu and Z. Lu, *Int J Mol Sci*, 2020, **21**, 2808.

3 W. Shi, A. K. Friedman and L. A. Baker, *Anal. Chem.*, 2017, **89**, 157–188.

4 S. Shekar, D. J. Niedzwiecki, C.-C. Chien, P. Ong, D. A. Fleischer, J. Lin, J. K. Rosenstein, M. Drndić and K. L. Shepard, *Nano Lett.*, 2016, **16**, 4483–4489.

5 Y.-L. Ying and Y.-T. Long, *J. Am. Chem. Soc.*, 2019, **141**, 15720–15729.

6 C.-C. Chien, S. Shekar, D. J. Niedzwiecki, K. L. Shepard and M. Drndić, *ACS Nano*, 2019, **13**, 10545–10554.

7 A. Nicolaï, M. D. Barrios Pérez, P. Delarue, V. Meunier, M. Drndić and P. Senet, *J. Phys. Chem. B*, 2019, **123**, 2342–2353.

8 A. Fragasso, S. Schmid and C. Dekker, *ACS Nano*, 2020, **14**, 1338–1349.

9 C. A. Merchant, K. Healy, M. Wanunu, V. Ray, N. Peterman, J. Bartel, M. D. Fischbein, K. Venta, Z. Luo, A. T. C. Johnson and M. Drndić, *Nano Lett.*, 2010, **10**, 2915–2921.

10 S. M. Gilbert, G. Dunn, A. Azizi, T. Pham, B. Shevitski, E. Dimitrov, S. Liu, S. Aloni and A. Zettl, *Scientific Reports*, 2017, **7**, 15096.

11 M. Graf, M. Lihter, M. Thakur, V. Georgiou, J. Topolancik, B. R. Ilic, K. Liu, J. Feng, Y. Astier and A. Radenovic, *Nature Protocols*, 2019, **14**, 1130–1168.

12 S. J. Heerema, G. F. Schneider, M. Rozemuller, L. Vicarelli, H. W. Zandbergen and C. Dekker, *Nanotechnology*, 2015, **26**, 074001.

13 K. Liu, J. Feng, A. Kis and A. Radenovic, *ACS Nano*, 2014, **8**, 2504–2511.

14 A. J. W. Hartel, S. Shekar, P. Ong, I. Schroeder, G. Thiel and K. L. Shepard, *Analytica Chimica Acta*, 2019, **1061**, 13–27.

15 J. Feng, K. Liu, R. D. Bulushev, S. Khlybov, D. Dumcenco, A. Kis and A. Radenovic, *Nature Nanotechnology*, 2015, **10**, 1070–1076.

16 W. H. Pitchford, H.-J. Kim, A. P. Ivanov, H.-M. Kim, J.-S. Yu, R. J. Leatherbarrow, T. Albrecht, K.-B. Kim and J. B. Edel, *ACS Nano*, 2015, **9**, 1740–1748.

17 T. Gilboa, C. Torfstein, M. Juhasz, A. Grunwald, Y. Ebenstein, E. Weinhold and A. Meller, *ACS Nano*, 2016, **10**, 8861–8870.

18 J. Larkin, R. Y. Henley, V. Jadhav, J. Korlach and M. Wanunu, *Nature Nanotechnology*, 2017, **12**, 1169–1175.

19 M. Graf, M. Lihter, D. Altus, S. Marion and A. Radenovic, *Nano Lett.*, 2019, **19**, 9075–9083.

20 C. Raillon, P. Granjon, M. Graf, L. J. Steinbock and A. Radenovic, *Nanoscale*, 2012, **4**, 4916–4924.

21 C. Plesa and C. Dekker, *Nanotechnology*, 2015, **26**, 084003.

22 J. Zhang, X. Liu, Y.-L. Ying, Z. Gu, F.-N. Meng and Y.-T. Long, *Nanoscale*, 2017, **9**, 3458–3465.

23 R. Luo, F. J. Sedlazeck, T.-W. Lam and M. C. Schatz, *Nature Communications*, 2019, **10**, 998.

24 S. Jia, H. Luo, Q. Gao, J. Guo, J. Su, J. Meng and X. Wu, 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019, pp. 1–5.

25 M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp and P. A. Pevzner, *PLOS Computational Biology*, 2017, **13**, e1005356.

26 A. D. Carral, C. S. Sarap, K. Liu, A. Radenovic and M. Fyta, *2D Mater.*, 2019, **6**, 045011.

27 A. Barati Farimani, M. Heiranian and N. R. Aluru, *npj 2D Materials and Applications*, 2018, **2**, 1–9.

28 C. E. Shannon, *Bell System Technical Journal*, 1948, **27**, 379–423.

29 H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, 2nd edn., 2003.

30 C. Bandt and B. Pompe, *Phys. Rev. Lett.*, 2002, **88**, 174102.

31 M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, J. Q. Trojanowski, R. Jakes and M. Goedert, *Nature*, 1997, **388**, 839–840.

32 U. Dettmer, A. J. Newman, V. E. von Saucken, T. Bartels and D. Selkoe, *Proc. Natl. Acad. Sci. U.S.A.*, 2015, **112**, 9596–9601.

33 S. Plimpton, *Journal of Computational Physics*, 1995, **117**, 1–19.

34 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33–38.

35 M. Riedl, A. Müller and N. Wessel, *Eur. Phys. J. Spec. Top.*, 2013, **222**, 249–262.