A Route toward Protein Sequencing using Solid-State Nanopores Assisted by Machine Learning

Andreina Nohemi Urquiola Hernandez[1], Adrien Nicolaï [1,*], Christophe Guyeux[2] and Patrick Senet[1]

[1] Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Université de Bourgogne Franche-Comté

[2] Laboratoire FEMTO-ST, UMR 6174 CNRS-Université de Bourgogne Franche-Comté

adrien.nicolai@u-bourgogne.fr

Abstract

Solid-State Nanopores made of 2-D materials such as MoS_2 have emerged as one of the most versatile sensors for single-biomolecule detection, which is essential for early disease diagnosis (**biomarker detection**). One of the most promising applications of SSN is DNA and protein sequencing, at a low cost and faster than the current standard methods. The detection principle relies on measuring the relatively small variations of ionic current as charged biomolecules immersed in an electrolyte traverse the nanopore, in response to an external voltage applied across the membrane. The passage of a biomolecule through the pore yields information about its structure and chemical properties, as demonstrated experimentally particularly for DNA molecules. However, **protein sequencing using SSN remains highly challenging** since the protein ensemble is far more complex than the DNA ensemble [1]. In this work, we performed extensive unbiased all-atom classical Molecular Dynamics simulations to produce data of translocation of biological peptides through single-layer MoS_2 nanopores. Peptide made of 12 different amino acids from the different families (non polar/hydrophobic, polar/neutral, basic and acidic) were chemically linked to a short polycationic charge carrier. First, ionic current time series were computed from MD and peptide-induced blockade events were extracted and characterized using structural break detection. Second, clustering (unsupervised learning) of ionic current subdrops and duration using Gaussian Mixture Model was applied. Using this technique, we demonstrate that each amino acid presents a large diversity of ionic current characteristics, however, **charged amino acids** were distinguished from the others. These promising findings may offer a route toward protein sequencing using MoS_2 solid-state nanopores.

Solid-State Nanopores

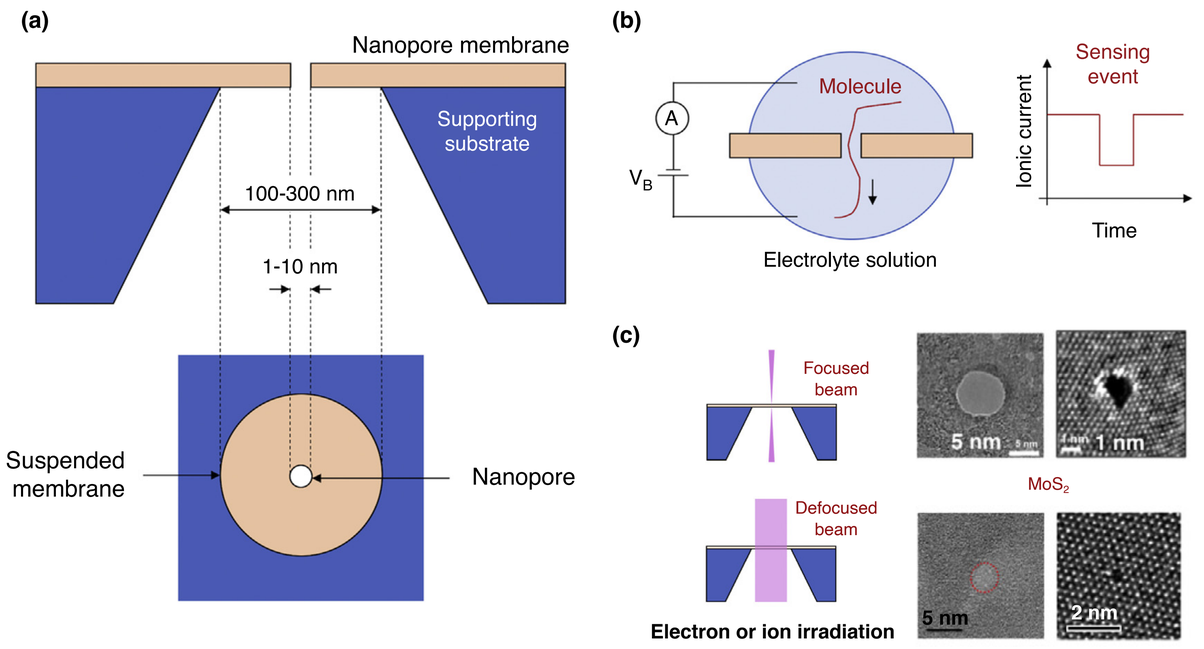


Figure: Schematic of typical nanopore sensor chips and nanoporous membranes [2]. (a) Cross-sectional and top view of a silicon-based or 2D material membrane with a nanopore suspended over a circular aperture in a supporting substrate (e.g. silicon, glass). (b) Voltage-induced translocation of an elongated charged biomolecule through a single nanopore immersed in an electrolyte solution (left) giving rise to ionic current changes, representing a sensing event (right). (c) Controlled fabrication of single 2D nanopores and ordered nanopore arrays using electron or ion irradiation on MoS_2.

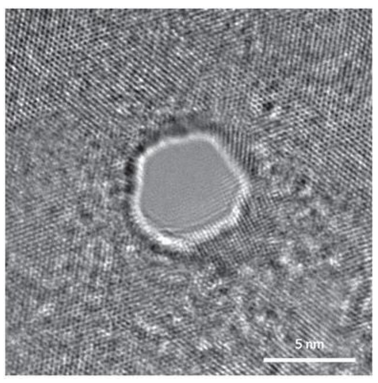
all-atom classical Molecular Dynamics in explicit solvent

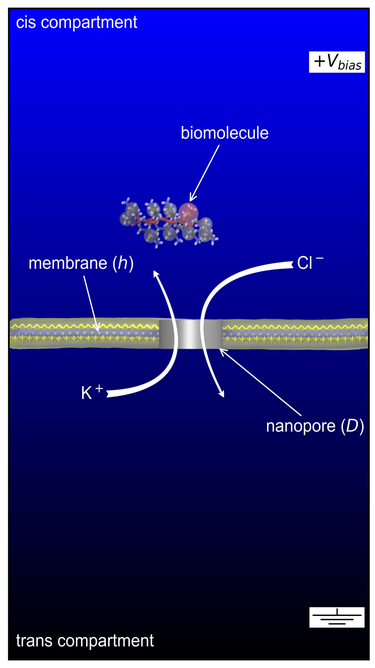
- GROMACS software package (2018.2)
- Machines: AMD EPYC 7302 @ 3Ghz (2 processors, 16 cores/processor)
- Size of the system: 100,000 atoms - Total simulation time: 150 μs
- Scaling: 150 ns / day on 256 cores - Total CPU time: 6 millions of hours

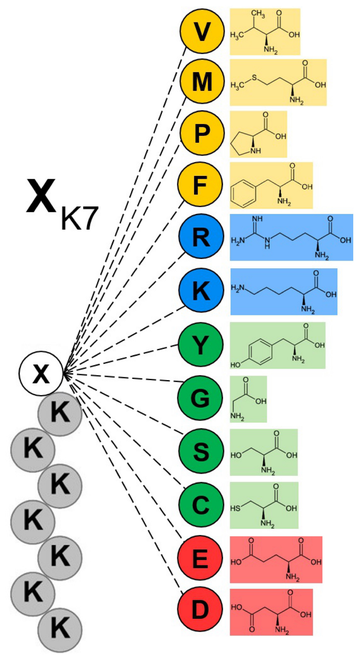
single-layer MoS_2

D = 1.3 nm

h = 0.3 nm







Results

Ionic Current Time Series

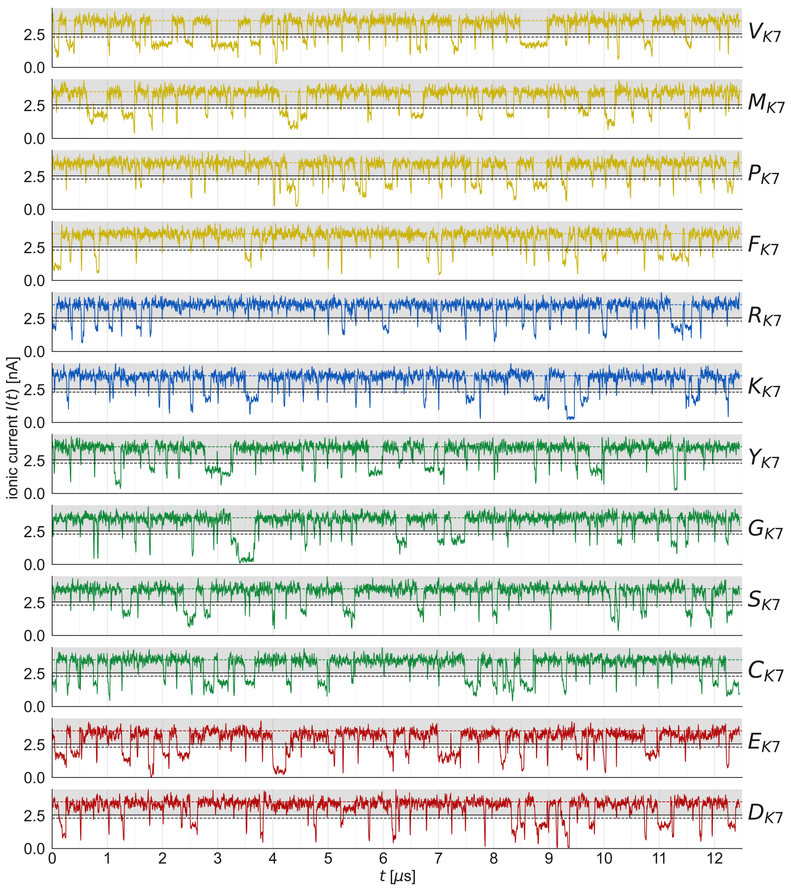


Figure: Ionic current (nA) *vs.* time (in μs) recorded during MD simulations.

Big data: statistical analysis of fingerprints

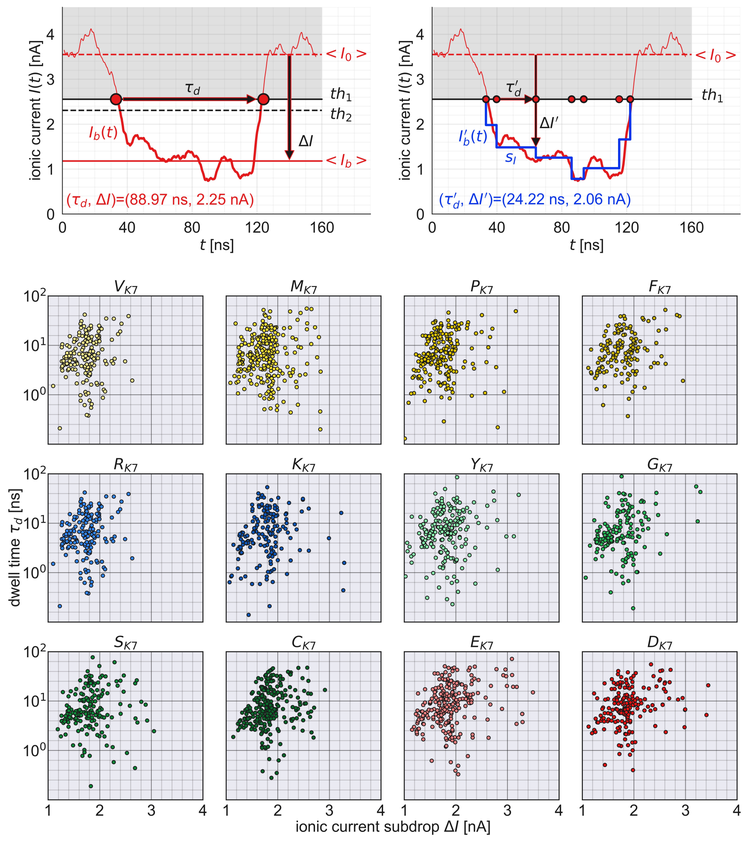
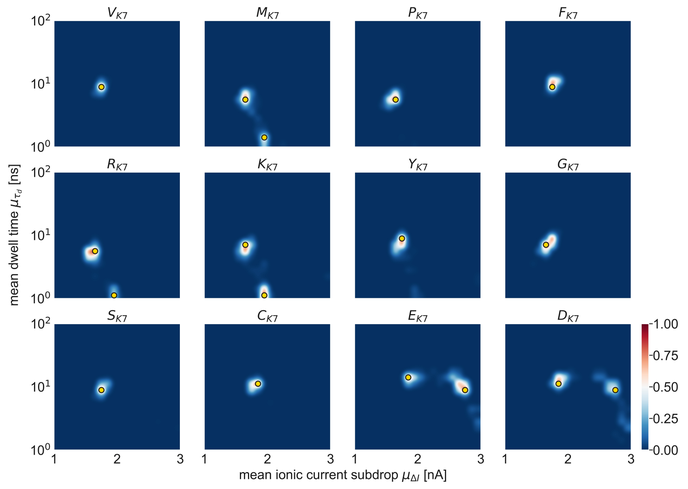


Figure: Top panels: Detection and characterization of peptide-induced blockade events. Bottom panel: Ionic current subdrops (in nA) *vs.* dwell time (in ns) detected for each amino acid during MD.

Unsupervised Learning: GMM Clustering



	V_{K7}	M_{K7}	P_{K7}	F_{K7}	R_{K7}	K_{K7}	Y_{K7}	G_{K7}	S_{K7}	C_{K7}	E_{K7}	D_{K7}
V_{K7}	100											
M_{K7}	6	100										
P_{K7}	5	76	100									
F_{K7}	77	2	2	100								
R_{K7}	1	76	84	0	100							
K_{K7}	11	90	60	5	67	100						
Y_{K7}	79	43	36	59	25	44	100					
G_{K7}	65	57	58	49	37	57	88	100				
S_{K7}	95	4	3	90	1	8	74	60	100			
C_{K7}	42	0	0	82	0	2	29	25	68	100		
E_{K7}	0	0	0	6	0	0	0	0	4	13	100	
D_{K7}	2	0	0	30	0	0	1	2	20	57	79	100

Figure: Top panel: 2-D PDFs of Gaussian means for subdrops and dwell time extracted from Gaussian Mixture Model clustering. Bottom panel: Matrix of similarity between 2-D PDFs. Values are given in percent..

References

[1] A. Nicolaï and P. Senet. Challenges in Protein Sequencing using MoS_2 Nanopores, **2022**. In Single Molecule Sensing Beyond Fluorescence. Nanostructure Science and Technology. Springer, Cham.

[2] G. Danda and M. Drndic. Two-dimensional nanopores and nanoporous membranes for ion and molecule transport. Current Opinion in Biotechnology **2019**, 55: 124–133.

Acknowledgements

The simulations were performed at DSI-CCuB (Université de Bourgogne). The work is part of the project NANO-NEURO-MED (2019–2022) and SEPIA (2021-2024) 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 and 2021/2027 programs.