XIII Argentine Congress of Bioinformatics and Computational Biology
XIII International Conference of the Iberoamerican Society of Bioinformatics
III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

Accurate Detection of RNA Modifications on Nanopore Sequencing Data at Signal-level with Machine Learning

Servi, L. [1,2] ; Petrillo, E. [1,2] ; Stegmayer G. [3].

[1] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Fisiología, Biología, Molecular, y Celular, Buenos Aires, Argentina

[2] CONICET-Universidad de Buenos Aires, Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), C1428EHA, Buenos Aires, Argentina

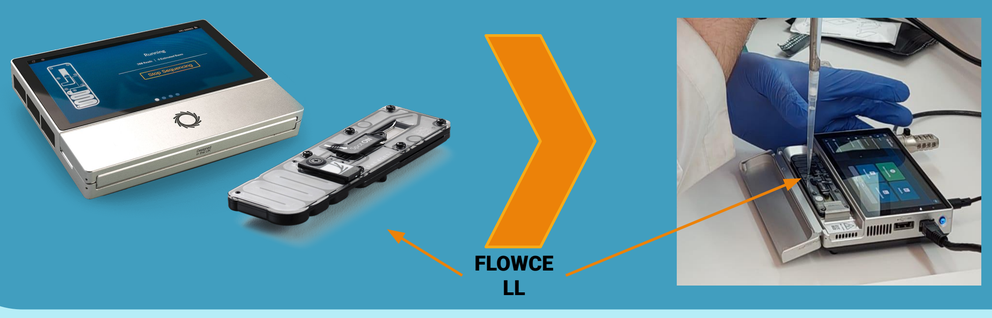
[3] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina

INTRODUCTION

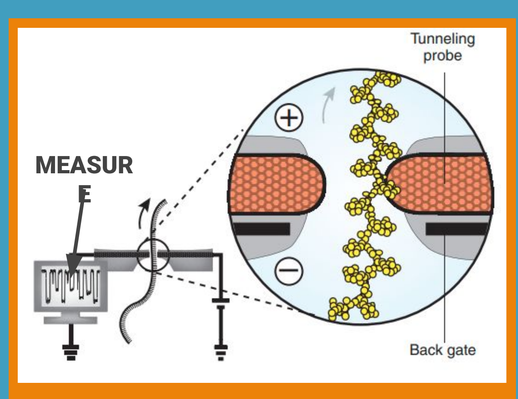
Accurate RNA modification identification is vital for understanding gene expression regulation and its implications in biotechnology. This study focuses on precise RNA modification detection using Oxford Nanopore Technology (ONT) sequencing data. RNA modifications play a central role in gene expression and post-transcriptional processes, underscoring their critical importance in diverse fields. Traditional methods often oversimplify by relying on reading errors as indirect indicators of RNA methylation, resulting in lower accuracy and a lack of specific modification details.

MATERIALS AND METHODS

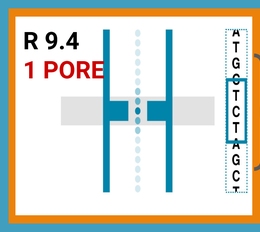
ONT enables comprehensive analysis from its raw electrical signal data. With Direct RNA Sequencing (DRS), numerous RNA features can be extracted. Each consumable flow cell contains 2048 pores actively incorporating nucleotide chains during approximately 48 hours of sequencing.

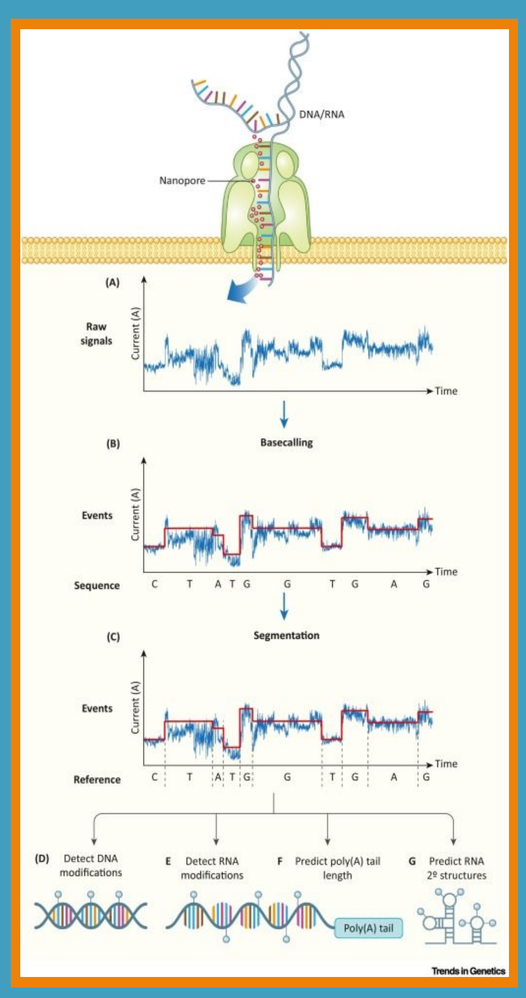


This device detects electric current changes from nucleic acids passing through the pore, measured in picoAmperes (pA) over time. A deep neural network, like Guppy, processes the signal to generate the nucleotide sequence in a step known as "basecalling."



Each pore covers approximately 5 nucleotides as they pass through, forming a signal from a k-mer instead of a single base.

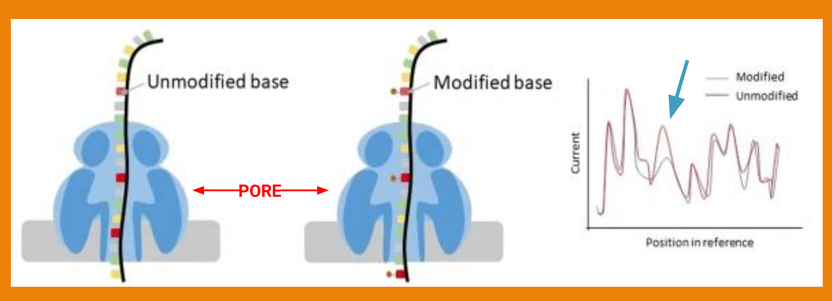


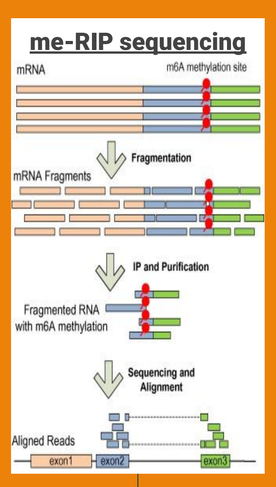


PROPOSAL

Our approach focuses on the development of a Machine Learning (ML) model based on Natural Language Processing (NLP) models, harnessing the electrical current generated during sequencing and taking advantage of basecalling errors as fundamental features.

The modified bases, when passing through the pore, generate a distorted signal in relation to the unmodified base.





For training, we will employ:

- ***me-RIPseq*** **data + Direct RNA Sequencing (DRS)**
- **Raw read signal from Nanopore**
- **Regular basecalled read (no modified nucleotides)**
- **Error rate from basecalling (of each base)**

Basecalled read (fastq with ID)

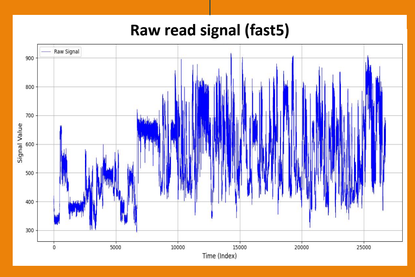
6dc1ebda-d1a3-42a7-a0f3-f9b5908ff7ad

AUACCAAACCAGAGAAACAAUACAAUGUG
AGAAUCACAGAUUACAGAGAGCGAGAGAG
AUCUGACGGACGAAGC

Error per position (base) in basecalled reads

	contig	position	reference_kmer	read_index	strand	event_index	event_level_mean	event_stdv	event_length	model_kmer	model_mean	model_stdv	standardized_level
729697	Chr3	22900066	TCATT	186573	t	2241	86.77	4.829	0.02224	NNNNN	0	0	inf
729698	Chr3	22900068	ATTGT	186573	t	2240	80.81	0.899	0.00564	ACAAT	84.13	3.02	-0.91
729699	Chr3	22900068	ATTGT	186573	t	2239	80.3	1.731	0.00498	ACAAT	84.13	3.02	-1.09
729700	Chr3	22900069	TTGTT	186573	t	2238	86.81	2.159	0.01228	AACAA	87.82	3.18	-0.26
729701	Chr3	22900070	TGTTA	186573	t	2237	89.98	2.223	0.00764	TAACA	93.83	3.45	-0.92
729702	Chr3	22900071	GTTAA	186573	t	2236	90.59	1.558	0.00432	TTAAC	90.31	3.04	0.09
729703	Chr3	22900072	TTAAT	186573	t	2235	84.06	1.725	0.0073	ATTAA	85.3	2.46	-0.42
729704	Chr3	22900073	TAATT	186573	t	2234	92.53	1.914	0.00631	AATTA	101.02	5.53	-1.26
729705	Chr3	22900073	TAATT	186573	t	2233	96.4	1.525	0.00432	AATTA	101.02	5.53	-0.69
729706	Chr3	22900073	TAATT	186573	t	2232	99.32	2.842	0.00198	AATTA	101.02	5.53	-0.25

Raw read signal (fast5)



TRAIN

Deep Neural Network

NEW Basecalled read (+ modifications)

6dc1ebda-d1a3-42a7-a0f3-f9b5908ff7ad

AUACCAm6ACCAGAGAAACAAUACAAUGΨ
GAGAAUm6ACACAGAUUACm6AGAGAGCGI
GAGAGAUCΨGACGGICGAAGC

CONCLUSIONS

Our proposal is a novel approach to the precise detection of RNA modifications with a ML model based on NLP and sequencing data. The accurate identification of RNA modifications will undoubtedly enhance our knowledge of gene regulation and open up new avenues for biotechnological applications.

REFERENCES

- Pagès-Gallego, M., de Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. Genome Biol 24, 71 (2023). https://doi.org/10.1186/s13059-023-02903-2
- Zhong ZD et al. Systematic comparison of tools used for m6A mapping from nanopore direct RNA sequencing. Nat Commun. 2023 Apr 5;14(1):1906. https://doi.org/10.1038/s41467-023-37596-5
- Matthew T Parker et al. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification eLife 9:e49658 (2020) https://doi.org/10.7554/eLife.49658
- https://github.com/nanoporetech/bonito
- https://distill.pub/2017/ctc/