A comparative method for the detection of various DNA and RNA modifications based on nanopore sequencing current signals

Ivan Vujaklija, dipl. ing.

mentor: Prof. Mile Šikić, PhD
University of Zagreb Faculty of Electrical Engineering and Computing

SVEUČILIŠTE U ZAGREBU
Fakultet elektrotehnike i računarstva

1. Introduction

It is widely recognized that epigenetic (DNA) and epitranscriptomic (RNA) modifications, play key roles in regulating major cellular process. Thus, precise detection of epigenetic and epitranscriptomic markers is of an outmost importance both for basic as well as for applied research and as such it is essential for making a transition towards both more precise and more personalized medicine. Since DNA/RNA modifications are not affected by nanopore sequencing, nanopore sequencing coupled with subsequent analysis of signals by computational methods is an attractive alternative to classical experimental methods for detecting epigenetic and epitranscriptomic modifications.

2. Problem statement

Although, more than 40 epigenetic and more than 170 epitranscriptomic RNA modifications have been reported [1], only a small fraction of them can be detected by currently available experimental methods.
Supervised methods (especially deep learning) have recently shown great success in many different applications but they have major limitations in this particular setting. Namely,
(i) They are unable to detect novel/unknown modifications since they require training datasets. **(ii)** Even if the training dataset is available it has to be of sufficient size and quality. This is a major challenge due to the limitations of currently used experimental methods. **(iii)** Supervised learning is sensitive to overfitting which is of particular concern having in mind the unprecedented variability of in vivo biological data, across species, cell lines and environmental conditions.

3. Methodology

Due to the above mentioned limitations of supervised methods in this particular problem setting, we decided to use an unsupervised approach. Nanosequencing dataset used in our study was taken from [2].

Our methodology consists of six steps:

(i) PCR amplification of control dataset which removes modification markers; **(ii)** nanopore sequencing of control and native datsets; **(iii)** alignment of control and native reads to the reference sequence (e.g. genome); **(iv)** resampling of nanopore signals; **(v)** statistical testing and finally, **(vi)** classification step

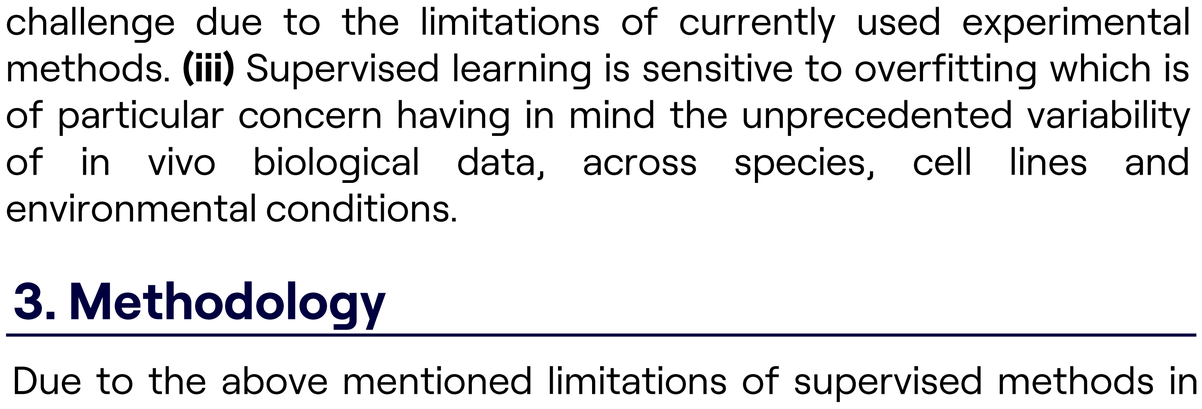

Figure 1. Flowchart of the proposed framework. Only key steps are shown

4. Results

Preliminary results (Precision-Recall curves) on ribosomal RNA of *E.coli* and *S.cerevisiae* are shown in the figures below. Due to short lengths of ribosomal subunits, results obtained on four ribosomal subunits (two of *E.coli* and two of *S.cerevisiae*) were combined into single test dataset, and are shown in Figures 2 & 3.

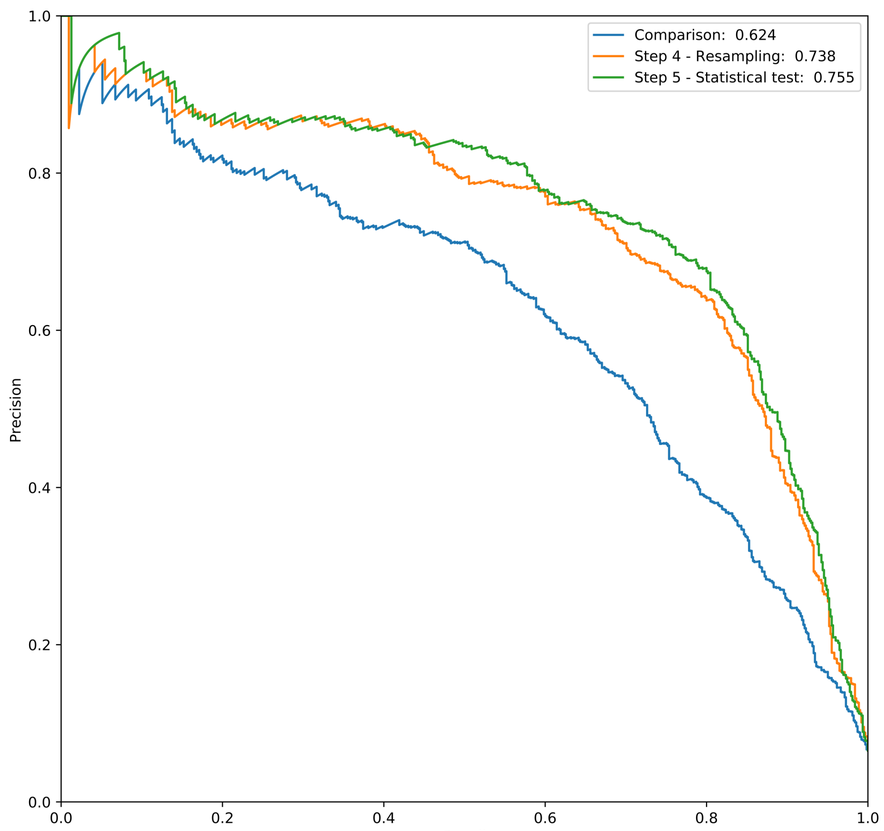

Figure 2. Precision-Recall curves obtained on rRNA datset. Coverge ~50 nucleotides

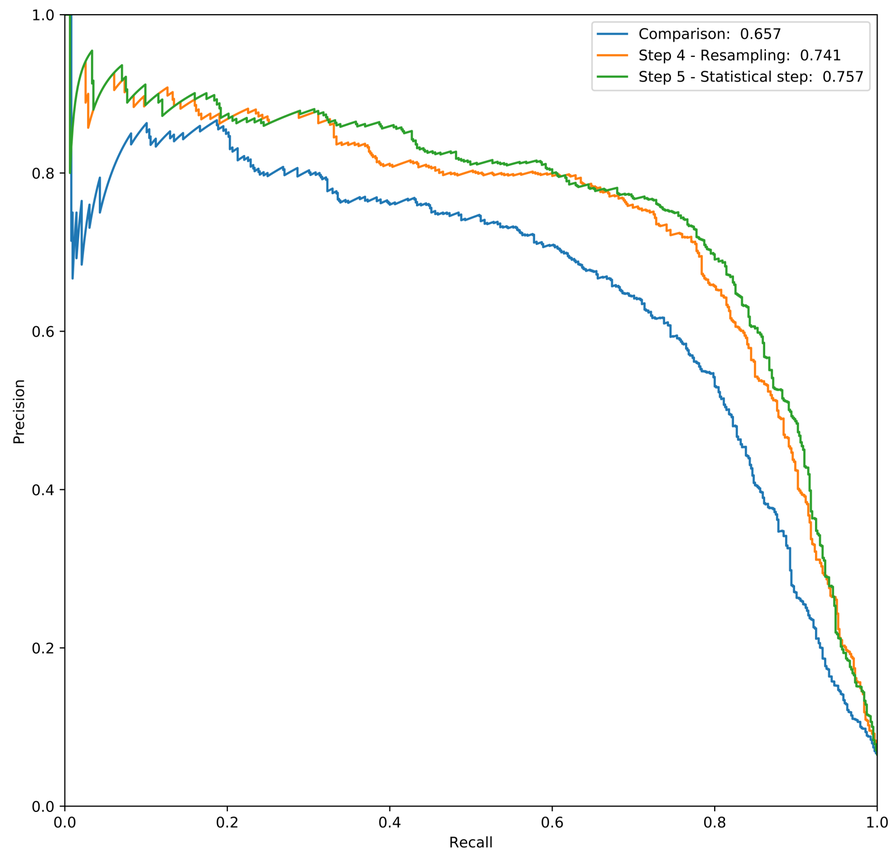

Figure 3. Precision-Recall curves obtained on rRNA datset. Coverge ~100 nucleotides

5.Conclusion

Here we present initial results of our unsupervised approach for the detection of a wide spectrum of epigenetic and epitranscriptomic modifications. As shown in Figures 2 & 3, substantial improvement was obtained on the proof-of-concept *E.coli* and *S. cerevisiae* ribosomal RNA datasets.

References:

[1] Jenjaroenpun P et al. Decoding the epitranscriptional landscape from native RNA sequences. Nucleic Acids Res. 2021 Jan 25
[2] Stephenson W et al. Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. Cell Genom. 2022 Feb 9

Contact:

Ivan Vujaklija, dipl ing.
ivan.vujaklija@fer.hr
Tel. Broj 091 799 2901