Properties of Current Signals in Nanopore Sequencing

V. Plocková[1] and K. Sedlář[1]

[1]Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Czechia

E-mail: xplock00@vut.cz, sedlar@vut.cz

Abstract—Oxford Nanopore technologies brought new and revolutionary technology in the field of DNA sequencing. Their sequencing device measures changes in the electric current flowing through pores together with DNA. This work aims to describe differences between raw signals produced by various sequencing kits and sequencing flowcells while sequencing several different bacteria. Two datasets combining five different organisms, two sequencing kits, and two types of flowcells were used to analyze various statistical parameters that would be suitable for the description of current signals gathered from nanopores.

Keywords— current signal, Oxford Nanopore Technologies, sequencing, statistical parameters, clustering.

1. INTRODUCTION

Nanopore sequencing is a hot topic of bioinformatics, currently being mentioned in a large number of scientific articles. In 2014, the company Oxford Nanopore Technologies (ONT) released its first portable nanopore sequencing device, thus enabling a revolution in sequencing, mainly due to the ability of DNA sequencing anytime and anywhere without the need for a laboratory. Nanopore sequencing has the potential to offer cost-effective genotyping, high mobility for testing, and fast real-time sample processing [1]. The principle of this technology is quite simple. First, you need to prepare the library and then place the sample in the sequencing device. After the voltage is applied, the DNA molecule begins to pass through the nanopore and begins to generate ion current. Current changes correspond to individual nucleotides of DNA. It allows fast sequencing of long individual DNA molecules [2]. This technology also offers an important tool in the fight against antimicrobial resistance. Usually, raw currents are immediately decoded into DNA sequences in a process called basecalling. A majority of studies sees basecalling as a black box using artificial intelligence and neural networks without working with signals themselves [3]. Here, we are dealing with the properties of raw signals, which is a neglected topic of nanopore sequencing.

2. MATERIALS AND METHODS

Data for this article were provided from a database of sequences at Department of Biomedical Engineering, FEEC, BUT. The raw data were stored in FAST5 format. Basecalling was done using Guppy [4] to get FASTQ files. For subsequent analysis, two datasets were created.

The first dataset was created using a single organism *Clostridium beijerinckii* to find out if the calculated statistical parameters of different signals, sequenced from one organism, differ.

The second dataset consisted of eight samples containing data from five organisms. The samples were selected to represent as many types of sequencing kits (LSK-Ligation Sequencing Kit vs. RBK-Rapid Barcoding Kit) and flowcells (MinION vs. Flonge) as possible. The dataset was created from the following organisms: *C. beijerinckii* (CB; MinION, LSK), *Klebsiella pneumoniae* (KP_1, MinION, RBK), *K. pneumoniae* (KP_2; Flonge, RBK), *K. pneumoniae* (KP_3; Flonge, LSK), *K. pneumoniae* (KP_4; MinION, LSK), *Pantoea agglomerans* (PA; MinION, LSK), *Schlegelella thermodepolymerans* (ST; MinION, LSK), and *Janthinobacterium lividum* (JL; MinION, RBK).

The first thing that needed to be done was signal preprocessing. The signals in FAST5 files are compressed, i.e., originally measured signals in pA are further converted and stored as 16-bit integer

values, so it was necessary to convert the signals back to picoampere values. The next step was median filtering, despite that ONT states on its website that signals no longer need to be further filtered [2].

The signals turned out to be disturbed by impulse noise, see Fig 1. If the data were not filtered, the following analysis would be affected by outliers and calculations could indicate erroneous results. This could lead to bias in the cluster analysis. To filter signals a median filter was used, the window length was set to five. This window size was selected because a larger window could filter out individual k-mers of length five that correspond to individual current levels [5].

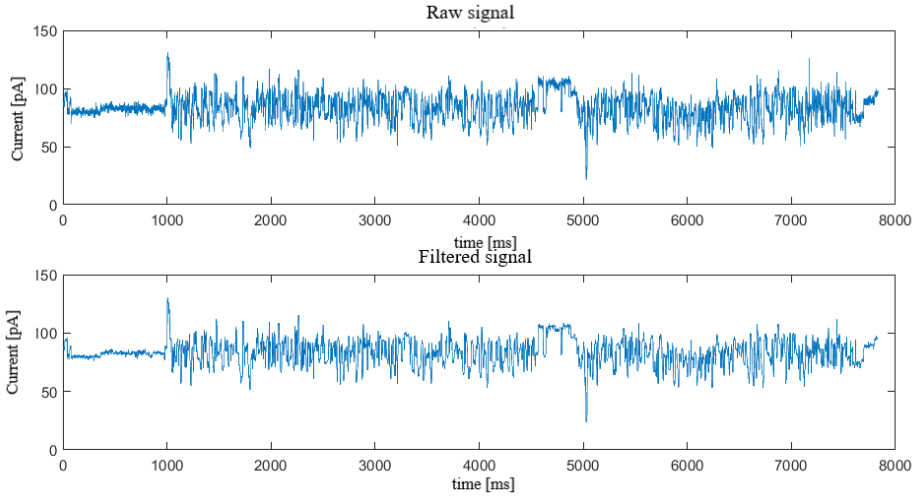

Figure 1: Signals before and after preprocessing

Then signals were described by chosen statistical parameters and statistical tests listed below were performed. The last step was hierarchical clustering. First, the distance matrix (or p-distance) was calculated when the Euclidean distance was chosen for calculation of the distance matrix. Futhermore, for mean and basecalled sequences hierarchical clustering using the c-link method was performed.

3. RESULTS AND DISCUSSION

Each signal in the dataset was described by statistical parameters. These are the mean, variance, standard deviation, coefficient of variation, skew coefficient, sharpness coefficient, and Hjorth descriptors, which include activity, complexity, and mobility. The data were plotted using box plots, see Fig 2 and Fig 3. (other parameters has not been shown)

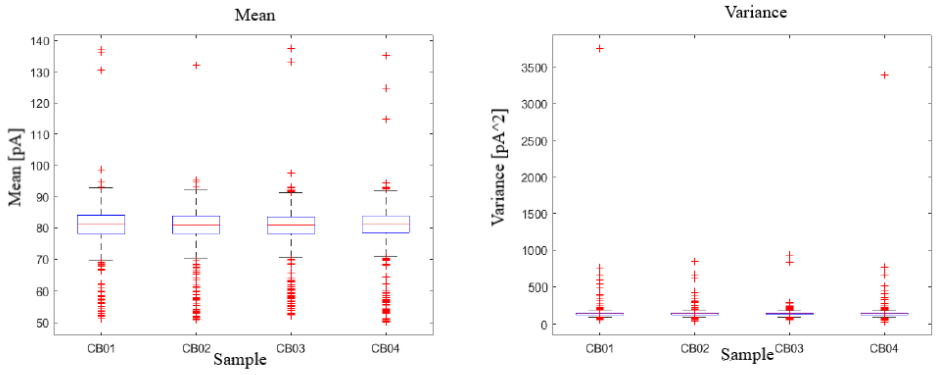

Figure 2: Mean value and variance of signals of the first dataset

Significant differences between the Fig 2 and Fig 3 can be seen. While the values of mean and variance for the first dataset do not differ significantly, the boxplots overlap, for the second dataset, we can see that each organism produces different signals, and even sequencing kits and flow cells produce different signals and are distinguishable from each other. This demonstrates that filtered nanopore current signals could be used to distinguish between various organisms and various sequencing kits and flowcells.

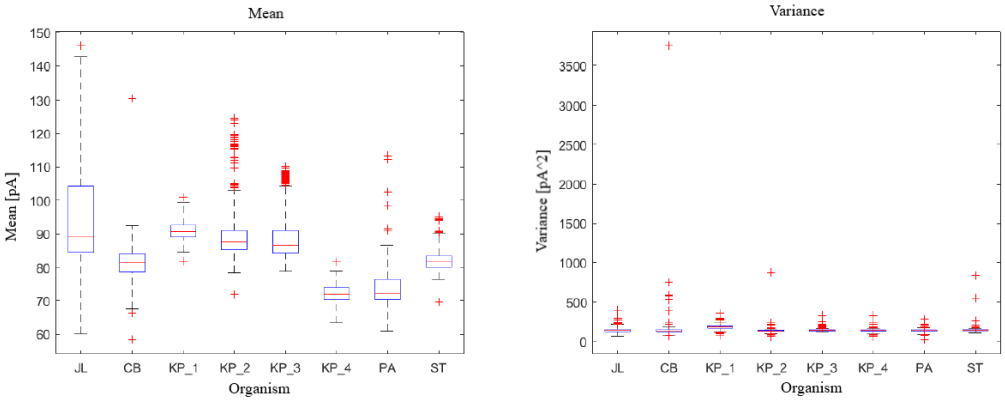

Figure 3: Mean value and variance of signals of the second dataset

In order to make the evaluation objective, statistical tests were calculated. The first – Shapiro-Wilk to find out if the data has a normal distribution and what statistical test to use next.

Since the data were not normally distributed, the Kruskal-Wallis test was used. This test is just an intermediate step before Tukey's test, in order to exclude parameters unsuitable for further analysis. In Table 1 you can see the results for selected statistical parameters for the first and the second datasets. The results show that for the first dataset, no statistical parameter can distinguish the samples from each other, but the results for the second dataset were already more satisfactory – most parameters could distinguish one sample from at least one other

Table I: The result of Kruskall-Wallist test for the first and second dataset

Statistical parameters	p-values of the first dataset	p-values of the second dataset
mean value	0,2799	0
standard deviation	0,2687	0
coefficient of variation	0,2349	0
skew coefficient	0,3823	2.4889e-198
sharpness coefficient	0,0283	4.1757e-168
variation	0,2687	0
mobility	0,2071	0
activity	0,2687	0
complexity	0,7325	0

Tukey's test was then performed (results has not been shown). The results for the first dataset are that the individual sample pairs do not differ from each other. In the second dataset, the samples are best distinguished by the coefficient of variation, mobility and mean. Other parameters with satisfactory results are skewness and complexity because the sample KP_1 to KP_4 is one organism, so they should be difficult to distinguish from each other and this is not entirely the rule. The only difference is the different library preparation procedure and that the samples were sequenced on different flowcells. The comparison of this parameter with standard clustering of sequences is shown in Fig 4, you can see differences between results of cluster analysis. In the basecalled sequences we can see the correct evaluation of the *Klebsiella* samples, only the KP_3 sample clustered far from the other samples from the same organism. This may be due to samples sequenced on the Flonge flowcell that can give poor results. But if we compare it to clustering of mean parameter where *Klebsiella* samples (and not only *Klebsiella* ones) are clustered completely differently, this leads us to the idea that sequencing kits and flowcells also need to be considered when a following analysis and basecalling are performed.

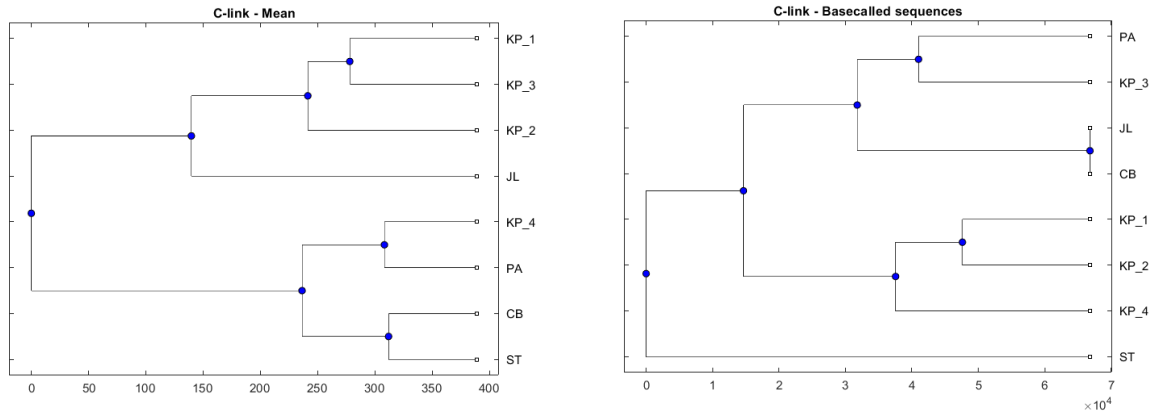

Figure 4: Example of clustering basecalled reference sequence and one of the statistical parameters (in this in it is mean)

4. CONCLUSIONS

The analysis of the first dataset showed that the signals are species-specific. It does not matter which signals we choose. Although they represent random parts of a genome, their signals properties are the same. This is important for the classification of organisms directly from the raw signals without the need of basecalling. The same analysis was repeated on the second dataset containing five organisms, two different sequencing kits, and two flowcells. Some parameters were able to distinguish among organisms and devices better than others, for example, mean value. Some results brought a surprising finding that even in the case of sequencing one organism we can get completely different results. This is most obvious for the parameter of mobility, where various samples of the organism *K. pneumoniae* show significant differences. The possible explanation is, as mentioned above, the use of various ONT tools. This may play a role in the following decoding of signals by different neural networks and, if this fact is considered, the accuracy of ONT technology might be theoretically improved. In conclusion, we can say that raw signals can distinguish organisms, but they don´t have to always be completely correct, and it will probably be necessary to filter the signals and eliminate outliers.

ACKNOWLEDGMENT

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] J. Miten. “Bioinformatic Analysis of Nanopore Data,” in Nanopore sequencing: An introduction, vol 1. Word Scientific., 2019, pp. 147-158. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/9789813270619_0009

[2] “How nanopore sequencing works.” Nanoporetech.com Available: https://nanoporetech.com/how-it-works

[3] R. R. Wick, L. M. Judd, and K. E. Holt, “Performance of neural network basecalling tools for Oxford Nanopore sequencing”, *Genome Biology*, vol. 20, no. 1, 2019. doi 10.1186/s13059-019-1727-y

[4] Oxford Nanopore Technologies, Oxford, United Kingdom. *Guppy Protocol*. (2020). [Online]. Available: https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revt_14dec2018

[5] F. J. Rang, W. P. Kloosterman and J. De Ridder, „From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy“, *Genome Biology,* vol. 19, no. 1, 2018