DOI:10.34918/87330

Tomasz STRZODA[1,*], Lourdes CRUZ-GARCIA[1], Mustafa NAJIM[2],
Christophe BADIE[2], Joanna POLAŃSKA[1,*]

Chapter 4. NO-MAPPING MAPPING OF OXFORD NANOPORE LONG READS

4.1. Introduction

Understanding the DNA code has been one of mankind's most important tasks for years [1]. There are many reasons for this, such as the development of new treatments and therapies [2, 3]. The problem has led to discoveries and advances in the field of science, bringing with it various sequencing techniques and the concept of mapping (i.e. aligning reads with a reference sequence, making it possible to find where a DNA read originates [4]). Historically, sequences have not been long strings. Depending on the generation and platform, they averaged between tens and hundreds of base pairs (bp). The situation has changed slightly with Third-Generation Sequencing (TGS) for which average lengths are mostly reported in thousands of bp [5]. Among these, we can distinguish companies such as Pacific Biosciences and Oxford Nanopore Technologies differing in their approach to reading the sequence of nucleotides [6]. The latter company bases its innovative idea on nanopores, i.e. proteins placed in a membrane through which a single DNA/RNA strand passes. Based on the changes in current arising from passing a string of nucleotides, we are able to read individual letters of the sequence. Such a process is characterized by the absence of restrictions on the read length, resulting in so-called “long reads” [7].

TGS opens up new possibilities and brings with it new problems to solve. One of them is the search for a specific sequence fragment in a relatively short time.

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland.
[2] Cancer Mechanisms and Biomarkers Group, Radiation Effects Department, Radiation, Chemical & Environmental Hazards, UK Health Security Agency, Didcot, UK.
* Corresponding authors: tomasz.strzoda@polsl.pl, joanna.polanska@polsl.pl.

Such a fragment could be a specific gene that acts as a marker for a screening test that determines the dose of ionizing radiation absorbed by a person. Finding the marker itself is not an easy task and is a separate problem. Based on the literature [8, 9], it can be concluded that such a marker potentially is the FDXR gene, which was used for the present study.

Despite the ever-growing list of sequence alignment software, it still ties the task of searching for a nucleotide fragment to using a computer to search all possible read locations on a given reference, which translates into computation time. In this paper, we wanted to replace the alignment program with a machine learning model, thus narrowing down the task to searching for a specific sequence/gene. We base our idea on methods known from Natural Language Processing (NLP) [10]. Using appropriate methods, we encoded the entire sequence in a first step and then focused on the problem of narrowing our encoding to the *n* initial sequences and observed how this affected the efficiency of our model. In this way, we showed that it makes sense to apply NLP methods to bioinformatics tasks and how far we can move with the encoding to obtain satisfactory efficiency.

4.2. Materials and Methods

4.2.1. Data

The data consisted of Oxford Nanopore sequencing data from three repetitions of the biological experiment: A, B, C. The HT1080 cell line was used. In each repetition, the cells were exposed to a radiation dose of 10 Gy. RNA was then extracted and libraries were prepared using a direct RNA kit. Finally, six samples of epithelial cells were available, where three of them were not irradiated (Control-A, Control-B, Control-C) and another three were samples 24 hours after exposure to a 10 Gy dose (24h-A, 24h-B, 24h-C). After sequencing, 8.5 million long-read RNA sequences were obtained.

4.2.2. Data visualization

Knowing the characteristics of the data is fundamental in making further decisions. For this purpose, visualization using various plots is used, which helps humans to understand the data, detect possible errors and anomalies. In this case, to check the distribution of read lengths of each sample, it was decided to generate a violin plot [11]. This made it possible to compare the density of the distribution and generally interpret the data.

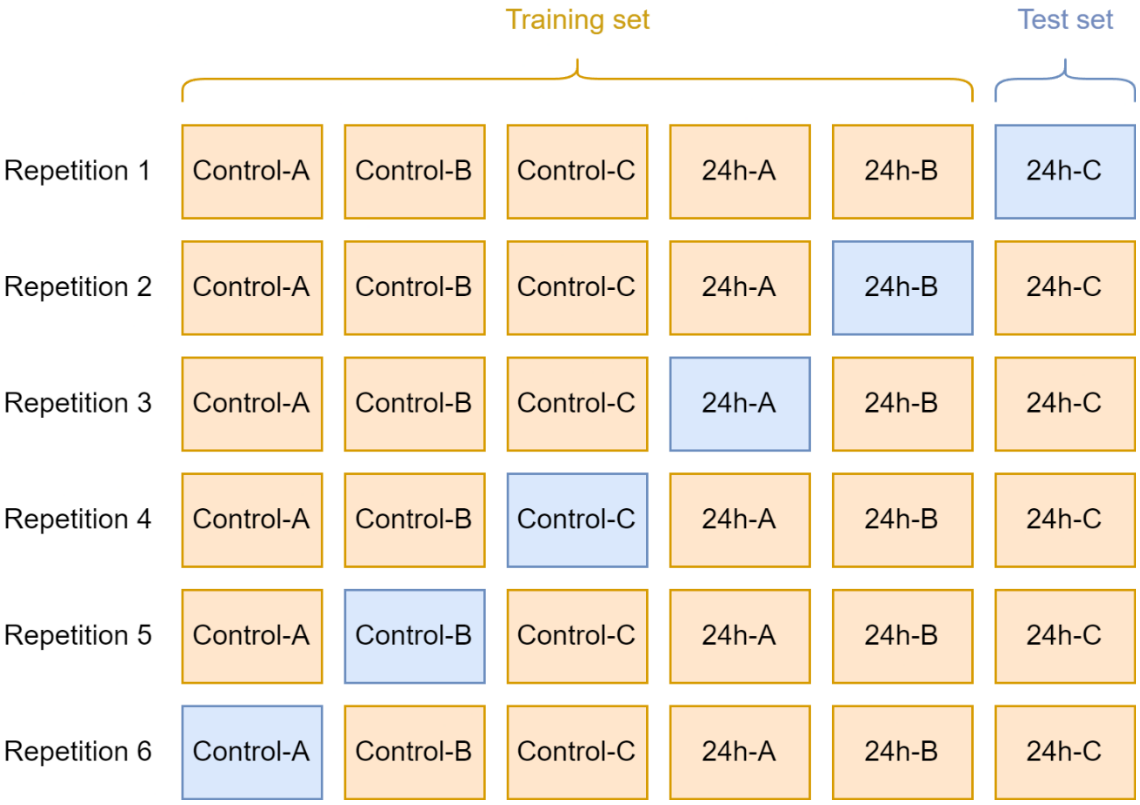
4.2.3. Data preparation

In order to prepare the data for further use, it was decided to perform filtering to remove short reads. It was assumed that such reads could be the result of various mistakes and would not be useful for the present study. For this task, it was decided to use the Empirical Cumulative Distribution Function (ECDF) plot that was generated for each sample. By analysing the data distributions shown therein, a threshold value was chosen to ensure that the beginning of the plot line on its right-hand side was characterised by a steep slope to the horizontal axis, while at the same time not excluding too many reads from further analysis.

The final step was to divide the reads into those that could potentially belong to a given genome fragment (the FXDR gene) and those that do not. For this purpose, the minimap2 [12, 13] was used, which is responsible for matching the sequence data with the given reference sequence. This allows us to find the likely location of each read. In this case, the FDXR gene was chosen as the reference sequence.

4.2.4. Experimental design

The research was based on a machine learning model, which was responsible for determining whether or not a given sequence could potentially originate from the FDXR gene. A neural network consisting of three dense layers. The first two contained fifty neurons and ReLU activation functions, while the last contained one neuron and a sigmoid function. The training process was based on ten epochs and used Leave-One-Out Cross-Validation (LOOCV) to assess the performance of the predictive models [14]. Each time, one of the six samples was set aside as a test set, while the remaining five samples were used for training. Such a process was repeated six times, thus ensuring that each sample was used exactly once as a test set (Fig. 1).



Repetition 1	Control-A	Control-B	Control-C	24h-A	24h-B	24h-C
Repetition 2	Control-A	Control-B	Control-C	24h-A	24h-B	24h-C
Repetition 3	Control-A	Control-B	Control-C	24h-A	24h-B	24h-C
Repetition 4	Control-A	Control-B	Control-C	24h-A	24h-B	24h-C
Repetition 5	Control-A	Control-B	Control-C	24h-A	24h-B	24h-C
Repetition 6	Control-A	Control-B	Control-C	24h-A	24h-B	24h-C

Fig. 1. Leave-One-Out Cross-Validation for performed experiments
Rys. 1. Walidacja krzyżowa Leave-One-Out dla przeprowadzonych eksperymentów

In order to overcome the problem of converting long sequences of nucleotides into a computer-friendly form, techniques known from natural language processing (NLP) were used. Firstly, the long sequences were presented as six-letter words (k-mers) with an offset equal to one nucleotide. This means that a single word consists of two codons. Sequences of four words (four-grams) were then created and such subsets formed the final dictionary with which each long-read was encoded. The method used is called "bag-of-words" [15] and is widely applied in the field of NLP.

The present study focused on encoding n initial nucleotides and observing the impact on the final performance of machine learning models. In order to select N (the set of considered n), quartiles were calculated as one measure of the position of the observations (the string lengths). Then a certain Δ step was selected to increase the size of this set. In addition, the other end of the sequence was considered, for which the same experiment was performed. In this case, the order of nucleotides read was from right to left.

Accuracy was used as a measure of model evaluation. All long-reads belonging to the test set were encoded identically as during the model training step (same n and same end of the sequence). It is noteworthy that even for the encoding of the whole reads, the strings for the start and end of the sequence are not identical. The difference is in the order of the nucleotides: from left end to right end, or from right end to left end.

4.3. Results and discussion

4.3.1. Data visualization

Fig. 2 shows the distribution of read lengths of each sample. As can be seen, all distributions are bimodal and right-skewed. The long tail indicates a relatively small number of reads greater than 3,000 bp. The exact number of reads per sample is shown in Table 1.

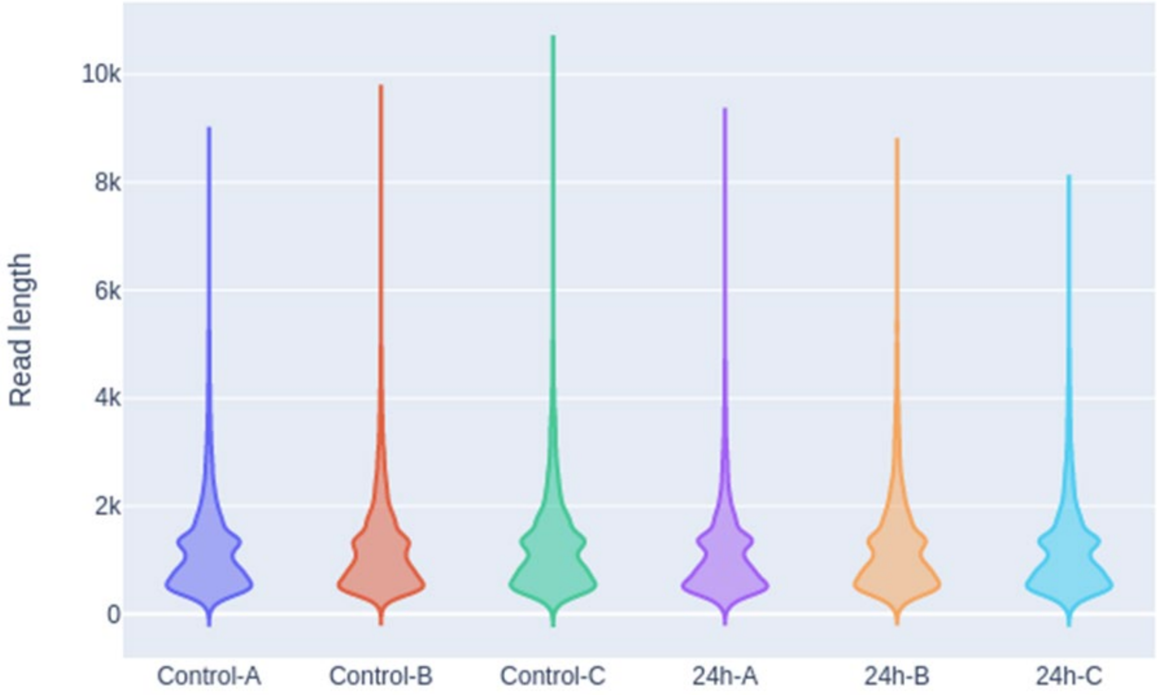


Fig. 2. Distribution of read lengths of individual samples
Rys. 2. Rozkład długości odczytów poszczególnych próbek

4.3.2. Data preparation

Plots of the empirical distribution function for the control and irradiated sample can be seen in Fig. 3. Based on these, a threshold value of 500bp was chosen. That said, all reads with a length greater than this value were not rejected. In the end, nearly 7.2 million long-reads remained. The exact numbers of long-reads per sample before and after filtering are shown in Table 1.

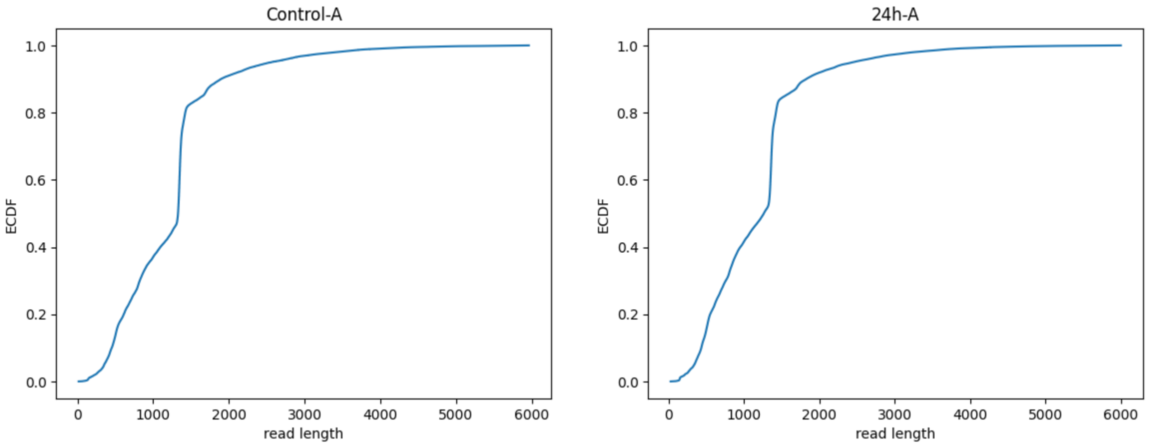


Fig. 3. Empirical cumulative distribution function plots for two samples
Rys. 3. Wykresy dystrybuanty empirycznej dla dwóch próbek

Table 1

Number of long-reads per sample

Sample	Not filtered	Filtered			
	Number of reads	Number of reads	*Q1*	Median	*Q3*
Control-A	1,232,364	1,040,293	784	1231	1692
Control-B	1,665,661	1,410,171	780	1204	1702
Control-C	1,254,760	1,069,828	802	1263	1720
24h-A	1,670,162	1,356,675	745	1146	1556
24h-B	1,859,656	1,570,158	786	1229	1721
24h-C	862,354	718,141	772	1209	1648
Total	8,544,957	7,165,266			

By performing the alignment process, the long-reads were divided into two groups. According to the aligner, the majority (7,128,352) were given the label “unmapped”. This means that no potential location was found for them on the reference. The remaining reads (36,914) according to the aligner could potentially originate from the FDXR gene. Such a large difference in group sizes is as expected, because this task focused on only one among the entire set of genes.

4.3.3. Experimental design

The LOOCV approach provides a good method for validation, but involves the training of several artificial intelligence models [14]. Considering also the number of total reads (nearly 7.2 mln), which on average contain about 1,400 nucleotides and which have to be transformed and encoded, such a process is time-consuming and complex in a memory context. Therefore, it was decided to randomly select 4,000 sequences potentially derived from the FDXR gene and 12,000 sequences unmapped to this gene during the training phase (hereinafter referred to as "approach 1"). The whole study was then repeated, but this time increasing the number of reads to 10,000 and 30,000 respectively (hereinafter referred to as "approach 2").

As mentioned earlier, the *N* set containing the numbers of *n* first encoded nucleotides was selected based on the calculated quartiles (summarised in Table 1) and the selection of a certain Δ step. That said, $N \in \{800; 1700\}$ were chosen, corresponding roughly to *Q1* and *Q3* of each sample. Next, the step $\Delta = 300$ was adjusted, thus obtaining the final $N \in \{200; 500; 800; 1100; 1400; 1700; \text{whole sequence}\}$.

Table 2

Model evaluation (approach 1) – accuracy with 95% confidence interval

Number of nucleotides (n)	Start of the sequence [%]	End of the sequence [%]
Whole sequence	98.31 (98.07; 98.55)	98.31 (98.13; 98.48)
1,700	96.15 (94.99; 97.31)	97.13 (96.33; 97.92)
1,400	95.59 (94.13; 97.05)	96.84 (95.85; 97.83)
1,100	94.86 (93.44; 96.28)	96.19 (94.92; 97.45)
800	93.99 (92.33; 95.65)	94.99 (93.43; 96.54)
500	92.54 (90.45; 94.63)	92.69 (90.66; 94.72)
200	90.17 (87.24; 93.09)	87.08 (83.35; 90.81)

Table 3

Model evaluation (approach 2) – accuracy with 95% confidence interval

Number of nucleotides (n)	Start of the sequence [%]	End of the sequence [%]
Whole sequence	98.58 (98.37; 98.79)	98.62 (98.50; 98.73)
1,700	96.96 (95.90; 98.02)	97.99 (97.54; 98.44)
1,400	96.67 (95.81; 97.56)	97.68 (97.04; 98.31)
1,100	96.10 (94.98; 97.22)	97.22 (96.45; 97.99)
800	95.36 (94.03; 96.69)	96.33 (95.26; 97.40)
500	94.12 (92.42; 95.81)	94.64 (93.03; 96.24)
200	91.90 (89.57; 94.22)	89.34 (85.95; 92.72)

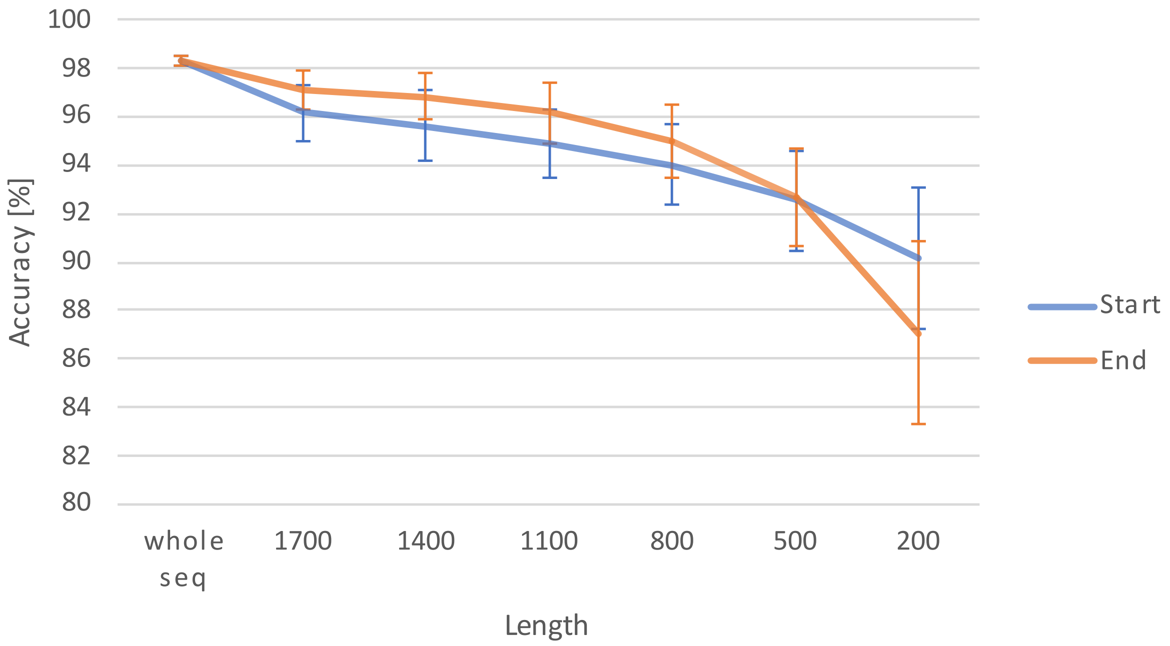


Fig. 4. Visual comparison of models' accuracies with 95% confidence intervals (approach 1)
Rys. 4. Wizualne porównanie dokładności modeli z 95% przedziałem ufności (podejście 1)

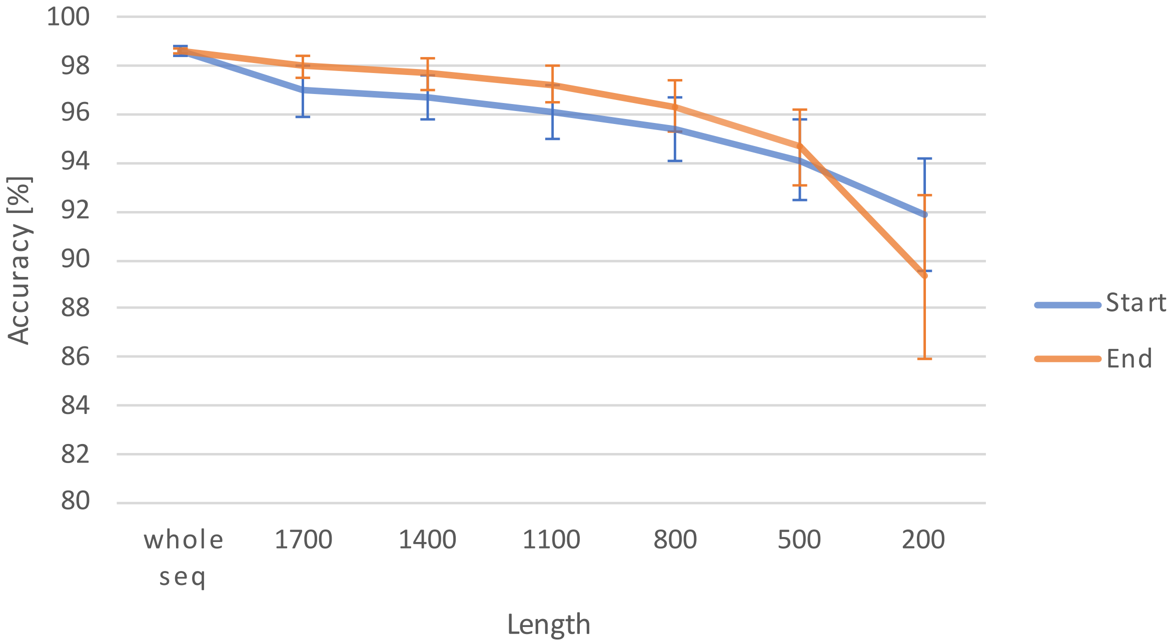


Fig. 5. Visual comparison of models' accuracies with 95% confidence intervals (approach 2)
Rys. 5. Wizualne porównanie dokładności modeli z 95% przedziałem ufności (podejście 2)

Model accuracies presented as confidence intervals with $\alpha = 95\%$ are shown in Table 2 and Table 3 for approach 1 and 2, respectively. In addition, Fig. 4 and Fig. 5 show a decrease in accuracy with a decrease in the number of *n* initial encoded nucleotides.

Referring to Table 3, the approach used to generate words and encode whole sequences gives a high accuracy of 98.62% thus showing the potential of applying methods known to NLP to bioinformatics. The value decreases with decreasing *n* of the initial nucleotides encoded and reaches a minimum value of 91.90%/89.34% for *n* = 200. In general, the second end of the sequence is more efficient, as can be seen in Fig. 5. Only in one case does the first end have a higher accuracy.

Table 2 shows that despite the training set being more than three times smaller, the accuracy of the models is marginally worse than the models trained on “approach 2”. The nature of the changes remains similar, which is brilliantly shown in Fig. 4. The reduction of the training set results in faster training of the classifier, leading to a smaller dictionary size needed to encode the sequence.

1.4. Conclusions

In summary, our research shows that methods known from NLP can be successfully implemented into the analysis of Oxford Nanopore long-reads and can serve as the alternative to classical mapping approaches. Encoding a sequence in this way, training a classifier and then using it for prediction achieves an accuracy of 98.62%. The research confirmed the conjecture that encoding only the initial *n* nucleotides is associated with a decrease in classifier performance. For the initial 200 nucleotides, this is a decrease of up to 10% compared to encoding the whole sequence. As *n* decreases, the standard error increases. Such results can be satisfying when computation time is crucial.

In addition, when comparing the results of both ends of the sequence with each other, higher accuracy was obtained for the second end in almost every case. Furthermore, when comparing the results of the models trained on the two different datasets, the accuracy remains roughly similar despite more than three times fewer long-reads in approach 1.

Acknowledgements

Partially financially supported by IDEA project (CRCE/CAP/20-21/406).

Bibliography

1. Collins F.S., Morgan M., Patrinos A. “The Human Genome Project: lessons from large-scale biology”. Science 300.5617 (2003): 286–290.
2. Boldogkői Z., et al. “Long-read sequencing–a powerful tool in viral transcriptome research”. Trends in microbiology 27.7 (2019): 578–592.
3. Chen Zhiao, Xianghuo He. “Application of third-generation sequencing in cancer research”. Medical Review 1.2 (2021): 150–171.
4. National Research Council. “Mapping and sequencing the human genome” (1988).
5. Kchouk Mehdi, Jean-Francois Gibrat, Mourad Elloumi. “Generations of sequencing technologies: from first to next generation”. Biology and Medicine 9.3 (2017).
6. Weirather Jason L., et al. “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”. F1000Research 6 (2017).
7. Branton Daniel, David W. Deamer. Nanopore sequencing: an introduction. World Scientific, 2019.
8. Cruz-Garcia Lourdes, et al. “In vivo validation of alternative FDXR transcripts in human blood in response to ionizing radiation”. International Journal of Molecular Sciences 21.21 (2020): 7851.
9. O’Brien Gráinne, et al. “FDXR is a biomarker of radiation exposure in vivo”. Scientific reports 8.1 (2018): 684.
10. Ofer Dan, Nadav Brandes, Michal Linial. “The language of proteins: NLP, machine learning & protein sequences”. Computational and Structural Biotechnology Journal 19 (2021): 1750–1758.
11. De Coster Wouter, et al. “NanoPack: visualizing and processing long-read sequencing data”. Bioinformatics 34.15 (2018): 2666–2669.
12. Li Heng. “Minimap2: pairwise alignment for nucleotide sequences”. Bioinformatics 34.18 (2018): 3094–3100.
13. Li Heng. “New strategies to improve minimap2 alignment accuracy”. Bioinformatics 37.23 (2021): 4572–4574.
14. Wong Tzu-Tsung. “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation”. Pattern Recognition 48.9 (2015): 2839–2846.
15. Qader Wisam A., Musa M. Ameen, Bilal I. Ahmed. “An overview of bag of words; importance, implementation, applications, and challenges”. 2019 international engineering conference (IEC). IEEE, 2019.

NO-MAPPING MAPPING OF OXFORD NANOPORE LONG READS

Abstract

Sequence mapping is a time-consuming and crucial step when processing DNA-/RNA-seq data. The purpose of the study is to evaluate the effectiveness of machine learning models in replacing traditional targeted mapping for long-read sequencing data. Oxford Nanopore sequences consisted of three pairs of epithelial cell samples, where one of each pair was irradiated with 10 Gy. A total of 8.5 million long-read RNA sequences were available. Our approach proposed an alternative to classical bioinformatics solutions by using techniques known from natural language processing (NLP). At the beginning, long strings of nucleotides were expressed as base word sequences. During this research, a word length of six nucleotides and an offset of one nucleotide were used. Several sequence encoding scenarios were considered: whole sequence, first 1700 only, first 1400 only and so on until only first 200 nucleotides. In addition, the second end of the string is included. As a sequence classifier, a neural network with a dense 3-layer architecture with 50, 50 and 1 neurons, respectively, was chosen. A Leave-One-Sample-Out-Cross-Validation scheme was applied. The best results with an average accuracy of 98.62% were obtained for whole sequences, as expected. Reducing the sequence representation to 200 nucleotides resulted in a decrease to 89.34% for the same end of the sequence. These results confirmed the potential of using NLP methods in bioinformatics.

Keywords: sequencing, long reads, classification, natural language processing, oxford nanopore technology