Nanopore Sequencing Simulator for DNA Data Storage

1st Eva Gil San Antonio
Laboratoire I3S, UMR 7271
Université Côte d'Azur, CNRS
Sophia Antipolis, France

2nd Thomas Heinis
Department of Computing
Imperial College London
London, United Kingdom

3rd Louis Carteron
Department of Computing
Imperial College London
London, United Kingdom

4th Melpomeni Dimopoulou
Laboratoire I3S, UMR 7271
Université Côte d'Azur, CNRS
Sophia Antipolis, France

5th Marc Antonini
Laboratoire I3S, UMR 7271
Université Côte d'Azur, CNRS
Sophia Antipolis, France

***Abstract*—The exponential increase of digital data and the limited capacity of current storage devices have made clear the need for exploring new storage solutions. Thanks to its biological properties, DNA has proven to be a potential candidate for this task, allowing the storage of information at a high density for hundreds or even thousands of years. With the release of nanopore sequencing technologies, DNA data storage is one step closer to become a reality. Many works have proposed solutions for the simulation of this sequencing step, aiming to ease the development of algorithms addressing nanopore-sequenced reads. However, these simulators target the sequencing of complete genomes, whose characteristics differ from the ones of synthetic DNA. This work presents a nanopore sequencing simulator targeting synthetic DNA on the context of DNA data storage.**

***Index Terms*—DNA data storage, nanopore sequencing, sequencing simulator, image coding.**

I. Introduction

Storing digital data is becoming a challenge for mankind due to the relatively short lifespan of devices. At the same time, the amount of digital data on the planet is expected to reach more than 175 billion Terabytes (TB) in 2025. Most of this data is seldom accessed and is referred to as cold (e.g old photos stored by users on Facebook). Unfortunately, all storage media currently used for archiving cold data (hard drives or magnetic tapes) suffer from two basic problems. Firstly, the storage density improvement rate is 20% per year at best, which is significantly behind the 60% cold data growth. Secondly, current storage media have a limited lifespan of five years (hard drive) to twenty years (magnetic tape). Since data is often stored for a significantly long period (50 years or more), it must be periodically migrated to new devices, leading to a huge waste of equipment and energy and thereby increasing the cost of storage. As a consequence, the search for new efficient ways to store digital information able to keep up with the current needs has become of great interest as it is the case of DNA, the carrier of heredity in living organisms. Theoretically, it is possible to store 455 million TB, i.e. the amount of data that can be stored in 45.5 million 10TB capacity hard drives, in 1 gram of synthetic DNA while still allowing for longevity of several centuries even in harsh storage environments. DNA is a complex molecule corresponding to a succession of four types of nucleotides (nts), adenine (A), thymine (T), guanine (G) and cytosine (C). It is this quaternary genetic code that inspired the idea of storing digital data in DNA which suggests that any digital information can be encoded in a quaternary alphabet sequence A, T, C, G. At the same time, the important advances made in the field of biology today allow the construction of DNA sequences in vitro thanks to molecular synthesis, as well as the reading of any DNA strand (oligo) carried out by special machines called sequencers. The storage of data on DNA is an emerging field of research which is very promising but also very complicated to implement. Indeed, it is subjected to various constraints imposed by the biochemical procedures of synthesis (writing) and sequencing (reading). On the one hand, the process of DNA synthesis is relatively expensive (several dollars per synthesized oligo containing of the order of 200 to 300 nts) and on the other hand, DNA sequencing is susceptible to errors and can introduce insertions, deletions or substitutions of nucleotides into the decoded DNA strands. Reducing the high cost of synthesis requires efficient compression of the data before encoding. In addition, the code generated must be robust to the errors introduced by the sequencing in order to allow its decoding, in particular by taking into account constraints linked to biology.

Since the release of nanopore sequencers, this technology has become more and more popular thanks to its affordability, small size and speed, which make it suitable for real-time applications. More precisely, nanopore-based sequencers measure the changes in the electrical conductivity as DNA strands pass through the pore. This electrical signal is then translated into a sequence of nucleotides in a process known as basecalling. Despite of all the advantages that sequencers such as the MinION [1] offer, it has a major drawback as

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863320.

it remains an error-prone process. This constitutes the main challenge when using this device in the context of DNA data storage, compromising the decodability of the data.

In the past years, several works have introduced sequencing simulators aiming to ease the implementation of new algorithms targeting nanopore-sequenced data. Such simulators allow testing while developing new tools thanks to their speed, low cost and high throughput. Commonly, simulators generate noisy reads using a model error profile extracted from experimental data. The introduction of the errors can be done directly by modifying the bases of the DNA sequences [2]–[4] or by simulating the electrical signals and allowing the basecaller to introduce the errors while translating it into a sequence of nucleotides which provides a more realistic scenario [5], [6]. The main challenge when using these simulators for DNA data storage applications lies in the fact that their error models are generated from the sequencing of complete genomes and thus, longer reads than in the case of synthetic DNA, whose length is limited to 300 nts. This work constitutes a first proof of concept of a new nanopore sequencing simulator addressing synthetic DNA in the context of DNA data storage. To our knowledge, this is the first nanopore simulator specifically designed to model the full DNA storage channel, including synthesis, storage, PCR amplification and sequencing.

In section II we introduce the simulator and its capabilities. Section III describes how the error rate of nanopore sequencing has been estimated for the specific case of synthetic DNA and how it has been used to parameterize the simulator. Simulation results illustrating the performance of the simulator in comparison to the ground truth are reported in Section IV.

II. Simulator

Our simulator models errors in all phases of DNA storage, i.e., synthesis, storage, PCR (for amplification) and sequencing. It takes as input the sequences encoding the information and returns the sequences incorporating simulated errors. We discuss the errors and configuration of each phase in the following. Error probabilities to configure the simulator can either be determined experimentally or be taken from related work [7].

A. Synthesis

DNA synthesis is a linear process, i.e., one nucleotide is added after another (in the 5'-3' direction). Errors occur during the physical assembly of the nucleotides. As such, errors are evenly or uniformly distributed across the synthesised sequence, meaning that an error is as likely to occur on the first nucleotide as it is in the middle of the sequence. Our simulation consequently passes over the whole sequence and at each nucleotide decides whether an error should occur based on the error probability.

Different errors such as deletions (absence of a nucleotide or accidental, early termination), insertion (additional nucleotide) or substitution (a different nucleotide than intended is added) can occur. The simulator is configurable in terms of likelihood of an error occurring and, if an error occurs, the likelihood of the type of error (and in case of a substitution, the likelihood of each type, e.g., A substituted by G).

Although insertions and substitutions are uniformly distributed across synthesised sequences, deletions are tail favoured. The simulator hence compensates for that by simulating an error with the same likelihood across nucleotides of the sequence, but if an error occurs, the chances for a deletion are higher at either end of the sequence.

B. Storage

Errors can also be introduced during storage. Increased humidity or temperature can drastically shorten the lifespan of DNA. Experiments with protocols for accelerated ageing by way of increased temperature (to speed up decay and thus simulate a storage time of multiple half-lives) have been carried out to understand the sources of errors.

The decay of DNA is modelled like standard radioactive decay. However, instead of removing nucleotides, the bonds between nucleotides are simply broken resulting in broken sequences. The broken sequences are no longer readable as the forward and reverse primer are no longer located on the same strand.

Our simulator uses a derivation of the Arrhenius equation with its values from studies on dated fossils [8] to model a breakage/decay event on a sequence. The only parameter needed to be configured is the storage duration in years.

Simulating errors in storage starts with the sequences resulting from the synthesis simulation. The process is iterative, meaning that in the event a sequence is fragmented due to decay, both (or all) fragments are added back to the pool of sequences, meaning that they can be broken again.

C. PCR

Polymerase Chain Reaction (PCR) is a method widely used to amplify sequences, i.e., to rapidly make millions to billions of copies of the sequences before sequencing. PCR is typically done in multiple cycles and at every cycle, the number of sequences is doubled, i.e., two identical sequences are produced for every sequence.

Standard experimental protocols suggest to run 40 PCR cycles. Doing so in a simulation quickly renders the simulation computationally intractable as it produces too many . It is, therefore, necessary to be able to reduce the size of the PCR output to a constant number. By taking a uniform random sub-sample after each PCR cycle, the simulator keeps a constant size in addition to a general representation of the error distribution. The number of PCR cycles in the simulator is configurable.

The downside of taking a random sub-sample is that there will be a bias towards the initial sequences for PCR phases with small cycles. Due to the exponential nature of PCR, the sequences generated during the first cycles will have a disproportionate representation in the sub-sample compared to the later cycles. Increasing the number of cycles will help reduce the bias and help arrive at a more randomly distributed final sub-sample.

D. *Sequencing*

The approach to modelling sequencing is very similar to modelling the synthesis phase. More specifically, the same errors as in synthesis can occur, insertions, deletions and substitutions, due to misreads can occur. Generally, across different sequencing technologies, the distribution of errors across sequences are uniform - as is the case for synthesis.

We thus implement simulation of sequencing the same as synthesis but the probabilities for substitution, deletions and insertions errors and their respective transitions (for substitutions) need to be configured.

III. Estimation of the nanopore sequencing noise

Although several works have provided studies about the error rates introduced by nanopore sequencing, most of them focus on the sequencing of long DNA strands (thousands of bases) as this technology targets the sequencing of complete genomes. Additionally, state of the art simulators expect a genome as reference, which will be subsampled following some nanopore read length distribution model and then corrupted by adding errors in form of substitutions, insertions and deletions. On the contrary, our work uses synthetic DNA, limiting the length of the oligos to 300 nts at most and making the sampling step unnecessary. As some studies state than the errors introduced by the nanopore sequencers affects dramatically both ends of the DNA strands [9], it remains unclear how the length of the input sequences affects the performance of the MinION. Therefore, to adapt the characteristics of the noise introduced by the simulator when sequencing short DNA strands, we have computed the noise rates from the nanopore-sequenced reads storing two images.

A. *Experimental data*

After being stored in sealed capsules for two years, we have sequenced the DNA strands which store 2 different images (see figures 2(a1) and 2(b1)) of size 128 by 128 pixels and 120 by 120 pixels representing a total amount of 662 and 875 oligos respectively. All the oligos had a length of 91 nts (without considering primers, which are special sequences required by the sequencer). In both cases, 11 oligos contained only headers encoding important information about the characteristics of the image and the parameters of the encoding. The rest of the oligos contained the encoded data itself. Figure 1 depicts a schema of the format of the oligos.

The synthesized pool of oligos was amplified with PCR and sequenced using an ONT MinION sequencer (SQK-LSK109 sequencing kit and MinKNOW 3.6.8 software). Base calling was performed in 4000 event batches using Guppy 3.2.10. The sequencing step led to a total amount of 3395789 raw reads. For the decoding of the stored data, sequencing adapters were removed using Porechop [10]. The trimmed reads were then filtered by length and clustered according to their identifier, headers and offset (see figure 1). Once clustered, a consensus sequence was obtained from each cluster using an algorithm based on majority voting, dividing each oligo into its different codewords and selecting as consensus the most frequent one. A detailed explanation about the decoding can be found in [11]. Figures 2(a2) and 2(b2) show the resulting decoded images from this wet-lab experiment.

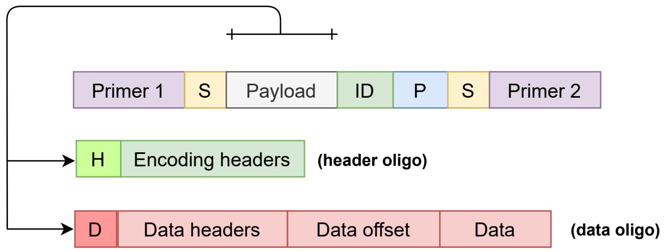


Fig. 1. Format of the oligos - All oligos contain primers that are needed for the sequencing: S denotes the sense nucleotide which determines whether a strand is reverse complemented when sequenced. P is a parity check nucleotide while the ID is an identifier of the image so to be distinguished from other data that may be stored. The payload can either contain encoding headers only which hold information about the image characteristics and the encoding parameters used (header oligo), or it can contain some data headers and an offset to denote the position and nature of the data field that follows (data oligo).

B. *Error rates*

For the estimation of the error rates, we first mapped each read to its reference using minimap2 [12] and computed the Levenshtein distance between each read and its reference considering only those reads that could be unequivocally mapped to a reference. Nanopore adapters were not considered when computing the distances. In the same way, we estimated its three different components (substitutions, insertions and deletions):

- Total error rate: 0.0686
- Substitution rate: 0.0253
- Insertion rate: 0.0179
- Deletion rate: 0.0255

In addition, we computed the error rates for the oligos encoding each image independently but no significant variation was found. Figure 3 shows the distribution of the different noise components.

C. *Parameterization of the simulator*

As described in the previous section, the synthesized DNA strands for our wet-lab experiment had a length of 138 nts and they were stored for two years in a sealed capsule. Considering that the error rate of DNA synthesis is almost negligible when the synthesized oligos do not exceed 300 nts length and that the capsule prevents its contact with water and oxygen keeping the DNA intact during the storage period, the only significant source of error is the process of sequencing. Therefore, in our simulations we only considered sequencing noise.

More precisely, the estimated rates of substitutions, insertions and deletions from the experimental data (see section III-B) were used as target rates. As these rates were computed without considering the nanopore adapters, they were not included in the input sequences of the simulator. Finally, the coverage was selected so to match the coverage from the experimental data.

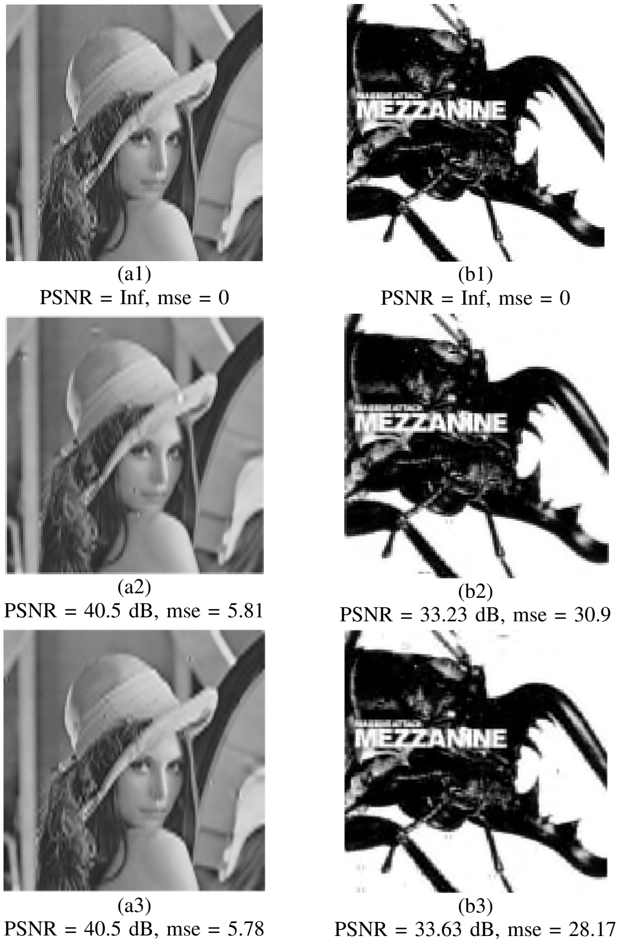
IV. Comparison of the results

We tested the performance of the simulator by running 50 realisations for the error rates provided in section III-B and averaged the results. The reference sequences used to feed the simulator were the same ones as the ones used in our wet-lab experiment (see section III-A).

In average, the simulations led to a total amount of 3396770 reads with the following error rates:

- Substitution rate: 0.0242
- Insertion rate: 0.0169
- Deletion rate: 0.0255

For the decoding of the simulated reads we followed the same process we used to decode the reads from our wet-lab experiment [11]. Figures 2(a3) and 2(b3) depict an example of the reconstructed images from one simulation run.



(a1)
PSNR = Inf, mse = 0

(b1)
PSNR = Inf, mse = 0

(a2)
PSNR = 40.5 dB, mse = 5.81

(b2)
PSNR = 33.23 dB, mse = 30.9

(a3)
PSNR = 40.5 dB, mse = 5.78

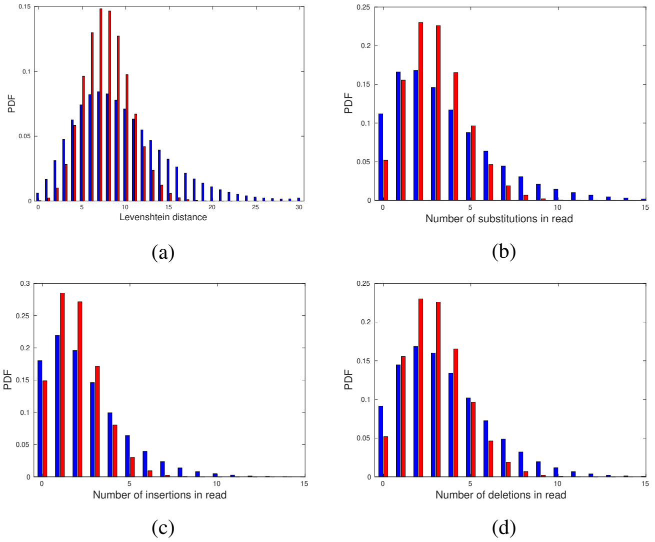
(b3)
PSNR = 33.63 dB, mse = 28.17

Fig. 2. Visual results of the decoded data - (a1) and (b1) correspond to perfectly decoded images. (a2) and (b2) correspond to MinION sequencing and our novel consensus algorithm based on Majority Voting in codewords [11]. (a3) and (b3) correspond to an example of the decoding of simulated reads.

In average, the decoded images from the simulated sequenced reads provided a Peak Signal-to-Noise Ratio (PSNR) of 40.23 dB and 32.7 dB with a Mean Squared Error (MSE) of 7.62 and 41.56 respectively. The results obtained from the simulations are comparable to the experimental ones in terms of PSNR, MSE as well as the visual quality of the decoded images for all the runs.

We have also compared the distribution of the errors introduced by the MinION and the simulator. Figure 3 depicts the Probability Density Function (PDF) for the different error types.

It is important to note that the only parametrization required for the simulations is the average error rates of insertions, deletions and substitutions. Therefore, while the mean error rate of the simulator can be controlled, the standard deviation of the error rate can vary. This fact explains the reason why the variability of the error in the experimental reads is higher than the one computed by the simulations. Nevertheless, this difference in the standard deviation does not have a significant impact on the reliability of the simulator results.



(a)

(b)

(c)

(d)

Fig. 3. Comparison of the Probability Density Function (PDF) for each kind of error between the reads form the wet-lab experiment (blue) and simulated reads (red) - (a) Edit distance, (b) substitutions, (c) insertions and (d) deletions.

V. Conclusions

This work is a very first demonstration of the potential of a new simulator that models the full DNA data storage channel. Although in this study we have only assessed the performance of the nanopore sequencing module (mainly due to the lack of experimental data), future works will focus on the evaluation of the rest of the existing modules in this simulation tool. Nevertheless, these first results are highly promising as they are comparable to the experimental ones in terms of PSNR, MSE and the visual quality of the decoded images, proving the capability of the simulator to reproduce the errors introduced during nanopore sequencing in short synthetic DNA strands.

VI. Acknowledgements

The synthesis of the oligos was done by Twist Bioscience. The capsules for long-term storage of the synthetic DNA were provided by Imagene company, located in Evry, France. Sequencing was carried out by IPMC/CNRS (Institut de Pharmacologie Moléculaire et Cellulaire) in Sophia Antipolis, France.

REFERENCES

[1] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The oxford nanopore minion: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, no. 1, p. 239, 2016.

[2] E. A. G. Baker, S. Goodwin, W. R. McCombie, and O. M. Ramos, "Silico: a simulator of long read sequencing in pacbio and oxford nanopore," *BioRxiv*, p. 076901, 2016.

[3] P. C. Faucon, P. Balachandran, and S. Crook, "Snaresim: synthetic nanopore read simulator," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2017, pp. 338–344.

[4] C. Yang, J. Chu, R. L. Warren, and I. Birol, "Nanosim: nanopore sequence read simulator based on statistical characterization," *GigaScience*, vol. 6, no. 4, p. gix010, 2017.

[5] Y. Li, S. Wang, C. Bi, Z. Qiu, M. Li, and X. Gao, "Deepsimulator1. 5: a more powerful, quicker and lighter simulator for nanopore sequencing," *Bioinformatics*, vol. 36, no. 8, pp. 2578–2580, 2020.

[6] C. Rohrandt, N. Kraft, P. Gießelmann, B. Brändl, B. M. Schuldt, U. Jetzek, and F.-J. Müller, "Nanopore simulation–a raw data simulator for nanopore sequencing," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1–8.

[7] R. Heckel, G. Mikutis, and R. N. Grass, "A Characterization of the DNA Data Storage Channel," *Scientific Reports*, vol. 9, no. 1, Jul 2019.

[8] M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, M. T. P. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce, "The half-life of dna in bone: measuring decay kinetics in 158 dated fossils," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1748, pp. 4724–4733, 2012. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2012.1745

[9] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, "Improved data analysis for the minion nanopore sequencer," *Nature methods*, vol. 12, no. 4, pp. 351–356, 2015.

[10] R. W. Porechop, "https://github.com/rrwick/Porechop https://github.com/rrwick/Porechop," 2018.

[11] E. Gil, M. Dimopoulou, M. Antonini, P. Barbry, and R. Appuswamy, "Decoding of nanopore-sequenced synthetic DNA storing digital images," in *2021 IEEE International Conference on Image Processing*, Anchorage, United States, Sep. 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03254404

[12] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.