Data storage in synthetic DNA: A brief analysis of two simulators

Eduardo Ferreira[1,2]
uc2021218018@student.uc.pt

Luis A. da Silva Cruz[1,2]
lcruz@co.it.pt

[1]Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, PT
[2]Instituto de Telecomunicações, Coimbra, PT

Abstract

Data storage in synthetic DNA is being considered as a solution to the problem of preserving the ever increasing digital information being created by humans. Unfortunately the biochemical processes involved in this type of storage result in chemical degradations and errors that need to be characterized to devise measures able to minimize the loss of information. Since the cost of synthesising, storing and sequencing DNA are high, research in DNA storage has to use simulators that can emulate those processes, simulating the errors and degradations that occur in real-life situations. In this paper we describe some experiments with two DNA storage simulators, showing how they are used, indicating their best use scenarios, and reporting their performance in terms of simulation accuracy.

1 Introduction

The wide use of data intensive services like social networks, archival of video and image generated by different sectors calls for more efficient storage media. Synthetic DNA molecules can be used to store large amounts of information in small physical volumes and so it is seen as a possible solution for cold archives. Since it's still expensive to conduct wet-lab DNA storage experiments there is a need for the use of simulators to evaluate the errors introduced during the storage process. The development of the simulators and the consequent improvement of their precision in modelling synthesis, storage and sequencing errors can lead to the development of error correction techniques enabling reliable DNA data storage. There are multiple factors causing DNA storage errors that have been discovered experimentally [1] [2]. As an example the use of sequences longer than 300 nucleotides increase exponentially the error rate, and so the sequence length should be about 200 nucleotides. Another factor is the amount of GC sub-sequences in the full sequence as GC content in excess of 50% increases the error rates. At the storage phase, changes in temperature and pH or exposure to moisture and UV light might cause DNA degradation with possible destruction or change in the stored information. Useful simulators must emulate faithfully these effects. Some of the technologies considered in this paper were Illumina [11] and Oxford Nanopore Technology (ONT) [13] for sequencing, ErrASE [5] for synthesis, Pwo [6] for PCR and storage in living organisms [12]. In this paper we report our experiments using NanoSim, a sequencing simulator and MESA a more versatile simulator that can simulate the entire process of DNA storage and readout as shown in Fig. 1. To verify the accuracy of the error simulations we used several reference datasets obtained when sequencing the human genome using the Nanopore technology [9] [4]. Due to time and space constraints we report results for only one reference dataset.

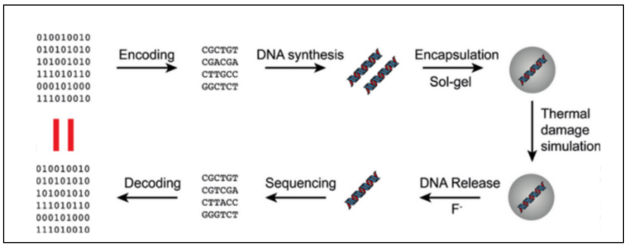


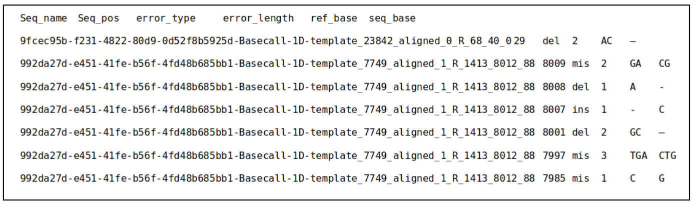
Figure 1: End-to-end DNA storage process (Figure 2 [2])

2 Simulators

The process of storing information in DNA requires the use of multiple techniques, each of them having multiple technologies with different characteristics. Since there are multiple technologies, simulators that are able to represent each of them are important assets that allow for a better understanding of the biochemical degradation and errors wich are introduced in DNA. This paper focuses on two of those simulators, with them being NanoSim, a simulator that emulates sequencing using ONT, and MESA, a simulator capable of emulating multiple processes such as sequencing, syntesis, storage and PCR and multiple technologies in each of them.

2.1 NanoSim

Nanosim [14] was designed to simulate the error introduced when using ONT sequencers. The simulator emulates the errors introduced by basecalling using the models that are trained through the input data. NanoSim works in two stages, in the first one, the characterization, the simulator analyses the reference data introduced and creates error profiles that are later used in the second stage, the simulation, to introduce the errors. The first stage allows the simulation to be conducted following patterns observed when using ONT sequencing making it so the error introduction is accurate. The simulator can simulate three separate modes, the genome, transcriptome and metagenome modes, with each of them having different data input and output. In this paper the only mode that was evaluated was the genome mode as the datasets for the sequencing of the human genome using the ONT were available [9] [4]. The simulator was designed so that it can be expanded to accommodate the evolution of the Nanopore technology. The output of the genome mode is a FASTA [10] file that indicates where in the sequence the errors were introduced, the error type, error length, the base sequence and the altered sequence. An example of the output is shown in Fig. 2. This information allows for an easy error analysis.



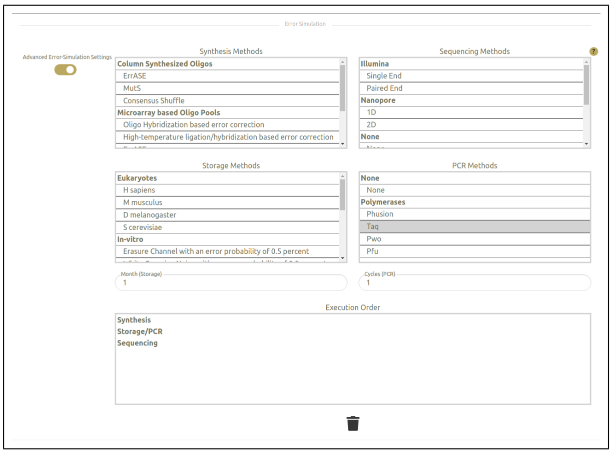
```
Seq_name  Seq_pos   error_type    error_length   ref_base  seq_base
9fcec95b-f231-4822-80d9-0d52f8b5925d-Basecall-1D-template_23842_aligned_0_R_68_40_0 29   del  2    AC   –
992da27d-e451-41fe-b56f-4fd48b685bb1-Basecall-1D-template_7749_aligned_1_R_1413_8012_88 8009 mis  2    GA   CG
992da27d-e451-41fe-b56f-4fd48b685bb1-Basecall-1D-template_7749_aligned_1_R_1413_8012_88 8008 del  1    A    -
992da27d-e451-41fe-b56f-4fd48b685bb1-Basecall-1D-template_7749_aligned_1_R_1413_8012_88 8007 ins  1    -    C
992da27d-e451-41fe-b56f-4fd48b685bb1-Basecall-1D-template_7749_aligned_1_R_1413_8012_88 8001 del  2    GC   –
992da27d-e451-41fe-b56f-4fd48b685bb1-Basecall-1D-template_7749_aligned_1_R_1413_8012_88 7997 mis  3    TGA  CTG
992da27d-e451-41fe-b56f-4fd48b685bb1-Basecall-1D-template_7749_aligned_1_R_1413_8012_88 7985 mis  1    C    G
```

Figure 2: NanoSim output example.

2.2 MESA

The MESA DNA simulator [7] [8] is a highly configurable DNA simulator that can emulate a varied selection of different technologies and processes necessary for information storage in synthetic DNA. The options include but are not limited to selecting the specific processes that are going to be simulated with the possibility of only simulating a few of them or even only one. It is also possible to select the specific technology to be used in the process and in the case of storage and PCR the amount of time and PCR cycles as shown in Fig. 3. There is also the possibility to personalize the error rates of the different methods and add conditions that change the error rates like adding sequences that are undesirable or changing the parameters for GC content as shown in Fig. 4.

All of these features allow for a personalized error simulation, suited for the needs of the user, within the limitations imposed in the nucleotides max number, which is set do 4000. Even with the constraints that it presents, this simulator allows for the evaluation of DNA storage as usually you wouldn't want long sequences in order to keep the error rates low. One negative aspect of the simulator is the incapability of simulating larger sequences and files due to the sequence size limit.



Synthesis Methods
Column Synthesized Oligos
ErrASE
MutS
Consensus Shuffle
Microarray based Oligo Pools
Oligo Hybridization based error correction
High-temperature ligation/hybridization based error correction

Sequencing Methods
Illumina
Single End
Paired End
Nanopore
1D
2D
None

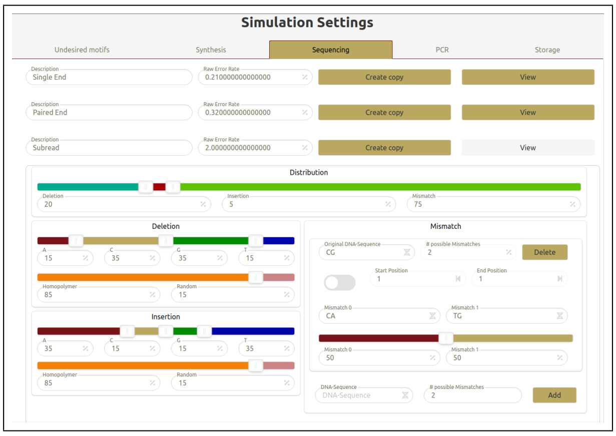
Storage Methods
Eukaryotes
H sapiens
M musculus
D melanogaster
S cerevisiae
In-vitro
Erasure Channel with an error probability of 0.5 percent

PCR Methods
None
None
Polymerases
Phusion
Taq
Pwo
Pfu

Month (Storage): 1
Cycles (PCR): 1

Execution Order
Synthesis
Storage/PCR
Sequencing

Figure 3: MESA simulator error simulation settings. [8]



Description: Single End — Raw Error Rate: 0.210000000000000
Description: Paired End — Raw Error Rate: 0.320000000000000
Description: Subread — Raw Error Rate: 2.000000000000000

Distribution: Deletion 20, Insertion 5, Mismatch 75

Deletion: A 15, C 35, G 35, T 15; Homopolymer 85, Random 15

Insertion: A 35, C 15, G 15, T 35; Homopolymer 85, Random 15

Mismatch: Original DNA-Sequence CG, # possible Mismatches 2; Start Position 1, End Position 1; Mismatch 0 CA, Mismatch 1 TG; Mismatch 0 50, Mismatch 1 50

DNA-Sequence: DNA-Sequence, # possible Mismatches 2

Figure 4: MESA simulator - sequencing type Subread settings.

3 User experience

The simulators are quite different when it comes to usability and user friendliness but are both relatively easy to use. NanoSim is very well documented and is accompanied by instructions for use that range from the installation to the analysis of the output. It can handle files of large size and different extensions (FASTA[10] and FASTQ[3]). Due to the large amount of time involved in training, of the order of one hour, the use of NanoSim is advised only in simulating long sequences where the training costs are recouped. If a pre-trained simulator is used this problem does not arise and the simulator can be used efficiently for sequences of any length. Regarding the output files, the output of the simulation stage is written in a way that makes it easy to understand what sequence it is referencing. The simulator is also highly configurable in either of the stages and has pre-trained models that are capable of being used to run simulations with a good adaptability to the data. All factors considered, this simulator is a useful and user friendly tool for the simulation of sequencing using the ONT, being better suited for larger files. MESA is used via a proprietary website making simulating small sequences (less than 4000 nucleotides) easy. There are options to select what kind of processes are being simulated, to configure error rates to suit the user needs and it can simulate all ranges of DNA storage. On the other hand when working with bigger sequences or large files this simulator does not have the capacity to handle them via the website and the documentation does not explain how to change the parameters of the simulator so it can handle larger files. The time it takes to run the sequences is reasonable, taking about two minutes to process a sequence of a 1000 nucleotides with ErrASE synthesis [5], Illumina Paired End sequencing [11], 24 months of storage using H sapiens [12] and 30 PCR cicles using Pwo [6]. All thing considered this simulator is user friendly and supports useful ranges of simulation parameters but can't handle larger sequences.

4 Conclusions

In conclusion both simulators operate well and are effective simulating errors according to the parameterization and are user friendly and efficient, even if less suited to some use-cases. Our analysis suggests that both simulators are capable of simulation within the parameter ranges supported, with MESA being more versatile in the technologies it can emulate, being a good option for DNA storage error simulation.

Acknowledgements

This work was partly supported by Instituto de Telecomunicações, FCT UID/EEA/50008/2013.

References

[1] Marc Antonini and Touradj Ebrahimi. JPEG DNA Exploration. *https://jpeg.org/jpegdna/documentation.html*, April 2023.

[2] Marc Antonini, Luis Cruz, Eduardo da Silva, Melpomeni Dimopoulou, Touradj Ebrahimi, Siegfried Foessel, Eva Gil San Antonio, Gloria Menegaz, Fernando Pereira, Xavier Pic, António Pinheiro, and Mohamad Raad. DNA-based Media Storage: State-of-the-Art, Challenges, Use Cases and Requirements version 8.0. *https://jpeg.org/jpegdna/documentation.html*, April 2022.

[3] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 12 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp1137.

[4] Miten Jain, S Koren, J Quick, AC Rand, TA Sasani, JR Tyson, AD Beggs, AT Dilthey, IT Fiddes, S Malla, H Marriott, KH Miga, T Nieto, J O'Grady, HE Olsen, BS Pedersen, A Rhie, H Richardson, AR Quinlan, TP Snutch, L Tee, B Paten, AM Phillippy, JT Simpson, NJ Loman, and M Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*, 2017. doi: 10.1101/128835. URL https://www.biorxiv.org/content/early/2017/04/20/128835.

[5] Sriram Kosuri and George M Church. Large-scale de novo dna synthesis: technologies and applications. *Nature methods*, 11(5):499–507, 2014.

[6] Peter McInerney, Paul Adams, and Masood Z Hadi. Error rate comparison during polymerase chain reaction by dna polymerase. *Molecular biology international*, 2014, 2014.

[7] Mosla. MESA - Mosla Error Simulator. *https://github.com/umr-ds/mesa$_d$na$_s$im*.

[8] Mosla. Mesa DNA simulator. *https://mesa.mosla.de/*, 2019.

[9] Nanopore Whole Genome Sequencing Consortium. Nanopore Reference Human Genome. *https://registry.opendata.aws/nanopore/*, 2016.

[10] William R Pearson. *FASTA Algorithm*. John Wiley Sons, Ltd, 2005. ISBN 9780470015902. doi: https://doi.org/10.1038/npg.els.0005255.

[11] Melanie Schirmer, Rosalinda D'Amore, Umer Z Ijaz, Neil Hall, and Christopher Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC bioinformatics*, 17:1–15, 2016.

[12] Way Sung, Matthew S Ackerman, Marcus M Dillon, Thomas G Platt, Clay Fuqua, Vaughn S Cooper, and Michael Lynch. Evolution of the insertion-deletion mutation rate across the tree of life. *G3: Genes, Genomes, Genetics*, 6(8):2583–2591, 2016.

[13] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. *F1000Research*, 6, 2017.

[14] Chen Yang, Justin Chu, René L Warren, and Inanç Birol. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience*, 6(4):gix010, 02 2017. ISSN 2047-217X. doi: 10.1093/gigascience/gix010. URL https://doi.org/10.1093/gigascience/gix010.