Efficient classification of DNA reads for robust decoding of data stored in synthetic DNA

Iulia Mitrica, Xavier Pic, Eva Gil San Antonio, Melpomeni Dimopoulou, Marc Antonini

Efficient classification of DNA reads for robust decoding of data stored in synthetic DNA

Iulia Mitrica, Xavier Pic, Eva Gil San Antonio, Melpomeni Dimopoulou, Marc Antonini
Laboratoire I3S, Université Côte d'Azur, CNRS

Contact information: Iulia Mitrica, Euclide B, 2000 Rte des Lucioles, 06900 Sophia Antipolis France
Email: mitrica@i3s.unice.fr, mitrica.iulia@gmail.com

Main objective

We propose to a simple yet efficient artificial neural network-based solution to classify DNA reads into noiseless or noisy.

Context and perspectives

- The need for new ways for storage: the exponential growth of digital data and the limited capacity of conventional storage devices
- DNA storage offers numerous advantages: an extended capacity, which could potentially allow storing the digital universe in less than 20 grams of DNA, and a high longevity of hundreds to thousands of years.

Constraints

The biological processes that are involved in DNA data storage are error-prone procedures that may degrade the quality of the decoded information.

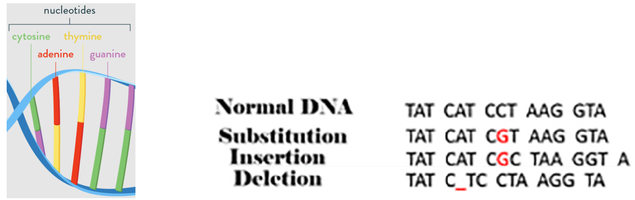

Figure 1: What is a DNA's nucleotide and how noise affects the DNA.

1 Overall scheme of DNA data storage

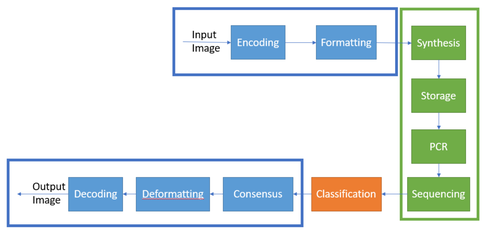

Figure 2: DNA data storage processing steps.

Chemical processes
Synthesis
Storage
PCR
Sequencing

Non-chemical processes
Encoding/Decoding
Formatting/Deformating
Consensus

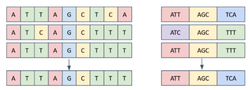

Figure 3: Examples of consensus: left: nucleotides level, right: word level [2].

2 Proposed method

2.1 Fully connected neural network

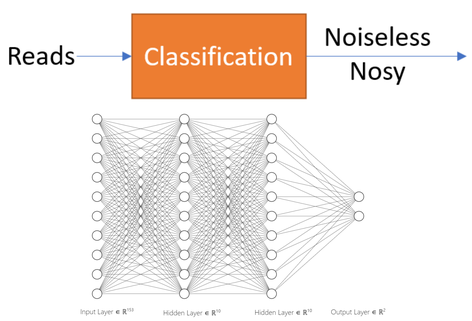

Figure 4: Top: Close look of the proposed method in the overall scheme of DNA data storage; Bottom: Fully connected network architecture, composed by an input layer of 153 neurons corresponding to the length of oligos, two hidden layers of 10 neurons each and an output layer of two neurons representing the classification classes, noiseless or noisy.

2.2 Training characteristics

The training is accomplished using cross validation with stocastic gradient descent and with sigmoid activation in hidden layers and softmax on the output layer. The learning rate is adaptive with an initial value of 0.01. The batch size is of 512 oligos. Several trainings for each experiment are performed, the number of iterations being kept into bounds.

2.3 Data

- Kodak dataset - 24 uncompressed images of 768 x 512 pixels.

3 Experimental results

3.1 Test setup and classification performance

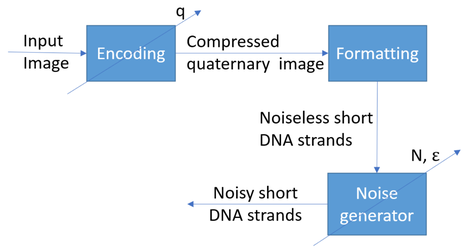

Figure 5: Method to generate synthetic oligos.

Table 1: Generation of the training and evaluation datasets for each test case. The last column express the accuracy obtained on the evaluation dataset.

	Noise generator - solutions		
Case	**For Training**	**For Evaluation**	Acccuracy[%]
I.	Imperial College London[3]	Imperial College London	99.98
II.	Imperial College London	WETLAB	81
III.	MESA[5]	MESA	98.99
IV.	MESA	WETLAB	98.04
V.	WETLAB	WETLAB	99.96

3.2 Quality performance comparisons for nanopore-R2C2 sequencing

Figure 6: Decoded WetLab on *child*, from left to right: original, reference **PSNR =** 41.07 **dB** [1, 4], proposed **PSNR =** 29.49 **dB** (case IV, consensus method in [2]), proposed **PSNR =** 40.92 **dB** (case V, consensus method in [2]), proposed **PSNR =** 47.43 **dB** (case V, trained by altering the model with a level of noise of Levenshtein distance smaller or equal with two, consensus method in [2]).

Figure 7: Decoded WetLab on *parrots*, from left to right: original, reference **PSNR =** 34.12 **dB** [1, 4], proposed **PSNR =** 29.84 **dB** (case IV, consensus method in [2]), proposed **PSNR =** 35.63 **dB** (case V, consensus method in [2]).

4 Conclusions and future works

- We propose a robust classification method that was evaluated on a considerable set of simulated and real experiments.
- Our proposed low-complexity reads filtering method is able to obtain good visual quality performances in comparison with a far much more complex solution.
- We plan to extend experiments for a less noisy sequencing method using Illumina.

References

[1] Melpomeni Dimopoulou and Marc Antonini. Image storage in DNA using Vector Quantization. In *EUSIPCO 2020*, Amsterdam, Netherlands, January 2021.

[2] Eva Gil, Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, and Raja Appuswamy. DECODING OF NANOPORE-SEQUENCED SYNTHETIC DNA STORING DIGITAL IMAGES. In *2021 IEEE International Conference on Image Processing*, Anchorage, United States, September 2021.

[3] Eva Gil San Antonio, Thomas Heinis, Louis Carteron, Marc Antonini, and Melpomeni Dimopoulou. Nanopore Sequencing Simulator for DNA Data Storage. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, International Conference on Visual Communications and Image Processing (VCIP), 2021, pp. 1-5, pages 1–5, Munich, France, December 2021. IEEE.

[4] Eugenio Marinelli and Raja Appuswamy. Onejoin: Cross-architecture, scalable edit similarity join for dna data storage using oneapi.

[5] Michael Schwarz, Marius Welzel, Tolganay Kabdullayeva, Anke Becker, Bernd Freisleben, and Dominik Heider. MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors. *Bioinformatics*, 36(11):3322–3326, 03 2020.