DECODING OF NANOPORE-SEQUENCED SYNTHETIC DNA STORING DIGITAL IMAGES

Eva Gil, Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, Raja Appuswamy

DECODING OF NANOPORE-SEQUENCED SYNTHETIC DNA STORING DIGITAL IMAGES

Eva Gil San Antonio[1], Melpomeni Dimopoulou[1], Marc Antonini[1], Pascal Barbry[2], Raja Appuswamy[3]

[1] Université Côte d'Azur, I3S, CNRS, UMR 7271, 06900, Sophia Antipolis, France
[2] Université Côte d'Azur, IPMC, CNRS, UMR 7275, 06560 Sophia Antipolis, France
[3] EURECOM, 06410, Sophia Antipolis, France

ABSTRACT

Digital media explosion has led to an exponential increase of the amount of data generated worldwide and the need for new means of storage able to keep up with the current growth of digital information has become a critical challenge. During the last decade, DNA has been proven to be a potential candidate thanks to its biological properties allowing to store information at high density (215 petabytes in 1 gram) for centuries. In previous works we have presented an end-to-end storage workflow specifically designed for the efficient storage of images onto synthetic DNA and proven its feasibility in a wet-lab experiment in which sequencing was performed using the Illumina machine. In this work we are studying the sequencing using rather the MinION sequencer on the same data after being stored in a sealed capsule for two years. MinION is a very promising sequencer although introducing a much higher error rate in the process of reading. In this paper, we propose a solution to deal with the MinION sequencing noise allowing to recover the original stored data.

Index Terms— DNA data storage, Nanopore sequencing, image coding, error detection, consensus finding

1. INTRODUCTION

The exponential growth of digital information threatens to exceed the capacity of conventional storage devices, challenging our ability to store all the data generated day after day. For the past years, big companies like Facebook started building data centers to fulfil their storage needs, which has an extremely high financial and environmental impact. 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 data storage. DNA is a complex molecule corresponding to a succession of four types of nucleotides (nts) A, C, T and G and is the support of heredity in living organisms. Its biological properties allow to store information at a high density for thousands of years (an example of that is the decoding of the DNA of a woolly mammoth that had been trapped into permafrost for 40,000 years [1]).

Very roughly, the general workflow for DNA data storage can be described as depicted in figure 1. Any kind of input data can be stored into DNA as long as it is encoded first into a quaternary representation using the 4 symbols of the DNA (A, C, T and G). This sequence is then biologically synthesized in a lab and, under the right conditions, it can be stored for long periods without any loss of information. Whenever the stored data needs to be retrieved, it can be read using some special machines called sequencers. This process is called sequencing. Finally, the initial data can be decoded by following the inverse process of the encoding. However, the biological processes of DNA synthesis and sequencing are error-prone and introduce some major constraints, adding extra complexity to the encoding and decoding steps. As it is the case of the error introduced during DNA synthesis, which is not significant when the length of the synthesized strands do not exceed 300 nts and it increases exponentially for longer DNA strands. Consequently, the encoded sequence has to be cut into shorter fragments, which are called oligos, which contain some general headers and a payload. The payload of the oligos can contain encoded data (data payload), or only headers regarding the image characteristics and encoding (header oligos) as depicted in figure 3.

Since the release of nanopore sequencing devices, some works about DNA storage have included it in their workflow, exploring ways to overcome the higher error rate that this technology introduces. [2] proposes a pipeline which integrates random access and generates the consensus sequence by combining different existent multiple sequence alignments (MSA). In [3] Takahashi et al. present an end-to-end automation of DNA Data Storage in which nanopore reads are progressively filtered. After, the remaining reads are decoded and the corrupted ones discarded. One of the latest works [4] addresses the high error rate in the nanopore reads by integrating a Viterbi error correction decoder with the basecaller and using convolutional codes.

In [5] we introduced a method for the specific encoding of digital images into DNA which includes compression to control the DNA synthesis cost (DWT and quantization of each produced subband) and a biologically constrained encoding that respects the restrictions imposed by the process of DNA sequencing. For a detailed explanation of the encoding algorithm, readers can refer to the aforementioned publication. The performance of the proposed encoding algorithm was tested in a biological experiment, in which the encoded

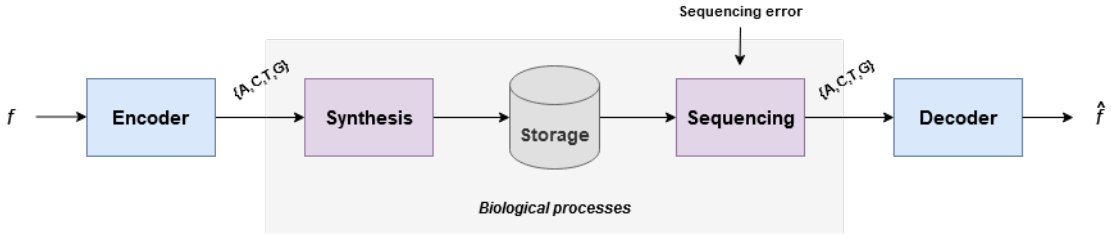

Fig. 1. General workflow for DNA data storage.

and formatted sequences were synthesized and stored in special capsules that allow long preservation of the DNA. For the decoding, the DNA strands were sequenced using the Illumina Next Seq machine [6], allowing a perfect reconstruction of the stored images. After two years of storage, the synthesized oligos have been sequenced with Nanopore which is a very promising sequencer although introducing a much higher error rate in the process of reading.

In this paper, we propose a decoding method to deal with the MinION sequencing noise allowing to recover the original stored data.

2. DNA SEQUENCING

DNA sequencing is the process of reading DNA strands, providing as a result a quaternary sequence. There are many sequencing machines currently available but two of the most widely used are Illumina and Nanopore sequencers. Although they share the same goal of reading DNA strands, those sequencers are based on different technologies which provide each of them with different assets. Despite the high accuracy provided by the Illumina Next Seq sequencing machine [6] which has proven to achieve a perfect decoding of the data in previous works [5], its high cost and low speed constitute two major drawbacks for the use of this device for DNA data storage applications. Aiming to overcome those obstacles, we tested the feasibility of including nanopore sequencing in our decoding workflow by performing a second sequencing of the data synthesized in [5] after being stored for 2 years in a sealed capsule [7] preventing its contact with oxygen and water and allowing its preservation.

For the new experiment, we used the MinION nanopore machine [8]. The speed, small size and affordability of this user-friendly sequencer makes it suitable for real-time applications, bringing DNA data storage one step closer to reality. As stated earlier, despite the advantages introduced by this technology, it also has one major drawback concerning its accuracy, ranging from 95% to 97%, which it is much lower compared to the one provided by Illumina Next Seq.

Prior to the sequencing step and aiming to add extra redundancy to deal with the errors that it introduces, the initial oligos are replicated into many copies thanks to a biological process called Polymerase Chain Reaction (PCR). Hence, the result of the DNA sequencing is a pool of reads which includes many noisy copies of each reference sequence.

2.1. Comparing the efficiency of the sequencers

After being stored in a sealed capsule for two years, we have sequenced the DNA strands which store 2 different images 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.

Due to the low error rate introduced by Illumina Next Seq and the nature of the errors, the selection of the most frequent ones as the most reliable led to a perfect reconstruction of the data. It is important to note that with the Illumina machine we were also able to fully retrieve the header oligos. The results are depicted in figures 5(a1) and 5(b1).

However, when using the MinION sequencing machine, due to the high error rate introduced by nanopore, the decoding is not that trivial and the aforementioned approach is no longer reliable. Instead, we decoded by following the process explained in the next paragraphs.

As mentioned in section 2, the output of the sequencing step is a pool of reads containing many noisy copies of the initial oligos. This noise comes in the form of substitutions, insertions and deletions of nucleotides, which affects dramatically both ends of the DNA strands. The resulting reads will also contain the adapters needed for nanopore sequencing, which were removed using the library Porechop [9]. In addition, the introduction of insertions and deletions at different rates creates significant variations in the length of the output reads. Thus, the next step of the decoding phase is to clean the data, discarding those reads that are highly corrupted due to the noise and will not contribute to the improvement of the results. Consequently, reads are filtered by length, keeping only those reads whose size belongs to the interval $L \pm 10$ nts, being L the expected size of the reads.

The second step corresponds to the retrieval of the reads corresponding to the data we want to decode as the pool of reads does not only contain one image but several as well as other kind of data. To do so, we need prior knowledge about

the identifier of the stored data encoded in some header field of the oligos, the ID field (see figure 3) and its position. As a consequence of the errors introduced by sequencing, the position of the identifier might be shifted so in order to retrieve as many reads as possible, we look for the identifier not only in the original position but within a range around it.

Once the reads which correspond to the data we aim to decode have been retrieved, they are clustered according to their headers. All the reads with non-decodable headers are discarded as they cannot be assigned to any cluster.

The last step before the decoding of the data is the selection of the most representative sequence of each cluster. One of the most widely used algorithms for consensus finding is based on majority voting, assigning to each position inside the sequence the most frequent symbol along the cluster.

Finally, the quaternary sequence obtained after consensus is transformed back to its initial representation to reconstruct the stored information.

The results are depicted in figures 5(a2) and 5(b2). Even though we were able to retrieve reads corresponding to all the reference oligos, the information decoded from the header oligos that contain important parameters regarding the decoding was corrupted, compromising the decoding of the rest of the data. To allow the decoding, we make the assumption that those parameters are known to the decoder. Although this might not be a realistic scenario, the synthesized oligos had been encoded to be read by a more accurate sequencer (Illumina Next Seq) and thus, those header oligos did not need stronger protection to be correctly retrieved. One solution to this problem when sequencing with MinION which has a much higher error rate would be protecting those important fields with the use of error correcting codes as for example error correcting DNA barcodes [10, 11]. The above results prove that despite the assumption of knowledge of the header oligos there is still too much noise corrupting the decoded data. Therefore, it is clear that we are in need of applying a more sophisticated consensus finding algorithm.

3. A NOVEL DECODING METHOD FOR NANOPORE SEQUENCING

As shown in the section 2 it is clear that the nanopore sequencer introduces much noise in the visual quality of the decoded images. In this section, we propose an advanced decoding method which takes advantage of the encoding proposed in [5] in order to improve the quality of the results. This algorithm was tested on the same data presented in section 2.1. In the following paragraph we briefly describe the algorithm that has been used for the encoding of the wet-lab experiment presented in [5].

3.1. Constructing the codewords

The encoding algorithm used to translate the input values into a quaternary code is inspired by the restrictions imposed by the biological procedures of DNA sequencing mentioned in section 1. Those constraints involve avoiding homopolimers and keeping the GC content between 40% and 60%. The main idea is the creation of codewords by selecting elements from the following dictionaries:

- $\mathcal{D}_1 = \{AT, AC, AG, TA, TC, TG, CA, CT, GA, GT\}$
- $\mathcal{D}_2 = \{A, T, C, G\}$

Codewords of an even length l, are constructed by selecting $\frac{l}{2}$ pairs from dictionary $\mathcal{D}_1$. Codewords of an odd length l, are constructed by selecting $\frac{l-1}{2}$ pairs from $\mathcal{D}_1$ also adding a symbol from $\mathcal{D}_2$ at the end of the codeword. To ensure that the code does not create homopolymers, dictionary $\mathcal{D}_1$ omits the pairs AA, TT, CC, and GG and to keep the GC content within an acceptable range, the pairs CG and GC are also excluded. Although the fact that this encoding algorithm does not contain all the possible permutations of the four DNA symbols (A, C, T and G) could be considered a drawback in terms of coding potential, the a priori knowledge about the words which are not considered in our code can be used to achieve a better decoding by adding some sort of error detection and correction during the decoding phase, as it will be further described in the following section. Figure 2 depicts an example of a 3-nt codebook where the red words are omitted according to the above code construction algorithm, ensuring that the concatenation of the codewords will respect the biological constraints of DNA sequencing.

AAA	AAT	AAC	AAG	ATA	ATT	ATC	ATG	ACA	ACT
ACC	ACG	AGA	AGT	AGC	AGG	TAA	TAT	TAC	TAG
TTA	TTT	TTC	TTG	TCA	TCT	TCC	TCG	TGA	TGT
TGC	TGG	CAA	CAT	CAC	CAG	CTA	CTT	CTC	CTG
CCA	CCT	CCC	CCG	CGA	CGT	CGC	CGG	GAA	GAT
GAC	GAG	GTA	GTT	GTC	GTG	GCA	GCT	GCC	GCG
GGA	GGT	GGC	GGG						

Fig. 2. All the possible permutations of the four DNA symbols for creating codewords of 3 nts. Our algorithm for the construction of the codeword excludes all the codewords in red.

3.2. Proposed method for the decoding

As shown in figure 2, the above codec is using a dictionary of 4-ary words that are constructed using known pairs of symbols. Consequently, according to this algorithm there are some words that are excluded from the codebook. In case of a sequencing error (insertion deletion or substitution), it is probable that a correct codeword is transformed into one of those words that are not included in the code, thus denoting an error. This fact can be used for improving the consensus finding algorithm so to provide better estimation of the correct oligos.

In this section, we propose a new implementation of this algorithm which is based on the same principle of majority voting but acts on DNA codewords rather than single nucleotides (see figure 4).

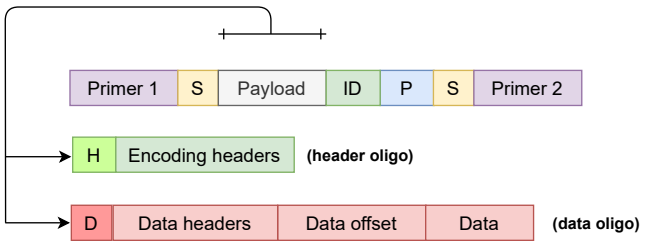

Fig. 3. Format of the oligos - All oligos contain primers that are needed for the sequencing: S denotes the sense nucleotide which determines whether a strand is reverse complemented when sequenced. P is a parity check nucleotide while the ID is an identifier of the image so to be distinguished from 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).

The consensus by codeword algorithm is applied to each cluster of reads and is briefly described as follows:

1. Divide each oligo into the different words
2. Sort words by frequency
3. Select as consensus the most frequent decodable word (i.e. the most frequent word that exists in our dictionary)

With this new consensus we allow an extra step of error correction to find a better consensus by ensuring that the final estimation does not contain undecodable words.

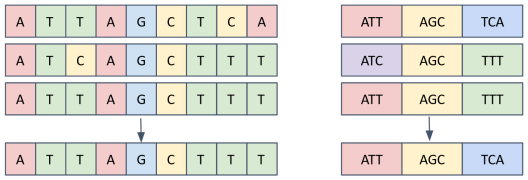

Fig. 4. Comparison of two methods for consensus finding. Left: majority voting on single nucleotides. Right: majority voting on DNA code-words, "TTT" is non-decodable as it does not exist in our dictionary (see figure 2), therefore, it is discarded when building the consensus even though it appears with a higher frequency.

For both images, the PSNR had a significant improvement of around 20 dB when using the proposed method for consensus finding compared to the previous results, leading to a notable improvement on the visual quality of the reconstructed images (see figures 5(a3) and 5(b3)).

4. CONCLUSIONS

This work is a very first demonstration of the potential of the proposed decoding method which takes advantage of the encoding that has been introduced in our previous works for the storage of images into synthetic DNA. To this end we have sequenced the data encoded and synthesized 2 years ago using two different sequencing technologies: Illumina Next Seq

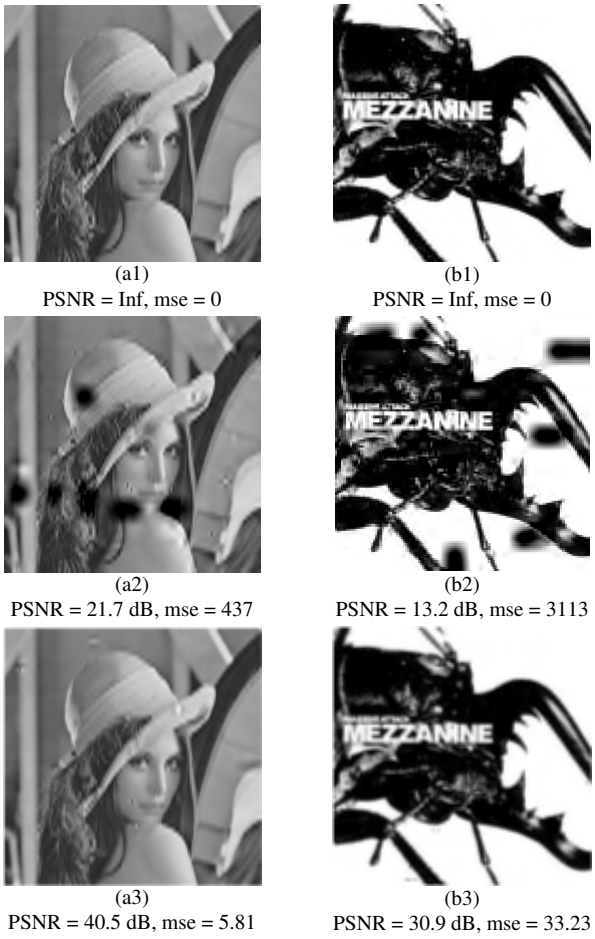

Fig. 5. Visual results of the decoded data - (a1) and (b1) correspond to Illumina sequencing and consensus based on the selection of the most frequent reads. (a2) and (b2) correspond to MinION sequencing and simple consensus based on Majority Voting in single nts. (a3) and (b3) correspond to MinION sequencing and our novel consensus algorithm based on Majority Voting in codewords

(accurate but slow and expensive) and MinION (real-time, user friendly and affordable but error prone). Our results prove that the proposed decoding can significantly improve the quality of the reconstruction. It is important to denote that even though the decoding is not perfect, this proposed method was proven to be very promising for the extremely high error rate of Nanopore sequencing and the results might be even improved by adapting the encoding algorithm to this sequencing technology and by further strengthening the error correction method.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Ellen M Prager, Allan C Wilson, Jerold M Lowenstein, and Vincent M Sarich, "Mammoth albumin," Science, vol. 209, no. 4453, pp. 287–289, 1980.

[2] SM Hossein Tabatabaei Yazdi, Ryan Gabrys, and Olgica Milenkovic, "Portable and error-free dna-based data storage," Scientific reports, vol. 7, no. 1, pp. 1–6, 2017.

[3] Christopher N Takahashi, Bichlien H Nguyen, Karin Strauss, and Luis Ceze, "Demonstration of end-to-end automation of dna data storage," Scientific reports, vol. 9, no. 1, pp. 1–5, 2019.

[4] Shubham Chandak, Joachim Neu, Kedar Tatwawadi, Jay Mardia, Billy Lau, Matthew Kubit, Reyna Hulett, Peter Griffin, Mary Wootters, Tsachy Weissman, et al., "Overcoming high nanopore basecaller error rates for dna storage via basecaller-decoder integration and convolutional codes," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 8822–8826.

[5] Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, and Raja Appuswamy, "A biologically constrained encoding solution for long-term storage of images onto synthetic DNA," in EUSIPCO, 2019.

[6] I Illumina, "An introduction to next-generation sequencing technology," 2015.

[7] Kevin Washetine, Simon Heeke, Camille Ribeyre, Camille Bourreau, Corinne Normand, Hélène Blons, Pierre Laurent-Puig, Claire Mulot, Dominique Clermont, Maha David, et al., "Dnashell protects dna stored at room temperature for downstream next-generation sequencing studies," Biopreservation and biobanking, vol. 17, no. 4, pp. 352–354, 2019.

[8] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson, "The oxford nanopore minion: delivery of nanopore sequencing to the genomics community," Genome biology, vol. 17, no. 1, pp. 239, 2016.

[9] Ryan Wick. Porechop, "https://github.com/rrwick/Porechop," 2018.

[10] Daniel Ashlock and Sheridan K Houghten, "Dna error correcting codes: No crossover.," in 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2009, pp. 38–45.

[11] Eva Gil San Antonio, Mattia Piretti, Melpomeni Dimopoulou, and Marc Antonini, "Robust image coding on synthetic dna: Reducing sequencing noise with inpainting," in ICPR, 2020.