Viability of Synthetic Nucleotides in Digital Storage

Author: Munsu Hwang

Date of Submission: September 22, 2013

Location: Information and Communications Technology Building

Supervisor: Dr. Reda Alhajj and Dr. Jon Rokne

Project Duration: May 1st, 2018 – Aug 31th, 2018

Introduction

As a consequence of the digital age, the amount of data produced annually is increasing exponentially, and our contemporary storage media do not seem to be up to the task. To deal with this deluge of data, researchers have recently turned to the original repository that nature has provided us: DNA. DNA presents itself as an appealing alternative for information storage because of its incredibly dense nature which stems from its extraordinarily small size, making it spatially efficient. Traditional data storage media are expected to last a few decades, while thousands of years old DNA have been sequenced such that the stored information was retrieved. Erlich and Zielinski (2017) describe the "DNA Fountain algorithm" which used a novel strategy to encode various digital items and recovered them with no errors, with the expected information density of 215 Petabytes per gram.

The realization of DNA as a storage molecule coupled with present attempts to synthesize artificial nucleotides therefore presents exciting new possibilities for the future of DNA digital storage. The extension of the genetic alphabet with new nucleotides has the self-evident benefit of increasing the bits per nucleotide ratio, and therefore further enhancing the information density of DNA. Another consequence of using synthetic nucleotides is that shorter strands of oligonucleotides will be able to encode the same amount of information, which will aid to storage fidelity since synthesizing longer oligonucleotides increases the probability of error in the resultant product.

This study uses the P and Z nucleotides from Benner's group as model synthetic nucleotides to test their applicability for digital storage. Using the reported error rates of P, Z and natural nucleotides associated with PCR, and sequencing, a program was built to simulate the molecular processes between the encoding and decoding of oligonucleotides using the DNA

Fountain algorithm. The goal was to determine whether synthetic nucleotides would be viable for DNA digital storage.

Methods

The DNA Fountain encoder was altered to incorporate P/Z nucleotides when encoding a file into a list of oligonucleotides. This was done by modifying DNA Fountain to represent P and Z nucleotides as G and C nucleotides, respectively. Since the encoder filters oligos containing homopolymer runs with lengths greater than four nucleotides by rejecting them , P and Z were strategically used to break up (by replacement) these homopolymer runs to lengths of three or less, allowing more oligos to pass the screening.

The simulation was built to mimic the molecular processes of synthesizing, amplifying and sequencing a given encoded set of oligos. It uses the reported error rates for the three steps listed above found in literature (. However, Benner's group has only published the error rate of PCR using P/Z nucleotides under optimal conditions, with no information about the other two steps. As a replacement, error rates of synthesis and sequencing of natural nucleotides (A, C, G, T) was used instead, but for both steps a range of error rates was presented for these processes and the highest error rate was chosen for the simulation. This was to show that if the rates of P/Z nucleotides were comparable to (or better than) the worst rates of natural nucleotides, then the synthetic nucleotides could be used for data storage. Furthermore, data addressing the probabilities of each types of error (substitutions, insertions and deletions) was lacking. To circumvent this problem, all errors were indicated by a new nucleotide N. The DNA Fountain attaches error correcting codes to the encoded oligos to detect any errors when attempting to recover the file. If the decoder detects an error present in a sequenced oligo, it is simply rejected,

meaning that simulating the type of error is not essential compared to maintaining the probability of an error occurring. This justifies the usage of the generalized error nucleotide N.

The simulation requires four parameters: number of PCR cycles, number of copies of each oligo, average oligo coverage (μ) and the size parameter (r) for sequencing. PCR rounds were set at 10 rounds (same as in the original DNA Fountain article) with μ = 10 and r = 2, which has been described as being average values for the two parameters. Under these conditions, the simulation was applied to two files to determine the minimum copies of oligos (the fourth parameter) required for consistently successful decoding attempts. After roughly establishing the minimum required copy number, ten trials of two different copy numbers per file size were ran.

Results

File size (bytes)	Alpha	Copies of oligos	Average dropout (%)	Success rate (%)
1024	0.65	2	2.88	60
1024	0.65	3	2.88	90
5120	0.40	5	2.93	80
5120	0.40	6	2.60	100

Table 1. Results of the simulations with the specified conditions. Data collected (two right-most columns) were assembled from ten trials under the listed conditions (first three columns). Alpha is specified during the encoding, not for the simulation. It determines the number of oligos the encoder should produce. For example, if alpha = 0.65, it encodes 1.65 x (file size / 8) oligos.

Alpha was set to minimize the copies of oligos without exceeding 1.0, since the cost of synthesizing oligos de novo would be more dependent on the number of copies. Essentially, alpha was chosen to minimize the number of oligos needed to synthesize while minimizing the theoretical cost of the endeavor. It was shown that increasing the copy number also increased the success rate of decoding for the trials under the listed conditions. With the given values for μ (=

10) and r (= 2), the expected dropout rate was 2.78% for all trials, but three out of four simulation runs had an average dropout higher than the expected value.

Discussions

Two key assumptions were made when constructing the simulation. One was that the worst reported error rates for synthesis of oligonucleotides would be comparable to the error rate of synthesis using P/Z nucleotides. The second assumption was that next generation sequencing technologies were available to sequence DNA strands containing P/Z nucleotides. Out of three major groups working on their own synthetic nucleotides, P/Z were chosen because technique for sequencing DNA oligo containing P/Z were available. However the technique utilizes Sanger Sequencing, which makes its use for our purposes highly impractical since the pool to be sequenced will contain hundreds of thousands of oligos. Under these assumptions, the data suggests that synthetic nucleotides can be used to store digital data in conjunction with natural nucleotides. Although we took the highest error rates available, the encoded data could still be recovered consistently with fairly low (3 for first file, 6 for the second) copies of oligos compared to the original DNA Fountain study, which synthesizes thousands of copies per each encoded oligonucleotide. Even if the real error rates of P/Z nucleotides were higher than the one supposed, increasing the number of copies of oligos could make up for extra errors as shown by the data. In addition, the error rates are implemented in the simulation such that every nucleotide has an equal chance (the addition of natural and synthetic nucleotide error rates) of a mistake occurring to it. This in tandem with the generalized error nucleotide causes the simulation to treat the encoded oligo in the same manner as an oligo created by an encoding scheme that properly utilizes the expanded genetic alphabet to increase its information density, even though P/Z was used conservatively in our encoding scheme.

Considering the above, I have concluded that the lack of sequencing techniques comparable to next generation sequencing is what prevents the viability of synthetic nucleotide usage in digital storage, rather than their error rates. Even after a pessimistic implementation of the simulation (highest error rate, generalized error nucleotides, choices of parameters, etc) the decoder was able to recover the original files with low copy numbers of the encoded oligos.

Knowledge Gained

For me, the theme of this research project has been embracing uncertainty. I was initially hesitant in tackling the black box that was research before this summer. Partaking in this project granted insight on developing and executing my own independent project, however, making future research far more approachable. My first important lesson was that plans change. Looking back at the first abstract I conjured up for consideration for this Award, the methods of achieving my research objectives are now unrecognizable in comparison. Another crucial realization was that an all-encompassing knowledge of a subject is not necessary for progress, and might possibly hinder it. Part of the encoding process for DNA Fountain was attaching error correction codes to the oligonucleotides and in an attempt to understand all aspects of the algorithm, I spent many hours in order to grasp how the error correction codes worked. In hindsight, I would have been better off to just comprehend what it did but not how it worked, as I ended up never having to fiddle with the error correction codes.

Acknowledgements

I would like to thank the PURE committee for funding this summer research project. The insight gained from it was invaluable, and my experiences will no doubt serve me well in the future. In addition, I thank Dr. Reda Alhajj and Dr. Jon Rokne for agreeing to supervise me and taking the time to address my questions about the project, and about research in general.