**A robust encoding and decoding algorithm for DNA based data storage**

Golam Md Mortuza1, Shoshanna Llewellyn1, Jorge Guerrero2, Kelsey Suyehira1, William L. Hughes3, Reza Zadegan2\*, Tim Andersen1\*

1Department of Computer Science, Boise State University, Boise, ID, USA,   
2Department of Nanoengineering, Joint School of Nanoscience and Nanoengineering, North Carolina A&T State University, Greensboro, NC, USA  
3Micron School of Materials Science & Engineering, Boise State University, Boise, ID, USA

Higher data density, longer information retention and lower energy consumption make DNA as a promising data storage medium. Here we present an encoding and decoding algorithm for DNA based data storage. By incorporating a novel mapping scheme and a translation stage into the previously reported DNA Fountain algorithm [1], we handled the following biological constraints: i) removing start codons, ii) avoiding repeating nucleotides, iii) excluding longer repeating sequences, and iv) maintaining near 50% GC content. Because of this integration, our algorithm is more robust and can handle a larger amount of error compared to the previous report. Additionally, our algorithm is more efficient compared to previous approach; *e.g.*, for a 10MB file it is two and 400 times faster for encoding and decoding, respectively. As proof of concept, we were able to store 13,170 bytes of information into 604 DNA sequences; each 250 nucleotides long, read the DNA pool using illumina sequencing, and successfully retrieve the original file with 100% accuracy. Further in-silico analysis proved that our algorithm handles files regardless of their content or size.

[1] Y. Erlich and D. Zielinski, “DNA Fountain enables a robust and efficient storage architecture,” Science, vol. 355, no. 6328, pp. 950–954, Mar. 2017, DOI: 10.1126/science.aaj2038.