Encoding Movies and Data in DNA Storage

Naveen Goela
Technicolor Research
175 S. San Antonio Rd., Suite 200
Los Altos, CA, 94022, USA
Email: naveen.goela@technicolor.com

Jean Bolot
Technicolor Research
175 S. San Antonio Rd., Suite 200
Los Altos, CA, 94022, USA
Email: jean.bolot@technicolor.com

***Abstract*— Focusing on error-correction methods and codes, a systems level design is presented for encoding movies and digital information in DNA storage. A source of data (e.g., movies, audio) is compressed, efficiently encoded with redundant information, modulated, and stored in multiple DNA oligonucleotide strands. The goal is to decode the source from the DNA reliably even in the presence of diverse errors introduced by DNA synthesis, PCR amplification, and DNA sequencing processes.**

I. Introduction

Deoxyribonucleic acid (DNA) contains the genetic program for the biological development of life. Fundamentally, DNA may be used also as a compact *storage medium* for petabytes of encoded information [1]–[3]. The potential benefits of DNA storage include: (1) Extremely high-density beyond the order of terabytes in 1 gram of DNA; (2) Stability at moderate temperatures; (3) Biological replication via PCR-amplification; (4) Biological search and indexing via primers; (5) Biological editing and re-encoding of segments via enzymes [4].

II. A System Design For DNA Storage

An end-to-end system for DNA storage begins with bits of information (e.g., a movie file) which are encoded, synthesized, and stored in multiple DNA oligonucleotide strands. The DNA strands must be sequenced, assembled, and decoded in order to reconstruct the original source reliably. Figure 1 depicts the complete storage cycle. The system is comprised of the following components: (1) Source data in bits; (2) Encoding mechanism including all error-correction codes and modulation from bits to nucleotides; (3) DNA synthesis of multiple DNA strands; (4) PCR-amplification of DNA pools; (5) DNA archival storage; (6) DNA sequencing; (7) Merging and assembling multiple DNA strands; (8) Demodulation and decoding of all codes for reliable recovery.

III. Error Correction Codes

In DNA storage, several types of errors occur including: (1) Insertion, deletion, substitution errors within oligonucleotides; (2) Missing DNA strands; (3) Synchronization errors across multiple oligonucleotides with the same address; (4) Low coverage and amplification yields for certain DNA segments; (5) Structural error patterns introduced by synthesis arrays and sequencing machines. Within the storage system, it is possible to define "DNA channels" which may have only approximately-known Shannon capacities in contrast to standardized, precisely-mapped wireless communication channels.

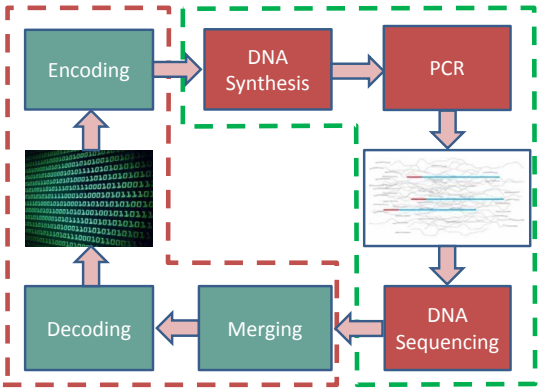

Fig. 1. Block diagram of a DNA storage system.

To engineer a feasible system, multiple levels of hybrid error-protection are necessary. Each oligonucleotide is equipped with an address code which is a unique identifier. Digital payload information is stored across multiple oligonucleotides, protected by modern two-dimensional array codes. While accuracy in retrieval is paramount, such codes must be efficient to reduce overhead costs in DNA synthesis and sequencing.

IV. Conclusion

DNA storage continues to be an emerging, innovative, and viable technology as the cost of high-throughput DNA synthesis and DNA sequencing decreases rapidly. Ideas such as biological editing of DNA [4], computing, search, and indexing provide focus for cutting-edge research.

Acknowledgment

This research was conducted in collaboration with Prof. G. M. Church, and Harvard University's Church Laboratory.

References

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, p. 1628, Aug. 2012.
[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, Jan. 2013.
[3] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: Technologies and applications," *Nature Methods*, vol. 11, pp. 499–507, May 2014.
[4] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *CoRR*, vol. abs/1505.02199, 2015.