

The Persistence of Microbial Memory: An Exploration of the Potential of Bacteria as an Information Storage Medium

James Myslinski^{1,2}, Sarah Katoski^{1,3}, Vanessa Funk¹, Stephanie Cole^{1,2}, Michael Kim¹, Frank Kragl¹, Matthew Lux¹
¹Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD, ²Excet, Springfield, VA, ³SAIC, Abingdon, MD,



Through synthetic biology, this project aims to exploit the use of DNA synthesis to encode large amounts of digital information into *E. coli*. Subsequent sequencing and accurate retrieval of this data will illustrate the feasibility of using living microorganisms for the storage and transferring of digital storage media.

Background

The number of microbial deoxyribonucleic acid (DNA) basepairs contained in a single human gut is roughly equal to the total bits of digital information found in the entire world today. Recent assessments of DNA-based information storage methods have theorized the data storage potential of DNA to range even further beyond today's global information levels, with improvements in DNA storage density potentially reaching an astonishing 2.2 petabytes per gram if DNA^{1,2} (that's approximately 14,000 50-gigabyte Blu-ray discs). In addition to data storage, the project offers a novel approach to study *in vivo* mutation rates in *E. coli* because, unlike genomic DNA, the encoded DNA should have no function and thus no mutational bias. We have leveraged well established techniques from information theory and communications to develop encoding/decoding software that ensure robustness of the DNA to mutation through error correcting codes and RAID strategies. Efforts to clone the encoded DNA into *E. coli*, passage the cells in a continuous flow environment, and sequence the cells over time for decoding and mutation analysis are ongoing.

Information Storage	Bytes	Grams	Information Stability	Robustness to Natural World	Self-replicating?
Digital Data in the World	10 ²¹	10 ¹²	High	Typically Low	N
Microbial DNA per Human	10 ²⁰	10 ³	Medium	Mixed	Y
Human DNA per Human	10 ²²	10 ⁵	High	High	Y
<i>In vitro</i> DNA Storage	10 ¹³	10 ⁰	Very High	Low	N
Bacterial DNA Storage	??	??	??	??	Y

DNA-based Data Storage:

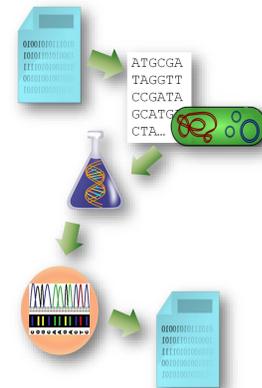
- Technological advances have dramatically increased our ability to sequence and synthesize DNA
- Information density of DNA is remarkable
- What about storing data in cells?

Stability of non-functional DNA:

- Important question for evolutionary and synthetic biology

METHODS

- Task 1:** Develop encoding / decoding algorithm
- Task 2:** Develop *in silico* continuous flow simulator
- Task 3:** Create control and data strains
- Task 4:** Passage control and data strains
- Task 5:** Sequence strains; assess mutation rates



DNA for data storage:

- Recent work has demonstrated the potential of storing digital data in DNA *in vitro*^{1,2}
- Storage in bacteria allows cheap replication of data and the ability to store information in the natural environment
- Potential applications include high density data storage, covert transmission of information, and genomic tagging of high risk or engineered organisms with key information

Mutation rates of "functionally-neutral" DNA:

- Studying *in vivo* mutation rates is challenging because of functional bias (e.g. lethal mutations will never persist in a population)
- Encoded DNA should have little to no cellular function, and thus provides a new way to approach the problem
- We hypothesize that mutation rates for encoded DNA will be higher than genomic DNA

Competition between phenotypically identical cells:

- New evidence for the key role of human gut microbiome is driving a need to better understand microbial competition
- Even competition between phenotypically identical strains can impact outcomes³
- Phenotypically identical cells with encoded DNA allow us to probe the impacts of carrying large amounts of functionally-neutral DNA on bacterial competition

Encoding Data in DNA

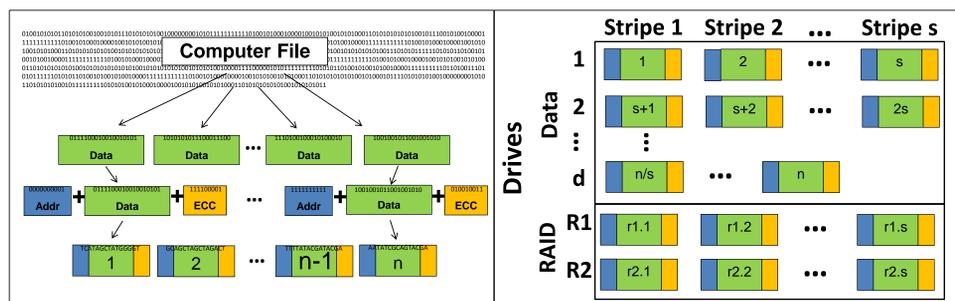
Encoding from digital data to DNA relies on well-established techniques from information theory. Importantly, Error-Correcting Codes (ECC) are used to correct mutations and Redundant Arrays of Independent Disks (RAID) are used to correct large-scale loss of data.

The process is:

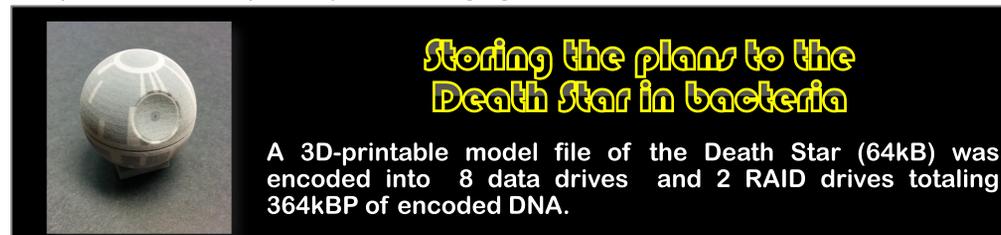
1. Binary data is divided into "blocks"
2. Address data added to each block
3. ECCs added to each block
4. Blocks converted to DNA via Table 1
5. Blocks are arranged into data drives
6. RAID drives are computed

Binary	00	01	10	11
Base Pair	A	T	C	G

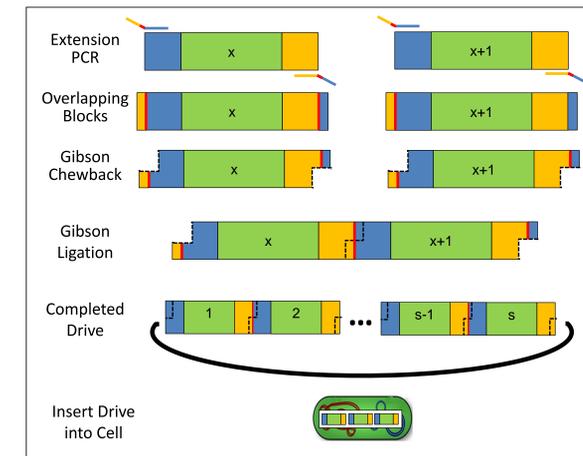
Table 1: Conversion between binary and base pair



Block sequences can then be ordered directly from a DNA synthesis company. To reduce synthesis complications, the sequences are first screened for problematic repetitive elements; if found, random encryption keys are applied, essentially shuffling the sequences in a reversible way, until all block sequences pass screening algorithms.



Drive Assembly

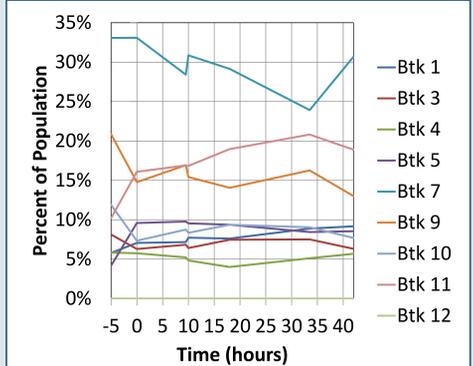


Blocks are currently being assembled by extension PCR and Gibson assembly techniques. First, blocks are PCR-amplified with extensions containing a stop codon and a sequence complementary to the neighboring blocks. In addition, a Bacterial Artificial Chromosome (BAC) vector is similarly amplified with extensions complementary to the first and last blocks. Next a Gibson assembly ligates the overlapping blocks in a single step. The fully assembled drive is then transformed into

competent cells. Assembled drives also each contain a unique barcode sequence developed for the barcoded spore project. These barcodes allow for monitoring relative levels of each drive in mixed population growth.

Strain Competition

Preliminary results of a mixed population of phenotypically identical cells in a chemostat environment. Each strain is *E. coli* DH5α with a plasmid containing a genetic barcode designed to be uniquely detected qPCR. The strains were not normalized before mixing. The figure shows that each individual strain maintains a fairly stable proportion of the population.



Relative error due to assay, sampling, and experimental variability are not quantified. The first time point is during batch growth, 5 hrs before initiation of continuous flow growth; remaining timepoints are at 0, 9, 10.5, 18, 33.5, and 42 hrs in continuous flow conditions. Flow rate was 6.72 L/hr at 19.2 L.

Conclusions

Cells containing encoded DNA are currently being assembled and the recovery of encoded data will subsequently be evaluated. Sequencing data will also shed light on mutation rates for functionally-neutral DNA. Preliminary data suggest that mixed pools of phenotypically identical *E. coli* strains remain stable in a chemostat environment. Future work will explore these questions further.

ACKNOWLEDGEMENTS

This project was funded by the FY14 ECBC Research & Technology Directorate ILIR Program. Work presented is the opinion of the authors and does not necessarily reflect the official policy of the US Government or US Army. This poster is UNCLASSIFIED and has been cleared for public release.

