A Storage and Search Demonstration with DNA-Encoded Text

Laurel Provencher
Catalog Technologies
USA
laurelp@catalogdna.com

Swapnil Bhatia
Catalog Technologies
USA
sbhatia@catalogdna.com

Sean Mihm
Catalog Technologies
USA
sean@catalogdna.com

Abstract - DNA-based data storage (DDS) holds promise to deliver a paradigm shift for long-term, secure storage of data. To tap into this potential, methods must be developed to produce data-encoding DNA molecules with cost- and time-effective processes. Combinatorial synthesis of DNA molecules from prefabricated fragments of DNA offers a solution to this challenge. We are developing a DNA-based platform combining encoding algorithms, high-throughput synthesis, post-synthesis processing, sequencing, decoding algorithms, and DNA computing architectures into a unified system. DNA datasets encoding images and literary works have been successfully created and translated back into conventional data files containing the entire original set of data or a targeted subset of data. In this work, we demonstrate the ability to search for specific molecules encoding a specific word in a DNA dataset encoding the complete text of multiple literary works.

Keywords – DNA, sustainability, storage, search

Conference Topics – Sustainability, From Theory to Practice.

I. Introduction

DNA is the densest known information storage medium capable of supporting a diversity of operations including writing, reading, copying, and certain massively parallel models of computation. DNA is several orders of magnitude more resilient to natural degradation over time than other extant storage media, with a lifetime in the range of 1000s of years. It can be stored in a dry form requiring minimal space and little or no cooling. DNA is also amenable to a wide array of useful chemical methods that scale favorably in cost and energy requirements with the length and diversity of DNA sequences. Technologies for automating DNA synthesis, quantification, purification, sequencing, and chemistry have improved exponentially in capacity, performance, and cost in the past two decades [1].

These observations have led to the emerging field of synthetic DNA-based data storage and computing (DDSC) and the exploration of DNA as the information carrying medium underlying a digital data platform. When successfully implemented, our approach to DDSC will offer a novel option for archiving data at petabyte scales. Platform development will encompass strategies enabling energy efficient options for periodic information extraction and massively parallel computation.

We are seeking input from digital archiving professionals to learn how conventional archiving processes could be re-imagined with DDSC. As performance and scale of DDSC improves, questions around the design of the archiving ecosystem become more important. We encourage this community to help define the minimal requirements that must be met by a DDSC archiving solution.

II. Combinatorial Synthesis Strategy

Most approaches to encoding data into DNA rely on a direct translation between binary source alphabets and quaternary DNA alphabets. For example, “00” → “A”, “01 → “T”, etc. They require the synthesis of a completely new DNA polymer, base-by-base, to produce the molecular dataset. This is infeasible at

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121123

scale without innovations addressing difficult chemistry and physics challenges.

We have developed a unique DNA data storage scheme which encodes data using a collection of disjoint sets S_0, S_1, S_{n-1}, each set containing distinct DNA molecules which we call components. Each component in S_i is designed such that it can concatenate with any component in an adjacent layer. Together, the cartesian product of the sets S_0 x S_1 x … S_{n-1} defines a combinatorial space of DNA molecules ("identifiers") that can be constructed by concatenation of components. We impose an order on each component set and extend this to a lexicographic order on the combinatorial space. We may then treat this combinatorial space as a linear address space. To write a bit value of "1" at an address, we assemble the corresponding DNA identifier using its constituent components and to write a bit value of "0" we do not assemble the identifier corresponding to that address.

A primary advantage of our scheme is that writing relies on rapid self-assembly from a small, fixed set of components, a process amenable to fast self-assembly chemistry, parallelization, and high-throughput automation, rather than base-wise sequential synthesis. Given n component sets each of size c, the size of the combinatorial space defined increases exponentially with the number of layers (c^n) and multiplicatively with the number of components (c x n) with only additive increase in component library size. Thus, the approach is highly scalable.

III. Writing and Reading Data

To assemble identifiers from DNA components correctly, efficiently, and at high throughput, we have prototyped the Shannon system. This print engine contains an array of inkjet printheads to dispense any combination of up to 114 different DNA components as well as reagents necessary to covalently link components. A substrate is continuously fed underneath the printhead array and different combinations of components are overprinted into specific locations to create droplets containing unique sets of DNA fragments. The substrate moves from the printer array into an incubator chamber, then through a collecting mechanism which combines the droplets into a collection vessel [7].

Ordered assembly of identifiers from components is achieved through specific sequence design of the components and the intrinsic base-pairing behavior of DNA. As a material, DNA normally exists as a double-helix structure composed of two polymeric (chain-like) molecules twisted around each other. Each monomer (link of the chain) is one of the four possible nucleotides designated as A, T, C, or G. When the two polymers, or strands, wrap around each other to form a double helix, pairwise bonds form between the complementary nucleotides: A:T and C:G [3].

Each of the 114 DNA components we use to build identifiers are made of a central "barcoding" region of double-stranded DNA surrounded by single-stranded "overhang" regions. The nucleotides in the overhang regions are specifically designed to only complement DNA components from the appropriate adjoining layer of the assembly. When the overhang regions pair together perfectly, the ligase enzyme can create a covalent bond to permanently link the components.

The collection sample from the Shannon system, containing a highly diverse pool of assembled DNA molecules, is processed via a set of standard biochemical lab procedures to concentrate, isolate, and make copies of the successfully assembled DNA identifiers that encode each byte of data. This primary dataset can be split and stored in multiple locations as a liquid or dry sample.

A key tool for accurately reproducing either an entire DNA dataset or a targeted portion of a DNA dataset is the Polymerase Chain Reaction (PCR). This well-established technique uses heating/cooling cycles and a polymerase enzyme to disassociate the two strands of a DNA double-helix and build new strands of DNA complementing each of the separated strands. PCR can exponentially amplify specific DNA molecules in a sample by using specific paired 'primer' sequences in the reaction. Each primer is a short piece of single stranded DNA that complements a unique sequence in each of the separated strands of the molecules targeted for amplification. The polymerase will only make a copy of DNA if it finds a primed region to start building the complementary strand [4].

The DNA components used to create identifiers in our encoding scheme are organized in a hierarchical structure that allows replication of

specific subsets of data via PCR. By performing one or more rounds of PCR with specifically designed primers, molecules representing specific elements of the data can be selectively amplified. This strategy allows us to access fractions of the dataset at different levels, as if accessing a specific file in a directory composed of multiple layers of folders. Alternatively, we can target amplification of identifiers representing a specific word and its position in an ordered string of words.

Once a dataset or subset of the dataset has been specifically amplified via PCR, the DNA sequence of the identifiers can be read with established DNA sequencing platforms normally used for life-science applications [5]. Sequencing the DNA returns results indicating the order of A, T, C, and G nucleotides in a large random subset of DNA molecules present in the amplified sample. The DNA sequences are passed through a decoding algorithm to determine the word and its position in the full text file by determining the molecules' address in the combinatorial tree created during the initial encoding step.

IV. Proof of Concept

To demonstrate the capabilities of our platform, we encoded the complete text of eight of Shakespeare's tragedies, totaling 208,183 words. We then demonstrated a search over this DNA-encoded dataset for a specific query word. Importantly, the time and cost of our search strategy was independent of the size of the dataset. Our approach is founded on targeting and isolating identifiers corresponding to the specific query word. This approach is independent of the size of the dataset as all molecules are targeted in parallel in one step. Therefore, we expect our approach to scale with data size without a commensurate linear increase in the number of steps. We expect our approach will use fixed resources and steps when searching datasets containing up to 100s of millions of words of text.

V. Conclusions

In addition to benefits associated with data density and durability, DNA-encoded data holds potential to enable a new paradigm for performing parallel operations on large datasets. In our example of search and retrieval from encoded text, the efficiency of the operation is governed by the chemistry of molecular interactions. Unlike conventional computing, as the size of the dataset increases, time and energy required for the molecular interactions remains almost constant and enables larger datasets to be processed in the same amount of time.

To illustrate the fundamental differences in scaling a text search with molecular data vs. conventional data, we will use an analogy. Imagine that every word in a text file is represented by a fish. The fish is composed of multiple segments, some of which represent the position of the word in the file and some of which represent the actual word. In this example, the segment with the word of interest is identifiable because it is magnetic.

To perform a word search analogous to conventional computing, each fish must be individually examined to determine if it has a magnetic segment. Thus, one can imagine sending each fish through a narrow pipe and asking whether or not it adheres to a magnetic sensor. As the number of fish in the 'dataset' increases, so does the amount of time it takes to interrogate the complete population for the 'magnetic' word. The only way to speed up the search is to increase the number of pipes and sensors. Likewise, with conventional computing, the only way to scale data processing is to increase the number of individual processors to accommodate larger datasets.

A different way to identify all fish with a magnetic segment would be to drop one or more magnetic probes as 'fishing lines' into the pool. All magnetic fish will be attracted to the magnetic 'lures', with a speed that depends on the strength of the magnetic field and the physical distance between any individual lure and target fish. As non-magnetic fish will pose minimal interference, the total population of fish can scale considerably without significantly scaling the amount of time necessary to conduct the search. This fishing lure approach is similar to DNA-based computing where molecular interactions between complementary DNA strands behave analogously to nano-range highly selective programmable magnets.

To our knowledge, this is the first demonstration of a search mechanism working directly on raw molecular data. The ability to search data without first translating it from DNA back to conventional code means that the entire archive need not be read back into conventional computers. Rather, a

selection process may be used so that the cost of sequencing the DNA data is only expended on the portions of interest. This consideration, together with the anticipated benefits of DNA as a compact and energy efficient solution for long-term data storage, makes combinatorial DNA synthesis an exciting potential solution for digital archiving.

1. REFERENCES

[1] Dong Y, Sun F, Ping Z, Ouyang Q, Qian L. DNA storage: research landscape and future prospects. National Science Review. 2020.

[2] Bhatia, S. Turing Meets Watson-Crick: A Massive Data Storage Platform for Extreme Longevity, Density, and Replicability. Extremely Large Databases (XLDB) April 30- May 2 2018: Stanford, CA, USA.

[3] Ferry, G. The structure of DNA. Nature 575, 35-36. 2019.

[4] Smith, M. Polymerase Chain Reaction (PCR). National Human Genome Research Institute. 2023. https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction

[5] Goodwin, S., McPherson J.D., McCombie W. R. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016.

[6] Bhatia, S, Gildea, K. A combinatorial writing scheme for high throughput DNA data storage. USENIX File and Storage Technologies. February 22-24, 2022: Santa Clara, CA, USA.

[7] Roquet N, Bhatia SP, Flickinger SA, Mihm S, Norsworthy MW, Leake D, Park H. DNA-based data storage via combinatorial assembly. 2021. bioRxiv doi: https://doi.org/10.1101/2021.04.20.440194