Workshop Results

By David Markowitz

This document summarizes the gating research challenges and paths to the development of practically useful DNA storage technology that were identified by participants in a recent IARPA/SRC workshop on molecular information storage. Workshop results were highlighted in Nature and IEEE Spectrum.[1,2]

1. Introduction

The goal of the envisioned DNA Storage program is to develop a fundamentally new storage paradigm that encodes bits using molecular scale polymers (e.g. DNA, RNA or synthetic polymers), which can be densely packed to achieve extraordinarily high volumetric information density. To achieve this goal, methods and tools need to be developed for synthesizing, storing and reading information from synthetic polymers with low latency and high bandwidth. These tools will then be integrated to produce an exabytescale archival storage technology that reduces the power, cost and physical footprint requirements of today's "cold storage" systems by three orders of magnitude.

A number of recent studies have demonstrated the feasibility of molecular information storage[1,2]. Furthermore, a large and highly collaborative research community has developed around this problem since 2012, with contributors from academia (10+ universities in the US and Europe), the biotechnology industry (Illumina, Pacific Biosciences, Twist Bioscience, Gen9), the semiconductor industry (Intel, Micron), and the software industry (Microsoft, Autodesk). Three publicly disclosed academic-industry partnerships now exist, sponsored by Micron, Microsoft, and Semiconductor Research Corporation (SRC)[3],[4].

2. Technical Discussion Summary

To guide discussion during the DNA Storage workshop, the workshop participants were asked to consider how one might build a molecular "hard drive" with the following performance characteristics:

- Total storage capacity: one exabyte,
- Write bandwidth: one petabyte/day, at an operating cost of $1k/petabyte,
- Read bandwidth: ten petabytes/day (1% of total archive), at a cost of $1k / 10 petabytes,
- Infrastructure cost, power and physical footprint requirements: Three orders of magnitude smaller than the requirements of today's exascale cold storage data centers:

Parameter	Current mainstream technology	Target
Infrastructure Cost	$1B / exabyte	$1M / exabyte
Power	200 MW	200 kW
Footprint	1,000,000 sq. ft / exabyte	1,000 sq. ft. / exabyte

The consensus by workshop attendees was that practical, exabytescale molecular information storage is achievable in 5 years. Evidence to support this view is provided below for three key technical areas: polymer synthesis, sequencing, and operation of a molecular storage system.

2.1 Scalable polymer synthesis

Workshop attendees identified three technologies that could deliver this capability. Two methods are specifically targeted at DNA synthesis, because DNA is a well-understood polymer and biology has already evolved machinery for manipulating it. The third method targets non-natural polymers that are simpler and potentially more easily manipulated than DNA. The consensus at the workshop was that these methods can be scaled up to achieve high throughput using dense arrays of nanometer scale wells, which would allow millions to billions of long polymers to be synthesized in parallel.

1. Parallel DNA synthesis using phosphoramidite chemistry. This chemical synthesis method has been the workhorse of synthetic biology for almost 30 years. Although it is slow (adding each nucleotide to a growing strand of DNA takes 200 seconds) and only capable of synthesizing short strands (<150 bases in length), this method has been parallelized to synthesize thousands of strands at once on a single chip[5]. The CTO of Twist Bioscience, Bill Peck, who developed the hardware underlying this technique, commented at the workshop that this method could be improved to synthesize millions or even billions of oligomers in parallel on single chips, and that, with enough synthesis machines running in parallel, it would be capable of writing a petabyte/day.

2. Parallel DNA synthesis using enzymatic reactions. This method exploits biology's own machinery for synthesizing DNA, as an alternative to synthetic chemistry. The bacterially derived "Taq polymerase" enzyme has a maximum synthesis rate of 1,000 nucleotides/second, which is *five orders of magnitude faster* than the phosphoramidite chemistry approach. This approach has the added advantage of being able to synthesize long strands of DNA, potentially millions of bases in length, which could reduce the number of parallel synthesis reactions that are required to achieve target write bandwidth. Both George Church's group at Harvard and biotech startup Molecular Assemblies in San Diego have demonstrated the ability to synthesize DNA strands in a controlled manner using this approach. Parallelization is still required, but a clear path to success exists.

3. Parallel chemical synthesis of non-natural synthetic polymers. This method uses new techniques for controlling the chain growth of non-natural polymers (such as polystyrene) to incorporate copolymers (such as maleimide) at precisely defined locations along the chain, thereby implementing a binary sequence[6]. The main advantage of this approach is that it allows for the use of polymers that have more desirable properties than DNA for archival storage (e.g. higher information density, stronger molecular bonds between adjacent copolymers, etc.). Although this synthesis method is currently as slow as the phosphoramidite chemistry approach, it can be accelerated by optimizing the polymer chemistry. Likewise, parallelization is still required, but a clear path to success exists.

<u>Challenges for achieving scalable polymer synthesis</u>

Workshop attendees identified the following research challenges that must be addressed in order to achieve scalable polymer synthesis:

Major Challenge 1: Fabricating nanoscale systems that maximize the number of wells per chip.

Major Challenge 2: Optimizing the synthesis process to reduce cost.

Getting the cost of synthesis down to $1k/petabyte will require improved parallelization, as highlighted above, as well as improved speed and reliability. During the workshop, the CTO of Twist Bioscience claimed his company (which uses the phosphoramidite method) can already synthesize one terabyte of DNA per day at a cost of $\$10^{-12}$/base. Absent other costs, this would permit the synthesis of one petabyte/day

for $4k; however, in practice, this number would be much higher due to the additional cost of constructing, operating and maintaining 1,000 synthesis machines. (These additional costs are the reason why Twist currently charges its customers a price of $0.03/base, rather than $10^{-12}/base.) Further innovations in both the chemistry and parallel fabrication methods are required to make this synthesis goal achievable with very few machines at low cost.

Major Challenge 3: Manipulating polymers on solid-state media

Although secondary to the first two challenges above, workshop attendees also identified the need to improve current methods for localizing and manipulating molecules on solid-state media (via precision mechanochemistry or electrochemistry) as an additional challenge that must be addressed to achieve scalable polymer synthesis.

2.2 Scalable polymer sequencing

1. Sequencing By Synthesis (SBS). This technique, currently marketed by Illumina, is the dominant method for "next-generation sequencing" of DNA. SBS uses a DNA polymerase enzyme to gradually synthesize a copy of the reference DNA to be sequenced, and uses optical methods to "read" each base immediately after it is appended to the duplicate strand. In practice, this must be done in parallel on many copies of the same reference DNA strand in a well, and the method is only able to sequence short strands of DNA (<150 bases). However, by splitting a genome into many short strands and then sequencing each strand on a chip with thousands of wells, Illumina technology is able to sequence 50 human genomes (~75 GB of data) per machine per day at a price of $1000/genome. The CTO of Illumina, Mostafa Ronaghi, noted at the workshop that there are many options for increasing the read bandwidth of SBS by multiple orders of magnitude if we're willing to tolerate some errors (see section 2.3 below).

2. Single Molecule RealTime Sequencing (SMRT). This technique, marketed by Pacific Biosciences (PacBio), is similar to SBS, both in price/base, and in that it uses DNA polymerase to synthesize a copy of a reference strand, and optically reads the identity of each base immediately after it is appended to the duplicate strand. However, unlike SBS, this method does not require the DNA to be amplified before sequencing; it is compatible with long sequences of up to 100k bases; and it is very fast (300 ms/base). Although the accuracy of SMRT sequencing is only ~85%, this is not a problem for DNA storage applications (see section 2.3 below). PacBio has demonstrated the ability to sequence 150k DNA sequences in parallel on a single chip, yielding ~10GB of data/day. The CTO of PacBio, Steve Turner, noted at the workshop that their newest chips are yielding 100GB/day. This is still a factor of 10^5 away from 10 petabytes/day, but with further optimizations, it may be achievable in 5 years.

3. Nanopore Sequencing. This technique, marketed by GeniaRoche and Oxford Nanopore Technologies, passes an ionic current through a nanoscale hole (typically a channel protein) and measures changes in current as a polymer passes through the nanopore or near it. These changes in current can then be mapped back to constituent molecules of the polymer sequence. This technique is capable of much faster sequencing speeds than other methods, at a cost of much higher error rates. (However, the speed can be adjusted by the user to achieve a target error rate.) Workshop attendees universally agreed this method has the greatest potential to achieve the 10 petabyte/day sequencing target, first of all because it does not require bulky optical components, and second because there is a clear path to combining it with nanofabrication technologies (via collaboration with the semiconductor industry) to achieve parallel sequencing of billions of polymers.

Challenges for achieving scalable polymer sequencing

In the context of a coordinated effort to develop practical molecular storage technology, the requirements faced by this technical area will be highly dependent on the synthesis approach. This highlights the first major challenge for sequencing:

1. Major Challenge: Optimizing sequencing technology for compatibility with synthesis.

Workshop participants noted that any sequencing technology would probably need to operate directly on the storage media produced during the synthesis step, for the sake of operational efficiency. (The logic being that it would be challenging to transfer a billion spatially organized polymers from one chip to another with any degree of reliability.) Therefore, close coordination will be required between engineering staff on both the synthesis and sequencing side of a development effort to ensure that both teams are optimizing their technologies for use with the same media.

Workshop participants also noted that certain sequencing methods are much better suited to certain synthesis approaches. For example, because SBS involves amplification of DNA, this sequencing method would likely require larger wells than those required by SMRT or nanopore technology. Larger wells mean fewer wells per chip, which means longer sequences would be required to maximize the amount of data on each chip. Enzymatic DNA synthesis is the only method that has the potential to produce long sequences, so this would be the best complement to SBS in a storage context. To ensure that synthesis and sequencing methods are appropriately matched, sequencing offerors to the MIST program will need to identify the constraints that their methods impose on the organization of storage media, and suggest how their methods will be optimized for compatibility with specific synthesis approaches.

2. Other Challenges:

Workshop attendees identified the need to improve speed and reduce error rate as additional prerequisites for achieving scalable polymer sequencing. However, as will be highlighted in the "operation" section, high sequencing error rate is not likely to be problematic for reliable information retrieval.

2.3. Operation of a DNA storage system

As with many information technology disciplines, the development of encoding and decoding methods for molecular information storage is a fast-moving field. Participants in the molecular storage workshop highlighted three recent advances that have been demonstrated by these groups:

1. Error Correction:

   Current methods for writing and reading synthetic polymers have a nontrivial error rate (e.g. for DNA, this is on the order of 1% per nucleotide). To mitigate this issue, multiple groups have adapted existing or developed new error-correcting codes to ensure that reliable decoding is possible. Early methods (from 2012) redundantly encoded information across multiple strands, which allowed rare single nucleotide errors to be identified and corrected[7,8]. More recent work employed Reed-Solomon coding, which uses part of the coding portion of a polymer to store parity symbols, and has better error correction properties than simple redundancy[9].

2. Random access:

In 2015, random access molecular storage was first demonstrated using codes adapted from classical storage systems[10]. In this work, DNA sequences were endowed with address strings that allowed specific strands to be targeted for sequencing, while also providing error-correction capabilities. In 2016, researchers from Microsoft and University of Washington used a variant of this approach with two levels of addressing to implement a more sophisticated random access molecular storage system[11]. This architecture is structured as a key-value store, in which each key is mapped to a unique primer sequence, and the corresponding value is distributed across multiple strands containing the same primer target. Random access of a target value can then be achieved by sequencing all strands with the right primer target. Each strand within the same primer group is also indexed by a unique address that supports decoding and error correction, as in [ref 5].

3. Controllable redundancy.

The work by Milenkovic group[10] also demonstrated a new "XOR" encoding scheme that offers controllable redundancy, enabling different types of data (e.g., text and images) to have different levels of reliability and density. This allows for more efficient storage, by enabling data that require high reconstruction reliability (e.g. spreadsheets, server logs) to be stored with high redundancy, and other data that are more tolerant of reconstruction errors (e.g. JPEG images) to be stored with low redundancy. Microsoft recently announced that it is purchasing 10 million DNA strands from Twist Bioscience to test this first-generation controllably redundant coding scheme at scale.

Challenges for operating a molecular storage system

Workshop attendees identified the following research challenges that must be addressed in order to achieve an operational molecular storage system:

Major Challenge 1: Optimizing the code for the channel and the data to be stored.

Methods for synthesizing and sequencing polymers are not perfect, nor are the polymers themselves. Any code that is used to store and retrieve information from a molecular storage system must be robust to strand breakage, "burst" deletions of multiple consecutive bits, transpositions/reversals of portions of a polymer sequence, and potentially other factors, as well. Understanding the properties of the "channel" that is used to encode/decode information (via synthesis and sequencing) will be necessary to design error-correcting codes that are optimized for the media (e.g. via logical or physical redundancy, multiplexing data across wells, etc). Multiple workshop attendees cited their recent published work in this area, and expressed enthusiasm that these problems are solvable within 5 years.

Codes must also be optimized for the data to be stored. The optimal code for storing an image, which can be compressed and reconstructed approximately, may be different from the optimal code for storing server logs, which must be reconstructed exactly. Tailoring a coding scheme to optimally handle multiple data types is an open problem.

Major Challenge 2: Addressing in the limit of exabytes of data.

Randomly accessing information that is stored across billions of polymers on a chip is a nontrivial problem for two reasons. First, each polymer strand must have a unique address that can be identified unambiguously using an imperfect sequencer. This mandates a large address space, which could limit the amount of available coding space on a polymer. Second, for paradigms that use primers for random access to data, it may be vital that each sequence have an address that doesn't appear in a similar form within the data-containing portion of any other polymer sequence. This becomes challenging in the limit of exabytes of data.

Major Challenge 3: Achieving target energy consumption metrics.

Workshop participants from the information theory community noted that encoding data is computationally cheap, but decoding tends to be quite expensive. The computational burden posed by decoding information from billions of polymers in parallel could require computing resources that do not fit into a small room, and use far more than 200 kW of electricity. To solve this problem, researchers in this technical area will need to identify or develop codes that are efficiently decodable, and identify ways to exploit chips that the storage industry has already developed for solving decoding problems rapidly. The latter approach will require innovation in computers and software, since analog-to-digital and data processing are likely to be the most energy intensive step of the decoding process.

---

[1] "Tech Companies Mull Storing Data in DNA", IEEE Spectrum, 20 Jun 2016 http://spectrum.ieee.org/biomedical/devices/tech-companies-mull-storing-data-in-dna

[2] "How DNA could store all the world's data", Nature, 31 Aug 2016, http://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496

[3] http://www.scientificamerican.com/article/techturnstobiologyasdatastorageneedsexplode/

[4] http://www.src.org/program/grc/semisynbio/semisynbioconsortiumroadmap/

[5] E. M. LeProust, B. J. Peck, K. Spirin, H. B. McCuen, B. Moore, E. Namsaraev, M. H. Caruthers, "Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process", Nucleic Acids Res. 38 (2010) 252240.

[6] R. K. Roy, A. Meszynska, C. Laure, L. Charles, C. Verchin, J-F. Lutz, "Design and synthesis of digitally encoded polymers that can be decoded and erased", Nature Communications 6 (2015) 7237

[7] G. M. Church, Y. Gao, S. Kosuri, "Next-generation digital information storage in DNA", Science 337 (2012) 1628.

[8] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA", Nature 494 (2013) 77.

[9] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, W. J. Stark, " Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes", Angewandte Chemie 54 (2015) 2552

[10] S. M. Yazdi, Y. Yuan, J. Ma, H. Zhao, O. Milenkovic, "A Rewritable, Random-Access DNA-Based Storage System", Scientific Reports 5 (2015) 14138

[11] J. Bornholt, Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, , 2016, March. "A DNA-based archival storage system", in: *Proceedings of the 21st International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 637649). ACM.