Summary Report

Technology Working Group Meeting on future DNA synthesis technologies
(September 14, 2017, Arlington, VA)

Bryan Bishop, kanzure@gmail.com
Nathan Mccorkle, nathan.mccorkle@intel.com
Victor Zhirnov, victor.zhirnov@src.org

2017-10-22

Participants:

1. Bryan Bishop / LedgerX
2. Brian Bramlett /Twist BioSciences
3. Sachin Chalapati / Helix Works
4. George Church / Harvard
5. Bill Efcavitch / Molecular Assemblies
6. Fahim Farzadfard / MIT
7. Randall Hughes / UT Austin
8. Devin Leake / Ginkgo Bioworks
9. Henry Lee / Harvard U
10. Qinghuang Lin/IBM
11. Andrew Magyar / Draper Lab
12. David Markowitz / IARPA
13. Nathan Mccorkle / Intel
14. Hyunjun Park / Catalog DNA
15. Bill Peck / Twist Biosciences
16. Michel Perbost /QIAGEN
17. Nimesh Pinnamaneni / Helix Works
18. Marc Pelletier/ DoDo Omnidata
19. Kettner Griswold / Harvard
20. Hua Wang / Georgia Tech
21. Victor Zhirnov /SRC
22. Howon Lee / Harvard

Table of Content

Executive Summary

This document summarizes the paths to the development of practical DNA data storage technology that were identified by participants in a recent IARPA/SRC workshop on DNA synthesis technology development.

The consensus of the workshop attendees was that DNA synthesis can be scaled up to significantly higher throughput and densities. The fundamental research questions of the synthesis based on *phosphoramidite chemistry* have been solved for more than 30 years. All of the primary chemistry and technological components are in place, but the remaining task is to navigate or organize the industry such that engineering efforts can be applied to the devices in order to scale up the throughput/write speed of DNA synthesis for data storage. Some of the different avenues are discussed below, followed by techniques that are in a much earlier research stage but could conceivably yield better performance eventually.

Piezoelectric-based and other deposition techniques for inkjets and pulse jets.
In the printing industry, inkjets for density scaling has stopped when the dots per inch (dpi) became greater that the human eye resolution. Instead, micro nozzle firing rate was optimized for replacement of traditional printing press technologies. These same nozzle technologies have been applied to phosphoramidite chemistry for DNA synthesis. The relevant factors are both nozzle density and nozzle firing rate, among other details. Firing rate is limited by heating problems, which could be solved by adding in cooling elements. The advantage of inkjets and pulse-jets is that they are based on MEMS technologies, which are within the capabilities of the semiconductor industry.

Electrochemical detritylation (for acid generation) on nano-/micro-electrodes.
In these methods, an array of nano- or microelectrodes is placed on a silicon substrate, and these electrodes are used during the synthesis cycle to generate an acid local to the growing polymer on the electrode surface. Usually, buffering reagents are used to prevent contamination between different electrodes on the surface between wells or array spots. The advantage of electrochemical detritylation is that these technologies are relatively easy to fabricate using semiconductor fabrication technologies.

Enzyme-assisted approaches

- Use terminal deoxynucleotidyl transferase (TdT)
- Players: Molecular Assemblies, Cooper Union-2014[1], etc.

Enzyme-only approaches

- Rather speculative at this point
- Early-stage academic research in progress

Library approach

- Deposition of 8mers from a library of 4^8 oligos; each deposited item is ligated
- Not phosphoramidite chemistry

Many common themes for improvement can be found in different DNA synthesis technologies. While not every item immediately improves all techniques, they do each tend to have a broad impact on performance. These themes include the importance of microarray resolution or density, spatial light modulation, addressability of nano-/micro-electrodes, chemical stabilization, reaction monitoring, and others.

1. **DNA as ultimate information storage (Victor Zhirnov/SRC)**

More than just for business and industry, information has been the social-economic growth engine of civilization since the very beginning. Information production correlates with social well-being and economic growth. It is instructive to quantify the information produced in different times. In around 300 BC the world population was about 200 million, and the total amount of information stored worldwide was ~10^{11} bits. And today, we have about 8 billion world population (40× increase) and stored 10^{23} bits of information (10^{12}× increase). Production and use of information has grown from 1000 bits per capita in 300 BC to 10^{13} bits per capita today. (Fig. 1).

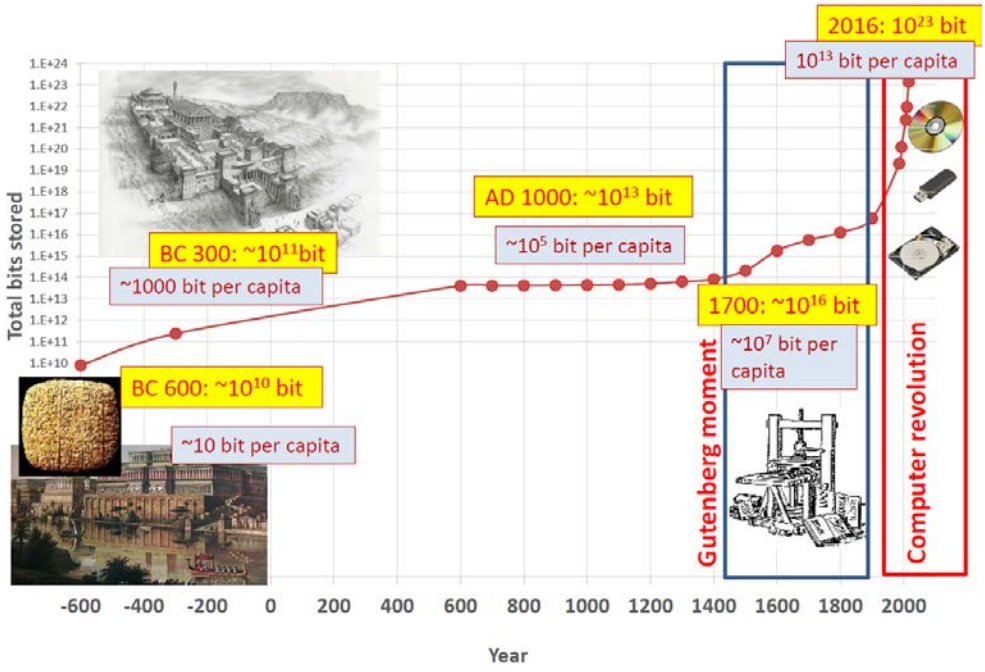


Fig. 1.Information along with Energy has been the Social-Economic Growth Engine of civilization.

By 2040 the conservative estimate the worldwide amount of stored data is 10^{24} bits, medium projection is ~10^{26}, and the high estimate is ~10^{29} bits as shown in Fig. 2 (these estimates are based on research by Hilbert and Lopez[2]).

> Question: " How many of those are "pictures of cats", i.e. "useless information"
>
> Answer: In my view, there is no such thing as unneeded information. Sometimes we save bits just to save time. The trend of growing information production has been here for thousands of years-- so it is difficult for me to see how this would stop. Why do we need more information? My argument is that it's just the nature, it's the driver of our civilization in particular. There's no useless information. That's my position.
>
> Comment (DM): There's a separate IDC study from around the same time that tries to distinguish between all information and probably useful information. It's maybe no more than 20% of all information. The definition of what's probably useful is *who is the stakeholder*. Some people value all the photos of cats they take. For example, the intelligence community would like to save all the cat pictures because you never know which photo will be used to determine which cat set off the bomb.

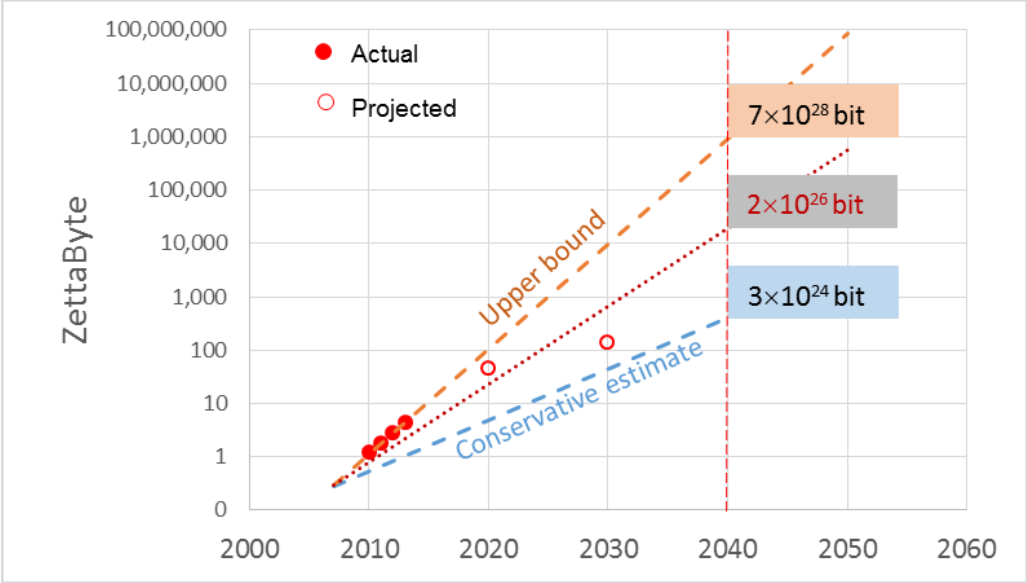


Fig. 2. Estimated and projected global memory demand – including a conservative estimate and an upper bound [3]

What materials resources will be needed to sustain developments of information technologies?

Today's electronic information storage technologies depend on chips made from silicon, e.g. flash memory. Physics-based analysis based on Heisenberg uncertainty principle suggests that 10 nm is close to the scaling limit of flash memory, corresponding to the weight of one bit in flash memory ~1 pg. Thus, storing all data in flash would require more than 10^{14} kg of wafer-grade silicon in 2040 on the aggressive estimate. The projected supply of single-crystal wafer grade silicon in 2040 is at best 10^8 kg. Despite its natural abundance, the silicon resources usable for semiconductor applications are in fact limited, and once mined cannot be replenished. Thus, there is a risk of depletion of basic materials resources needed to sustain developments in information technologies. In order to mitigate this risk, a new materials base needs to be utilized for future data storage hardware that would enable 1000□□higher information densities and support sustainability through recycling, bio-degradability etc.

Nature's solution to the storage problem is DNA. In flash memory, we have at best 10 nm per bit, and here we have 0.34 nm per 2 bits in DNA. DNA has an information storage density that is several orders of magnitude higher than any other known storage technology: 1 kg of DNA stores 2×10^{24} bits, for which $>10^9$ kg of silicon flash memory would be needed. Thus, a few tens of kilograms of DNA could meet all of the world's storage needs for centuries to come.

Future ultra-high density storage systems can be envisioned that are based on biological principles and use biological materials as building components: DNA, RNA, and other polymers.

DNA can store information stably at room temperature for hundreds of years with zero power requirements, making it an excellent candidate for large-scale archival storage[4]. Also, DNA is an extremely abundant (see Box 1 below) and totally recyclable material.

Box 1. Some facts about DNA

- Sedimentary DNA released per year from rivers alone: is 10^7 kg[5]
  - Equivalent of 10^{31} bits.
- Total live DNA in the biosphere is 10^{37} bases[6]
  - Equivalent of 10^{38} bits.
- Information stored in the human body: 6 Gbits per cell times 10^{12} cells equals 6×10^{19} bits.
  - Only the human genome, not including the bacteria (human genome is only 1% of the total DNA content of the human body).

The DNA synthesis technology working group meeting is part of a bigger activity of the SemiSynbio Roadmap, looking for the convergence of semiconductors and synthetic biology. This project seeks to establish a 10-year vision for the semisynbio technologies, including roadmapping system capabilities to enable production of a prototypical DNA hard drive by 2023. The idea of a DNA-based portable storage system is explored with 500 terabytes of searchable data, approximately the size of the Library of Congress web archive. The target is a 5-10 year horizon on a practical product.

Based on a thorough analysis of the topic, it appears that DNA synthesis is currently a critical bottleneck for realization of practical DNA storage technology. Today, our ability to read DNA is considerably better (both cost and throughput) than our ability to write it. How can we address the DNA read-write discrepancy? This is the main question for the DNA synthesis technology working group.

> Comment (DM): On the DNA synthesis cost shown, we should distinguish the actual process cost and the commercial price. With an array you can write something that is in the millionths of a dollar. We addressed this as an IARPA-SRC workshop last year, and with help from George and a few others there, if you're synthesizing a million oligos in parallel, the total cost is like $100 total. I'll make that distinction in my slides. The commercial prices are not the prices from oligos from chips. It's high quality DNA used for genes. As a comparison with Moore's Law, the transistor curve is only for transistors in chips.

Considering the concept of DNA storage it might by useful to look in the recent past. Best available storage technology in 1982 was the IBM 3350 Direct Access Storage Facility (magnetic disk drives). One unit of the size of a size of a big washer machine (Fig. 3) stored 635 megabytes, and it cost was $70,000 USD in 1982. Of course, today we can store much more in a keychain flash stick. And the cost reduction is equally impressive: For example, storage cost for 128 gigabits was reduced from $14 million in 1982 to $20 with today's technology. All this progress has been achieved within our lifespan.

Fig. 3. IBM 3350 Direct Access Storage Facility.

2. IARPA's perspective on molecular information storage (David A. Markowitz)

IARPA's mission is to invest in high-risk/high-payoff research that has the potential to provide the U.S. with an overwhelming intelligence advantage over our future adversaries. IARPA tries to create transformative new capabilities by funding companies and labs all over the world. IARPA's programs are designed to have clear goals, that are measurable, ambitious and credible.

Today, the world is creating data at a much faster rate than storage technologies can handle. There is a risk that within 10-15 years, buying exponentially more storage capacity will become prohibitively expensive (and potentially impossible due to silicon wafer supply). Thus, large data consumers will soon be forced to throw away an exponentially increasing fraction of available data. For some data consumers, that might not be a problem. For example, only a small fraction of the personal photos we take every day might be used in the future. But for the intelligence community, it is often difficult to know in advance, which information is will be useful later, e.g. for understanding terrorist attacks. Depending on the user, it could be really important to save a proportional amount of the information you collect.

Currently, there are three main paradigms for data storage (Fig. 4): Optical (e.g. Blu-ray), Magnetic (HDD & Tape), and Solid-State (e.g. Flash). Their bit features are already close to the physical limits of scaling and further improvement in storage density can be achieved only through 3D integration. However, even in the case of an extreme 3D packing the potential for improvement is limited.

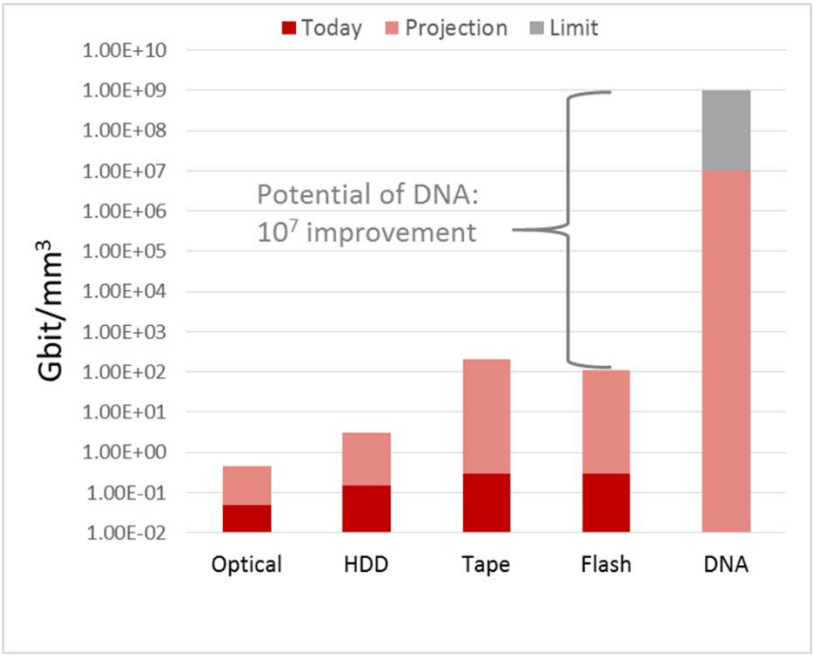


Fig. 4. The volumetric information density of conventional storage media vs. DNA

Molecular media offers far greater more potential for scaling exponentially, e.g. 10^7 above the best expectations for flash or tape storage. In addition to DNA, there are other options for molecular media, such as peptides or non-natural synthetic polymers. For example, peptides are 3x smaller monomer units, and they get you 15 bits/unit instead of 2 bits/unit in DNA.

Organizations that consume large volumes of data typically implement a 3-tiered storage environment, consisting of "cold," "warm" and "hot" storage layers that are optimized for specific access patterns and performance characteristics (Table 1):

- Cold storage is used for long-term archival of data that are rarely accessed and to which rapid access is not required. This is always the largest tier, and is the most appropriate target for immediate technology improvements to close the Storage Gap.
- Warm storage provides high availability and low latency access to frequently used data.
- Hot storage is optimized to provide instantaneous access to small amounts of data.

Table 1. Performance Characteristics of Media In Different Storage Tiers

Storage Tier	Typical Media	Longevity*	Latency	Bandwidth	Price/GB
Cold	Magnetic Tape	10-30 years	50,000 ms	0.3 Gb/sec	$0.004/GB
Warm	HDD	3-5 years	10 ms	6 Gb/sec	$0.05/GB
Hot	Flash	<3 years	0.00001 ms	80 Gb/sec	$0.3/GB

*Longevity = unpowered data retention time (Note: Longevity of tape is counterbalanced by the need for routine refreshes to scrub corrupted data, replace faulty units, and refresh technology. It is typical for organizations to replace all tape media on a 3-5 year schedule.)

Resource requirements of large-scale cold storage can be quite considerable. For example, the 10-year total cost of ownership (TCO) of a 1 exabyte magnetic tape archive could easily reach $1.2B (this estimates is based on the Library of Congress costs for building/maintaining 14 petabytes of archival storage; to our knowledge, a 1 exabyte tape archive doesn't exist yet).

Using molecular media (e.g. DNA), a 1 exabyte storage system could fit on a tabletop and have a 10-year TCO in the $10M-$20M range.

In 2016 IARPA and SRC asked, weather there is a credible path to developing molecular information storage (MIST) technologies that can inexpensively scale to meet future storage needs? And the answer is positive, according to the April 2016 IARPA/SRC workshop on DNA-based Massive Information Storage[7]. IARPA is continually evaluating the current state of MIST technology, and the question now is what new capabilities do we have to develop to make this technology practically useful at scale?

To make some observations, all previous demonstrations of MIST technology have been proof-of-concept using life sciences equipment; there is no commercially available MIST system today.

A number of recent studies have shown that DNA can support scalable, random-access and error-free information storage[8,9,10]. This year, a method, DNA Fountain, was presented, which approaches the theoretical maximum for information stored per nucleotide.[11] This work also demonstrated efficient encoding of information—including a full computer operating system—into DNA that could be retrieved at scale after multiple rounds of polymerase chain reaction (currently, it costs $7000 to synthesize the 2 megabytes of data in the files, and another $2000 to read it). State-of-the-art operating system is a 2016 DNA-based archival storage framework from Microsoft and the Univ. Washington that supports random access from a DNA key-value store.[12]

Two major categories of technical challenges remain:

- Physical Media: Improving scale, speed, cost of synthesis and sequencing technologies.
- Operating System: Creating scalable indexing, random access and search capabilities.

The key challenges are in improving performances beyond the life sciences industry. In the life science industry applications require perfect synthesis and perfect sequencing, while scale, throughput and cost are secondary considerations. For data storage, high read and write error rates can be tolerated, and information encoding schemes can be used. In this application, scale and throughput and cost are primary considerations.

Current MIST workflows take weeks to write and then read data, due to reliance on life sciences technologies that were not designed for use in the same system. Table 2 shows an example workflow for DNA write-read cycle. It is too slow and costly to support exascale archival data storage. Solving this problem will require: (i) Substantial reductions in the cost of DNA synthesis and sequencing, and (ii) Deployment of these technologies in a fully automated end-to-end workflow.

Table 2. Example workflow for current MIST systems

Step	Time Required	Cost
1. Encode files to be archived as a set of oligonucleotide sequences in software	Seconds	Negligible
2. Contract with a DNA synthesis company to synthesize many copies of the desired oligos	Multiple weeks from order to receipt of synthesized DNA	1 x 10^{-4} dollar/base in 2015 (~$3k for 100MB)
3. For each file of interest, identify primers needed to pull it from the archive	Seconds	Negligible
4. Order primers from DNA synthesis company	Days to weeks, depending on the number of primers	A few dollars
5. Hybridize primers to oligos in the data archive; isolate hybridized DNA using gel electrophoresis	Hours	Labor
6. Sequence oligos that were isolated from archive	Hours, if on-site; Days otherwise	1 x 10^{-9} dollar/base in 2015
7. Decode retrieved files from oligo sequences	Seconds	Negligible

As a motivational problem, let's say we want to write 1 TB/day for $1,000, then read it back just as quickly. Using numbers from [11':

- Encoding 2,146,816 bytes requires 72,000 oligomers of length 152 nucleotides each
- That's 10,944,000 bases, or 5.09 bases/byte; 1 terabyte should require 5.09×10^{12} bases
- Writing 1 terabyte/day for $1,000 implies cost/base = 1.9×10^{-10} $/base, write/read speed = 1.1×10^{7} bytes/s

It is instructive to compare these numbers with the DNA synthesis/sequencing cost plot in Fig. 5. On important caveat is that the 'cost' parameter in this plot are commercial prices, and the likely true cost of array synthesis (based upon communications at the 2016 workshops[7]), is much lower, about 10^{-6} dollars per base. However for 10^{-10} dollars per base we still need a 4 orders of magnitude improvement, which is very nontrivial task. Sequencing is more in line with what we need.

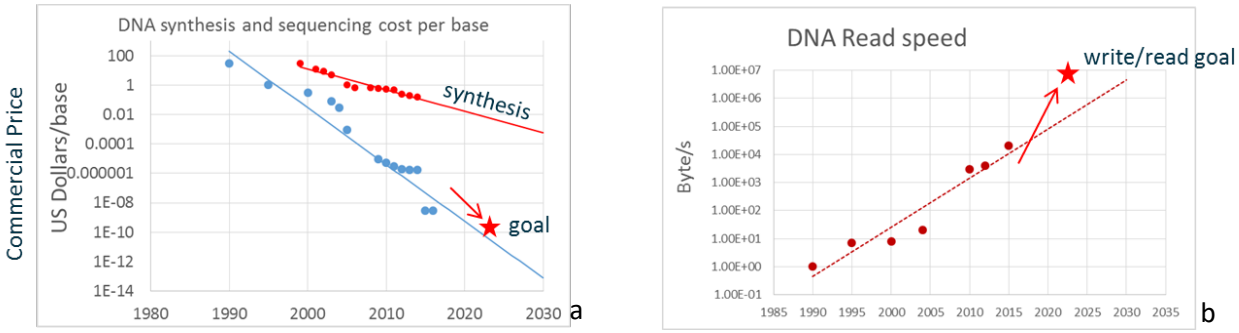


Figure 5. (a) DNA synthesis/sequencing cost, (b) DNA read speed .
(based on Carlson curves[13])

On the throughput side, we need a write speed of 10^7 bytes/s. The current best DNA synthesis speed, is ~1 kByte/s using chemical synthesis where about 1 base per 144 seconds can be written. This needs a considerable improvement

> Comment (WP): That depends on how many machines you're running in parallel. If you put 1 million in a machines in a room, you get to your goal. I could build 40 machines and we can do a terabyte a day.
>
> Comment: With the archival tapes, time to first byte can be order of minutes. For molecular archive, you can do massively parallel, if I can synthesize a lot of primers in parallel, I can search in parallel. Using molecular media offers a lot of potential advantages over tape for data access for more rapid access.
>
> Comment: Throughput can be linearly scaled to some degree. What about speed? If we want a terabyte for under $1k and, and a terabyte per day, so where would you put speed in that kind of hierarchy of goals? I view that as something equally important as reducing cost. If I'm collecting a lot of sensor data, I need to write it as quickly as it's coming in. So that's one use that is relevant to the intelligence community. Even if I can't read it back soon.
>
> Comment: Throughput can be scaled linearly with more machines, however we would need a warehouse of machines to write a terabyte/day. You want to optimize for throughput to avoid that situation. You might want to be able to synthesize a lot of data pretty compactly in a smaller form factor. If I want to label cargo with DNA barcodes where I encode provenance of information, you might want a handheld thing, and then you want a warehouse around it, whatever you do in reduction wise, we're going to put it in warehouses anyway. It's going to be a giant building with its own power supply.
>
> Comment: On the speed, in the spirit of having multiple data... We can store most of that in tape and then slowly write it to data, it could be that speed could be a little less prioritized.
>
> Reply: As a near-term stopgap solution, potentially, yeah. Up to the limit of the bandwidth of each transfer. At some point you have a limit of filling the tape pool, and you need to drain it. As long as we write 10^7 kg per year, we should be fine.
>
> Reply: That's not necessarily unreasonable, if you look at e.g. the example of phosphoramidite chemistry: Right now, there's a large commercial incentive to bring down the cost of phosphoramidites. You need to inject kilograms of oligo into a patient in those applications. If you are talking about nucleotides for biological purposes, there's a question of supply chain, but there's a lot of scalable methodologies for kg scale. dNTPs right now you just normally take a DNA substrate, you exonuclease it and get a bunch of dNTP and you get your triphosphate by doing a cascade of reactions by doing different kinds of cascades; so it could be expensive in the initial phases, but it's a biological substrate as are other things, so the kilograms are not out of the question.

What engineering challenges must be solved to improve synthesis throughput and reduce cost for data storage applications? What new technical approaches have the greatest potential to address these challenges? IARPA usually uses a portfolio of approaches to manage risk. And we need to understand what the current level of maturity is of each approach.

Other Questions/Concerns

Biosecurity. The question about biosecurity is always asked in discussion of DNA synthesis with government stakeholders. If we make DNA synthesis cheap and broadly accessible, then how do I stop anthrax? We got this with 3d printing-- what's to stop people from printing guns at home?

> Comment: Use alternative nucleotides that make it difficult to translate anything dangerous.
>
> Comment: If you're considering current next gen sequencing, you really want short pieces of DNA from a sequencer, even if you're making massive quantities of DNA rapidly, but they're short strands, you're not giving an advantage that someone is going to be doing something threatening. They could ligate them, you could buy pieces today and ligate them. So these tools are available today, and this type of tool won't be transformative to the bad actor in the biosecurity space depending on how it's designed.

For the data storage application, it's probably a less of a concern. I don't want to make presumptions that the current standard of synthesizing short oligos of 150 bp or less is what will be a winning approach for MIST. You can do 1000 bp even with errors. I agree that higher error rate is possibly helping to reduce the biosecurity. We need to have all these answers prepared for non-experts in the national security space who are tracking the progress here.

> Question: We were talking about projections and price per base per for synthesis; $\$10^{-6}$ is the number for gene arrays. Does the previous work on DNA storage only use 100-200 bp long fragments? That's not a gigabase molecule. What are the requirements, -- let's say we're stuck with 100-200 bp, is that acceptable, is that usable? The true cost per megabase DNA is several dollars per base pair because of all the labor and assembly. The projections might look different if we used long DNA numbers instead of the shortest possible.
>
> Answer (DM): Whether you have short or long oligos, like 1k bp or more, it all comes down to what is suitably matched to the device you use for manipulating the biomaterial and for reading it out. So certain sequencing approaches could be better suited for short and long molecules. The cost estimates were for oligos generated for the life sciences industry, which are short. At this point, we don't know of what length would be the optimal solution for data storage. From a design perspective, there might not be compelling reasons to make long molecules. Most of our information tech is based on small packets that is easy to recover or transmit again. If you have one giant chromosome, it becomes physically easy to damage and lose it and all those types of things. With current understanding, I would choose smaller fragments instead of long fragments.
>
> Comment (MP): If you have three letters instead of four you have different system. You don't care about double strand, so maybe three letters would be enough. It's not functional, thus no biosecurity risks.
>
> Comment: Regarding the write speed instead of throughput -- are there methods that allow controllable error rates to modulate throughput? Consider an oligo with an absolutely perfect address, and then the data block is lossy. If I'm trying to encode, for example, a video, I can tolerate a high error rate, but the address needs to be perfect. So a library of oligos, pools of addresses maybe. Can we increase the bit rate by expanding the genetic alphabet? Then you need greater veracity to manipulate the constituent molecule. What new challenges does that create?

Supply chain for the reagents. What's the current cost of reagents for various synthesis approaches? Are there changes required in supply chain? If we want 10^{-10} dollars per base or less, we need essentially

free reagents. How do we drive down that cost? The cost of waste disposal from some chemical synthesis methods can be high and that could contribute to a lot of overhead, does this limit the overhead of scalability of chemical synthesis?

To conclude, my view is that for a program in this space the first phase should focus on synthesis challenges. Are there aspects of DNA synthesis R&D that definitely need to be constrained or guided by the requirements of the later steps in a storage and retrieval MIST pipeline? What are the risks of just focusing on synthesis as a first-order problem?

3. DNA synthesis: Current status and future trends (Randall Hughes)

3.1. Timeline

1970s - synthesis of tRNA gene (20 person years of effort)
1980s - Caruther's elucidation of the phosphoramidite chemistry
1984 - Bruce Merrifield nobel prize in chemistry for peptide synthesis chemistry[14]
1987 - automation of DNA synthesis phosphoramidite chemistry (ABI and others, 200 man-years of effort)
1990s - microarray technologies (such as Agilent)
2000s - CombiMatrix electrode array approach
2007 - mycoplasma genome from Venter's group
2016 - Yeast genome synthesized (Sc2.0)
2025 - Human genome synthesized?

3.2. Phosphoramidite chemistry

The basis of chemical DNA synthesis is the phosphoramidite and the phosphoramidite chemistry cycle. This is based on a protection-deprotection scheme where after removing a protecting group, the next monomer can be coupled to the previous monomer. Under some variations of the chemistry, there is next a capping step. The cycle repeats until the desired molecule is completely constructed. The chemistry is performed "in bulk". Though it is possible in principle (and in certain practices) to construct only a single DNA molecule at a time, it is most common to synthesize 10^{10} to 10^{13} molecules at a time if not more.

The major advance in phosphoramidite chemistry was protecting all the potential reactant motifs on the nucleotide to prevent and to control and direct the chemistry. The first step in the synthesis cycle is to remove the DMT, which is covering the 5-prime hydroxyl. Once this hydroxyl is free of the protecting group, it's now available for chemistry, and it can react with the next monomer in the chain.

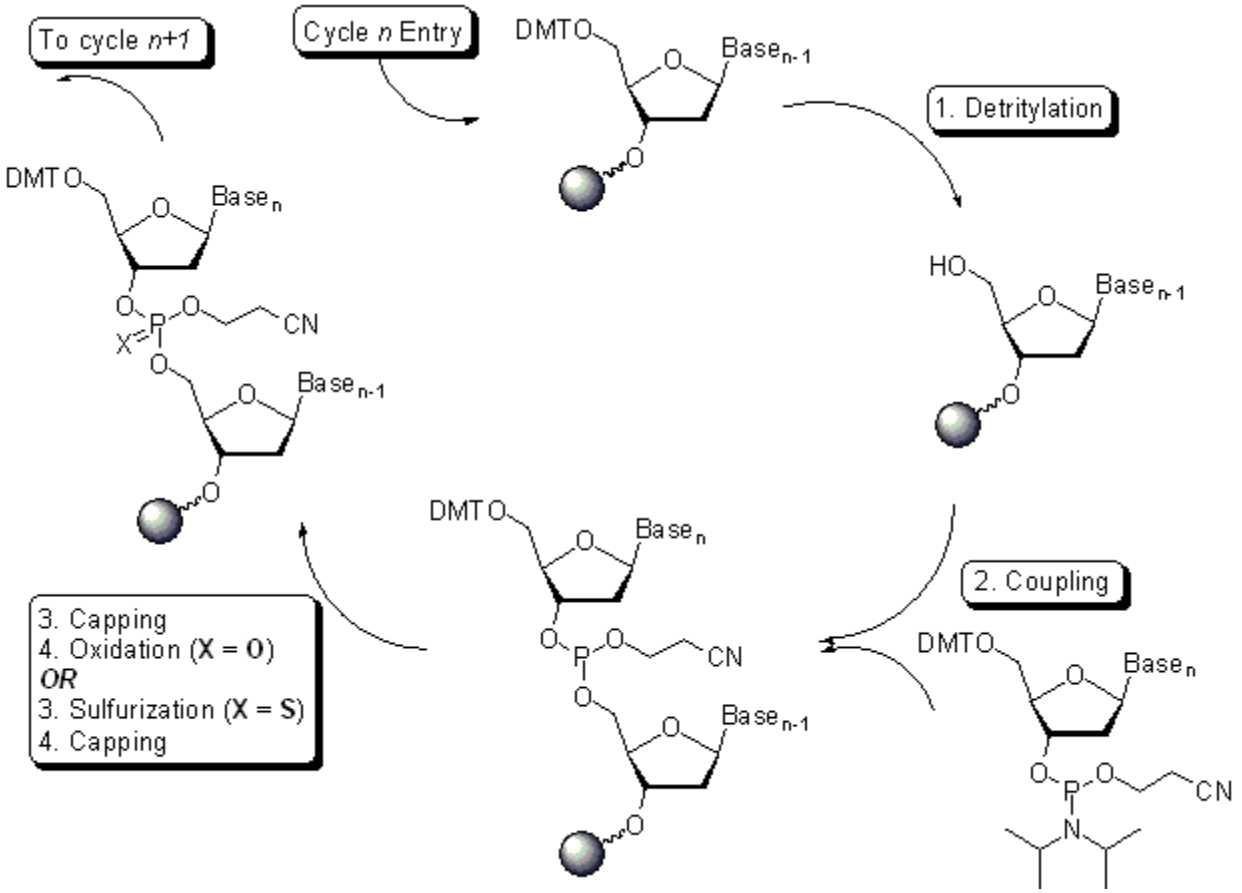


Figure 6. DNA synthesis using the phosphoramidite chemistry cycle.

The **efficiency** of this chemistry is very high, with yields of 99%. For a desired (targeted) sequence of length N, the yield rate calculation is $(0.99)^n$. The current upper limit in error-free synthesis has been approximately 150 to 200 base pairs (bp) since the late 1980s.

The coupling efficiency is 99.5%. As the oligo length increases, the yield of the reaction decreases due to that fraction of a percent loss of efficiency for each step. This places an upper limit or ceiling on the size of the oligos that can be synthesized with minimal errors.

There are many critical parameters. The workshop participants were particularly interested in the time per cycle, which has been about 1.5 to 2 minutes since the 1980s. One participant remarked that 1 gram of product could be synthesized simultaneously in cycles that were each 1 minute 30 seconds long. It was remarked that nobody has really focused on the underlying chemical kinetics of the reaction to optimize the parameters and understand the necessary parameters of the reaction chemistry. Because of the high number of different variables involved, optimization is a laborious task. To optimize the engineering of DNA synthesis devices, it may not be required to completely understand the kinetics of the underlying reaction, although such research activity has the potential to dramatically lower costs, as well as provide more insight into the origin of synthesis errors (deletions, additions, side reactions, and other errors). For example, one participant remarked that to improve the yield and quality of the reactions, they were operating at 55 degrees celsius (which is greater than the usual temperature most often used), however this practice was not adopted by others.

Another cost-sensitive component of the chemistry is the wash step, usually using large quantities of acetonitrile. In fact, one DNA synthesis company is known to distill its output waste to recover acetonitrile as a cost saving measure. The amount of necessary acetonitrile varies based on the physical principles underlying the implemented synthesis technology. For example, oligonucleotides synthesized in confined microchannels will naturally require far less acetonitrile by volume than methods where multiple cm^2 of surface area needs to be washed with acetonitrile.

Total efficiency vs coupling efficiency vs yield vs length vs error rate

Origin of length limitations: purely error-related, or some other underlying reason that 100 kbp molecules are not synthesized regularly?

- deprotection efficiency
- capping efficiency
- incorporation efficiency
- real-time analysis of oligonucleotide during the synthesis reaction?
- G's had lowest coupling efficiency (at Illumina); total incorporation efficiency (not just G's) was 98.5%

The kinetics of the reaction also includes details such as diffusion and mixing. In low volume reaction chambers, such as in micro- or nano-fluidic confined spaces, the diffusion time can be reduced. With CPG methods (not microarray-based method), diffusion has to occur inside of the pores. MIcroarrays and surface-based synthesis is more efficient than CPG because there is a lesser requirement for diffusion time.

Some steps of the phosphoramidite chemistry are monitored in practice. For example, when the DMT is cleaved from the oligonucleotides, the product is colored and this coloration can be detected by DNA synthesizers, which can be used to monitor the quantity of DMT removed during that step.

3.3. DNA synthesis technologies

A number of different DNA synthesizers have been produced, some that have been "mass produced" and others that are more closely associated with one-off research projects.

The most widespread deployed DNA synthesizer type is the column-based CPG synthesizer, which was originally commercialized in the 1980s by multiple manufacturers. Controlled-Porosity Glass (CPG) offers a high surface-area to volume material which silanol (Si-O-H) acts as a starting point for polymerization to begin from. A linker molecule composed of a SiOH-reactive group and a distal nucleoside is first attached to the SiOH, and subsequent steps proceed to work on the nucleosides with phosphoramidite chemistry. After many rounds of extension, the oligos are cleaved from the solid-support, and can be isolated, purified, and potentially sequenced for correctness.

- 96-well plates -- still in use, although technologically deprecated by microarrays.
- CombiMatrix oligoarray - using electrochemical detritylation and electrode arrays
- Agilent - inkjet printhead DNA synthesizer printing onto a surface array. High feature density--up to 1 million features (different oligonucleotides) per chip or array. Each oligo has a maximum length of about 200 nucleotides. In the Agilent method, there are only 3 steps in the

phosphoramidite chemistry cycle. Since this method involves physically printing the reagents, there is no requirement for a capping step.

- Microfluidic (nanofluidic) "picoarray" oligonucleotide synthesizer, commercialized by LC Sciences. This uses a light-activated version of the phosphoramidite chemistry, using a photogenerated acid.

Waste comparison between Mermade 192 and a CombiMatrix synthesizer -- 96-plate in the Mermade 192 will generate 15 liters of waste. In the CombiMatrix synthesizer, an array with 90,000 features (90,000 unique spots with different oligonucleotide sequences) will generate 10 liters of waste.

From the perspective of creating DNA molecules that are meant to be used in biological systems, there is a requirement that the error rate of the total DNA molecule be extremely low. A single deletion error is known as a "frameshift mutation" and can trivially cause a protein to be specified incorrectly, which in turn becomes useless in a biology project. As a consequence of this requirement for a low error rate, it is very common to use DNA sequencing methods to verify the integrity of a short (up to 200 bp) DNA molecule. After the integrity has been verified, the molecule is "stitched" or "ligated" together into a larger assembly. These assemblies are often put through the same sequencing and validation process before being combined with similar molecules (and so on) to make even longer DNA molecules. This sequencing and validation process imposes a tremendous cost on DNA synthesis for biological applications, where errors are intolerable.

In some DNA synthesis techniques, the total synthesized amount of DNA is a low, sometimes as low as a single molecule. In these situations, amplification techniques are necessary to bring the total quantity of DNA up to a useful quantity. In some techniques, such as surface arrays, the DNA is amplified directly on the surface of the chip.

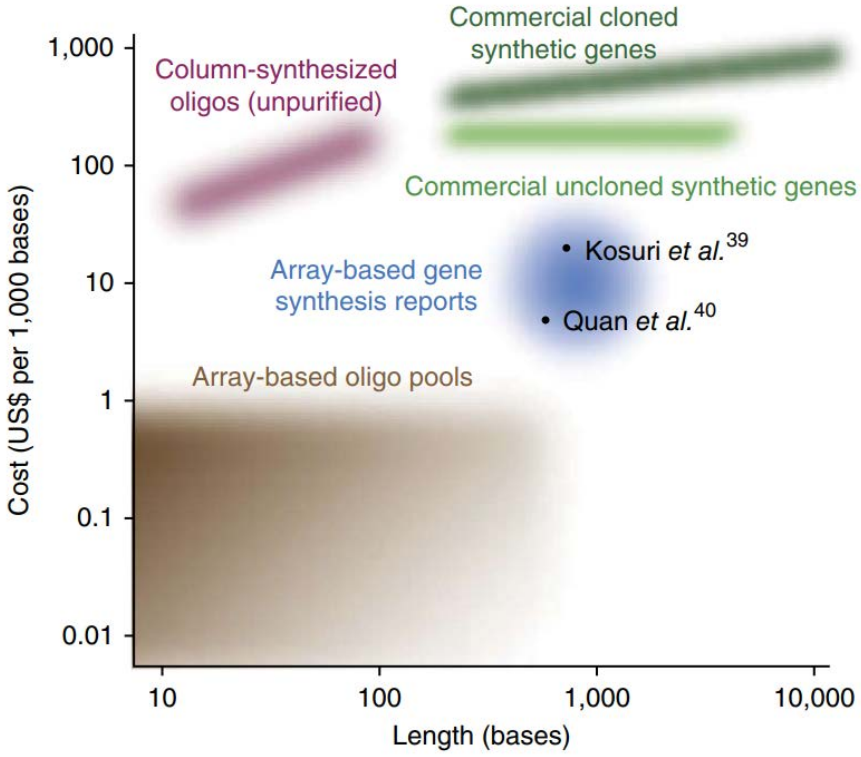
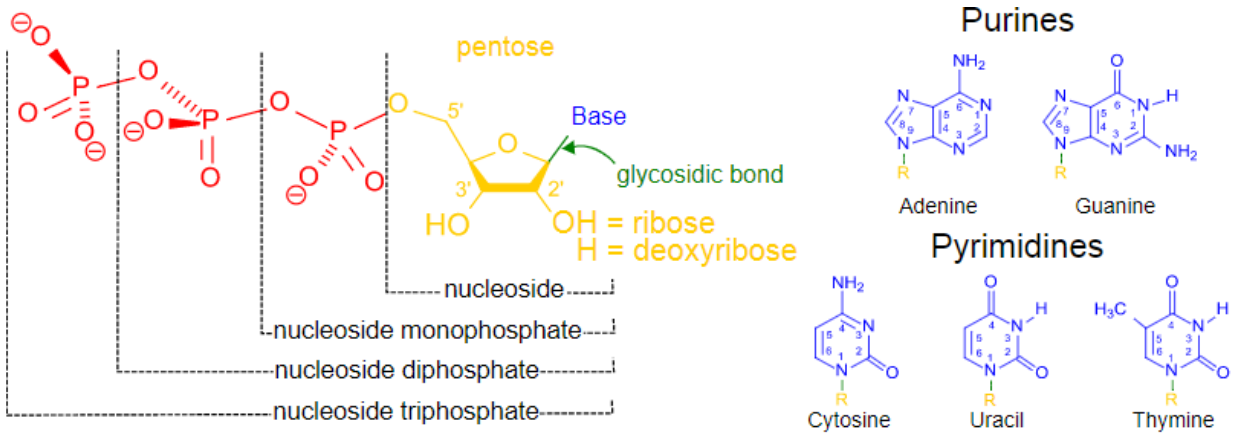


Figure 7. DNA synthesis using the phosphoramidite chemistry cycle[15].

In biological systems, such as the natural replication of DNA by DNA polymerase and other enzymes, the total error rate is often on the order of 1 error per 1000 bp or even lower. A contributing factor is that many DNA polymerase (and related) enzymes have an error checking module that causes the polymerase enzyme to return to the previous nucleotide and attempt another incorporation. The existence of high fidelity low error biological methods of DNA replication indicates that it may be possible to achieve much lower error rates in custom DNA synthesis.

1. **Enzymatic techniques**

Enzymes are attractive because they can be used in water, avoiding harsh organic solvents. . They often require ATP for energy input, and produce ADP as waste. Some examples of potentially useful enzymes include: terminal dioxy transferase (TdT), poly A polymerase, engineered DNA polymerase, wild-type polymerase (ex. error-prone PCR) , recA-like proteins, ligase, cas9/CRISPR (see below for a more extensive list).



A nucleotide is a term to describe any nucleoside with at least one phosphate group attached.
More extensive list of interesting enzymes or enzymatic systems:

1. restriction nucleases
2. endonucleases
3. exonucleases
4. zinc finger nucleases (ZFNs)
5. proof-reading mechanisms
6. polymerase
7. terminal deoxytidyl transferase (TdT)
8. recombinase, strand invasion, etc.
9. Cre recombinase
10. Tyr/Ser site-specific recombinase (SSR)
11. zinc-finger recombinase
12. site-specific recombinase
13. CRISPR/Cas9
14. dCas9 (deactivated Cas9, such as without dsDNA double strand breaking activity)
15. Cpf1
16. fCas9
17. TALENs (transcription activator-like effector nucleases), TALE proteins, ..
18. error-correction enzymes (like MutS)
19. non-homologous end-joining (NHEJ)

20. homologous recombination
21. flippase
22. DNA ligase IV, ligases
23. phage integrase (such as λ phage integrase (Int))
24. phiC31 integrase
25. retroviral integrase
26. recA (a recombinase that doesn't have a specific "core site", but still scans dsDNA for complementarity) (see also Uvsx protein from T4)
27. methyltransferase
28. reverse transcriptases
29. DNA repair pathways, error-prone DNA repair pathways
30. bacterial immune systems
31. chimeric recombinases, (various other protein chimeras should be investigated, various fusion protein stuff too)
32. homology-directed repair (HDR) of double-stranded DNA breaks (DSBs)
33. homing endonucleases
34. programmable endonucleases (zinc finger nucleases, TALE nucleases, cas9, fCas9)
35. meganucleases (such as I-CreI meganuclease)
36. targetrons
37. group II introns (mobile ribozymes that invade DNA), such as 3BWP
38. cytidine deaminase
39. histone methyltransferases
40. protein methyltransferases
41. LwaCas13a (previously C2c2), an RNA-guided RNA-targeting CRISPR-Cas effector Cas13a
42. casposons and self-synthesizing DNA transposons (usually including a polymerase and an integrase among others)

As noted in the above list, fusion proteins should be given a special note. The fusion of two proteins into one longer protein does not always end with function of either starting protein. But in the cases where both functions are retained, if enzymes are involved in a series of reactions, often an increase in reaction rate is seen throughput is seen. In addition, sometimes new functionality can arise which is comprised of a synergy between the pre-fusion enzymes. For example, with the “base editor” fusion protein of cytidine deaminase and dCas9, authors (Alexis C. Komor, et. al) were able to achieve not only precision homing with an “guide” gRNA “primer”, but were also able to convert C:G basepairs to T:A basepairs near the gRNA binding site.

4. Thoughts fair: New technical approaches for future DNA synthesis

4.1. George Church's Thought Fair

In addition to direct synthesis of molecules, the concept of polymer editing was suggested as an alternative regime to consider. DNA and related polymers can be edited in many ways, such as by proteins that cut, insert, bind to specific sequences, copy, recombine, methylate, de-methylate, and many other operations. These proteins need to be useful in both reading (sequencing) and writing (synthesis). In fact, although there are direct chemical techniques for constructing and modifying DNA molecules, it is conceivable to have an entirely enzymatic approach (using biochemistry) to constructing DNA molecules with desired sequences.

A table of gene editing enzymes was presented. This included: group 2 introns, targetrons (mobile ribozymes that invade DNA), zinc finger nucleases, zinc finger recombinases, CRISPR-Cas9, cas1, cas2, cas9, fCas9, TALENs (TALE nucleases), meganucleases, Cpf1, restriction nucleases, endonuclease, exonuclease, flippase, invertase, integrase, recombinases, methyltransferase, etc.

In vivo biological memory systems using cas1 and cas2 (Shipman 2017 and Science July 2016). These memory systems are regulated using small molecules such as TET and lac. In another project, cas9 was used to encode cellular lineage information for the purposes of studying developmental biology.

In other work, Zamft has used a number of different cations (such as calcium) with various polymerases to induce mismatches such that the timing of a calcium pulse in a neuron can be encoded by DNA. This has been previously proposed by Marblestone et al. to be used in the future as "molecular ticker tape", to record the timing of molecular events in situ and later recover the DNA to get specific information about action potentials over the history of the cell.

From a data surveillance perspective, it may be more cost-effective to manufacture the camera and the data storage at the same time. A proposal was offered that instead of recording with electronics, the concept would be to use biological organisms to record with their natural eyesight and record the information in situ inside DNA molecules. Later, the information could be recovered from the brain through DNA sequencing technologies to reassemble the original data. It was mentioned that a dragonfly has a range of 7000 km. This would require significant synthetic biology advances not only to construct the in situ molecular recording devices, but also advances necessary to program the dragonfly to go towards a certain destination or deploy certain behaviors. Behavior could be dynamically programmed using rhodopsins and modulated light, such as demonstrated in the field of optogenetics.

Other than light, there are chemical inputs that can be used, such as odors, TET, lac, calcium spikes, and others.

From an experimental perspective, the task would be to hook up an insect (such as a fly) to a fiber optic cable, flash an image, recover the neurons and attempt to recover the information from the brain a reliable way. The data is collected electrochemically in real time by the organism, and the data is then electronically recovered later using other techniques. An example of biological data storage was provided- the bee remembers the location of certain flowers, and even though it may not be using a single neuron to record that information, there is still information physically stored in the bee brain.

Biological organisms are probably better suited for low-cast data collection and storage, not high write rates. However, the parallelism of biology even in a single organism is quite high.

Using in vivo DNA synthesis, it is possible to make many kilograms of DNA for dollars, not millions of dollars. The research question becomes, how to record relevant information in the DNA using living cells in an organism? The manufacturing cost of organisms (such as bacteria or insects) is very low.

4.2. Brian Bramlett's Thought Fair (Twist)

A chart was given for tape data storage versus DNA data storage cost. The observation was made that tape currently offers more storage per cost than DNA using current techniques. This creates a funding gap between the current technologies (which are much more expensive than alternative data storage techniques) and where the technology could go using basic engineering and optimization of existing devices and techniques.
Generally the thinking of the workshop participants is that an improvement on the order of 10,000x to 100,000x can be achieved in DNA synthesis cost. Additionally, the floor on cost per base pair is expected to be somewhere around 10^{-16}.

The primary engineering targets to achieve these cost reductions is expected to be increased density and increased parallelization. The resolution limits have not yet been reached. "We haven't pushed them to near resolution limits. We haven't tried to get this kind of density. This is not rocket science. We know the physics and manufacturing. This density and parallelism will get us most of the way there and to the market entry point for cost. A lot of the other stuff is refinement, like material transport, kinetics and process control, system and storage architecture, then durability and waste recovery."

By increasing density, the total volume of reagent can stay the same (or perhaps even go down, depending on which synthesis regime is used) and the total number of unique molecules produced increases at the same time. So this demonstrates that there is a capacity increase while keeping a constant price.

An observation was offered regarding inkjet printheads. Twist Bioscience does not use inkjet printheads because colloquially these devices were engineered for human visual acuity and as a consequence their designs were not further optimized by industry. Instead, their term of art is "material deposition systems" and "pulse jets".

The expectation is that the current limiting factor in the development of this technology is investment. It would be helpful to have a "customer-investor" that is willing to tolerate high initial costs to strategically lower long-term costs. This could be an industry consortium or it could be the government, which may have special interests in high-density data storage technologies.

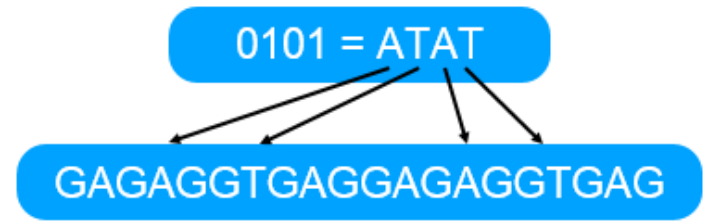
It was estimated that about $20 million is needed to do the engineering and development to get past the major hurdle to the point where DNA synthesis technology starts to look economically credible against tape storage solutions.

The biological industry does not demand high-error DNA synthesis because their projects are completely intolerant of a high error rate. So the credibility push may have to come from outside the biotech industry in order to find that customer-investor or strategic investor that wants to push this technology to the point where it starts to be more competitive than tape storage.

4.3. Helixworks Molecular Storage System (MoSS) "Annealing based DNA assembly"

A DNA assembly and data storage coding scheme was presented. The strength of this scheme is that it works with today's macroscale laboratory synthesis techniques, and uses DNA hybridization for error-correction.

The system begins with a 20-mer oligo storing 4 bits of information using only 2 bases, A and T, interspersed within stretches of G and C for good annealing.

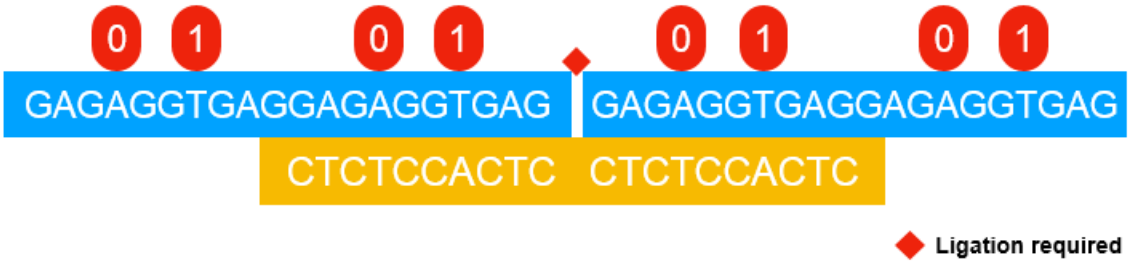


A & T bases = 0 & 1 of binary; Stretches of G & C – Annealing; 4 bits per oligo

Because only two bases (A and T) are used for data encoding and order is important, a 4-mer oligo library contains 16 unique possible sequences (2^4). Because 4-mers are considered too small for hybridization and ligation (at least ezymatically) and to prevent self-priming, linkers/adapters on either

side of the the data is required. This doubles the number of unique oligos, to end with a library of (16) forward and (16) reverse 20-mer oligos. Each oligo in the forward set has a stop codon for biosafety, but while the reverse oligos lack this, they lack a CAAT sequence or TATA box (DNA to RNA transcription promoters in many organisms) and thus are assumed to be bio-inactive..

Combining 3 oligos (or pools of oligos) with i.e. (2) forward and (1) reverse at a time allows hybridization to occur, at which point the adjacent two are joined by ligase or Click-chemistry. This is done in a recursive step-wise manner, where a ligated oligo is fed into the next hybridization/ligation cycle as one of the 3 oligos.



Using this system with current commercial synthesis, data could be stored for $50k/GiB and written at a rate of 10 bytes/minute. It requires stringent temperature and concentration conditions, and thorough washing between steps.
This system seems to assume that short oligos, like those standard today, will be unable to scale up in length or down in reaction molarity, but could be made much cheaper.
Prior art shows that using synthetic oligos with forward and reverse sequences, along with PCR and other enzymes, is able to take advantage of hybridization for error-reduction and assemble correct clones of up to 1.8 kilobases (https://doi.org/10.1093/nar/gkn457).

4.4. Bryan Bishop's Thought Fair

A brief overview of the engineering around inkjet printheads and pulse jet printheads was given. An observation was offered that DNA data storage can tolerate much higher error rates than biological applications of custom DNA. As a consequence, some amount of additional error in pulse jet technology can also be tolerated, such as faulty or entirely broken valves, as long as the total information transferred to the surface is high relative to the requirements of the information error encoding scheme used in the custom DNA molecules.
Industrial inkjet printheads have been known to have 2,500 or more micro-nozzles that dispense liquid volumes anywhere between 3 pL to 200 pL. Off-the-shelf industrial printheads have firing rates between 10 and 100 kHz. It was remarked by multiple participants that inkjet printhead technology was mainly developed for the purposes of meeting resolution limits on human visual acuity for viewing printed media. However, the actual limits of this technology are unrelated to the targets that the industry was developing for.

Hire firing rates have only been briefly explored in the academic literature, with rates as high as 1 MHz. Hire firing rates interact with problems related to liquid-air interfaces as well as overheating of the printhead, which negatively impacts performance unless cooling elements can be introduced into the printheads.

The presenter also gave a brief overview of a hobby-scale "garage biology" project for DNA synthesis based on POSAM (http://bioinformatics.org/pogo/). In this scheme, the overall throughput of the device is lower than the targets sought by the working group participants. The advantage of a hobby-scale inkjet synthesizer, besides personal educational value, is that the components can be freely licensed using open-source licenses, which reduces the transacton costs of follow-on innovation. The

original POSAM printer was constructed in parts at a cost of less than $50,000 USD. This provides a base system that can be indefinitely expanded with optimized components for low cost, while staying within the reach of amateurs and hobbyists or researchers that wish to operate their own equipment.

An analysis of whether matter deposition (such as inkjets or pulse jets) can realistically scale better than CMOS electrode arrays for electrochemical detritylation remains as an open research question.

More speculatively, some thoughts were offered regarding a draft of an upcoming review paper on tethered polymerase enzymes that can be optically or electronically controlled for specific sequences.

4.5. Nathan McCorkle's Thought Fair

Computer engineers have expected to read and write single memory elements for a long time, aside from high-reliability applications like RAID. At a reaction scale of 100 pico Mols, which is quite small by average commercial DNA synthesis scales, you still have 10^13 molecules. Enabling a DNA based hard-drive is foreseeably possible with today's *known* reaction kinetics and fabrication techniques, reading single molecules via nanopore sequencing is becoming common today, and DNA write via downscaling from 10^13 molecules at a time to single molecule synthesis. Economic gain can easily be had with an integrated semiconductor and nanofluidic handling system. It has been shown that solid-state nanopores with atomic-thickness metal electrodes can be fabricated with high-volume manufacturing (HVM) techniques including Reactive Ion Etch (RIE) and Atomic Layer Deposition (ALD). It has been shown that single molecules of DNA can be moved around nanofluidic channels using electrophoresis. Using nanopores with atomic-thickness enables sequencing or single-monomer dispensation dosing via tunnelling current. It it thus not difficult to imagine delivering single monomers to an enzyme or other active chemistry.

- transistors - many 1000s of atoms per logic element (14nm gate)
- DNA - about 15 atoms per nucleoside (monomer of ssDNA)

liquid phase synthesis seems really interesting, it enables you to operate on the synthesis reaction in ways solid phase doesn't. Such as sorting the molecules by size/mass, to determine if synthesis proceeded as expected, during synthesis instead. This real-time feedback in-situ is not done today with solid-phase synthesis, and error-detection happens at a much later time in a different laboratory device.

Nanofluidics enable us to get away from bulk reactions (100 pico mols == 10^{13} molecules) and enable the controlled and coordinated manipulation (1) and analysis (2) of single molecules.

Even without nanofluidics, just being able to check during synthesis "did it get added" would be a valuable but not unfeasible enhancement. You only need to add length discrimination to the overall flow of things. If an A was supposed to be added to a 10-mer, and we enumerate the lengths of the synthesis oligos, we should expect an 11-mer, any 10-mers need to be re-cycled in the addition step (or discarded). You could imagine this done with liquid-phase synthesis where after each step you perform electrophoresis on the DNA a pre-calibrated length of time in the X direction, then siphon it off in the Y direction.

There are numerous ways to sense single molecules besides nanopores, one is the Ion sensitive field effect transistor (ISFET). These detect protons, and are useful for watching a polymerase which for each nucleotide added has a reaction product of a proton. The ISFET amplifies this single proton event into a cascade of many electrons rushing through an electronic logic network. There is also Surface Enhanced Raman Spectroscopy (SERS) which can be combined with micro and nanofluidics, as well as nanopores. Scaling down will increase our parallelism and throughput. Nanofluidic handling will solve chemical transfer loss and reduce the number of redundant molecules being manipulated. Integration with electronics will reduce signal loss which traditionally copes with this by increasing the number of redundant DNA molecules.

Citations:

1. Fabrication of sub-20 nm nanopore arrays in membranes with embedded metal electrodes at wafer scales
   https://dx.doi.org/10.1039/c3nr06723h
2. DNA translocation through short nanofluidic channels under asymmetric pulsed electric field
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4000398/
   Supplement (videos): http://aip.scitation.org/doi/suppl/10.1063/1.4871595
3. Silicon radio-frequency planar nanofluidic channels
   http://ieeexplore.ieee.org/abstract/document/6697337/
4. High speed nanofluidic protein accumulator
   http://pubs.rsc.org/en/Content/ArticleLanding/2009/LC/b823409d#!divAbstract
5. DNA tracking within a nanochannel: device fabrication and experiments
   https://dx.doi.org/10.1039/C1LC20075E
6. Entropic cages for trapping DNA near a nanopore
   https://www.nature.com/articles/ncomms7222
7. Electrophoretic manipulation of single DNA molecules in nanofabricated capillaries
   http://pubs.rsc.org/en/content/articlelanding/2004/lc/b312592k/
8. Direct Analysis of Gene Synthesis Reactions Using Solid-State Nanopores
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5154552/

4.6. Molecular Assemblies Thought Fair

Applied BIosystems was manufacturing DNA synthesizers in the 1980s. It has been observed that 150mers were synthesized back then and it's still roughly the standard length today and still state of the art. There is a generally poor understanding available of the physico-chemical limits of phosphoramidite chemistry. It would be useful if there was more basic research into the underlying synthesis chemistries. However, at the same time, it was remarked that building DNA inside of an organic solvent might be a non-sequitur that warrants more investigation into enzymatic approaches to DNA synthesis, both hybrid chemistry plus enzymes and enzyme-only.

The lack of interest in DNA synthesis from an economic perspective might be because therapeutics and primers are the dominant application of the technology in industry, and custom DNA for synthetic biology purposes has remained in little demand. As long as the cost of DNA remains high, there not be as much demand from researchers to design projects that require large quantities of custom DNA.

The observation offered is that the bottleneck is all around the underlying chemistry and enzymes. The cost reduction drivers have included parallelization and high density surface arrays, which are downstream engineering concerns that are not strongly associated with the chemistry.

Instead, Molecular Assemblies has been investigating the use of Terminal Deoxynucleotidyl Transferase (TdT) and template-independent polymerases as a method of synthesizing custom DNA. This is a hybrid chemistry and enzyme scheme. In the near future, it may be possible to eliminate the terminators from the chemistry scheme. One scheme investigation involves polymerase extension, followed by removing the reactants, followed by removing the terminator. The ideal scheme would have only a single yield-limiting reaction.

One idea that was proposed was to do enzymatic synthesis directly on a plasmid. The reason that this is economically interesting is because in other synthesis technologies that might reach a near-zero reagent cost per step, there is still a requirement to do amplification which has an associated cost as well. With an enzymatic approach, there can be a reduction in the number of downstream handling steps.

Bit space increasing modifications were mentioned as a possibility-- such as using alternative polymer chemistries where instead of 4 nucleotide types there would be 40 nucleotide types. However, turning these into phosphoramidites is an expensive operation. The enzymatic approach could accommodate an 8 nucleotide type system, or perhaps higher.

Phosphoramidite chemistry took 100s of man years. It was remarked that the total amount of time spent on enzymatic synthesis at Molecular Assemblies so far is about 3 man years.

Molecular Assemblies has hired its own organic chemists. They make their own precursors. In addition, Molecular Assemblies works on enzyme engineering to tailor the enzymes to improve their performance in the enzymatic synthesis scheme. The primary goal is to make TdT add one and only one nucleotide per step. The secondary problem is the terminator removal chemistry. The terminator must be compatible with the enzyme but this compatibility requirement makes it uneasy to remove from the oligonucleotide molecule.

One of the problems with TdT is that there are side reactions where it removes a nucleotide. In a highly confined liquid volume system, direct (optical) observation of fluorophore reaction byproducts would be necessary to detect additions or removals and to inform system decision regarding what the next reaction step and next reagents to deliver should be.

Enzymatic methods have the potential to require less resources and generate less waste. This is important in the context of a "black swan" event that occurred where acetonitrile prices went up due to the failure of several other components of the supply chain. "It was the downturn of the automobile industry, it was the tsunami, it was the Beijing Olympics, and it was hurricane Rita hitting the Texas coast. It knocked out a lot of our chemical production. Downturn of the automobile industry-- largest consumer of plastics that produce acetonitrile as a byproduct. Air quality standard shutdown during the beijing olympics. Tap off, no acetonitrile."

Traditional DNA synthesis methods have a mixed waste stream. A question was raised as to whether it's possible to have separate waste streams for the phosphoramidite chemistry technique, and then in particular whether the reagents in the enzymatic synthesis scenario can be reused. It is expected that the cost per enzyme (which is already exceedingly low, below 10^{-8} dollars per enzyme) could be amortized over the length of the polymer being constructed.

5. Implementation Strategies

5.1. DNA synthesis methods and technologies (list)

- inkjet DNA synthesis (using phosphoramidite chemistry)
- non-inkjet rapid deposition manufacture techniques ("toner transfer"?)
- photolithography phosphoramidite chemistry (DMD/micromirror array, LCD, LED matrix, etc.)
- column-based bulk synthesizers (the consensus was that these are exceedingly unlikely to meet the demands)
- electrochemical detritylation
- DNA methylation approaches (nanopore or otherwise)
- TdT methods (Molecular Assemblies, Cooper Union 2014, and others)
- enzyme-mediated phosphoramidite chemistry
- directed evolution of enzyme-stabilized phosphoramidite chemistry
- alternative polymer chemistries (Lutz et al., XNAs/PNAs, unnatural nucleotides, oligonucleotides functionalized with other properties, methylated DNA molecules, etc.)
- electronic, optical or resonance-based control of tethered polymerase enzymes or other proteins (such as methyltransferases, recombinases, gene editing proteins, etc.)
- oligo library methods (which can be applied to inkjet/deposition-based methods)
- microfluidics/nanofluidics/mems - usually this is an enabling platform, to which specific synthesis techniques can be applied

Broad trends:

- parallelization -- multiple separate write operations working simultaneously
- density -- more inkjet nozzles, more micro-/nano-electrodes, micro/nanofluidics
- throughput -- inkjet nozzle firing rate (overheating & cooling, fluid properties, etc.)

5.2. Some questions addressed at the workshop (with answers)

Question: Today, our ability to read DNA is considerably better (both cost and throughput) than our ability to write it:

Qa. Why has DNA synthesis technology lagged so far behind DNA sequencing. Are there fundamental reasons (such as the laws of physics and chemistry) or is it simply economics (e.g. lesser demand)?

Answer:

- Economics
- Need was stressed for "a big customer" (probably the government) to signal credibility
- For biology or synthetic biology applications, one of the requirements is that DNA molecules contain minimal errors. Sequencing is used to eliminate errors from the produced molecules, which adds to the total cost of the initiative. However, this requirement is far less relevant in the context of molecular information storage during the write stage.

Qb. What needs to be done to obtain a radical decrease in the cost of DNA synthesis? When would this be possible?

Answer:

- Minimize volume of reagents used
- Increase total throughput and parallelization
- Reduce the cost of technologies around DNA synthesis to achieve synthesis cost improvements

- Companies need to employ engineers to improve the technology, at the cost of not attending to their other main lines of business
- No shorter than 9 months

Qc. What needs to be done to obtain a radical increase in the throughput of DNA synthesis? When would this be possible?

Answer:

- Engineering task
- Increase working area (surface area)
- Increase density (microarrays, nanoarrays, etc.)
- Increase total spot count (microarrays, nanoarrays, inkjet valves, etc.)
- Needs engineering work and significant testing, no shorter than 9 months?

Question: What are ideal synthesis capabilities to have 5 years from now?

Qa. What applications would this enable?

Qb. What is the time horizon for miniaturized (e.g. portable) DNA synthesis technologies?

Answer

- participants mainly discussed large synthesizers stored in warehouses
- portable synthesizers could conceivably exist (microfluidics) but there's not a strong economic incentive at the moment for the industry to pursue this avenue (in fact, the market for DNA synthesizers is very small)

Question: The semiconductor industry has accumulated unique tools and experience that could be useful to DNA synthesis/storage research and development activities. What are some pathways for deployment of the expertise and infrastructure of the semiconductor industry in the DNA synthesis/storage domain?

Answer:

- inkjets
- nano-/micro-electrode arrays
- photolithography (used in some DNA synthesis techniques)
- focused ion beam etching of nanopores, high-density arrays

Question: Why DNA instead of alternative polymers?

Answer:

- Broad base of biotech industry and tools that can already work with DNA.

Question: Can DNA ever compete with SSD?

Answer:

- Yes, if enough R&D is put forth. With the correct memory hierarchy and system architecture.

Question: Can DNA ever compete with archival media?

Answer:

- Yes, it would probably just be an earlier, less parallelized version of the 'SSD' type goal. Or be a simpler architecture.

5.3. R & D Targets and Timelines

5.3.1. Plan of Action (high level)

- 3 months: complete extensive literature review of microfluidic/nanofluidic sequencing and synthesis (chemical and enzymatic)
- 6 months: completion of review document, finish development of some potential system architectures. Finish design of experiments to de-risk portions of a given system (some experiments/test-chips could contribute learning for multiple architectures)
- 9 months: determine universities/companies which will perform the experiments
- 1 year: check-in on experiments, re-evaluate system architecture possibilities, depending on if experiments have yielded any data
- 1.5 years: design fabrication procedure for an integrated proof-of-concept device
- 2 years: approach goal of proving concept of highly-dense microfluidic/nanofluidic synthesis device which is manufactured in high-volumes (HVM)
- 5 years: DNA hard-drive that competes with current SSD products
- Semiconductor fab technique based molecular handling system (assumes no enzyme engineering, no chemistry engineering, just debugging materials compatibility)

5.3.2. Technology Developments Roadmap

Microarray density and parallelization

- 1 year: exploration of scaling limits of picotiter plates.
- 2 years: manufacture of large picotiter arrays with millions of spots/cells
- 5 years: manufacture of large picotiter arrays with billions of spots/cells

Inkjet and other deposition printheads

- 1 year: exploratory engineering of printhead compatibility with phosphoramidite chemistry
- 2 years: optimization of printheads for phosphoramidite chemistry parameters
- 5 years: development of printheads with 50,000 nozzles
- 10 years: development of printheads with 1 million nozzles
- 15 years: development of printheads with 100 million nozzles

Electrochemical detritylation

- 1 year: investigation into electrode density and chemistry parameters
- 3 years: 200 nm gap between electrodes
- 5 years: 50 nm gap between electrodes

Enzyme engineering approaches

- 1 year: initial investigation into fusion proteins and available DNA synthesizing/editing enzymes
- 3 years: engineered proteins controllable by optical or electrical signals
- 5 years: broadening scope and toolbox of engineered proteins that have direct relevance to DNA synthesis, editing, methylation, etc.
- 10 years: first versions of engineered proteins that manufacture DNA according to digital signal input
- 20 years: total control of DNA synthesis by digitally-interfacing proteins

APPENDIX

A1. Implementation Strategies Discussion

(Lead: Devin Leake / Ginkgo Bioworks)

Program framework discussion:

- DL: The consensus of this group of experts is that significant improvement in DNA synthesis for data storage are needed. An initial goal would be to demonstrate write speed of 1 terabyte/day at the cost less than $1k/terabyte.
- Scale of the project: “Apollo project” (e.g. an ambitious end-to-end solution for data storage) vs. a set of less grandiose tasks (e.g. improvement in DNA synthesis?)
- We need a roadmap, and not just one point. What technologies should be under consideration now? There are several goals that need to be achieved simultaneously. We are focusing on DNA synthesis, so it seems that we are not focusing on an end-to-end solution at this point. Should we be? What is the disadvantage if we do this?
- At this point, nobody is architecting or attempt to architect an end-to-end solution, even on paper. So maybe we're not ready to attack that.
- Investment/risk management standpoint: The need for improvements is orders of magnitude greater on synthesis than on the sequencing side. This represents a significant risk. We probably don't want to spend all resources and not have anything deployable at the end. Is it better to spend, e.g., a third of that upfront to just de-risk synthesis? Does that maximize the remaining money?
- DM: We still want to deliver an exoscale read-write solution to the government. The reason to focus on tera-scale is that there are problems that you can solve and exist today, within government, where tera-scale is sufficient. Being able to synthesize a terabyte is sufficient.
- An integrated solution should include both reading and writing. Today, the perception is that the reading is owned by at least 80% by one company. So you can't write a whitepaper around an integrated solution without anticipating either a brand new read method, or some sort of collaboration and partnership.
- We are a small startup and we can throw a dart really far, because the people investing in us don't care. We're attempting that integrated system and process, without divulging too much IP. Our goal is to have that one device and replacing a hard drive that you can just plug in. We have the philosophy of a roadmap. I don't think you're hurting your budget by spending on synthesis. I think it's a logical process.
- DM: IP is a serious issue. When you operate, you launch the project, you will be walking on a lot of feet because there's a lot of IP around and you will probably have to use ideas from other people. Like open-source in software, could there be a shared licensing scheme, where you are guaranteed to get a license?
- The error rates specified in the last year workshop seem to be awful ambitious and not needed for data storage. I was told in the workshop to not specify error rates, and leave that to the coding theorists to figure that out. So this phase 1 would say you're going to do your project, you're going to store a petabyte,
- What about a DARPA model where you get multiple groups to work together and put together a common proposal? For an end-to-end solution? In the first workshop, there was clearly a division of the sequencing entities and they were there - and if you were not partnering with them, you were at a distinct disadvantage in terms of developing end-to-end solution. There are

models that might could work. DARPA iFAB with results owned by government and then distributed to the participants.

- DM: My thinking is that phase 1 would be on synthesis to de-risk that. And then phase 2 would be more collaborative.

Technical approaches discussion:

- What's the current state of the DNA synthesis landscape? Gas-phase radical chemistry could work. Radical reactions are very fast. For example, molecular beams were used to study this in the 80s. For the purpose of doing synthesis of a polymer, there's no precedent for that, people didn't chase that down. What were the relaxation trajectories for these molecules? They had an optical probe at where the beams would intersect. It's basically atomic molecular orbital physics. It's not as commonly done today because we have very good computers and some people prefer to do density functional theory for low-level quantum calculations to calculate these things. When you look at it as a technology where you're considering all the different methods that have been developed like with mass spectrometers and with other classes of making molecular beams, and radical mechanisms that have been discovered, it may be possible.
- DM: Just to be explicit about what I care about, whether or not feasibility has been demonstrated makes the difference between whether it's appropriate for funding through an IARPA program that is trying to deliver a new capability versus funding as an NSF project where it's a basic project.
- Twist can do it today. Is uses the chemistry that is 40 years old. It works. You can do it on an array in your sleep. All that they need to do is build a warehouse. Go get a warehouse that makes DNA. It would be very expensive to have 100s of machines in that space. If you want that top line number and it pains me to say this, they can probably deliver it.
- DM: The central tension is that should IARPA be putting its money into that right now? There's a molecular informatics program. NSF and SRC are putting in money into that. There might be other stakeholders putting in money into that. It helps make me a more common sales pitch to management to say this is ready for us.
- DM: Maybe as a benchmark, when you look at, for example, enzymatic synthesis, how much proof of concept is there, and does it compel you to say that could be an option? You can create a sequence of AAAAs using the enzymatic approach. I haven't seen any reports of sequence-controlled enzymatic DNA synthesis. The most I have heard is 10 minute cycle time, all nucleotides can be done, although some might be less efficient, and you're looking at 20 bp or less. Is that sufficient from a benchmark standpoint? What is an achievable goal? I would like to see quantitative evidence of what current capabilities are. We can say terabyte/day in 2 years, but if we are still at the point of only being able to do 20 mers with enzymatic approach and no clear path to go beyond that, is that 1 terabyte/day really achievable?
- What is missing here is polymer chemistry. DNA is a polymer. They can make 10,000mer. It might not be the kind of monomer you want. ABS, maybe. What can Jean-François Lutz do? His work looks promising. But one of the advantages of DNA are the 50 years of tools available. He has to to develop tools for himself.
- We are looking at DNA as the backbone, deoxyribose and phosphate - only because we are thinking of biology. If you remove the backbone, there was PNA but I hated it. You don't need to have a phosphorylase. You can make a very long polymer with some really simple molecule and 10,000 different groups, you can have base 20 or base 50. That's not an unreasonable thought. But then you're not taking advantage of the readout technologies. It's an integrated problem.
- What if it's biocompatible? Say you swap out the sugar phosphate backbone, if it has compatibility with the same polymerase, then sequencing. It's challenging, but not impossible.

But the transferase or TdT or a diester or something - you need it to have catalytic property that might not be easy to find. Polymerase is invoking sequencing by synthesis. You could also use a nanopore, but then there's a lot of challenges in nanopores. A lot of the initial direct sequencing bifurcated between sequencing-by-synthesis and fancier ways of reading off of DNA.

- With nanopore sequencing if you were going completely synthetic, you would have more options because you could introduce more differentiation between bases. You can copy DNA with DNA polymerase - is a really nice feature. While just starting to get into sequence-controlled polymer synthesis, there are no polymer template-dependent polymerases that I know of. So there's not good ways to copy that information. So the fact that you can take your whole DNA molecule and put in a primer and copy things out, is a really nice feature of DNA that would take a lot of engineering to get to in a new system.
- I have confidence in DNA as a molecule. Its advantages are density and so many others. Density, half-life it will last a long time, perhaps not an advantage over other polymers, but it lasts long enough. In the long run, DNA is going to be the molecule to answer the questions. When you are doing tape-to-tape backups, you have to upgrade your machines every 10 years. But DNA is going to be the same. If a polymer changes, then you have to use new machines and maybe there's a collapse of society. In a million years, they are going to be able to read DNA.
- There's biology vs in vitro question. In five to 10 years, if you were to make a bet of progression of this technology, how far into biology is this going to go? Based on your experience, do you see biology being a driver in this space? Do we really need DNA synthesis to archive information? What about editing DNA, or libraries of oligos? Like you take the human genome, and you introduce point mutations. For example, take the human genome, and then you edit a specific base pair on the genome, and that's new information. You just need to have all the addresses of the human genome. Do you see biology getting to the point where it can address these types of characteristics?
- Rather than doing bottom-up synthesis, you're doing some sort of top-down manipulation. For example Tim Lu's lab at MIT uses Cre recombinases to access a number of states and to make combinatorics work out so that you get a certain level of encoding. But you have to keep in mind that you are playing with physics so there is going to be some level of degeneracy that you must achieve to have distinct orthogonal recombinases. Maybe you do this sequentially. The recombinases are one of many possible ways to modify DNA. You can edit the DNA, like the MAGE technique. So perhaps you could introduce the mutations.
- It could be done in vitro, you could introduce precise mutation in vitro into a plasmid. We could modify every single place on a plasmid. Or we can consider de novo synthesis in vivo.
- Most organisms have a haploid step in their lifecycle. They have roughly gigabytes of data. Mostly without mistake. If we could do that with a computer, put 12GB in our DNA strand, one single strand, just one copy of everything.
- What about oligo libraries instead of de novo synthesis?
- If you had a gigabase of homopolymer "A", can you edit that whole thing in a day? If you are looking at homopolymers, it would be challenging. You would start with a known genome, like e coli genome.
- You don't really need to synthesize DNA if you take a cell and readout DNA. You can just fixate a primer where this data might be bits and pieces across the genome, and then you need the primers for the exact regions where the data could be assembled back. We don't need to synthesize the DNA, it's already there. It's easy to copy.
- Library of oligos: Before you use them, you amplify them, you keep your copies, you have a library that never gets depleted, and now you can do mix and match, I think that's what Bryan was talking about. And this would be the alternative to de novo DNA synthesis.

- What about size? We talked about 150 nt. With regard to biological approach, chemical, or enzymatic approach, what is an optimal size to go for? 150nt is within the scope of phosphoramidite chemistry without having loss of fidelity or information. Are there reasons to go longer?
- Do we know what the optimal size is? Do we have the back of the envelope calculations for that? For longer chains we will need deliver reagents to a specific things, have a deblocking step, a shearing across the surface of some kind, etc. There are biophysical limits when the DNA becomes long enough.
- John J. Kasianowitz at NIST has a good insight on the DNA length from the read perspective. Let's say I have this scheme. I don't want to get into the details of how to encode it, but you can say order of magnitude faster if you had a Reed-Solomon unique callable code that is very distinct in these spaces you are measuring. It comes out to like 50,000 or 100,000 something that you could maybe grab and move around. That you think should be able to fit through the nanopore in models that we have that are working. That's within the reading limits of what people are working. As soon as you have a polymerase involved here, or a helicase or something, you're talking about something that can run at a fairly large speed, the secondary question is what sort of electronics infrastructure is needed to decode it? To make calls of symbols rather than bases. 50-100k nt can be done. I don't know if there's an advantage to it. So the reason why you would do that, let's say parallelism or throughput.
- DM: The information theorists said leave that decision to them. This is determined by the coding scheme you use. The theoretical maximum best coding framework is like 99% of the sequence is address, and then 1 base is actually coding for information. I forget why this was true. Their opinion was that a 10k nt is not ideal, if you want to do error correction for example. If I say we're going to do phase 1 of the program, and you need to make 500mers in such and such amount of time. That's arbitrary. It might not be optimally matched to a given coding scheme.
- You can only put so much current through it. If you are actuating with photons, electrons, there's a number of different things to consider. At some point you decide this is the threshold of parallelism that you are going to do at each pixel. How much are you going to do in one synthesis run? There is an upper limit at the top of the flow cell. At some point it's not feasible to think about how you might collect that, and start making a new batch. The reason would be that you are Manhattanizing your DNA synthesis because there's only so much real estate, and that's when length becomes key. DARPA does this all the time. They put out this proposal for pandemic proposal and response. And they say things like, make this virus and have this mutation rate. With these three questions, we talked about size, we talked about throughput, and I'll add in cost. It didn't really impact the program and it's money that was picked up. The goals changed a little bit, but it still met the overall objective.
- Chemistry aside, we're building equipment with screwdrivers today but we're in a room of semiconductor technology focused folks that do things in planar fashion and we get a great economy of scale by doing so. What are the touchpoints that semiconductor know-how or expertise can lend to this endeavor? Further than just saying instrument or equipment, are there specific examples on what has been done in semiconductors and if we could marry that with biology or chemistry? We talked about miniaturization this morning.
- There are things we know we can do in semiconductor fabrication. We know we have to do multiple chip runs to explore the process control space. It could be prohibitively expensive, unless we have a market that is demanding that.
- What about combimatrix approach, electrochemical detritylation - what are the scaling limits there?

- What about DMD micromirror arrays and photolithography? The chemistry is similar. It's the chemistry of the protecting group that is the limiting factor. You get 96% deprotection and you're happy. It is an investment area.
- Inkjet, line up a billion nozzles, and just go. What's the state of the art in parallelism and hertz? I think it would give you a good idea of the best way to actuate. A billion nozzles in a row is not competitive with current production rates. With inkjet you can remove steps, like capping, and secondly you remove washes between steps. This seems compelling differentiating point.
- On arrays, you realize that the cost is not the phosphoramidite. It's in the washes. That's where the cost is. You put nanodrops, they are nothing. And then you have to wash the whole surface, that's where the money is.
- But we also don't know what are the hard limits. People buy screens with certain pixel density - there's a limitation on what lumens we need on each pixel, and what sort of lumens we need to actuate a photochemistry. Or acid generation - we know there are theoretical drawbacks, but there are other chemistries that could be used for deblocking. From the practitioner perspective of biochemistry and chemistry, if I am going to make a decision about what I'm going to map to, what are the hard limits on the parallelism?
- I can tell you how many spots I need to get a really high resolution fab of my DNA array. I need to know is it physically possible to put that many amperes into my chip? How many can I put in before I start seeing cross-talk or things that are actually affecting it? And finally, what are the yields on massive parallelism? For example, there's a reason why when DARPA went for gigapixel imaging they split it up into many chips. These are fundamental questions that need to be answered. This is more easily answered by SRC people than biologists.
- VZ: What I Was going to propose later, but I'll do it now, to form a couple of task forces. For example, one would be exploring tradeoffs on the size. What is the optimal size for this in the best case scenario? And we need to explore the possible options for advantages or disadvantages on sizes. What I would like to have as an outcome of this event is a documentation of ideas. Even numbers on comparing DNA and current e-storage. Make everything transparent. As soon as we have it, it becomes much easier.
- VZ: Another task force to explore limits of miniaturization, how small do you want to go? How much leverage can we get from the semiconductor people?
- VZ: Henry from Church's lab suggested a number associated with the cost of waste disposal. To compare it to the waste we produce in semiconductor industry - there's a lot of waste. When we're concerned about waste in DNA synthesis, well we should compare it to the waste to make pure silicon and so on.
- Enzymatic approaches: is there any waste that can't go down the city water drain?
- BE: Most of it can go down the city drain in the case of Molecular Assemblies. They are relatively innocuous agents. It's vastly different from phosphoramidite chemistry.
- This is one of the most interesting and detailed questions that was posed, which isa comparison between cold storage and warm storage and accessing. We haven't talked a lot about accessing information and some of the advantages and disadvantages. What are some of the limitations with these types of approaches or at least considerations as we think about the technologies used to develop it?
- We've been talking today about write-once read-once, not that there's a problem with that. We should consider the possibility of accessing. There have been some good papers - like using enrichment strategies, PCR design, but I think there's some interesting things that are brought up. Exact content versus similar content. Do we see limitations with DNA being employed for that?

- I think one of the problems is that, talking from the chemistry and biochemistry side of the room, I feel like I don't know what the state of the art is in storage system engineering. How do they index what is what, without searching through the entire data set? How do databases do access patterns? There has been a lot of information and gained from developing that.
- This is a very complex problem. My specialty is synthesis and sequencing. I can give ideas here and there, how it interacts I don't know about. In my case, the question today is synthesis. How many molecules do you want? How long do they have to be? And how many do you want?
- It would be helpful to have people with direct data storage experience. The way we see it, DNA data storage fits into a cache system. And how does this fit into algorithms and when to put things into DNA storage and when to warm it up and start reading? This can help solve latency. You want schematics for each of the steps and make sure the output from what you're synthesizing is matching whether it is accessibility or some kind of read capability. We don't replace tape or silicon, we augment it. We're sorting out how this fits into that system.
- When you look at all the potential technologies, there are many things you could do. On the DNA sequencing world, Illumina went and started on sequencing, at first it would cost $1 million, and then single molecule sequencing was supposed to kill Illumina, but we improved what we knew to the point where we were almost killing everyone else. The company displaced the market by improving what is known. We can think about something elegant, but at the end of the day we have something work.
- The problem that everyone has is that whenever you take any individual component, you have a higher-dimensional tradeoff problems that lead to “manhattanization”. Maybe there's some smart people that know each individual components and say yeah I have thought about these x permutations and pruned it down. But even he might not know what's the current state of the art and how you do indexing and the database searching and queries that are clever. But not all of us have the background in computer science to say we could incorporate that into the encoding scheme. How to discriminate that at the DNA level?
- What do you mean by “Manhattanize”? There was a single customer in the Manhattan project. It was the US government. It was a national priority. So if the intelligence community thinks this is an important problem, then they need to create a Manhattan project. And they didn't bet on one solution. They bet on several solutions because it was that important. Now, Dow Chemical said they would charge $1 for their piece. Molecular Assemblies won't' do that. But that's what it's going to take. It is going to require government coordination. AT&T developed technology that was their own monopoly, and their customer was the US government. I agree customers from the government could be huge. National intelligence, NASA could be customers. Start something in the government, put together a proof of concept, and then it's translated out to others. I think there's a market out there. We're not aiming our business plan.
- We have supercompetent people in this room. We have some knowledge to make some initial estimates, and have people check these results. We should put together some numbers, and then engage them. We need pretty high level overview, and how much can you leverage from existing database and search technology?
- Look at DNA sequencing, on the gel, there was like 30 lanes and people started to go for the human genome after that point. If you think about it, it was pretty crazy. Nobody could envision that 15 years later you could do 5 humans per week per machine. You have to start somewhere. You should try to do not a terabyte, but a megabyte, and then you do, like the first computer.
- DM: If you go by who has already shown they can synthesize at scale, then there's a clearly plausible path for getting to the tera scale, that leaves those who are only in commercial operation already. If you open up the scope for funding, to ideas that are much earlier stage of development where a clear proof of concept has not been demonstrated yet, how do you within

phase 1 of the program, achieve a shared set of evaluation of criteria where you can say who is performing well against a goal. One way to do this is not to prescribe you have to synthesize this much material, at this cost and this amount of time. When your synthesis is demonstrated with an encoding framework of your choosing, the cost has been to X, the throughput has to be Y. That establishes the common basis by which to evaluate everyone. You don't have to demonstrate 1 terabyte in day, but an effective throughput where if I am synthesizing 5 oligos in parallel. And make progress towards ambitious goals, and still focus on synthesis. Once you have identified likely winners with that approach, then in the next phase you integrate with sequencing in an end-to-end solution.

- Q: So it has to be a method where you see no bottleneck so far?
- DM: Right, it doesn't have to be demonstrated, but it must absolutely be plausible path to scaling up. So column-based would be thrown out. You should have some exploratory angles. If DNA synthesis is the only point at all, then it would not take advantage of design aspects to try to integrate everything. I think there's a value to re-designing the circuit. We're in the process of re-designing a transistor and then building an integrated circuit. We have a roadmap that we think is - we haven't seen the impossibilities, we haven't seen the uninvented aspects. We don't know the unknown unknowns, we're taking a step back from where molecular biology has been for 35 years and trying to reinvent it.
- The beauty of this approach is that you can eliminate incremental ideas. Like ideas where it doesn't get you where you want to go. There must be a clear roadmap.
- There was discussion about what are the limits of certain technologies. We know what the limits of certain technologies are. We have 14nm-10nm, the limit of transistors. We also know that the limit of DNA is 1 molecule. If you have 1 good molecule, it's going to work. I want to start from 1 molecule. If I can't get 10 molecules, then I want 100 molecules or something. I don't want to start with 10^{13} molecules. I dislike bulk reactions.
- Alternative option, look at DNA writing using natural enzymes. DNA that is naturally available to us, and modify it with alternative methods, that would be an alternative avenue to explore the DNA synthesis aspect. Or oligo library pools.
- We need more access to information theorists, people who do data storage so we can find out their pain points and specs, and as a company of biology majors, it's hard to get access to those people. We're paying a lot of money to consultants. If we could use this group for this, that would great.
- Irene: I am probably your ultimate customer. David will build it, but Irene will use it. Take risks, go beyond the theoretical. I really need something beyond the theoretical. I need something that makes petabytes in this amount of time so that I can to go management and go to Congress and say this is why it's going to help the agencies. We talk to the White House and Congress. There are stakeholders in the government who are not biologists. I talk with many agencies, they are all interested in this technology, but they are waiting for this one person to put some money up front, and going to take the risk and show that this can get where it needs to go.
- NM: Since this is co-sponsored by SRC, I'd like to emphasize the capability that the semiconductor industry has achieved. In integrated circuits, we are at the level where we can have very fine control on a chip with millions or even billions of multiplexing and control of electrical fields and magnetic fields or electromagnetic field or even optical or thermal signals with very fine spatiotemporal resolution, spatial like 1 micron, and temporal like kilohertz definitely and maybe megahertz. This is a very capable platform, not just microfluidics, it's actuation, interfacing with biological entities and biomolecules. There could be something innovative here with how to interface this with the chemistries or enzymes we have, to facilitate DNA synthesis in the future.

- BP: I think I'm an advocate of some better defined granting government support for innovative high-risk research. DNA sequencing had a customer that was the NIH. We developed the DNA sequencer, NIH would fund innovative work. I would like to see clearly defined proposals that companies can write and seek funding.
- Kettner: As much as we try to keep cards to ourselves, it's a big step forward for these orders of magnitude and looking at the trade space and scoping the work. I don't think any one of us could claim to have full control over that. I am hopeful about the progress of DNA synthesis.
- I have two thoughts as this area of development. A consortium could help to develop standards, whether it's just an algorithm for how the data is encoded or stored... in the early days of this technology, there might be multiple platforms that want to interface with computers in the same way, so having early standards would be valuable. My second thought is that I have heard a lot of creative ideas, and I think that we should aim to the density benefits of DNA. If you have to have the DNA stored in low density, it's no longer high density data storage solution. If you can't get it compacted and and address it, then it's not a good approach to solve this problem. It might be a great DNA synthesis solution, but not a great data storage solution.
- How do we scale from where we are to a 1 million fold cost reduction? That's predicated on system architecture decisions, risk management where the early stages are obvious, but the later ones are perhaps extra or further down the pipe but have more risk or something. So laying out that structure from a systems approach, I don't have a recommendation, but I think it's critical for a program like this where you want an end-to-end solution and then you want to focus on synthesis. There are many ways to solve synthesis that are incompatible with a viable solution for data storage.
- NM: We need to have well-defined goals. In 2 years, what do you want? For just DNA synthesis, it's a question of too many ideas and not enough money. To some extent, the limitation we have is the miniaturization, which is now limited by our ability to get access to the tools, e.g. for 10nm feature size, we need millions of dollars to afford the instruments. You probably have to demonstrate a machine as big as this hotel at the beginning, but it's okay, we can go do it. And after that, working at the size of this hotel, and shrink it. Maybe you will find a solution faster than everyone else, but if we want to go small scale on DNA synthesis, we know how to do it, we just never had the chance, the money to get the instrument. Technically I don't think there's a problem.
- We need to laser focus on the big goal which is that we want “to put a man on the moon”. The footprint is victory. The goals for synthesis we discussed today are great, but there's a risk of missing some of the potential breakthrough idea. Maybe you don't need to synthesize the DNA at all, maybe use naturally occurring DNA. Using different kinds of nucleotide, so you can synthesize and make the sequencing earlier. We might have missed the alternatives like libraries, editing, etc.
- Can we just make progress by working on kinetic rates, what are the rates, what are the limits? We really don't know. Maybe, we should incrementally go into the fundamentals rather than the device approach.

Roundtable and Concluding remarks: Let's develop a plan of action.

- DM: Thank you all for coming. This has been extremely valuable for me. The things that would be helpful for me to have in the last hour is what is realistically achievable on the synthesis side in 1 year, 2 years, 3 years, and how much would it cost?
- This is an exciting time for biology. We are fond of the idea of really seeing the full potential for the biorevolution coming to life, and DNA synthesis is a big piece for realizing that potential. The DNA data storage is one component of it. It's enlightening to hear such a good group of people

makes it very clear that there are opportunities and ways to move forward, and there is a lot of ideas. Getting people to think of DNA as a medium to store data is a nontrivial task - the general public has difficulty thinking about biology in the first place. Thinking about goals over the years to come is a good approach. What is the first step that is going to get the most commercial traction possible? It might not be DNA as data storage, but we can get there. Something to get the public energized about DNA synthesis. That's where we want to see it go in the shortest amount of time possible. And get the whole market pushing and innovating beyond what we could imagine.

- SC, NP: We are very grateful to cross the Atlantic and interact with everyone here. As a startup, we are going to try it on a kilobyte scale. One of the biggest concerns is who is the customer. Intelligence community might be the customer. There needs to be an incentive for startups or larger companies. In startups, it's about survival. With a clear incentive, perhaps startups would be more willing to die for the cause. Perhaps the intelligence agency or others would front, I don't know if it's $100 million upfront, I think Google and Facebook and so on will put 10x more once they see an indication that this is going to be on the market.
- What could be the technology that takes DNA synthesis parallel to 3d printing form stacking layers of melted plastic at low resolution into stereolithography where you pull the Eiffel Tower out of a pool of amorphous substrate. If we can do that for DNA synthesis then that would great.
- My say is that first of all thank you all for making this very interesting and productive and useful meeting. Second, homework tasks for me and several of us. We had a lot of ideas, we need numbers. What are the fundamental limits? How can we bypass those limits? We need the numbers. After that, we can think about what we can do about it. This is homework for us, and maybe most of you, asking for help in different tasks. Once again, thank you very much. It's been extremely useful.
- One is that biologists and biochemists don't typically talk with information theorists and how do we facilitate those interactions in a way that is value-building.
- People have talked about building large devices the size of this room that can achieve arbitrary throughput goal. That requires a lot of people, right? What are the practical considerations of going beyond what one grad student can do at the bench to prove out this enzymatic synthesis approach in principle, to actually building a device at some scale? There are a number of you in the room with practical device building experience.
- VZ: I suggest we ask Hua Wang. He was working at Intel as design engineer and now he is putting living cells on silicon.
- HW: As mentioned previously about the current semiconductor technologies, there are a lot of things we could leverage. Just using the technology platforms already available. I think the question is, how much you can achieve in X years from ground zero? Let's say 2 years. From experiences with SRC, 1 year is exploration, and the acceleration of process happens in year 2 and year 3. It can take a minimum of 4 months to get something fabricated from the design phase to having it in hand, maybe 6 months, if there are multiple layers, like post modifications, interfacing with wet systems and so on. If I wanted to design a phase where you had adequate time to design a device, and then multiple times for fabricating it. I think it's probably about half a year, from starting from zero, it's about two plus month phase for design work. And a lot depends on how mature the concept is. And then 3-4 months is fabrication. This also depends on the technology you are doing. For bio-related work, we can use less advanced technologies, so fabrication can be made faster, like 2-3 months. This also depends on scale.
- Q: What if I gave you exclusive access to a fab?

- HW: If you basically schedule a group run in the foundries with high priority, sure. After that, post processing takes about a month, depending on what you're doing. How much are we talking about iterations? Let's say that I wanted a demonstration of some minimum throughput and some effective throughput for a synthesizer. With no constraints placed on how long it would take or the means it would then take to sequence the information. Is it a meaningful thing to say effective throughput of 1 terabyte/day, but the minimum real throughput would be like a megabyte/day. What complexity system to build that? Is 18 months enough? Between 1 megabyte and a terabyte is a substantial leap.
- HW: If I was to speculate about how to do that process, for first generation, I wouldn't want to use CMOS process, I Would use silicon, but fabricate metal on top to create structures that you could do chemistry experiments on. That would be a 6 month first phase, you then use the results to get a design, and the design is 6 months, and that might be your megabyte scale. You're not investing a lot into the complex design systems, just enough to show that it scales. And then you might be able to scale by another factor of a million maybe. I'm just thinking, like 18 months, yeah, 3 six month chunks.
- HW: And you can take advantage of the fact that things can be run in parallel for fabrication. We can use the existing chip as a platform to optimize it. In addition that if chips goes wrong depending on different levels, you can intercept it at different points to do re-fabrication.
- DM: All of this supposes you have a synthesis solution that is known to work, as trip of oligos of a reasonable length, more than 20 mers. So you have to budget in some basic R&D time for those techniques. Realistically, if I were to give you a pile of money tomorrow and say how long would it take you given infinite resources, given you have synthesized a 150 mer say, how long would it take you? Even just one oligo, like 150 mer, using enzymatic synthesis approach, or a simplified chemical synthesis approach, any emerging technique that you're tracking.
- BE: I've been through seven commercial product development cycles. I don't think you can do anything in less than 18 months. But that's always been really close to going into commercial production as opposed to a prototype. I think we could demonstrate enzymatic synthesis not for synthetic biology applications, because I consider the DNA storage application to have much less constraints. I'd like to think somewhere between 6 to 9 months. I'm not sure if that includes the hardware. But maybe at some low parallelism. Is that singleplex, or multiplex? Well, low multiplex. Like a 6-well, or something. First time you do something, you don't optimize for volume necessarily. I think, that's reasonable.
- And if you're trying to demonstrate proof of concept for a single-plex synthesis, you can probably manufacture the device the single well in which you would do that, in house, you might not do that at a fab, Photolithography mail order, you can also focused ion beams that do 5 nm etchings immediately, put in your silicon, pump down, blip the blame and you got a 5 nm spot.
- Many universities have access to these fabrication facilities. The problem is the time and the cost and the eventual scalability. For the small scale scientific research, you don’t need to be concerned about the manufacturing tools. You have to know about manufacturability to know that the processes you're using will transfer over. And depending on the chemistries, the primary issue is secondary modifications you have to do. In phosphoramidite chemistry, talking about organic reagents they are probably fairly compatible with semiconductor surfaces. PacBio had a lot of surface issues with enzymology on silicon. I think Twist had to do quite a bit of that too for a biosystem. I would worry about doing the hardware in parallel because I don't know what the secondary treatment requirements are going to be required, when you mix enzymes on a surface with silicon, that you don't anticipate. So it might be easier to think about that from a phosphoramidite perspective.

- HW: How many more months would you need, of the 6 to 9 months. There's the biochemistry issue and then the material science. I would probably do in parallel a compatibility study to figure out the processes that you need. Before even starting to think about feature size, scale etc. Because there could be some substantial secondary treatment process. Silicon etches very well. If you start growing cells on a microelectrode array, the cellular effluent could start to etch the chip. We did some study on the biocompatibility of semiconductor materials with cellular systems. While we did not cover molecules and enzymes, this is actually, surface chemistry is always a big problem. When we talk about interfacing semiconductor materials with the biology world, there's a lot of literature talking about this. Regarding DNA synthesis, there are many papers talking about how to maintain the functionality of the DNA polymerase when they are immobilized on a semiconductor surface, and people know those details. It is challenging, but it is doable. There is a lot of background materials on these topics.
- DM: Let's talk about resource requirements in terms of FTEs and equipment or reagent requirements. Say, I gave you 24 months for the full spectrum of techniques from least mature to most mature in 6 months, you develop the concept, and then 18 months to go make it work with microfabricated devices. What resource requirements are we talking about? Do we need enormous teams?
- VZ: Very good researchers are too busy already. They need a threshold, they are doing many things, and we need to set up incentives that make them put aside some other things and push this.
- HW: Assuming a big stack of money, and put it on the table, is it going to take 6 months to hire the staff for this sort of staff. Is everyone already in place? Finding the right people could take many months. Building a team from scratch, 6 months is short. Select the teams already with the capabilities or maybe encourage certain teams to work together and have complementary expertise.
- DM: Here's an idea: When IARPA announces a program, we use a broad agency announcement that lays out what we're looking for, what the deliverables are, what the proposals should look like. It's a variable time between then and when proposals would be due. We could have a proposers day which would be virtual, which might build awareness but wouldn't necessarily contribute to teaming.
- I think the total team would consist of 20 or 30 people. Electrical engineer, system engineer, analog engineer, CAD person, fluids dynamics computation person, photonics person, and then someone who can analyze the design, if we simulate a fluid traveling down here with some big molecule, what's going to happen, is it going to shear off when it rounds a corner, is the field strength of your electrophoresis going to be strong enough to move things around.
- VZ: Our goal here is to draw a roadmap for DNA synthesis and it will be part of the bigger semi synbio roadmap. So once again, it was a great pleasure, thank you very much. I hope to see many of you soon and I hope to communicate with all of you.

A2. Numerical estimates on DNA storage capabilities

Units

The DNA memory size is typically measured in grams (g), picograms (pg), number of nucleotides (nt), bases (b), or base-pairs (bp) with unit conversions of: 1 pg = 978 Mbase or 1 Mbase = 1.02×10^{-15} g for double-stranded DNA (ds-DNA) or 1 pg = ~1.9 Gbase for single-stranded DNA (ss-DNA) or RNA.[8] For conversion to binary units (bits), the information in megabases is multiplied by a factor of 2 (quaternary to binary conversion), e.g., 1 Mbase = 2 Mbits. Each of the 2 bits of equivalent binary information is

approximately 0.34 nm of length along a DNA "tape." The resulting weight of a DNA bit is m_{DNA} = 5.1´10^{-22} g/bit, which is 10 orders of magnitude less than the weight of a flash bit.

Volumetric storage densities

Double-stranded DNA iis 2nm wide and 3.4nm long for 10 basepairs. Assuming a *dense crystal* packing, this could yield a density of

$$(10 \text{ bp} \times 2\text{bits/bp})/(2\text{nm}\times 2\text{nm}\times 3.4\text{nm}) \sim 1.47 \text{ bits/nm}^3$$

Thus, theoretically the same amount of ds-DNA could be stored with a memory density of ~10^{21} bit/cm^3 when stored as a crystal.

Since in information reading and writing processes other components are needed, it can be speculated that the practical density of DNA memory would be close to that of living cells. Additionally, each storage node needs to be connected to other components to be functional – reducing the practical density of DNA memory.

The next level of bio-organization is the nucleosome, which contains about 146 bp of "data" + 80 bp of linker. DNA wrapped around a multi-protein core in about 11nm×11nm×5.5nm cube, corresponding to ~0.679 bits/nm^3.

The smallest known eukaryotic organism is unicellular green alga *Ostreococcus tauri (O. tauri)* with the mean cell length and with 970 and 700nm respectively[16] (the volume is 2.5×10^8nm^3=2.5×10^{-13}cm^3). The cell contains not only 12.6 Mbp[17] (25.2Mbit) but all the machinery to reproduce and harvest energy. The characteristic storage density is ~0.1 bits/ nm^3 or 10^{20} bit/cm^3.

Next, human white blood lymphocyte (a part of the immune cell) has the characteristic storage density of 0.046 bits/ nm^3 (10^{19} bit/cm^3).

To compare the above numbers with a current Solid State Disks, consider a packaged SSD that stores 1TB in 3 cubic inches:

$$0.33 \quad \text{B/in}^3 = 2.26 \text{ GB/cm}^3 = 18.1 \text{ Gbit/cm}^3 = 1.81\times 10^{-11} \text{ bits/nm}^3$$

Endurance

Experimental studies of ancient DNA have revealed that the DNA half-life near room temperature is 521 years[18]. Theoretical calculations also confirmed that DNA retention time can very large e.g. in comparison with memory (Fig. A1) [4]. It was emphasized in [4] that a drastic improvement in retention time occurs by changing the environment from wet to dry and that the effect of water on retention is greater than temperature

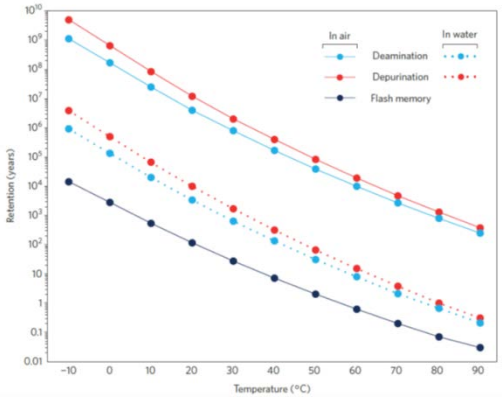


Fig. A1. Calculated memory-retention times for DNA flash memory [4].

References

[1] http://2014.igem.org/Team:Cooper_Union

[2] M. Hilbert and P. Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information", *Science* 332 (2011) 60

[3] Rebooting the IT Revolution: A Call for Action. SIA-SRC Report (2015): https://www.nsf.gov/crssprgm/nano/reports/2015-0901_RITR%20WEB%20version%20FINAL_39p.pdf

[4] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, W. L. Hughes, "Nucleic acid memory", *Nature Materials* 15 (2016) 366

[5] S. Overballe-Petersen, "Bacterial natural transformation by highly fragmented and damaged DNA", *Proc. Natl. Acad. Sci.* 110 (2013) 19860.

[6] H. K. E. Landenmark, D. H. Forgan, C. S. Cockell "An estimate of the total DNA in the biosphere", *PLoS Biol* 13 (2015) e1002168.

[7] SRC/IARPA Workshop on DNA-based Massive Information Storage, April 27 & 28, 2016, Arlington, VA. Workshop Summary: https://www.src.org/program/grc/semisynbio/semisynbio-consortium-roadmap/

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

[9] 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

[10] S. L. Shipman, J. Nivala, J. D. Macklis, G. M. Church, "CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria", *Nature* 547 (2017) 345

[11] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture", Science 355 (2017) 950

[12] J. Bornholt, L. Ceze, R. Lopez, G. Seelig, D. M. Carmean, K. Strauss, "A DNA-Based Archival Storage System," OPERATING SYSTEMS REVIEW 50 (2017) 637-649

[13] www.synthesis.cc/synthesis/

[14] https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1984/

[15] S. Kosuri and G. M Church, "Large-scale de novo DNA synthesis: technologies and applications", Nature Methods 11 (2014) 499-507

[16] C. Courties, et al., "Smallest Eukaryotic Organism", Nature 370 (1994) 255

[17] E. Derelle et al., "Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features", Proc. Natl. Acad. Sci. 103 (2006): 11647–11652.

[18] M. E. Allentoft et al., "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils", Proc. Royal Soc. B 279 (2012) 4724-4733