AUTO, SECURITY & PERVASIVE COMPUTING (/CATEGORY-MAIN-PAGE-IOT-SECURITY/)

DNA Edges Forward As Data Storage Option

Ability to store huge quantities of data is possible, but getting the cost down is challenging.

SEPTEMBER 28TH, 2022 - BY: **BERNADETTE TANSEY (HTTPS://SEMIENGINEERING.COM/AUTHOR/BERNADETTE-TANSEY/)**

At technology conferences back in 2015, scientist David Markowitz raised the idea that DNA could be adapted as a data storage material. The audience response wasn't all he had hoped for.

"They would laugh me off the podium," Markowitz recalls, but without rancor. Facing skepticism comes with his job at IARPA, the research arm of the U.S. intelligence community. The agency anticipates future challenges to national security, and explores futuristic solutions.

IARPA, short for Intelligence Advanced Research Projects Activity, takes risks on "approaches that seem barely plausible," Markowitz says.

In 2014, Markowitz had foreseen a big problem, not only for intelligence agencies such as the CIA, FBI, and National Security Agency, but also for other government departments, and even big tech companies. Many were generating exabytes worth of data, but the supply of data storage on that scale was becoming too scarce or too costly for government units, whose budgets could not expand exponentially to accommodate all the valuable information they were collecting.

DNA, nature's nanoscale information coding material, was a possible answer. In 2016, Markowitz designed a program to test the feasibility of DNA data storage as an alternative to conventional media such as hard drives and magnetic tape. Various research teams had already "written" binary code into synthesized DNA sequences.

Meanwhile, industry analysts had seen the global total of data creation break into the zettabyte scale and gain momentum, amid concerns that improvements in data density for conventional storage media were slowing, and that manufacturing output might not keep up with demand.

DNA could offer unparalleled data density. One cubic inch of dried DNA could contain 11.2 exabytes of data, the U.S. Government Accountability Office said in a recent report. That's nearly 2 billion times the capacity of a 5.7 gigabyte DVD, and enough room to store the contents of multiple enterprise data centers packed with hard drives.

But the GAO, guardian of U.S. government agency budgets, also estimated the current price tag of writing and reading

DNA-based data at $3,500 per megabyte, which is “millions of times more than silicon-based storage.”

Markowitz wanted to propel DNA data storage technology toward an exponential drop in cost, to de-risk it so that private companies would invest their own money to refine and commercialize it. The Molecular Information Storage (MIST) program he heads has funded contracts to support collaborations among biotechnology and semiconductor companies, academics, and government units with the many skills needed to advance the technology.

Both startups and well-established companies are working to shave costs significantly for two core processes — the synthesis of DNA strands to represent binary code, and the later sequencing of the DNA so it can be translated back into binary language.

Companies such as Twist Bioscience have stored hundreds of megabytes of data in DNA files by synthesizing custom sequences of DNA subunits, called nucleotide bases (or simply, bases). Each sub-unit carries one of four different bases — adenine, thymine, cytosine, or guanine.

Under one data encoding method, the base adenine represents 00, cytosine stands for 01, guanine for 10, and thymine for 11. A synthesized DNA strand reading ATTGC would translate back as the binary code 0011111001.

But a wide variety of coding schemes have been explored. For example, two young DNA synthesis companies, Iridia and Catalog, design DNA sequences of varying lengths that can each be assigned to represent a chosen row of binary digits. In DNA’s natural coding system, a series of three bases, such as AGA, leads to the addition of a specific amino acid, such as arginine, to a protein that’s being constructed within living cells.

Uses on the near and far horizons

The most likely near-term use for DNA is cold archival storage for data that needs to be read only rarely, such as movies that must be preserved for decades. Those archival backup copies represent a substantial share of the storage market, according to a recent analyst’s report co-sponsored by Twist.

Some companies working on the technology are already envisioning racks of DNA storage devices in data centers within a decade or so. Early experiments on methods of computing within DNA files are also under way.

“Every major player in the storage space is looking at this,” says Iridia CEO Murali Prahalad. “The view of this as pure Star Trek versus a reasonable possible achievement has started to pivot.”

Markowitz says the optimism among developers is “very realistic.” The MIST program won’t wrap up for at least two years, and it doesn’t release interim progress reports. But Markowitz sees progress on several fronts. For one thing, he has noted an uptick among companies spending their own R&D dollars on DNA data storage.

One of those companies is Twist, which synthesizes millions of customized DNA strands for biomedical company research and other industrial initiatives.

Twist co-founder and CEO Emily Leproust says she had been eager to focus on DNA data storage when the company first launched in 2013. But an investor told her the timing was 10 years too early. “He was right,” Leproust says. “So we invested a little bit every year.”

But this fiscal year, Twist plans to spend about $40 million — almost a third of its $130 million R&D budget — on DNA-based data storage.

If the timing is now favorable for this novel technology, it’s due in part to the decades of progress made by companies that have already industrialized nature’s toolkit of nanoscale biomolecules that organize and operate cells.

After academic scientists learned how DNA could be copied, cut, and pasted by naturally occurring enzymes, entrepreneurs bioengineered the genomes of microorganisms to mass- produce drugs in the early 1980s. They later applied such techniques to speed up DNA sequencing. The race to complete the first draft of the human genome, beginning in 1990, stimulated improvements in the sequencing process through automation, computer control and analysis, and multiple parallel assays in arrays of miniaturized wells.

As biomedical research teams mine the resulting flood of genomic information to develop new treatments, they soon could outsource chores like DNA sequencing and DNA synthesis to a rising group of companies that specialize in performing those techniques efficiently, at scale. DNA synthesis companies provide custom libraries of short DNA strands as test beds for drug research, but they also help clients cobble together synthetic genes, which can be incorporated into microorganisms designed to manufacture products like fuels and fragrances.

DNA data storage technology builds on the advances of those companies, whose tweaks can include tightening the interfaces between chip electronics and biological molecules undergoing operations in microfluidic chambers.

Twist synthesizes millions of DNA strands on silicon wafers integrated with CMOS chips. Twist's first chip held 1 million DNA synthesis sites, but the company is working on expansion to 256 million sites, and later to as many as 50 billion pieces of DNA. Scaling up is expected to lower costs for DNA data storage.

The emerging field has created a new market for novel chip designs and uses. Although DNA itself isn't a silicon-based storage material, DNA data storage systems aren't immune from current stresses on semiconductor supply chains, Markowitz says. The recent passage of the CHIPS and Science Act, which bolsters U.S. domestic semiconductor manufacturing, could help the cluster of DNA data storage developers based in the United States, he says.

Sequencing moves ahead

Sequencing giant Illumina, founded in 1998, also incorporates low-cost CMOS chips in its machines. In the early 2000s, the cost of sequencing an individual human genome was about $100,000. Illumina and other companies later reduced that price to $1,000 using next-generation technologies.

Markowitz says an Illumina CTO was among the first industry leaders to recognize the possible financial payoff of developing DNA-based data storage systems during an early IARPA workshop in 2016. Though Illumina's revenue topped $2.3 billion that year, the CTO said data storage at enterprise scale "blows our current market out of the water," Markowitz recalls.

Illumina's current sequencing method relies on the enzyme DNA polymerase, an all-around industrial tool that can make multiple copies of a DNA strand. To discover the sequence of a single strand of DNA, Illumina's optical scanners track the enzyme as it moves along that template strand and creates a new, complementary strand. It adds bases tagged with fluorescent dyes in various colors to differentiate them. If the template sequence is ATTGCA, the enzyme would synthesize the complementary strand TAACGT. DNA polymerase follows simple rules — A pairs with T, and G pairs with C.

As each base joins the new DNA strand, scanners detect its particular colored light. When the full sequence of the complementary strand is known, the sequence of the template strand can be deduced.

Illumina has invested in startups developing DNA-based data storage systems, and is among the companies working toward the next milestone, a human genome sequence costing $100.

The startup Ultima Genomics said in May that it had hit the $100 target, and later unveiled a collaboration with NVIDIA to use its GPUs for AI-enhanced DNA sequence analysis.

Another startup, Roswell Biotechnologies, unveiled a “molecular electronics chip” this year that is designed to significantly miniaturize the process of “watching” the enzyme DNA polymerase build a new DNA strand based on a template strand whose sequence is to be discovered.

Roswell’s chip uses a single DNA polymerase molecule as a sensor by integrating it into a semiconductor’s circuitry. The enzyme is attached to a molecular wire connected to nanoelectrodes at either end. An electrical current flows through the wire. As the enzyme builds a new DNA strand by adding a particular DNA base, resistance to the current changes in signal ways, identifying the base. The process repeats with each base added.

Millions of these nanoscale sensors could fit onto a single CMOS chip, making it possible to sequence a human genome in an hour for less than $100, says Barry Merriman, Roswell’s chief scientific officer. That technology could equip small point-of-care devices to report diagnoses based on quick genetic analysis. But the Roswell chip also could scale up further, dropping the cost of sequencing to $10 per human genome. That improvement also could lower the cost of reading DNA that encodes binary data, Merriman says.

Markowitz says he is now more confident that continued progress among sequencing companies will help DNA data storage systems in their bid to become more competitive. He’s also enthusiastic about the formation of an industry group that, like MIST, aims to catalyze progress in the field. Twist and Illumina joined with Microsoft and Western Digital in late 2020 to form the DNA Data Storage Alliance. The fledgling trade group, which fosters networking and product development partnerships, has now signed up 50 members, including imec and Dell Technologies.

Microsoft, operator of the Azure data storage service, is funding basic research on DNA data storage technologies through a partnership with the University of Washington. Imec also is supporting work in the DNA-based data storage field. Imec recently published its own roadmap to improve DNA sequencing through advanced semiconductor design.

The DNA Data Storage Alliance is publicizing its case that DNA, a substance that can be degraded by exposure to light, moisture, or micro-organisms, will nevertheless succeed as a storage medium. Under the right conditions, DNA can remain intact and readable for hundreds to thousands of years, the group says.

Synthesized, freeze-dried DNA can be embedded in glass inside metal vials topped up with an inert gas; it is suspended in fluid for sequencing, copying, and other operations. The alliance maintains that DNA-encoded data could become a cost-competitive option because it would not need to be transferred to a new storage medium every three to 10 years. With conventional storage, these file transfers are done routinely to avoid loss from the deterioration of materials, or because reading devices may no longer exist for the original storage vehicle.

DNA can be stored at room temperature, advocates say, so it could reduce energy demand compared with huge data centers that must be kept powered up and cooled down. While the initial cost of encoding data in DNA may remain higher than writing data into conventional media, the alliance argues, DNA could avoid decades of maintenance costs for long-term storage.

DNA may deliver other savings. Once a DNA molecule is synthesized to encode binary data, millions of copies can be made cheaply in a massively parallel process that relies on DNA polymerase. That process, called PCR or polymerase chain reaction, is behind the current sensitive tests for infection with the Covid virus.

The Synthesis challenge

The MIST program has laid out a multi-factor stretch goal for early DNA-based data storage systems. That milestone would be a system that can “write 1 terabyte, and read 10 terabytes with random access capabilities, per day, at a cost of less than $1,000, while consuming less than 1 kilowatt of power, and fitting on a table-top.”

Reducing DNA synthesis costs remains the toughest challenge, Markowitz says, while the need for hyperscale data

storage supply “has only become more urgent in the last six years.” In Twist’s current synthesis process, starter DNA strands are tethered to rows of sites on Twist’s silicon wafer, and bases are added to each DNA sequence one by one in an inkjet-like process orchestrated by the company’s proprietary software.

The system can simultaneously build a variety of different short DNA sequences, called oligonucleotides, each comprising 200 to 300 bases. Every strand includes an encoded “bar code” to identify it, so that all the short pieces of information can be assembled in the right order to make up the data file once the code is read.

Twist’s current synthesis method, phosphoramidite chemistry, suspends the DNA in solvents derived from fossil fuels. Some of the ingredients are toxic and explosive. As an alternative, Twist is also developing an enzymatic method of creating customized DNA strands, as have other companies such as DNA Script, Evontix, and Anza Biotechnologies. Enzymatic synthesis has the potential to create longer strands of DNA than chemical synthesis.

Twist wants the option to offer on-premise DNA synthesis to customers who may closely guard their data. Enzymatic synthesis in watery solutions would avoid bringing hazardous materials into customer data facilities. Twist may sell DNA-based data encoding instrumentation that enterprises could operate themselves. Or, Twist could offer to operate its system for the client on-site, says Steffen Hellmold, Twist’s business development executive for data storage. Hellmold estimates that Twist could market DNA-based data storage as a service, at a premium above commodity storage pricing, if the company could offer it at $25 to $50 per terabyte, or $50 to $100 at the high end.

In the future, Twist hopes to synthesize multiple petabytes of DNA-encoded data per month from data center racks, says Hellmold. He estimates that by 2030, the company could benefit from a gap between storage demand and the supply of conventional storage. DNA-based systems could capture one or two zettabytes of that unmet demand, he predicts.

For now, Twist is looking for earlier commercial opportunities. In the near future, the company plans to market a “Century Archive,” a DNA-encoded data set of about one gigabyte, written in one pass on a single semiconductor chip, Hellmold says. “That will get us off the ground.”

Customers could use these for backup copies of, say, the blueprints for a bridge or a power plant, Hellmold says. Health care systems required to preserve records such as X-rays for decades could store them in a tube of DNA tucked in a safe deposit box, he says.

But the ultimate aim for Twist is to provide enterprise hyperscale data storage for high-volume uses, such as security and surveillance information, legal archives, and data gleaned from sensors on driverless cars.

Research teams already are working toward future DNA-based systems with elements of an operating system. Random access to read only certain sections of a DNA file, which are mixed together with other DNA strands in a solution, can be achieved using a classic lab technique. The chosen DNA strands can first be tagged by stirring in short DNA probes designed to stick to a unique DNA sequence on the target strands, such as their “bar code.” Those DNA probes, called primers, spur DNA polymerase to latch on to the strands of interest and make multiple copies of them to be sequenced.

Twist CEO Leproust foresees a data center in a tube, with enzymes performing functions such as search, copy, and paste within the fluid memory space of DNA-encoded files. “We are witnessing the next evolution of the storage market,” she says. “We’re just scratching the surface.”

MIST works closely with both Twist and DNA Script, Markowitz says. The program’s goal isn’t to pick a sole winner, but to help build a robust ecosystem of competing providers. “That’s better for the government, for taxpayers, and for national security,” he says.

Although no DNA-based system yet exists to serve data storage needs at enterprise scale, companies are rolling out intermediate or related products. As one example, Markowitz points to DNA Script's line of benchtop DNA synthesis machines, now marketed for biological research.

"In the next couple of years, I expect a proliferation of products that will make DNA data storage accessible to customers who want to explore the value it could have in solving their problems," Markowitz says.

> **Related Reading**
> Los Alamos National Laboratory has developed a key technology that could one day pave the way towards DNA storage (https://semiengineering.com/manufacturing-bits-april-13/)
> DNA storage for TV shows; DNA error correction. (https://semiengineering.com/manufacturing-bits-aug-25/)

TAGS: ANZA BIOTECHNOLOGIES (HTTPS://SEMIENGINEERING.COM/TAG/ANZA-BIOTECHNOLOGIES/) CATALOG (HTTPS://SEMIENGINEERING.COM/TAG/CATALOG/) DELL TECHNOLOGIES (HTTPS://SEMIENGINEERING.COM/TAG/DELL-TECHNOLOGIES/) DNA DATA STORAGE ALLIANCE (HTTPS://SEMIENGINEERING.COM/TAG/DNA-DATA-STORAGE-ALLIANCE/) DNA SCRIPT (HTTPS://SEMIENGINEERING.COM/TAG/DNA-SCRIPT/) DNA STORAGE (HTTPS://SEMIENGINEERING.COM/TAG/DNA-STORAGE/) EVONTIX (HTTPS://SEMIENGINEERING.COM/TAG/EVONTIX/) IARPA (HTTPS://SEMIENGINEERING.COM/TAG/IARPA/) ILLUMINA (HTTPS://SEMIENGINEERING.COM/TAG/ILLUMINA/) IMEC (HTTPS://SEMIENGINEERING.COM/TAG/IMEC/) IRIDIA (HTTPS://SEMIENGINEERING.COM/TAG/IRIDIA/) MICROSOFT (HTTPS://SEMIENGINEERING.COM/TAG/MICROSOFT/) NVIDIA (HTTPS://SEMIENGINEERING.COM/TAG/NVIDIA/) ROSWELL BIOTECHNOLOGIES (HTTPS://SEMIENGINEERING.COM/TAG/ROSWELL-BIOTECHNOLOGIES/) TWIST BIOSCIENCE (HTTPS://SEMIENGINEERING.COM/TAG/TWIST-BIOSCIENCE/) ULTIMA GENOMICS (HTTPS://SEMIENGINEERING.COM/TAG/ULTIMA-GENOMICS/) UNIVERSITY OF WASHINGTON (HTTPS://SEMIENGINEERING.COM/TAG/UNIVERSITY-OF-WASHINGTON/) WESTERN DIGITAL (HTTPS://SEMIENGINEERING.COM/TAG/WESTERN-DIGITAL/)

Bernadette Tansey **(all posts) (https://semiengineering.com/author/bernadette-tansey/)**

Bernadette Tansey is a contributing writer at Semiconductor Engineering.