DNA-Based Digital Storage

ABSTRACT

Digital data generation is increasing exponentially, and estimates suggest that the amount of new data being generated each year has already surpassed the data storage capacity of current technologies. DNA presents an alternative technology for the storage of digital information due to several attractive properties. DNA is an extremely stable molecule, it takes up very little physical space, and since it makes up the human genome, the technology needed to read it back will never become obsolete. We discuss the practical and technical issues being addressed to further develop DNA synthesis technologies, enabling use of DNA as a digital storage medium. We also offer a perspective on opportunities at the intersection of the synthetic biology and information technology industries, highlighting how the convergence of research in these areas will accelerate discovery in both.

BIOLOGY'S DIGITAL MEDIA

Modern digital computers store, communicate, and operate on binary data, represented as ones and zeroes. These bits of information are associated with physical structures and signals, such as the electronic state of transistors or field orientation of magnetic materials. Nature also stores digital information, such as the genetic code in your cells, in the form of molecular polymers. In DNA, these polymers are built from a set of four small molecules known as nucleotides, or *bases* for short (Fig. 1). Instead of just two values available with binary data (one or zero), each base position in DNA can take one of four values (A, C, G, or T, representing the chemical name of the base), so each base is essentially the information equivalent of two bits (Fig. 2).

Each typical human cell contains a genome consisting of about 6 billion base pairs of double helical DNA, organized into 23 sets of chromosomes (3 billion base pairs of DNA correspond to the chromosomes in each half of the set). The DNA in these chromosomes encode about 1.6 gigabytes of information in total, per cell. Adding up all the

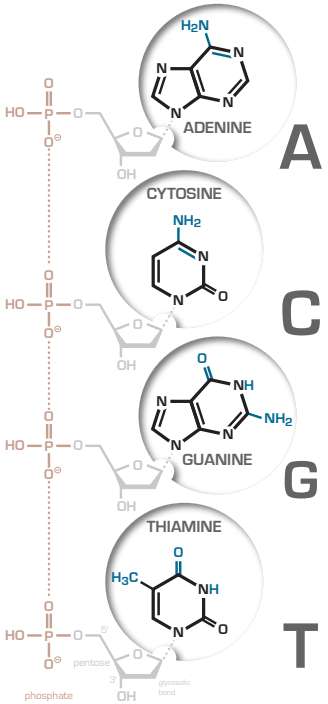

Figure 1. Chemical structure of DNA bases. Each base can encode two bits of digital data.

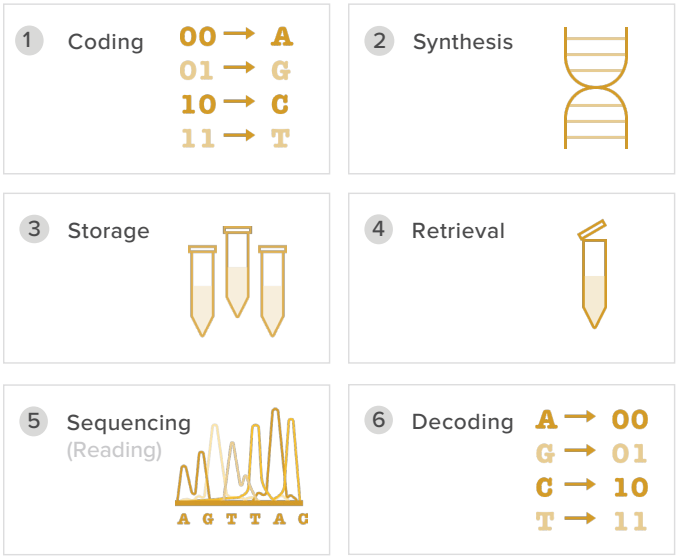

Figure 2. Steps involved in the encoding and decoding processes with DNA-based digital data storage.

cells in our body (Sender ***et al.***, 2016), this translates to approximately 100 zettabytes of information stored in our DNA—far more digital data than all humans combined have ever generated.

One of the most compelling characteristics of DNA as a digital storage technology is its theoretical data storage density potential, and therefore storage capacity. In the physical limit, two bits of data stored in DNA require at most 28 atoms, which includes both the structural backbone of DNA and the information-encoding base. At ≤14 atoms per bit for the entire medium, no other technology comes close to being able to store information so densely. In addition, as a *molecular* medium, DNA-based digital storage can be implemented in three dimensions—as a volume, rather than as a two dimensional surface like a disk, tape, or chip, which translates to DNA taking up much less physical space than other media.

DNA is also extremely stable, when stored properly (Zhirnov ***et al.***, 2016). In contrast to current digital data archival technologies which degrade over years and need to be replaced approximately every decade, DNA can preserve information for centuries, or even thousands of years. Even if DNA degradation does occur, redundancy and other error-correction schemes can compensate.

Finally, copying DNA is cheap and fast. Just as each of your cells has its own copy of your genome, which is replicated every time a cell divides, DNA information can be copied in minutes using common and well-established molecular biology procedures, with the resulting copies small enough to physically transport by postal service.

In order to make DNA data storage technology a reality, there are three major physical components: DNA writing, storage/retrieval, and reading. This white paper addresses the current states and future developments for each of these areas, along with trends in data growth and technology development.

History of DNA Data Storage

The idea of using DNA to store digital data was first proposed in the mid-1990s (Baum, 1995), with the initial proof-of-concept experiments demonstrating that information could be stored in DNA published at the end of that decade (Clelland ***et al.***, 1999). Research in this area continued over the next several years, but storing relatively large amounts of digital data in DNA was not achieved until over a decade later (Church ***et al.***, 2012 and Goldman ***et al.***, 2013) due to limitations in DNA synthesis and DNA sequencing technologies. These studies renewed interest in storing increasing amounts of information in DNA. A collaboration by researchers at the University of Washington, Microsoft, and Twist Bioscience has recently encoded 200 Mb of information in DNA, and retrieved this data with 100% accuracy, which is believed to be the largest DNA-based storage project to date.

With the proof-of-concept experiments showing that DNA is an effective medium for archival storage of digital data, research and development have now turned to improving the throughput of DNA synthesis technologies to make the production of synthetic DNA more economical.

Additional lines of research are focusing on applying improved encoding algorithms to the DNA storage strategies to increase the practical data storage density, to correct errors in the DNA synthesis or sequencing steps, and to compensate for the potential of DNA degradation.

TRENDS

Data Growth

Prior to focusing on DNA specifically, we'll first establish a general framework for digital storage. Starting in about 2005, we crossed an analog/digital threshold: more information was recorded and preserved using digital technologies (i.e., computing) than with analog (e.g., audio tapes, books, film). In addition, the growth rate of digital data generation is rapid, doubling every few years, such that we now regularly produce more information per time period than all previous periods cumulatively. By some estimates, beginning around 2010 (Fontana and Decad, 2016), a gap began to open between storage capacity and data generation (Fig. 3). That is, the total capacity of data storage technology manufactured was less than the amount of data being produced. By the year 2025, conservative estimates suggest that traditional data storage technologies will be able to store less than half of the digital data being generated.

Simply increasing the production of traditional silicon-based digital data storage media is not a sustainable solution. The Semiconductor Research Corporation has projected digital data storage demand to exceed worldwide silicon supply by 2–3 orders of magnitude by 2040 (Zhirnov ***et al.***, 2016). So the question is not whether new data storage technologies are needed, but rather, when.

Biotech & Computing

One of the most widely publicized success stories in the biotechnology industry is the development and advancement of DNA sequencing technology. In a couple of decades, the cost of sequencing (*reading* the DNA sequences) a human genome has dropped from about $3 billion to under $1000, and this trajectory puts the cost of reading DNA to cross under costs associated with accessing data archived in tape media in the near future, without targeted development for the DNA-based data storage application. Not only has DNA sequencing technology had a profound impact on life sciences, but the

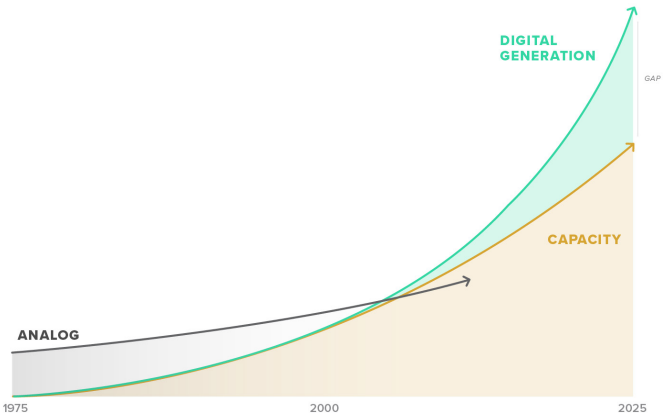

Figure 3. Estimated trends of analog and digital data generation and data storage capacity. More data started being stored in digital format that analog in 2005, and starting in 2010 the amount of storage capacity being manufactured was estimated to be less than the amount of data generated.

growth in sequencing data is itself a very large part of the overall trend of growth in digital data that must be stored (Stephens ***et al.***, 2015) —biologically-derived digital data is already having an impact on information technology.

DNA synthesis technologies (*writing* DNA sequences) are improving, but historically at a slower rate than sequencing. The reasons for this are straightforward. The widespread availability of sequence data is relatively new, as is the associated science for analyzing and understanding genomes. Therefore, there has been a lag between discovery and application, but new DNA synthesis technologies are changing this dynamic and opening up new markets for synthetic DNA in the process. Unlike sequencing, the application of technologies from the information technology industry (semiconductors, computing) to DNA synthesis has been limited. This is where Twist Bioscience's expertise in using silicon as a substrate for DNA synthesis begins to exploit the potential of these technologies.

Technologies borrowed from or inspired by biology (e.g., polymerase-based technologies) or from technologies developed for biology (e.g., DNA sequencing) can be greatly improved with the assistance of previous research and development in the information technology industry.

Neither DNA synthesis nor sequencing has previously had targeted development motivated by an IT application (such as digital storage), nor significant investment in the kind of system integration and high volume manufacturing processes common in the electronics industry. By developing these technologies specifically for biotech, the trend lines for cost and performance can be driven to surpass those of conventional technologies. In the process, the development of tools for synthetic biology will also be accelerated. Partnerships between the biotechnology and the computing/electronics industries are essential to realizing these technologies.

Twist Bioscience has already developed a novel platform for manufacturing synthetic DNA on a massively parallel scale. Instead of synthesizing DNA on previous generation solid substrates such as plastic or glass bead reactors, Twist's technology uses custom fabricated silicon wafers and synthesizes millions of unique oligonucleotide sequences in each synthesis run, with improved synthesis error rates and sequence uniformity over previous methods. Future generations of DNA synthesis technology are currently being developed specifically for the DNA-based digital storage applications to produce DNA on an even higher scale of throughput.

With sufficient continued focus on development, it is possible for DNA-based technologies to intercept and surpass the economics and capacities of existing electronic-based digital data storage technologies.

Boundaries

In terms of the medium itself, a theoretical lower bound cost per byte for DNA-based data storage can be derived from the price of reagents (chemicals) needed to add a new base to a DNA molecule. Without consideration for decreasing costs of the reagents themselves, we estimate material costs of less than a fraction of a penny per gigabyte to encode and recover stored information. This puts the economic potential of DNA-based storage well beyond projections for existing technologies, and with enough runway to support several cycles of technology improvements.

Likewise, the upper bound for performance is determined by the speed of adding a single base. Using examples from biology, many DNA polymerases, the enzymes which replicate DNA, synthesize around 1–10 kilobases per second. Semiconductor fabrication capabilities to drive toward molecular-sized reaction surfaces permits estimates of scaling synthesis reactions densely across a control surface, producing millions to billions of DNA synthesis sites per square centimeter. We see the potential for DNA-based storage to exceed bandwidths for writing disk and tape, but with dramatically smaller physical space and energy requirements and greatly increased stability. This also means that the impact to datacenter footprints scales with bandwidth requirements, rather than storage capacity, as overall capacity grows.

Both the theoretical cost of storage and retrieval of information stored in DNA, as well as the DNA synthesis throughput required to support use of DNA as a storage media are both achievable to make DNA-based data storage an economically feasible reality.

Progression

From an economic perspective, the value of density, capacity, and longevity provided by DNA data storage open potential markets in long-term, "cold" archival storage where effectively permanent preservation of information is critical. The ability to rapidly replicate large data archives and geographically distribute them extends the value of using DNA to replace traditional electronic data storage technologies. As the cost and capacity of DNA synthesis improve, DNA storage becomes increasingly competitive with existing technologies. As a second generation of technology, we anticipate that DNA storage systems in the short term will exceed the performance and cost characteristics of tape while retaining the advantages specific to DNA-based technology. From there, we expect another generation of technology to be mature enough to approach hard drive characteristics while maintaining the previous trajectory in cost advantage. Finally, as noted earlier, since the medium itself is not constrained to the

scale of the DNA read/write devices, physical space and energy required for storage remains small even as the technology for read and write evolves. This is unique when compared to technologies used in current of data centers.

Each major step will be aligned to DNA synthesis technology development. A series of technical challenges are being addressed by Twist Bioscience in sequence to reduce the time required to achieve each incremental step, as well as to reduce the cost associated with each step in development.

FUTURE IMPACT

Synthetic biology is an emerging industry, drawing comparisons to the industrial revolution, and is anticipated to have a greater impact in the coming century than computing had in the previous one. Bringing information technology and biotechnology together—building new biotech instrumentation and manufacturing tools by adapting semiconductors and electronics for biological applications—will serve to accelerate this revolutionary transition. The same technologies needed for DNA-based digital storage are directly adaptable to biological gene synthesis, and DNA synthesis more generally.

Motivating IT to contribute to technology development for computing applications will directly catalyze the development of new technologies and applications in biology. Twist Bioscience is positioned to lead the transformations in both industries by bringing them together.

REFERENCES

Baum EB. (1995) Building an associative memory vastly larger than the brain. ***Science***.268(5210):583-5.

Church GM, Gao Y, Kosuri S. (2012) Next-generation digital information storage in DNA. ***Science***.337(6102):1628.

Clelland CT, Risca V, Bancroft C. (1999) Hiding messages in DNA microdots. ***Nature***. 399(6736):533-4.

Fontana R, Decad G. (2016) Storage Media Overview: Historic Perspectives. Presentation. http://storageconference.us/2016/Slides/BobFontana.pdf

Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, Birney E. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. ***Nature***. 494(7435):77-80.

Sender R, Fuchs S, Milo R. (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. ***PLoS Biol***. 14(8):e1002533.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. (2015) Big Data: Astronomical or Genomical? ***PLoS Biol.*** 13(7):e1002195.

Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes WL. (2016) Nucleic acid memory. ***Nat Mater.*** 15(4):366-70.