3 Biotech (2018) 8:239
https://doi.org/10.1007/s13205-018-1246-7

REVIEW ARTICLE

DNA as a digital information storage device: hope or hype?

Darshan Panda[1] · Kutubuddin Ali Molla[1] · Mirza Jainul Baig[1] · Alaka Swain[1] · Deeptirekha Behera[1] · Manaswini Dash[1]

Received: 26 November 2017 / Accepted: 11 April 2018 / Published online: 4 May 2018

Abstract
The total digital information today amounts to 3.52×10^{22} bits globally, and at its consistent exponential rate of growth is expected to reach 3×10^{24} bits by 2040. Data storage density of silicon chips is limited, and magnetic tapes used to maintain large-scale permanent archives begin to deteriorate within 20 years. Since silicon has limited data storage ability and serious limitations, such as human health hazards and environmental pollution, researchers across the world are intently searching for an appropriate alternative. Deoxyribonucleic acid (DNA) is an appealing option for such a purpose due to its endurance, a higher degree of compaction, and similarity to the sequential code of 0's and 1's as found in a computer. This emerging field of DNA as means of data storage has the potential to transform science fiction into reality, wherein a device that can fit in our palms can accommodate the information of the entire world, as latest research has revealed that just four grams of DNA could store the annual global digital information. DNA has all the properties to supersede the conventional hard disk, as it is capable of retaining ten times more data, has a thousandfold storage density, and consumes 10^8 times less power to store a similar amount of data. Although DNA has an enormous potential as a data storage device of the future, multiple bottlenecks such as exorbitant costs, excruciatingly slow writing and reading mechanisms, and vulnerability to mutations or errors need to be resolved. In this review, we have critically analyzed the emergence of DNA as a molecular storage device for the future, its ability to address the future digital data crunch, potential challenges in achieving this objective, various current industrial initiatives, and major breakthroughs.

Keywords Digital data · DNA storage · Silicone pollution · Data crunch · DNA hard drive · DNA steganography · Data longevity

Introduction

The entire human race is driven by information, and in this technology-oriented era, information is power. Humans have a natural propensity for accessing more and more information in as little time and space, as possible. Storage of all the accumulated information for future reference is an inherent part of our intellectual evolution. When the prehistoric man realized the significance of data storage, he tried to preserve all that he could see in his surroundings, through cave paintings and engravings on stone tablets. Starting from rocks, bones, paper, and punched cards, we traversed into the time of magnetic tapes, drums, films, gramophone records, floppies, and so on. The modern method of data storage has evolved from optical discs including CDs, DVDs, and Blu-ray discs to portable hard drives and USB flash drives. As mankind's rate of knowledge acquisition has escalated to its current peak, increasing amount of data is being accumulated and various ways of data storage and retrieval have been invented. An unprecedented breakthrough was witnessed in 1928 when Fritz Pfleumer, a German–Austrian engineer, invented the magnetic tape. This opened a new avenue of data storage, making us capable of concentrating piles of information into compact spaces. As technology evolved, computers made increasingly capacious and efficient data storage devices possible, which in turn allowed for more sophisticated ways of its utilization. The subsequent revolutionizing invention that gave a cutting edge to the technology of data storage was the invention of an integrated circuit. The first integrated circuit, also called a computer chip, was a simple structure

✉ Kutubuddin Ali Molla
kutubuddin.molla@icar.gov.in

✉ Mirza Jainul Baig
mjbaigcrri@gmail.com

[1] ICAR-National Rice Research Institute, Cuttack, Odisha 753006, India

lacking the sophisticated engineering of similar modern devices. It was created by the American engineer Jack Kilby at Texas Instruments and was demonstrated on 12th September 1958. Kilby's chip was made of the semiconductor germanium (Ge). However, within months, another inventor named Robert Noyce, popularly called "the mayor of the Silicon Valley", created a silicon (Si)-based chip. Silicon is another semiconductor and is placed next to Ge in group 14 of the periodic table. The entire modern computing industry can trace its lineage back to this one chip, although modern chips are millions of times more complex. Today, our world is replete with integrated circuits, starting from computers to almost every modern electrical device including cars, television sets, CD players, and cellular phones. The use of silicon to manufacture the microprocessor of computers, our giant data storage devices, immediately influenced our data storage efficiency. It favoured the idea of a practical desktop computer possible. This immediately made the computer an important member of every busy office desk. However, in this era of digital dependency, the generation of massive amounts of data has created problems pertaining to long-term storage, high energy consumption, and pollution associated with the manufacture of silicon chips, which has motivated researchers to investigate alternative data storage media.

The nucleic acid, DNA, a natural information storage medium, gets immediate consideration owing to its extraordinary capability of safe storage of the tremendous amount of genetic information of nearly all the biological organisms in the world. The DNA as an alternative data storage media has recently become a subject of worldwide discussion, with technology tycoons aggressively coming forward to capitalize on this possibility. In this review, we critically analyze the emergence of the concept of DNA as storage media, its historical perspective, feasibility, recent breakthroughs, and challenges to overcome for becoming a marketable data storage media, and make an effort to find an answer to the question, "Is DNA as storage media a hope or a hype?"

The nuisance of silicon pollution

Gordon Moore, a coworker of Noyce, was the first to predict the rapid rate of advancement of digital technology. Within just a few years of the silicon obsession, he put forth his theory predicting that the number of transistors on a chip would double every 2 years on average (Moore 1998). Today, with the universal presence of silicon chips, Moore's law is practically felt everywhere. Although this may be making us "technologically smarter", it is also thoroughly deteriorating our environment. In the manufacturing process, the semiconductor industry uses large amounts of harmful chemicals including silicon tetrachloride, cyanide compounds, di- and trichloroethane, and arsine gas. Furthermore, the US Bureau of Statistics has revealed that about one-third of the workers in the semiconductor industry are suffering from an illness caused by silicon pollution (Chepesiuk 1999). In addition, a typical chip manufacturing factory makes around 2 million chips per month, consuming about 20 million gallons of water. Production of a single 32 MB dynamic random-access memory (DRAM) chip weighing about two grams requires an estimated 1600 g of secondary fossil fuel and 700 g of elemental gas (N_2), while about 1.5 kWh of electricity is consumed per square centimeter of a silicon wafer processed (Williams et al. 2002). Moreover, in addition to being environmental pollutants, the reserves of silicon and other non-biodegradable materials are limited and would be exhausted one day. All the methods of digital data storage available today have a limited lifespan and become unreliable after a specified amount of time. A CD or a DVD cannot sustain as a functional piece forever. Therefore, we urgently need to find a solution to help us store and access our precious data, with minimal impact on the environment and energy consumption.

The global data crunch

An increased usage of social networking and cloud computing has led to an exponential increase in the global digital data. Globally, the total digital information today amounts to 4.4 zettabytes (3.52×10^{22} bits). If we continue storing everything for instant access, the global memory demand will reach 3×10^{24} bits by 2040 (Zhirnov et al. 2016). To meet this demand, manufacturers would need ~ 10^9 kg of silicon wafers, while the total estimated supply is only ~ 10^7–10^8 kg. In addition, data storage density of silicon chips is limited. Presently, magnetic tapes which have high storage density but are much slower to read are used to store rarely accessed data. We would require 1 billion USD over 10 years if we wish to build and maintain permanent archives on magnetic tapes (Extance 2016). The European Organization for Nuclear Research, which currently leads the world in research on particle physics, presently stores 0.08 exabytes (8×10^{16} bytes) of Large Hadron Collider data. This data grows at 15 petabytes (15×10^{15} bytes) every year. Ten percent of this data is stored in disks, while the rest is stored in magnetic tapes, which start deteriorating within 20 years (Van Bogart 1995). Thus, potentially important information is lost due to the absence of better archival systems. Although flash memory is commonly used nowadays, its rapid degeneration leads to loss of data. The degradation of flash memory is not usually associated with age, but with the number of write cycles. The number of cycles of erasing old information and writing new data is directly proportional to the rate of degradation of the memory device. In addition, the high power consumption of DRAM is one of the major bottlenecks in modern computing. Hence, to find solutions

to the issues of digital data storage, alternative technologies and principles are subjects of innovative experimentation in major laboratories of the world. Over the past two decades, scientists worldwide are continuously trying to develop reliable and stable methods for storage of non-biological data on a medium that is dense, resistant to obsolescence, universal, and durable.

In the footprints of mother nature

Nature has already developed amazing molecules having high data storage capabilities. One of these is deoxyribonucleic acid (DNA), the molecular repository of biological information. It has an astonishing ability to store biological data. A diploid human cell has 23 pairs of chromosomes, which would extend to a length of 2 m of DNA if stretched end to end. However, the nucleus of a human cell, which contains the entire set of the human genome, is only about 6 µm in diameter. This is dimensionally equivalent to stuffing 40 km of hair-fine thread into a single tennis ball! These 23 chromosomes contain information coded from the four letters of A, T, G, and C, to store data for 20,000–25,000 proteins. A diploid genome (6×10^9 bp) can store 1.5×10^9 bytes or 1.5 gigabytes of data, which could otherwise be saved into two CDs (Grigoryev 2012). As DNA is able to encode two bits per nucleotide, its storage density and small size enable one gram of dry DNA to store 455 exabytes of information (Church et al. 2012). Also, it was reported that the character density (char/m^2) of a spore of the bacterium *Bacillus subtilis* (having a genome of 4.2 mega base pairs packed in DNA of diameter 1 µm) is twenty million times that of a 200 megabyte ZIP disk having diameter of 10 cm (Shrivastava and Badlani 2014). A gram of DNA contains 10^{21} DNA bases (A, T, G, and C), which can store 108 terabytes of binary data (in 0 and 1 forms). DNA sequences can contain more information than their binary counterparts because DNA with four bases has 4X representations possible for a character string of length X, while the binary system can contain only 2X times that information. In addition, data are stored in a volumetric fashion in a DNA molecule. This enables it to store more information, unlike other media which store data linearly.

DNA for long-term storage

DNA is one of the most robust biomolecules found in nature and has an extended shelf life with no attenuation in data. The complementary nature and ability to self-assemble during the formation of tertiary structure enables DNA strands to be folded arbitrarily into polygonal digital meshes (Benson et al. 2015), engineered into complex wireframe nanostructures (Zhang et al. 2015), and scaffolded to organized biological molecules (Yang et al. 2015). In addition, DNA is non-volatile and utilizes low energy for operation in living cells. In contrast to silicon, DNA does not require lithography, which makes it more cost-effective and enables it to serve as an alternative to devices with high-volume and high-density storage capacity (Kim et al. 2004a). These factors make DNA 10^8 times more efficient than flash memory. For example, a haploid human genome (3×10^9 bp long) can store 6×10^9 bits of information while the genome of the bacterium *E. coli* (5.44×10^6 bp) can store ~ 10^{19} bits of digital data per cm^3. If we could successfully employ DNA to store data, the entire current global information (3.52×10^{22} bits) could be packed in a 0.00352 m^3 box and ~ 1 kg of DNA would be sufficient to address the world's storage requirement in 2040 (3×10^{24} bits).

Nature does not compromise the integrity of genetic information. According to the calculations of Allentoft et al. (2012), the degradation rate of mitochondrial DNA in the bones of Moa (a flightless bird species that lived until *circa* 1300 CE in the forests of New Zealand) is 1 bp per 6,830,000 years at − 5 °C. Recently, the oldest known human DNA from 430,000-year-old fossilized teeth, and bone remains from Spain were sequenced (Allentoft et al. 2012). Similarly, genome of 700,000-year-old Pleistocene horse (Orlando et al. 2013) and 110,000 years old polar bear (Miller et al. 2012), a gene from Magnolia plant fossil from Miocene (180,00,000 years old) (Kim et al. 2004b), and the entire genome of Tyrolean Iceman (a 5300-year-old Copper age individual) (Keller et al. 2012) have been successfully sequenced. DNA is one of the most stable media for storage of data in nature. However, accurate retrieval of the sequence of ancient DNA is not easy, as it is often cross-contaminated with recent DNA (Hofreiter et al. 2001), is subject to chemical damages such as deamination (C>T mutation) and depurination, and is inevitably degraded by a small degree (Briggs et al. 2007). A pointwise comparison of DNA and conventional data storage media is presented in Table 1. The whole range of available digital data storage media (DVD, floppy disk, CD, Magnetic tapes, etc.) begins to lose their integrity within a few years. In contrast, DNA has a significantly higher longevity as a data storage molecule and can be easily amplified by polymerase chain reaction techniques to get the desired number of its copies. Hence, once the data are stored in DNA, multiple copies could be generated. Due to these reasons, the method of data storage in the strands of DNA has been considered to be "apocalypse-proof" by some researchers, because it is presumed that after a hypothetical global disaster in the future, the surviving generation would find all the global data safely retained in the DNA (Yong 2013). Moreover, DNA can be read as a code (encoded as a sequence of four nitrogen bases) in both directions, a property which ensures more chances of data

Table 1 A comparative account of conventional storage media with cellular DNA as data storage medium

Device	Data retention	Storage density	Power usage (watts/gigabyte)	Access time
Hard disk	10 years	~10^{13}	~0.04	7 ms
Flash memory	~10 years	~10^{16}	~0.01–0.04	5 ns
DRAM	~64 ms or less	~10^{13}	A few tenths of watt	60 ns
Cellular DNA	> 100 years	~10^{19}	< 10^{-10}	Slower than conventional media

Flash memory does not usually degrade because of its age, but rather because of the number of write cycles. The more you erase and write new information, the more quickly the memory device will start to degrade. Power usage by DRAM is one of the most major bottlenecks in modern computing

retrieval and improved latency. Furthermore, the presence of naturally occurring enzymes for reading and writing of DNA ensures that it will remain readable in the foreseeable future (Church et al. 2012). Because DNA is invisible to the human eye, it cannot be destroyed as easily as silicon chips. In a recent study, Grass and his co-workers translated 83 kB of information to ~5000 silica encapsulated DNA segments and used accelerated aging experiments to demonstrate that the data for simulated millennia can be archived on DNA and retrieved error-free even after 1 week of heat treatment at 70 °C (Grass et al. 2015).

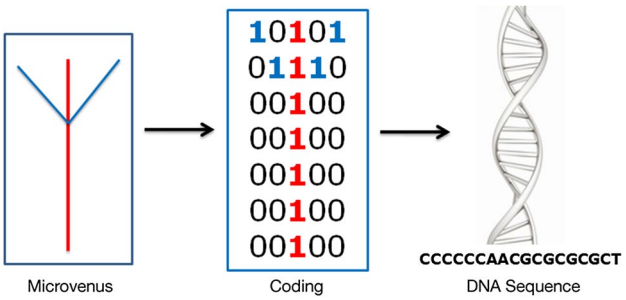

Fig. 1 Microvenus icon coding in synthetic DNA by Joe Davis

The collapse of Moore's law

Years ago, Michio Kaku, a theoretical physicist at the City College of New York, predicted the failure of Moore's law (which states that computer power doubles in every 18 months) within the next decade (Kaku 2012). The signs are now evident, as computers are not developing exponentially with traditional silicon technology. This presents a massive threat to the very foundation of Moore's law. Although Intel has already started using three-dimensional chips to mitigate this issue to a certain extent, Kaku has argued that the heat and leakage issues associated with silicon will greatly diminish its capacity in the near future. The advent of 5 nm processes for chip production will signal the end of silicon use, as smaller silicon chips tend to overheat. Hence, we urgently need to think of the alternatives that could replace silicon in the future. Several possibilities including protein computers, DNA computers, optical computers, quantum computers, and molecular computers are being proposed.

The journey so far

Long back in 1988, Joe Davis and a collaborator from MIT demonstrated the creation of a molecular artwork called 'Microvenus' (Fig. 1), wherein a small piece of synthetic DNA containing a coded visual icon representing external female genitalia and an ancient Germanic rune representing the female Earth was incorporated into live *E. coli* cells (Davis 1996). Another attempt to store digital data in DNA was made by Clelland et al. (1999), when they successfully stored encoded words in a segment of DNA strands in the form of microdots. They tried to encrypt the message using the four bases in a DNA strand and were the first to develop encrypted information packed in human genomic DNA. This mode of encryption was based on the principle that if the desired PCR primer sequences and the specific encryption key were known, then the required DNA could be readily amplified and analyzed to reveal the secret information (Fig. 2).

The next breakthrough in the field of digital data storage in DNA was achieved by a group of scientists from New York, USA, who developed the iDNA (information DNA) and used the Poly-primer Key (the primer base sequence) to access the information on the iDNA (Bancroft et al. 2001). The data stored and retrieved from DNA were the opening lines of Charles Dickens's 'A Tale of Two Cities'. A major achievement in the industrious approach to escape our dependence on silicon to store digital data was made in 2012 by Church and Kosuri at the Harvard Medical School in Boston, Massachusetts, and Gao at Johns Hopkins University in Baltimore, Maryland (Church et al. 2012). The team successfully encoded a 659 kb version of a book in several short strands of DNA, using binary digits '0' and '1'. In this approach, they designated 0 as A or C and 1 as T

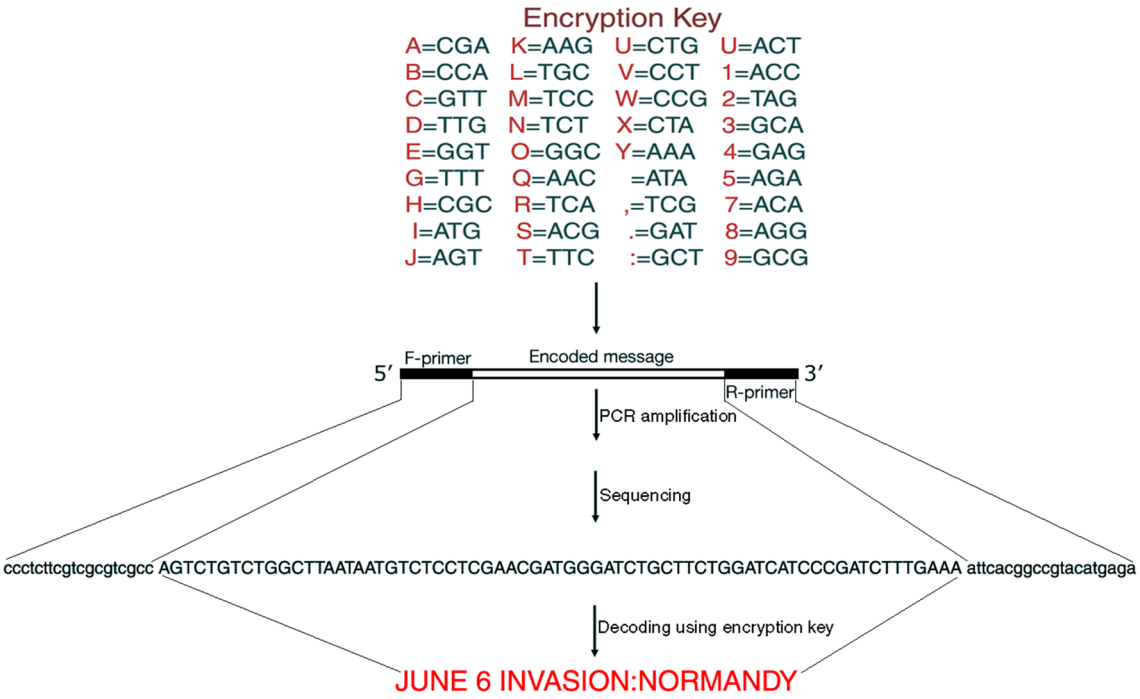

Fig. 2 The encoding scheme developed by Clelland et al. The scheme involved encryption of letters in DNA bases, encoding by synthesis of DNA, PCR amplification using specific primer pair, sequencing of the amplified product followed by decoding on the basis of primary encryption

or G. These bits were encoded onto 54,898 159-nt oligonucleotides (oligos), which were used to create ink-jet printed DNA glass microchips, followed by synthesis of the oligonucleotide library. The book was decoded by amplifying the library using limited-cycle PCR followed by sequencing. In this case, 100% recovery was obtained, with an error rate of 10 bits for every 5.27 million. The major advantages of this method were that it generated storage density of 5.5 petabits/mm^3 (100X synthetic coverage), avoided cloning and sequence verification, and used Next Generation Sequencing for both synthesis (encoding) and sequencing (decoding), which resulted in many fold cost reduction as compared to previous methods.

Another milestone in data storage on DNA was witnessed on 23rd January 2013, when the impeccable work of Nick Goldman and his team at the European Bioinformatics Institute was published (Goldman et al. 2013). Their creation of this remarkable information system was based on three objectives: to achieve feasibility, higher capacity, and lesser maintenance than existing storage media. The researchers encoded 740 kb of digital data, including Shakespeare's Sonnets, Watson and Crick's classic paper published in 1954, a JPEG color image, and an MP3 audio file containing an extract of the famous speech by Martin Luther King to strands of DNA, synthesized the data, and sequenced it to obtain 100% retrieval. They had initially used Huffman code for coding and subsequently converted it to DNA code in silico. Huffman code, a minimum redundancy code, is commonly used in computer science for data compression without loss. A simplified version of the steps involved in coding and retrieval of data in and from DNA is represented in Fig. 3. This data could be archived in DNA for millennia

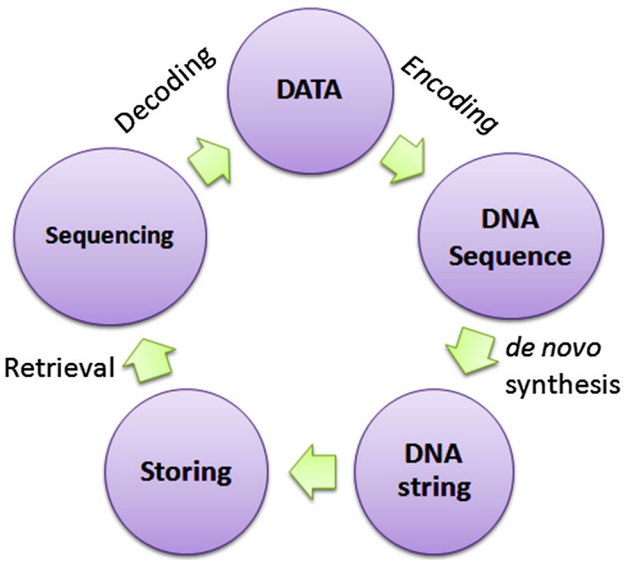

Fig. 3 DNA digital data storage: turning warehouse-sized storage to one-micrometer block of molecules (steps involved in using DNA as digital data storage)

under various adverse conditions and retrieved error free, as evidenced from the study by a group from the Swiss Federal Institute of Technology, Zurich (Grass et al. 2015). A recent study by Yazdi et al. (2015) further demonstrated that DNA could also be used in rewritable storage applications. More recently, scientists from the New York Genome Center and Columbia University, USA, developed a highly robust storage mechanism called 'DNA Fountain', and demonstrated the storage of 2.14×10^6 bytes of data, including a movie and a complete computer operating system, on DNA oligonucleotides (Erlich and Zielinski 2017). An updated timeline of information stored in DNA is represented in Fig. 4. In addition to the above studies for storage of information, several other studies have shown the successful use of DNA in steganography, a mode of encryption (Arita and Ohashi 2004; Khalifa and Atito 2012; Khalifa et al. 2016). The limitations of the

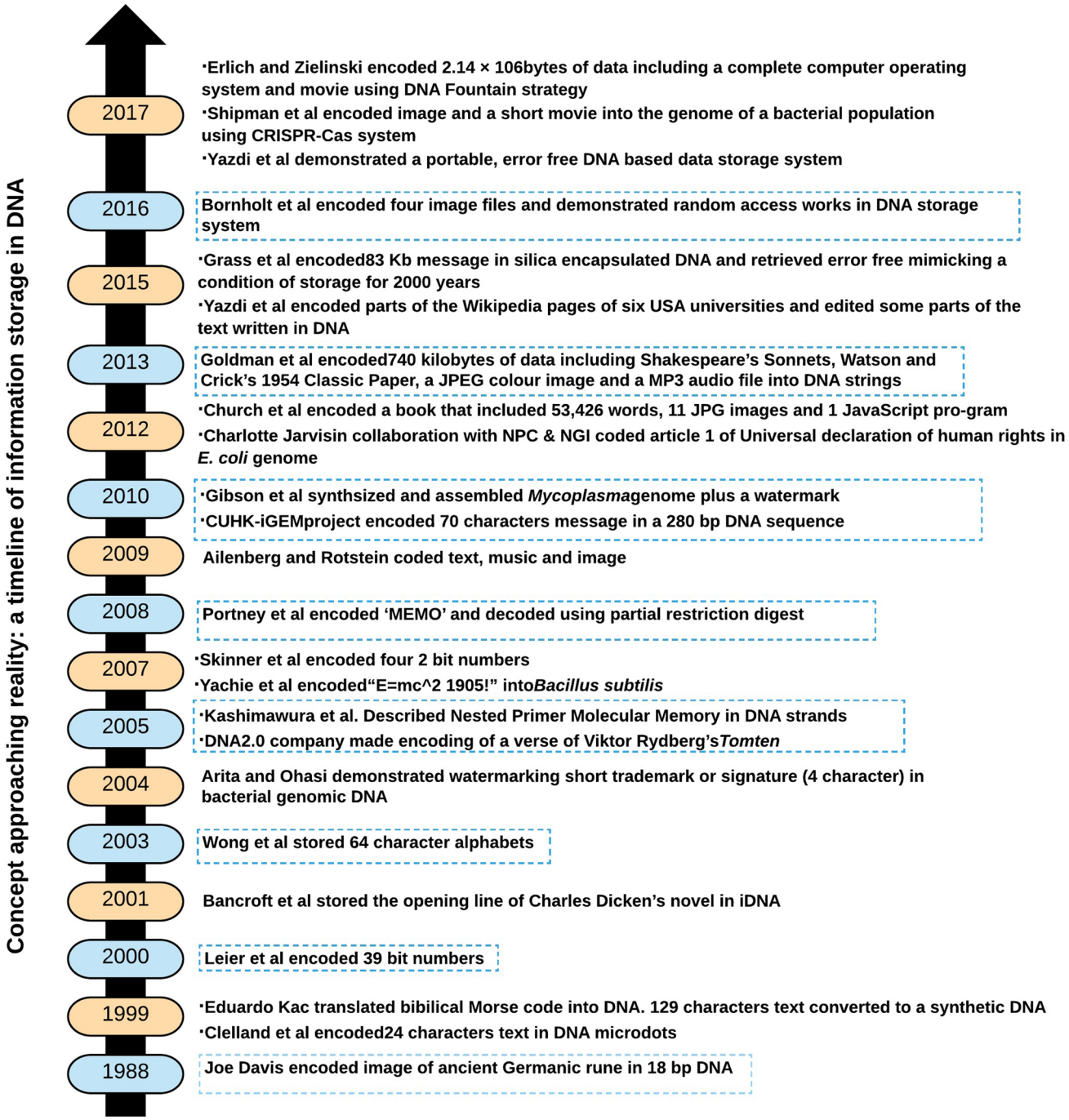

Fig. 4 A timeline of data storage in the nucleic acid. *NPC* Netherlands Proteomics Centre, *NGI* Netherlands genomics initiative. Full citations are available in the reference section

earlier systems included the requirement of laboratory expertise to read the encoded data, erroneous data recovery due to sequencing error and difficulties in synthesis and sequencing of homopolymer strings. To overcome these, Yazdi et al. developed a DNA-based data storage system that efficiently used portable nanopore sequencers to produce error-free reads with the highest rate/density of information. Their entire process was divided into two stages. The first stage involved encoding, in which compression, representation, conversion, encoding into DNA, and synthesis were carried out, followed by insertion of the homopolymer checks directly into the string of DNA. In the second stage, post-processing via sequence alignment and editing of homopolymers was carried out. Interestingly, in another novel approach, Shipman et al. (2017) successfully demonstrated the storage of information in DNA using CRISPR–Cas system to encode the pixel values into the genomes of a population of living bacteria.

Industrial initiatives

The efficacy of DNA storage to resolve the future issues involving data crunch has created novel possibilities in the economic sector, wherein several tycoons of the digital industry are collaborating with research laboratories around the globe to develop a feasible device which could be commercialized. Helixworks, an Irish startup company founded in 2015, has recently developed its patented DNA data storage technology, which is the first ever commercially available DNA data storage device (openMoSS). Furthermore, Helixworks has also designed the prototype of a DNA storage device, which is already on sale on the e-commerce retailer Amazon, by the trade name *DNADrive* (amazon.com). An open source application, available on the company's website, can convert any file into a DNA sequence in a '.moss' output format, and similarly, a valid openMoSS DNA sequence can be decoded back to the original file. Recently, Microsoft purchased ten million long oligonucleotides from Twist Bioscience to encode its digital data. In a recent joint collaboration, researchers at the University of Washington and Microsoft successfully encoded and decoded the video "This Too Shall Pass" by the band 'Ok Go', the Universal Declaration of Human Rights in over 100 languages, the top 100 books from Project Gutenberg, and the Crop Trust's seed database, encoding a total of 200 MB on the stands of DNA. This project was funded by the National Science Foundation (CSE, University of Washington). To add to these efforts, a leading memory chip manufacturing company Micron Technology, in collaboration with Boise State University, has invested in the development of DNA-based storage devices that can exceed the limit of silicon-based memory devices (Patel 2016).

Challenges

Considering the long data retention time and high storage density of DNA, it is apparent to predict that DNA has tremendous potential to become the strongest competitor to the semiconductor industry. However, it will not be easy to achieve that potential. Researchers will need to carefully address and overcome several challenges to develop user-friendly DNA-based data storage devices. The whole process of encoding information in DNA and subsequent retrieval of the required information is far more time consuming than conventional storage technology. For these reasons, DNA is more likely to face tough competition from optical, magnetic, or quantum techniques in the foreseeable future. The mechanization of various molecular processes associated with DNA has failed to perfectly mimic the natural processes. The presence of homopolymers, various sequencing errors, and errors due to lower access rate are some examples of this. A living cell has a precisely designed proofreading and DNA repair mechanism for the correction of various errors in the DNA, but such enzymatic correction mechanisms are not available in the artificial synthesis of DNA. Although the error-free synthesis, amplification, and sequencing of DNA cannot be achieved yet, a breakthrough was made by Blawat et al. (2016), who recently reported storage and successful error-free retrieval of 22 MB of digital data in synthetic DNA, using a forward error correction scheme. Due to its sensitive structure, DNA is prone to mutations under extreme conditions; hence, the chances of data alteration cannot be ignored. At the same time, it is quite difficult to synthesize long sequences of DNA de novo. Moreover, while conventional and popular storage systems such as hard drive and flash drive are more expeditious in erasing and rewriting data, this aspect has barely been dealt with for DNA data storage systems.

Cost involved

The major drawback of 'DNA-based data storage systems' is the cost involved in writing and reading data on nucleotide sequences. Cost of synthesizing DNA (writing/encoding) is higher than that of sequencing (reading/decoding). Out of the USD 12,660 spent during experiments for the creation of 739 kb of hard disk storage by Goldman et al. (2013) at the European Bioinformatics Institute, 98% was spent on the synthesis and only 2% on sequencing (Extance 2016). Under the DNA fountain scheme, Erlich and Zielinski (2017) spent USD 7000 to encode 2.14 MB data. Hence, DNA fountain costs about ~ USD 3500 per MB of data writing and another USD 1000 to read it (Service

2017). Recently, Yazdi et al. (2017) spent USD 2540 to first compress 10,894 bytes of data into 3633 bytes and then encode it in DNA of length 16,880 bp. If we compare this cost with that of a modern-day conventional hard drive, the cost of writing data in DNA is significantly impractical. Even under the DNA Fountain scheme which is presently the most cost-effective, the data storage equivalent to a modern 1 TB hard drive would cost about 7×10^7 times more in DNA. While the cost of DNA storage devices must be at par with their counterparts to make an impact on future markets, the current high costs of storing and reading per MB data makes DNA-based storage much more expensive in comparison.

Conclusion

Developing a novel method to store an astronomical amount of data in the modest double-stranded DNA molecule is no more a matter confined to the realms of science fiction. Like any revolution in technology, DNA-based digital data storage has to face major challenges in coming years to fulfill its tremendous potential. According to theories today, just a few grams of DNA can store all the information ever produced by mankind. To reach the commercial mainstream for data retrieval, DNA-based digital storage has to make several breakthroughs. Emerging technologies such as nanopore sequencing have facilitated a 50,000-fold reduction in the cost of DNA sequencing, from USD 31,250/MB in 2002 to USD 0.63 in 2016 (Shendure and Aiden 2012). An analysis has revealed that the cost to read and write information on DNA has reduced to ~USD 10^{-7}/bit and ~USD 10^{-4}/bit, respectively (Zhirnov et al. 2016). The renowned mathematician and genome scientist Nick Goldman once said: "As DNA is the basis of life on Earth, methods for working with it, storing it and retrieving it will remain the subject of continual technological innovation". Looking into the future, it could be predicted that to handle an enormous amount of data, a practical large-scale DNA archive would apparently need stable DNA management and innovative indexing solutions. This will result in a favorable paradigm shift in computing, as the aspect of data storage would be an integral part in the realization of the idea of DNA computing. This development will mutually benefit research and development in sequencing and synthesis technologies. However, the synthesis of long-stranded DNA molecules with low error rates desired for the process of data archiving will still take a considerable amount of time to reach practical levels of higher reliability. In addition, synthesizing long strands of DNA with high fidelity and then sequencing them back to accurately recover the information will involve a high-throughput laboratory and skilled manpower. With the advancement of technology in the coming years, the cost of writing on DNA and reading it will come down. However, DNA-based data storage is bound to face competition with the more traditional means of data storage which are ubiquitous, inexpensive, and user-friendly (for example, conventional large capacity disk storage costs as little as USD 0.05 per gigabyte today, down from USD 500,000.00 per gigabyte back in the early 1980s). Overall, this technology has already developed a momentum with the potential to transform our vision for the future of digital storage. It would not be farfetched to assume that DNA, if not by itself, then at least as a hybrid with silicon, could also do a fantastic job as storage media of the future.

Acknowledgements We thank Justin Shih, PSU, USA for his assistance with language editing. Critical input from Chitra Lele, UK is acknowledged. We acknowledge the funding from Indian Council of Agricultural Research, New Delhi.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

Ailenberg M, Rotstein OD (2009) An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques 47:747

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc R Soc Lond B Bio. https://doi.org/10.1098/rspb.2012.1745

Arita M, Ohashi Y (2004) Secret signatures inside genomic DNA. Biotechnol Prog 20:1605–1607

Bancroft C, Bowler T, Bloom B, Clelland CT (2001) Long-term storage of information in DNA. Science 293:1763–1765

Benson E, Mohammed A, Gardell J, Masich S, Czeizle E (2015) DNA rendering of polyhedral meshes at the nanoscale. Nature 523:441–444

Blawat M, Gaedke K, Huetter I, Chen XM, Turczyk B (2016) Forward error correction for DNA data storage. Proc Comput Sci 80:1011–1022

Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G (2016) A DNA-based archival storage system. ACM SIGOPS Oper Syst Rev 50:637–649

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci 104(37):14616–14621

Chepesiuk R (1999) Where the chips fall: environmental health in the semiconductor industry. Environ Health Perspect 107:A452

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

Clelland CT, Risca V, Bancroft C (1999) Hiding messages in DNA microdots. Nature 399:533–534

Davis J (1996) Microvenus. Art J 55:70–74

Erlich Y, Zielinski D (2017) DNA fountain enables a robust and efficient storage architecture. Science 355:950–954

Extance A (2016) How DNA could store all the world's data. Nature 537:7618

Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329:52–56

Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM (2013) Toward practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494:77–80

Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed 54:2552–2555

Grigoryev Y (2012) How much information is stored in the human genome? Technical report from BitesizeBio http://bitesizebio.com/8378/how-much-information-is-stored-in-the-human-genome/

Gustafsson C (2009) For anyone who ever said there's no such thing as a poetic gene. Nature 458:703

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001) Ancient DNA. Nat Rev Genet 2(5):353–359

iGEM C (2010) Bacterial-based storage and encryption device. http://2010.igem.org/Team:Hong_Kong-CUHK

Jarvis CNPC (2012) Blighted by Kenning. http://www.artforeating.co.uk/restaurant/indexphp?/blighted-by-ken/project-overview

Kac E (1999) GENESIS. http://www.ekac.org/geninfo2.html

Kaku M (2012) Physics of the future: How science will shape human destiny and our daily lives by the year 2100: Anchor

Kashiwamura S, Yamamoto M, Kameda A, Shiba T, Ohuchi A (2005) Potential for enlarging DNA memory: the validity of experimental operations of scaled-up nested primer molecular memory. BioSyst 80:99–112

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Stade B (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3:698

Khalifa A, Atito A (2012) High-capacity DNA-based steganography. IEEE Bio 76–80

Khalifa A, Elhadad A, Hamad S (2016) Secure blind data hiding into pseudo DNA sequences using playfair ciphering and generic complementary substitution. Appl Math 10:1483–1492

Kim C, Li M, Rodesch M, Lowe A, Richmond K (2004a) Biological lithography: improvements in DNA synthesis methods. J Vac Sci Technol B Microelectron Nanometer Struct Process Measurement Phenomena 22:3163–3167

Kim S, Soltis DE, Soltis PS, Suh Y (2004b) DNA sequences from Miocene fossils: an ndhF sequence of Magnolia latahensis (Magnoliaceae) and an rbcL sequence of Persea pseudocarolinensis (Lauraceae). Am J Bot 91:615–620

Leier A, Richter C, Banzhaf W, Rauhe H (2000) Cryptography with DNA binary strands. BioSyst 57:13–22

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina C (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci 109:E2382–E2390

Moore GE (1998) Cramming more components onto integrated circuits. Proc IEEE 86:82–85

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74–78

Portney NG, Wu Y, Quezada LK, Lonardi S, Ozkan M (2008) Length-based encoding of binary data in DNA. Langmuir 24:1613–1616

Patel P (2016) Scientific American https://www.scientificamerican.com/article/tech-turns-to-biology-as-data-storage-needs-explode/. Accessed 31 May 2016

Service RF (2017) DNA could store all of the world's data in one room. Science. https://doi.org/10.1126/science.aal0852

Shendure J, Aiden EL (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30:1084–1094

Shipman SL, Nivala J, Macklis JD, Church GM (2017) CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature 547(7663):345

Shrivastava S, Badlani R (2014) Data storage in DNA. Int J Elec Energy 2:119–124

Skinner GM, Visscher K, Mansuripur M (2007) Biocompatible writing of data into DNA. J Bionanosci 1:17–21

Van Bogart, JW (1995) Life Expectancy: How long will magnetic media last?. Council on Library and Information Resources

Williams ED, Ayres RU, Heller M (2002) The 1.7 kilogram microchip: energy and material use in the production of semiconductor devices. Environ Sci Technol 36:5504–5510

Wong PC, Wong KK, Foote H (2003) Organic data memory using the DNA approach. Commun ACM 46:95–98

Amazon:https://www.amazon.com/s/ref=nb_sb_noss_2?url=search-alias%3Daps&field-keywords=DNADrive

openMoSS: Open Molecular Storage System. https://openmoss.org/

Yachie N, Sekiyama K, Sugahara J, Ohashi Y, Tomita M (2007) Alignment-based approach for durable data storage into living organisms. Biotechnol Prog 23:501–505

Yang YR, Liu Y, Yan H (2015) DNA nanostructures as programmable biomolecular scaffolds. Bioconjug Chem 26:1381–1395

Yazdi SHT, Yuan Y, Ma J, Zhao H, Milenkovic O (2015) A rewritable, random-access DNA-based storage system. Sci Rep. https://doi.org/10.1038/srep14138

Yazdi SHT, Gabrys R, Milenkovic O (2017) Portable and error-free DNA-based data storage. Sci Rep 7(1):5011

Yong E (2013) Synthetic double-helix faithfully stores Shakespeare's sonnets. Nature. http://www.nature.com/news/synthetic-double-helix-faithfully-stores-shakespeare-s-sonnets-1.12279. Accessed 23 Jan 2013

Zhang F, Jiang S, Wu S, Li Y, Mao C (2015) Complex wireframe DNA origami nanostructures with multi-arm junction vertices. Nat Nanotechnol 10:779–784

Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes WL (2016) Nucleic acid memory. Nat Mater 15:366–370