Indian Journal of Animal Sciences **90** (2): 134–140, February 2020/Review Article

Big data management: from hard drives to DNA drives

AMBREEN HAMADANI[1], NAZIR A GANAI[2], SHAH F FAROOQ[3] and BASHARAT A BHAT[4]

Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar, Jammu and Kashmir 190 006 India

Received: 1 February 2019; Accepted: 16 July 2019

ABSTRACT

Information Communication and Technology is transforming all aspects of modern life and in this digital era, there is a tremendous increase in the amount of data that is being generated every day. The current, conventional storage devices are unable to keep pace with this rapidly growing data. Thus, there is a need to look for alternative storage devices. DNA being exceptional in storage of biological information offers a promising storage capacity. With its unique abilities of dense storage and reliability, it may prove better than all conventional storage devices in near future. The nucleotide bases are present in DNA in a particular sequence representing the coded information. These are the equivalent of binary letters (0 &1). To store data in DNA, binary data is first converted to ternary or quaternary which is then translated into the nucleotide code comprising 4 nucleotide bases (A, C, G, T). A DNA strand is then synthesized as per the code developed. This may either be stored in pools or sequenced back. The nucleotide code is converted back into ternary and subsequently the binary code which is read just like digital data. DNA drives may have a wide variety of applications in information storage and DNA steganography.

Key words: Big data, Coding, DNA drives, Storage

The world is rapidly going digital and digital information is accumulating at an astounding rate which in turn is straining our ability to archive and store it (Church *et al.* 2012). Digital technologies today continuously monitor physical environments and thereby produce massive amounts of data with unprecedented rapidity (Kamilaris 2017). The "digital universe" (all digital data worldwide) has increased to an amount greater than 16 zettabytes in 2017. Counting everything from astronomical images and journal articles to online social networking, the global digital archive will hit an estimated 44 trillion gigabytes (GB) by 2020, a tenfold increase over 2013. By 2040, if everything were stored in memory sticks, the demand would be 10–100 times the expected supply of microchip-grade silicon needed for the manufacture of memory sticks (Zhirnov *et al.* 2016). The world also is seeing a rise in e-waste, and there is a growing requirement of more and more energy. In future, it will become difficult to store the increasing amount of data.

Digital data storage

All world's data today is stored on magnetic and optical media. Digital data is stored on portable storage devices like hard drives, flash memories, solid state drives etc. The data generated on the internet is stored via servers in data centres. There are 509,147 data centres in the world today spanning an area of 585,831,841 sq. ft. (Miller 2011). Google alone has 15 data centres; 8 in the US, 2 in Asia, 4 in Europe, 1 in S. America and its data centre in South Carolina, United States alone covers an area of 2 lakh square feet. Current long-term archival storage solutions require refreshes, scrubbing of corrupted data, replacement of faulty units, refreshes in technology. World's highest capacity hard drive (3.5" of Seagate) today can store data of 60 TB (terabyte) (Singleton 2016). It would take an average of 1 billion such drives (88900 km) in order to preserve the entire world's data. As per Dr. Martin Hilbert, "If we were to take all the information and store it in books, it would cover the entire area of the US in 13 layers of books" (Stewart 2011).

Modern archiving technologies cannot keep pace with this exponential growth rate of digital data (Extrance 2016) and in near future, the task of storage will get more cumbersome, even after accounting for the predicted improvements in storage technologies. Therefore, if we are to preserve the world's data, it is essential that we seek significant advances in both storage density and durability. Nature may help us find a solution to this problem. Studies conducted over the last few decades suggest that systems in which quantum features play a prominent role may prove to be much more efficient than classical physical-dynamical analogues and biological cells have great information

Present address: [1]PhD Scholar (escritor005@gmail.com), Division of Animal Genetics and Breeding, Faculty of Veterinary Sciences and Animal Husbandry. [2]Director Planning and Monitoring (drnazirahmad@gmail.com), Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar, J&K. [3]Research Scholar (shah.faizanfarooq6@gmail.com), Central University of Kashmir, Jammu and Kashmir. [4]Researcher (basharatbhat09@gmail.com), University of Otago, New Zealand.

processing efficiency than the conventional systems employed for the purpose (Conrad 1990). Therefore, the alternative for storage lies in the creation of DNA Drives (Goldman 2013). DNA stores the information of all living organisms and therefore it may be used to store digital information as well.

DNA

A DNA strand, or oligonucleotide, is a linear sequence of these nucleotides. A single strand of DNA consists of a phosphate group, a deoxyribose sugar and nitrogenous bases. The nitrogenous bases are Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) (Calladine *et al.* 2004). The backbone of the DNA strand is made from alternating phosphate and sugar residues (Limbachiya and Gupta 2015). A nucleotide is the basic building block of DNA.

The two ends of a DNA strand, called the 5′ and 3′ ends, are chemically unlike. DNA sequences are represented conventionally starting with the 5′ nucleotide end. (Bornholt *et al.* 2016) and the interactions between different strands are predictable based on sequence. Two single strands bind to each other to form a double helix only when they are complementary: A in one strand aligns with T in the other, and likewise for C and G. The two strands in a double helix have opposite directionality (5′ end binds to the 3′ end the other strand), and therefore two sequences of DNA are "reverse complements" of each other (Alberts, 2003).

DNA as a unique computational element

DNA has extremely dense information storage (Ray 2019), enormous parallelism and extraordinary energy efficiency. It has the ability to conceal data in condensed form and also to allow this data to be copied when required, via self-propagation (Bancroft *et al.* 2001) and via various lab techniques.

With bases spaced at 0.35 nm, data density in DNA is greater than a million Gb/inch compared to 7 Gb/inch in archetypal high performance HDD. Unlike most digital storage media, DNA storage is not limited to a planar layer. Computer chips are "planar" storage devices as is obvious from their shape. The capacity of a computer chip can be improved by putting several layers of circuits in it, thus making it 2D. This, however, causes an additional problem of heat generation. The theoretical limit of storage capacity in DNA is above 1 EB/mm3 (eight orders of magnitude denser than tape), and 1 g of single-stranded DNA can accumulate up to 455 Exabyte. That goes on to say that 1 g of DNA can store about 455 billion GB. It is concluded from this that 1 g of single-stranded DNA, one can supposedly store an equivalent of 250 million DVDs (Church *et al.* 2012). Losick and Hoch found that the density to contain characters (char/m^2) in *Bacillus subtilis* bacterium (Genome size 4.2 Mega base pairs, with 1 µm diameter) spore is twenty million times greater that of a 200 Megabyte ZIP disk of diameter 10 cm. Table 1 indicates the comparison of DNA storage capacity with conventional storage.

The enormous parallelism of DNA is yet another exceptional ability of DNA. Parallel computing is that type of computation in which multiple calculations or execution of processes can be carried out simultaneously (Sudha and Valli 2017, Gottlieb and Almasi 1989). For DNA, the parallel computational ability is colossal and self-assembly properties of the DNA molecule contribute to this high degree of parallelism (Conrad and Zauner 1997). A single test tube can contain trillions of strands of DNA. Operation can be carried out on all strands in the tube in parallel which is an estimated (3×10^{14}) molecules at a time.

Table 1. Comparison of DNA storage capacity with conventional storage

	DNA	Conventional storage
Density	> million Gb/inch	7 Gb/inch
Dimension	3D	2D (>1 layers of circuit) (Castillo 2014)
Coding unit	< 1/2 nm	10 nm

DNA is also extraordinarily durable, especially under cool and dry conditions. DNA has been found to be readable despite degradation in non-ideal circumstances over millions of years. Scientists have been able to decipher DNA from mammoths and Neanderthals and have decoded horse genome from a bone trapped in permafrost for 700,000 years **(**Extrance 2016). DNA also has been recovered almost intact from the 6000 years old fossilized remains of Bison. Researchers have reconstructed the genomes of ancient humans from bones in a Spanish cave after more than 400,000 years (Rosenblum 2016). All this indicates that data can safely be preserved for centuries within DNA and recovered with near perfect precision.

It has also been seen that DNA can tolerate a wide range of temperatures (–800 to 800°C) (Shrivastava and Badlani 2014). Scientists who attached small double strands of DNA to the outer casing of a rocket discovered that it could survive temperatures soaring to greater than 1,000°C (Perry 2014). Table 2 shows the comparison between hard disk, flash disk and bacterial DNA.

Table 2. Comparison between hard disk, flash disk and bacterial DNA (Extance 2016)

	Hard disk	Flash memory	Bacterial DNA
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01– 0.04	$<10^{-10}$
Data density (bits per cm^3)	$\sim10^{13}$	$\sim10^{16}$	$\sim10^{19}$

Data storage within DNA

All digital files including movies, text, music etc. can be converted to a "genetic file" and stored as strands of DNA both *in vivo* and *in vitro*. However synthetic data storage is better than living vectors.

Data on digital devices is translated into the binary code and stored in memory spaces called memory cells within storage devices. The smallest unit of storage is a bit/ binary digit. A single file may contain millions of bits. To store information in DNA, the digital file's binary code is converted into the four-letter genetic code, which consists of As, Cs, Gs, and Ts that represent the chemical building blocks of DNA strands. By varying the order of the 2 base nucleotide pairs (A-T, G=C), one can encode all types of data. A single nucleotide can represent 1 bit per base (e.g. A or C for zero, G or T for one) (Church *et al.* 2012) or 2 bits of information in 3D (Castillo 2014). And over 3 billion years of evolutionary optimization of the machinery has enabled DNA to faithfully replicate this information (Ross *et al.* 2013, Schwartz 2012)

Coding scheme

The coding scheme for DNA comprises of binarization into the binary code and subsequently into the nucleobase code which is similar to a computer that stores data by changing the order of 1s and 0s. Source data in form of binary bits (0 and 1) may be either converted to tertiary/ ternary bit code (0, 1 and 2) or be encoded in base 4, producing a string of n/2 quaternary digits from a string of n binary bits. The quaternary digits (mapping 0, 1, 2, 3 to A, C, G, T respectively) or ternary digits can then be mapped to DNA nucleotides. Ternary bit code decreases the chances of encoding errors (Niedringhaus 2011) and is hence preferred. Many coding schemes have been reported for information coding in DNA. These include the Huffman code, the Comma Code, the Alternating Code (Smith *et al.* 2003), the Improved Huffman Code, Perfect Genetic Code, Cambridge Code, Reed-Solomon Code etc.

Each coding scheme has its own advantages and disadvantages and is used as per the scientists' requirements for data storage. An essential factor for consideration is the judicious use of nucleotide bases per character. It has been proven mathematically that the base to character ratio of around three is most optimum and economical for a coding system. For this reason, many researchers prefer to use the Huffman Coding Scheme (Huffman 1953). It is a base 3 representation which is applied extensively for lossless data compression in digital communication and data storage.

After the conversion, the digital data is encoded into the nucleobases of DNA. By changing the positions of nucleobases, A, T, G and C in all possible combinations, the tertiary code can be mapped onto the nucleobases codes. Each ternary digit maps to a DNA nucleotide based on a rotating code (Goldman *et al.* 2013). A rotating code is used to avoid repeating the same nucleotide more than once thereby avoiding homopolymers (repetitions of the same nucleotide). Homopolymers would otherwise significantly increase the chance of sequencing errors. For instance, if the previous base was A, then T would be used to represent 2, G would represent 1 an T would represent 0 but if the previous base was G, then 2 would be represented by C, 1 by A and 0 by T. There are similar substitution rules that cover all possible combinations of letters and numbers. Therefore, an identical digit sequence in the data will not be represented by a sequence of identical bases in the DNA. As a result, mistakes will be avoided (Goldman *et al.* 2013). Messages can be encoded in many ways. Selection is based on the need to avoid sequences that are difficult to read or write, e.g. extreme GC content, repeats, or secondary structure.

DNA synthesis

The write process for DNA storage records digital data into DNA nucleotide sequences. Then it synthesizes DNA molecules. DNA synthesis is a standard practice in biotechnology. To eliminate the need for construction of log DNA sequences, the bit stream is split into addressed data blocks. Repetitive blocks of nucleobases encoding the data are thus formed (Bancroft *et al.* 2001).

Payload (Strand encoding the information to be stored in DNA), to be stored is divided into data blocks, whose length depends on the desired strand length and the supplementary overheads of the format used. In order to facilitate decoding of the synthesized strands later, two sense nucleotides ("S") indicating whether the strand has been reverse complemented or not and primer sequences are added to the payload. These sequences serve as a "foothold" for the PCR process and allow the PCR to selectively amplify only those strands with a chosen primer sequence. Identification tags added to data objects to facilitate retrieval because DNA is stored within storage pools. Encoding data addresses, error detection codes, payloads, and the primer target sequences will produce final DNA sequences. A synthesizer is used for the manufacture of the desired sequence.

Even though the coupling efficiency for each step is higher than 99%, meagre errors often result in an exponential decrease of product yield with increasing length. Therefore, the size of oligonucleotides that can be efficiently synthesized to about 200 nucleotides (Church *et al.* 2012).

Data retrieval

Scientists use a simple key-value architecture for retrieval of data objects, where a put (key, value) operation associates value with key, and a get(key) operation retrieves the value assigned to key (Bornholt *et al.* 2016). The requirement for the implementation of key value architecture includes:

1. A function to map a key to the DNA pool (in the library) where the concerned strands that contain data lie
2. A mechanism that will selectively retrieve only desired strands from the storage pool and provide random access.

For this reason, a key is mapped to a pair of PCR primers and those primers are added to the strands. A mapping from keys is designed to unique primer sequences. Since usually, all strands for a particular object share a common primer, addressing therefore helps distinguish the strands with the same primer but different addresses. The key is used to get

the PCR primer sequences and therefore, to determine the pool in the DNA library where the resultant strands will be stored (Bornholt *et al.* 2016). At read time, those same primers are used in PCR to amplify only the strands with the desired keys and the resulting pool contains higher concentrations of desired strands. A sample from that pool will contain all of the desired strands.

Reading the data involves sequencing the DNA molecules and decoding the information back to the original digital data. The encoded DNA when sequenced, read back to tertiary and then to binary data will enable us to retrieve the data back. These technologies are similar to those used to map the human genome.

The read process takes as input a key. It uses the key to obtain the PCR primer sequences that identify molecules in the pool associated with that key. The storage system physically extracts a sample from the DNA pool that contains the stored data, but likely also includes large amounts of unrelated data. PCR thermocycler amplifies the sample and the PCR primers. The resulting pool is transferred to a DNA sequencer, which ultimately produces the digital data readout. The DNA synthesizer is used for producing the DNA strands that hold data payload (The string of nucleotides representing the data) and PCR primers that are used to amplify data during the random access read process. The read process removes a sample of DNA from the pool but the pools can easily be replenished/ resynthesized after read operations as per requirement because DNA is easy to replicate, and so if necessary.

DNA sequencing

There are several high-throughput sequencing techniques, but the most popular method among them is "sequencing by synthesis". Sequencing by synthesis uses DNA polymerase enzymes (Bornholt *et al.* 2016). The strand of interest serves as a template for the polymerase, which creates a complement of the strand. Importantly, fluorescent nucleotides are used for the synthesis process. Since each type of fluorescent nucleotide emits a different colour, it is possible to read out the complete complement sequence optically. PCR is a method for exponentially amplifying the concentration of selected sequences of DNA within a pool.

DNA synthesis and sequencing error rates are of the order of 1% per nucleotide. Sequencing errors therefore contribute significantly to the total errors: the sequencing error rate is on average an order of magnitude higher than other errors and sequences may degrade during storage. This may further decrease data integrity. A key aspect of DNA storage is to devise appropriate encoding schemes that can tolerate errors by adding redundancy (Bornholt *et al.* 2016).

Encoding schemes

If each bit of binary data is encoded in only one location of the output DNA strands, the data retrieval will rely heavily on the robustness of DNA. And if this is by chance lost, all the data will also be lost forever. In order to come up with a solution to this, several schemes have been employed for encoding data reliably and for retrieving information accurately. One such scheme is called the Goldman Encoding which splits the input DNA nucleotides into overlapping segments. This provides a fourfold redundancy for each segment. Each window of four segments corresponds to a strand in the output encoding (Goldman *et al.* 2013). This is the most successful DNA storage technique published so far. In addition, it offers a tunable level of redundancy which can be achieved by reducing segment widths and therefore repeating them more often in Input Nucleotide.

Another method of encoding incorporates redundancy by taking the exclusive-or of two payloads to form a third payload (Bornholt *et al.* 2016). Recovering any two of the three strands sufficiently recovers the third strands of the same length (for example, if the length of the overlapping segments were halved, the number of repeated strands would double.) The Goldman encoding scheme compromises density but provides high reliability. XOR encoding developed by Bornholt and co-workers, provides similar levels of redundancy to Goldman Encoding, but with reduced overhead. This encoding provides redundancy by a simple exclusive-or operation at the strand level wherein the exclusive-orof the payloads *A* and *B* of two strands, produces a new payload and so a new DNA strand. The address block of the new strand encodes the addresses of both the input strands that were used as the inputs to the exclusive-or. The original payload is distinguished from the exclusive-or strand using the high bit of the address.

The exclusive-OR (XOR), operator uses the symbol •". The following logic operation is performed by the operator: X •" Y = X Y' + X' Y. Exclusive-or or exclusive disjunction is a logical operation. It has a "true" output only when inputs differ from each other (one is true, the other is false) otherwise the output is "false". This could be taken as "A or B, but not, A and B". This encoding is just as reliable as the Goldamn encoding but theoretically, its density of is much higher. In Goldman scheme, the nucleotide repeats (up to) four times and in the scheme proposed by Bornholt and coworkers, the nucleotide repeats an average of only 1.5 times. However practically, the density may be lower due to the presence of overheads of addressing and primers that are constant between the two encodings.

It is important to note that all types of data do not require high-precision storage. This is especially true for every data structure (Huffman 1952, Guo 2016) and small errors in the payload may sometimes be tolerable at the cost of some decoding imprecision. Tunable redundancy allows the storage system to optimize the balance between reliability and efficiency as per the requirement and the importance of the data to be stored.

Storage

A DNA storage system consists of a DNA synthesizer which encodes the data to be stored in DNA. Pools of DNA are stored in a storage container with compartments. A DNA

Table 3. Pioneering researches in DNA data storage

Year	Institute/Scientist	Research
1988 (Gibson *et al.* 2010)	J.C. Venter Institute	· First demonstrated storing messages in DNA · 7920 bits of data were stored
1999 (Clelland*et al.* 1999)	K.T. Clelland and Coworkers	· Developed DNA based, doubly steganographic technique for secret messages · Encoded 23 character message camouflaged within DNA which was further concealed in microdot
2000 (Leier *et al.* 2000)	A. Leier and Coworkers	· Encoded three 9-bit numbers into DNA

Table 4. Recent researches in DNA data storage

Year	Institute/Scientist	Research
2015 (Grass *et al.* 2015)	R.N. Grass and Co-workers	· Encoded 83 kB message without error using Reed-Solomon code
2015 (Yazdi *et al.* 2015)	S.M.H.T. Yazdi and Co-workers	· Developed a method for rewritable random-access DNA-based storage. · Encoding was dictionary-based and focused on storage of text 2016 (Herkewitz 2016)
	Seth Shipman and Jeff Nivala	· Used segment of genetic data into bacteria carrying CRISPR/Cas system · Used a colony of *E. coli* to create a jumble of tiny hard drives which could pass information to the progeny making the colony an *in vivo* recording device
2016 (Langston 2016)	Microsoft/ University of DNA	· Encoded 200 MB data including War and Peace + 99 literary classics, Universal Declaration of Human Rights in more than a 100 languages, top 100 books of Project Gutenberg, Crop Trust's Seed database, HD music video (This Too Shall Pass) into the DNA
2016 (Bornholt *et al.* 2016)	James Bornholt and Co-workers	· Designed a simulator for DNA synthesis and sequencing
2016 (Bornholt *et al.* 2016)	James Bornholt and Co-workers	· Translated 151 KB data which included 4 image files in x.jpg format
2017 (Erlich and Zielinski 2017)		· Created 45,652 sequences of length of 120 nucleotides Erlich& Dina

sequencer that reads DNA sequences and converts them back into digital data. A typical data object maps to a very large number of DNA strands. The DNA "pools" have stochastic spatial organization and do not permit structured addressing, unlike electronic storage media.

Scientists are constantly finding novels ways to store data in the most reliable form. Researchers have recently encapsulated DNA in silica (glass) for the encapsulation of DNA in bones thereby creating a "Fossil Shell". Later they were able to separate this using fluoride solution and could recover all information, error free, even after storing the DNA at 70°C for 1 week (experimentally) which was equivalent to storing DNA in central Europe for 2000 years (Grass *et al.* 2015).

Applications

DNA can be used in National security for information hiding purposes and for data stenography because it is ultra-compact and will not degrade overtime. DNA steganography is a special type of cryptography and is much safer than ordinal cryptography. The encrypted message can be safely hidden from unwanted interference. It may also safely preserve the personal information of a person such as medical information and family history in their own bodies. Storage of archival documents may also be done using this technique. The present scenario is that compared with other forms of data storage, writing and reading to DNA is relatively slow. So, this approach would be better suited for archival applications. Storage of important scientific information may also be considered.

Cutting Edge Researches

Constraints

Today, neither the performance nor the cost of DNA synthesis and sequencing is viable for data storage purposes. However, exponential improvements have historically been seen in both. One of the major constraint is the cost of synthesis and sequencing. It takes $7,000 to synthesize the DNA to archive 2 megabytes of data, and another $2,000 to read it.

However, the costs of DNA synthesis and sequencing have been dropping at exponential rates of 5 to 12-fold each year. The rate of drop of the prices is much greater than electronic media at 1.6-fold per year (Carr and Church 2009). The cost reductions and throughput improvements of the DNA synthesis and sequencing technologies have been compared to Moore's Law in Carlson's Curves (Calson

Table 5. Recent cutting-edge researches

Year	Institute/Scientist	Research
2001 (Bancroft *et al.* 2001)	C. Bancroft and Coworkers	• Developed iDNA (information DNA) which was the name given to the encoded data) • Used a Poly Primer Key and 'Universal' Forward & Reverse primers for the purpose. • This yielded ordered fragments of iDNA
2003 (Wong and Foote 2003)	P. Wong and K. Foote	• Used *E. coli* & *Deinococcus radiodurans* as vectors. • Used 'safe sequences' foreign to bacteria for encoding (25 in10 billion). This was cloned - into a recombinant plasmid. • Used stop codons to protect the message within the DNA. • Chemically synthesized 7 DNA fragments with 57–99 base pairs (bp)
2010 (Gibson *et al.* 2010)	D. G. Gibson and Co-Workers	• Developed and *in-vivo* technique for storing data in DNA. • 1280 characters encoded in bacterial genome as "watermarks"
2012 (Church *et al.* 2012)	G. M. Church and Co-Workers	• Encoded 600 times more information in DNA than ever before. • Encoded his own Book (54,000-words, 11 images), 15.27 MB, 59 oligonucleotides. • Used an inkjet printer to embed short fragments of synthesized DNA on the surface of a tiny glass chip. • Made 70 billion copies of the DNA sequence • 2012 (Goldman *et al.* 2013) Nick Goldman and Co-workers · Encoded 739 kB message in DNA • Encoded 5 computer files, 154 Shakespeare sonnets (ASCII), MP3 Martin Luther King's "I have a dream", PDF version of a Watson & Crick Paper, Photograph of their lab and Huffman code

2014) and comparisons reveal that sequencing productivity is growing faster than Moore's Law. The human genome project, which ran from 1990 to 2003, costed about $3 billion then and today whole genome sequencing can be done about $1,000 (Bright 2016). Moreover, synthesizing a strand of DNA containing 100 million base pairs cost US $10,000 in 2001 but only 10 cents today (Castillo 2014).

Reading and writing in DNA is slower than in other media, which makes it less suitable for quick retrieval or data processing (Leo 2012). Retrieval process for DNA still is 6 orders of magnitude slower than a PC today (Shrivastava and Badlani 2014). Also, there is no auto correction mechanisms to correct errors during DNA synthesis and this may be a cause of reduced reliability (Shrivastava and Badlani 2014). Several types of errors are related with the current machines dealing with DNA.

However, technology is advancing at an enormous pace. Hand-held, single-molecule DNA sequencers are becoming available e.g. MinION, PromethION, SmidgION etc. These have the potential to vastly simplify reading DNA-encoded information. Other polymers or DNA modifications may also be considered to maximize writing, reading and storage capabilities (Benner *et al.* 2013).

Manipulation of nature is both a boon and a bane. Hence it is necessary for researchers to be vigilant about certain facts and most certainly more research is needed regarding the storage of data in DNA. DNA, if left out in the wild, could get incorporated into a living organism. This, though unlikely because cells tend to expel foreign DNA, cannot be completely ignored. The cell could produce proteins hitherto unknown (unlikely) or die, if this technique is used *in vivo*. Incorporation of DNA into the human genome would not increase individual knowledge because human bodies lack mechanisms to read this DNA and to move its information to the brain (Castillo 2014).

Conclusion

Data storage within DNA may hold the key to all data storage problems of the future and may eventually lead to breakthroughs in Science and Technology. With the rise in big data from new emerging realms of science and technology, like Artificial Intelligence, Machine Learning, Bioinformatics etc., DNA data storage devices may be crucial for storage and retrieval of data. With adequate amount of research and a thorough consideration of the risks involved in the use of DNA as a long-term storage device, DNA may prove to be a better option than all conventional means of storing data.

REFERENCES

Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P. 2003. DNA replication mechanisms. *Molecular Biology of the Cell (4th edition*). Garland Science. New York.

Anonymous. 2013. Where in the world is storage. http://www.idc.com/downloads/where_is_storage_ infographic_ 243338.pdf.

Bancroft C, Bowler T, Bloom B and Clelland K.T. 2001. Long-term storage of information in DNA. *Science* **293**: 1763–65.

Benner S A, Yang Z and Chen F. 2011. Synthetic biology, tinkering biology, and artificial biology. What are we learning? *Comptes Rendus Chimie* **14**(4): 372–87.

Bornholt J, Lopez R, Douglas, Carmean M, Ceze L, Seelig G and Strauss K. 2016. A DNA-Based Archival Storage System.

Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. p. 637–649**.**

Bright P. 2016. Microsoft experiments with DNA storage: 1,000,000,000 TB in a gram. ArsTechnica. https://arstechnica.com/informationtechnology/2016/04/microsoftexperimentswithdnastorage1000000000tbinagram/

Calladine C, Drew H, Luisi B and Travers A. 2004. *An Introduction to Molecular Biology for Non-Scientists*. pp. 1–17. Understanding DNA. Elsevier Academic Press, California, USA.

Carlson R. 2014. Time for new DNA synthesis and sequencing cost curves. http://www.synthesis.cc/2014/02/ time-for-new-cost-curves-2014.html.

Carr P A and Church G M. 2009. Genome engineering. *Nature Biotechnology* **27**: 115-62.

Castillo M. 2014**.** From hard drives to flash drives to DNA drive. *American Journal of Neuroradiology* **35**: 1–2.

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

Clelland C T, Risca V and Bancroft C. 1999. Hiding messages in DNA microdots. *Nature* **399**: 533–34.

Conrad M. 1990. Quantum mechanics and cellular information processing: The self-assembly paradigm. *Biomedica biochimica acta* **49**: 743–55.

Conrad M and Zauner K P. 1997. DNA as a vehicle for the self-assembly model of computing. *BioSystems* **45**: 59–66.

Erlich Y and Zielinski D. 2017. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**(6328): 950–54.

Extance A. 2016. How DNA could store all the world's data. *Nature* **537**: 22–24.

Gibson D G, Glass J I, Lartigue C, Noskov V N, Chuang R Y, Algire M A, Benders G A, Montague M G, Ma L, Moodie M M, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova E A, Young L, Qi Z Q, Segall-Shapiro T H, Calvey C H, Parmar P P, Hutchison C A, Smith H O and Venter J C. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**(5987): 52–56.

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

Gottlieb A and Almasi G. S. 1989. *Highly Parallel Computing*. Benjamin/Cummings Publishing Cooperation. California.

Grass R N, Heckel R, Puddu M, Paunescu D and Stark W. J. 2015. Robust chemical preservation of digital information on DNA *in silica* with error-correcting codes. *Angewandte Chemie International* **54**: 2552–55.

Guo Q, Strauss K, Ceze L and Malvar H. 2016. High-density image storage using approximate memory cells. *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*. Atlanta, Georgia, USA, 02–06 April. pp 413–426.

Herkewitz W. 2016. Scientists Turn Bacteria into Living Hard Drives. *Popular Mechanics*. http://www.popularmechanics.com/science/animals/a21268/scientiststurnbacteriaintolivingharddrives/

Huffman D. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE* **40**(9): 1098–1101.

Kamilaris A, Kartakoullis A and Prenafeta-Boldú F X. 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2017.09.037

Langston J. 2016. UW team stores digital images in DNA — and retrieves them perfectly. UW Today. http://www.washington.edu/news/2016/04/07/uwteamstoresdigitalimagesindnaandretrievesthemperfectly/

Langston J. 2019. With a "hello," Microsoft and UW demonstrate first fully automated DNA data storage. Microsoft.https://news.microsoft.com/innovation-stories/hello-data-dna-storage/

Leier A, Richter C, Banzhaf W and Rauhe H. 2000. Cryptography with DNA binary strands. *Biosystems* **57**(1): 13–22.

Leo R A. 2012. Writing the Book in DNA. Harward Medical School. https://hms.harvard.edu/news/writing-book-dna-8-16-12

Limbachiya D and Gupta M K. 2015. Natural data storage: a review on sending information from now to then via Nature. *Journal on Emerging Technologies in Computing Systems*. arXiv preprint arXiv: 1505.04890.

Miller R. 2011. How Many Data Centers? Emerson Says 500,000. Data Center Knowledge. http://www.datacenterknowledge.com/archives/2011/12/14/how-many-data-centers-emerson-says-500000/

Nguyen H H, Park J, Park S J, Lee C S, Hwang S, Shin Y B, Ha T H and Kim M. 2018. Long-term stability and integrity of plasmid-based DNA data storage. *Polymers* **10**(1): 28.

Niedringhaus T P, Milanova D, Kerby M B, Snyder M P and Barron A E. 2011. Landscape of next-generation sequencing technologies. *Analytical Chemistry* **83**: 4327–434.

Perry K. 2014. DNA can survive re-entry into Earth's atmosphere, *Telegraph Media Group Limited*. http://www.telegraph.co.uk/news/newstopics/howaboutthat/11256420/DNA-can-survive-re-entry-into-Earths-atmosphere.html

Ray S. 2019. DNA Data Storage. https://hackernoon.com/dna-data-storage-d0f0e93513b

Rosenblum A. 2016. Microsoft Reports a Big Leap Forward for DNA Data Storage. MIT Technology Review. https://www.technologyreview.com/s/601851/microsoft-reports-a-big-leap-forward-for-dna-data-storage/

Ross M G, Russ C, Costello M, Hollinger A, Lennon N J, Hegarty R, Nusbaum C and Jaffe D B. 2013. Characterizing and measuring bias in sequence data. *Genome Biology* **14**(5): R51.

Schwartz J J, Lee C and Shendure J. 2012. Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nature Methods* **9**(9): 913–15**.**

Shrivastava S and Badlani R. 2014. Data Storage in DNA. *International Journal of Electrical Energy* **2**(2): 120–24.

Singleton M. 2016. Seagate has built a 60TB SSD, the world's largest. https://www.theverge.com/circuitbreaker/2016/8/10/12424666/seagate-60tb-ssd-worlds-largest

Smith G C, Fiddes C C, Hawkins J P and Cox J P L. 2003. Some possible codes for encrypting data in DNA. *Biotechnology Letters* **25**: 1125–30.

Stewart J. 2011. Global data storage calculated at 295 exabytes. http://www.bbc.com/news/technology-12419672

Sudha P and Valli S. 2017. A study of parallel processing and its contemporary relevance. *International Journal of Computer Science* **5**(20).

Wong P, Wong K and Foote H. 2003. Organic data memory using the DNA approach. *Communications of the ACM* **46**: 95.

Yazdi S M H T, Yuan Y, Ma J, Zhao H and Milenkovic O. 2015. A rewritable, random-access DNA-Based storage system. *Nature Scientific Reports* **5**: 143.

Zhirnov V, Zadegan R M, Sandhu G S, Church G M and Hughes W L. 2016. Nucleic acid memory. *Nature Mater* **15**: 366–70.