Content available at: https://www.ipinnovative.com/open-access-journals

IP International Journal of Comprehensive and Advanced Pharmacology

Journal homepage: https://www.ijcap.in/

Review Article

DNA fountain: As storage system

Darshan Patel[1], Sohan Patel[2,*]

[1] *Smt. S.M.Shah Pharmacy College, Amasaran, Kheda, Gujarat, India*
[2] *Dept. of Pharmacology, SMT. S.M.Shah Pharmacy College, Amasaran, Kheda, Gujarat, India*

ARTICLE INFO

Article history:
Received 19-06-2021
Accepted 07-09-2021
Available online 28-10-2021

Keywords:
Binary code
Genetic code
Data storage
Technology
Encoding
Decoding

ABSTRACT

There are some new technologies introduced by scientist named “DNA Fountain”. This is a way to secureand keep data safely as long as possible. 455EB of data can be encoded in1 gm of single std. DNA. Atearly 20th century researchers did they effort to developed new technology for storing data it should be eco-friendly which does not produce any waste however that the development of that is very struggling andtedious but with the constant effort in this area make that possible. Castillo states that ‘all the informationin the entire internet could be located in device which is lesser than unit cubic inch. Goldman’s and someresearcher’s took a data file and that it was converted into the binary code and that they created effective andefficient relationship between binary code (0, 1) and genetic code (A, C, G, T) and then they synthesizedthe New DNA from the freshly made nucleotide sequences according to the binary code they also achievedsuccess however that, bigger problem is to retrieved the encoded data from DNA that problem solved bythe Bronholt. They developed efficient process for decoding the data from DNA finally, they achieve theall data which is encoded in DNA and also, they developed the calibrated and accurate method for that andthat technology named “DNA Fountain” Now a days that technology is on progress for more innovation inthis area.

1. Introduction

The journey of data storage is started from bones, rocks and Paper. Then this journey drift to punched cards, magnetic tapes, gramophone records, floppies. After this last of CDs, DVDs & Blu-ray discs & flash drivers came into Market.[1]

Rotating Discs are maintaining data for 3 to 5 years and tape is maintaining data for 10-30 years. Here we need some powerful data storage system which has potential to store more data as well as keep their data safely so, we need new storage system. All this storage devices are decay& destroyable & non-biodegradable material that spoil our environment. For increment of digital system for the purpose of generation, transmission & storage information is initially need for active & non destroyable digital media with (massive)large amount of digital data that has to store for future use. The demand for data storage is rapidly increasing day by day. The total information storage of entire world was around 2.7 ZB in 2012.every year the storage necessity is increasing by 50%.

The relic bones genetic material preserve for long time in addition that more Researcher’s works on DNA as a storage medium. DNA has an unbelievable storage capacity. The newly founded storage system named “DNA FOUNTAIN”. Castillo states that ‘all the information in the entire internet could be located in device which is lesser than unit cubic inch.[2] Some researchers said that DNAhas an amazing ability.DNA is extremely dense material with a great theoretical limit above 1EB/mm3 so, it has been observed long lasting with half-life of over 500 years in

* *Corresponding author*.
E-mail address: sohanpharma@gmail.com (S. Patel).

https://doi.org/10.18231/j.ijcaap.2021.023

harsh environment.[3] DNA consisting of adenine, guanine, cytosine & thymine (A, G, C, T). it is always paired of two A-T and G=c. It can be utilized for storing information in form of binary code.

The writing(input) process for DNA storage maps (encode) digital data into DNA nucleotide base sequences synthesize of related DNA nucleotide sequences synthesize of related DNA molecules & storage information. the reading (output) the data which is involved into the sequencingof the DNA molecules and also in decoding the information is retrieve back to the original digital data. Single nucleotide can represent 2bits of information.455EB of data can be encoded in1 gm of single std. DNA.[4] whole world produces information in one year to be stored in just 4 gm of DNA.[1] High memory space is offered by DNA as it is 3D structure. DNA offers readable & reliable &secure information for thousands of years, which can be extended to almost infinity by drying &protecting from o2 and h2o.[4] DDNA can stable a broader range of temperature (-800.c -800.c). the important fact that DNA is invisible to human eye. Ensures that DNA is secure &impossible to be harmed by living organism.[1] Many models of encoding which is used to encode data into DNA. In 1994 DNA based storage system was first introduced encoding and recovering a 23-character contain message.[5] In 2013 researchers was successfully recovered a 739 KB size of message.[6,7]

1.1. Some problems which is necessary to overcome

1. In past due to deficiency of technology DNA synthesis and sequencing was not perfect with full of error.Some 1% per nucleotide sequence can also degrade white stored, further compromising with data integrin.so, it gives an error full result.
2. The biggest problem was an access of data, randomly like our computer and hard disk done.

To read even a single byte of information from storage the entire DNA pool must be sequenced and decode. Which is very time consuming and costly. So, some researchers proposed method for random access that uses PCR to amplify only the desired data, it improved Sequencing towards that data. By This method both accelerates reads and ensures that an entire DNA pool need not be sequenced. They perform some lab experiments to check feasibility of their system in DNA. They performed various random access to read back only selected values. They further investigate in their design (method) using various computer stimulation to understand the error correction characteristic of different encoding schemes, access their overheads and make future based on technology trends. Growth in sequencing productivity eclipses even more's law.

1.2. Effort done in this field

They believed that the DNA storage system is last and golden key of a deep as well as large storage problems.

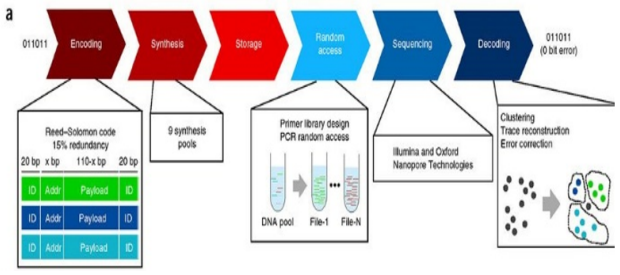


Fig. 1: Steps involved in dna storage

A DNA storage system consists of a DNA synthesizer that encodes the data into DNA Pool which is store into small compartment, and a DNA sequencer instrument that reads DNA sequences and convert them back into digital data (Figure 1). The basic unit of DNA storage has DNA strand that has roughly 100-200 nucleotide long which is capable of storing 50-100 bits total. The DNA strands was stored into "DNA pools" that have stochastic spatial organization and like hard disk and cd it does not permit structure addressing. Therefore, it is necessary to add the address itself into data stored in a strand.[8,9]

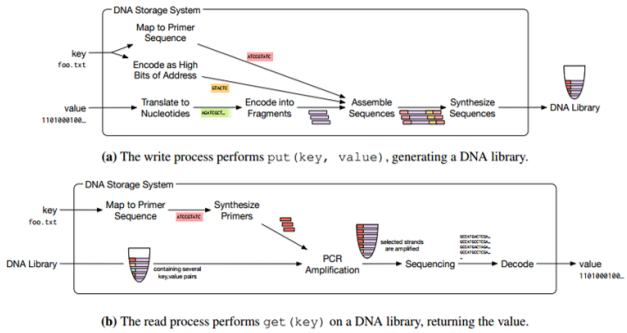


Fig. 2: Overview of a DNA storage system operation as a key value store.

Above figure is shown flowchart for input and output process of DNA storage in more detail. The write (input) process (Figure 2) needs key and value to store input. Key is useful for addressing and to determine the pool in the DNA library where the resulting strands storedand themultiple strands generated by the value. The primer target sequences, to produce final DNA sequence to be synthesized. The resulting DNA molecule is store into DNA library for future.The read (output) process, they needed to key. It is useful into obtained of the PCR primer sequence which was identify the key associated with DNA pool. The sample and PCR primer were sent to the PCR thermocycler, by

use of thermocycler they amplified the desired strands. The resulting pool were further processin the DNA sequences, which produced the digital data readout. In the reading process there are some losses of sample of DNA from the pool So, it reduced quantity of DNA But DNA was easy to replicated, and so the pools can easily be restoring after read operations if needed. Whole DNA pool can be re synthesize after reading process.

1.3. Continues efforts which is done by various scientist

1.3.1. J. bronholt et al.:

The nucleotide is main base of the Data storage system. Its organic molecule consisting of one base (A, C, G, T) and Sugar Phosphate. This storage system mainly based on these 4 bases.as the results of some famous scientist new approach to stored binary data in DNA. It was quite difficult but it possible by great effort of the scientist and nucleotide base pair. The Quaternary digit can then be mapped to DNA nucleotides by producing string of n/2 digit from binary bits. (ex. mapping 0,1,2,3, to A, C, G, T, respectively). For example, the binary string 011001 maps to the base 4 string 1201, and then to DNA sequence CGAC. However, the DNA sequences and synthesis are very complicated it arise manyerrors so, it requires a more careful encoding. Some error is eliminated or reduce by encoding binary data in base 3 instead of base 4.[7] to avoid the repetitions of same nucleotide they maps ternary digit to DNA nucleotide. This encoding avoids homo polymers-repetitions of the same nucleotide that significantly increasing the chance of sequencing error.[9]

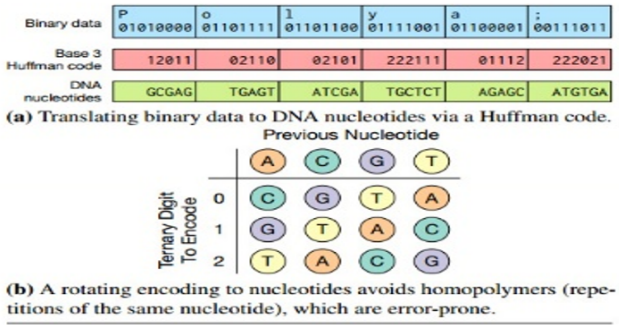


Fig. 3: Encoding a stream of binary data as a stream of nucleotides. A Huffman code translates binary to ternary digits, and a rotating encoding translates ternary digits to nucleotide.

Because base 3 is not a multiple of base 2, mapping directly between the bases would be inefficient: 6 ternary digits (36 =729) can store 9 bits of data (29 =512), but waste 217 possible states. Instead, they used a Huffman code[10] that maps each binary byte to either 5 or 6 ternary digits.

For example, the Huffman code maps the binary string 01100001

To the base-3 string 01112. The rotating nucleotide encoding Maps this string to the DNA sequence CTCTG. by the help of Huffman code, they map ASCII character to 5- digit string.

1.4. Data format

Another practical problem is that they don't have any synthesis technology to synthesize small length of sequences of nucleotides. Data is existing in hundreds of bits therefore cannot be synthesize as single strand of DNA. DNA pool do not perform spatial isolation, and so they addedsome keys which is irrelevant to a single read operation.Isolation of interested molecule and exiting DNA storage technology sequence entire pool which increase significant cost and time. To solve these two problems, they organized data in DNA in similar fashion to Goldman et al.[7]

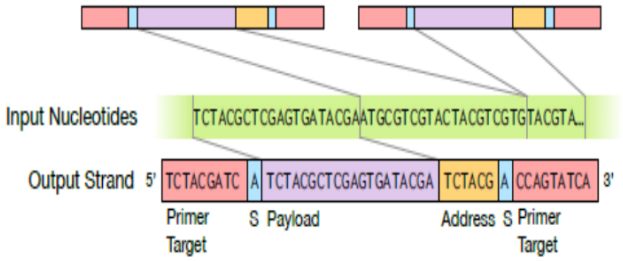


Fig. 4: Overview of the DNA data encoding format. After translating to nucleotides, the stream is divided into stands. Each strand containing payload from the stream, together with addressing information to identify the strand and primer targets necessary for PCR and sequencing.

Above Figure 4 shown the segment of nucleotide is divided into the block, which they synthesize a separate strand, so they get large storage capacity. Connect those strands with the identifying primers allows the read process to isolate the main interest of data molecule and so perform random Access. They add these different keys into our DNA sequence:

1. **Payload:** It is the sequence of nucleotides representing the data to stored is broken into data blocks, whose length depends on the desired length and additional overheads of format, to aid decoding, two sense nucleotide "s" indicate whether the strands has been reversing complemented.
2. **Address:** Each data block is containing addressing information to identify its location in the input data sequences. The address space is mainly two part first is high part of the address identifies the key a block is associated with. Second the low part of the address index the block within the value associated with that key. The combine address is padded to a fixed length and converted to nucleotide as described above. A parity nucleotide is added for basic error detection.

3. **Primers:** Each end of strands we attach the primer sequences. These sequences serve as "foothold" for the PCR process, and allow the PCR to selectively amplify only those strands with a chosen primer sequence.

1.5. Encoding system for storage

In previous section study about organization of DNA storage system and how they store information by blocking system. They store a data in to DNA by broken strands of nucleotide sequences. It relies on the robustness of DNA for durability because each bit of data is encoded in exactly one location in the output DNA. Some Early work done by scientist they used simpler encodings technique For example, Bancroft et al.[2,5] translate text to DNA by means of a simple ternary encoding: each of the 26 English characters and a space character maps to a sequence of three nucleotides drawn from A, C, and T (so exactly 33 = 27 characters can be represented). They successfully recovered a message of 106 characters, but this encoding suffers substantial overheads and poor reliability for longer messages.

1.6. Goldman encoding

Let's focus on an existing encoding proposed by Goldman et al.[7] shown in Figure. This encoding is divided DNA nucleotide into overlapping segment to provide four-fold redundancy for each segment. This encoding provides high reliability. The Goldman used this encoding to successfully recover a 739 Kb message. He uses this encoding as a baseline because that time it is most popular DNA techniquein addition, it offers a tunable level of redundancy, by reducing the width of the segments and therefore repeating them more often in strands of the same length.

1.6.1. The experiment

Goldman and his team are work done on high capacity, low maintenance storage of digital information in synthesized DNA. They encoded computer files sizes of 739 KB of hard disk storage by help of Shannon information[11] of 5.2*106 bits into a DNA code and synthesized these DNA sequenced it and redeveloped the original files with 100% accuracy.[8] A series of experiments and their results proves DNA storage to be a realistic technology for large scale digital archiving that may already be cost effective for low access.

They understand and study the other DNA storage approaches problem. They developed an in-vitro approaches that represents the information is stared as a long DNA molecule and encodes this using shorter DNA fragment as same as church et al.[6] They selected computer files and then encoded this into DNA. The five files comprised all 154 of Shakespeare's sonnets (ASCII text), a classic scientific paper.[12] (pdf format), a medium resolution color photograph of the European bioinformatics institutes (JPEG 2000 format), a 265 exert from martin Luther king's 1963 "I have a Dream" speech (MP3 format). They used Huffman code to study to convert bytes to base 3- digits (ASCII text), and that produce a total of 757,051 bytes (Shannon information (11, 13) 5.2*106 bits). These five files were represented by a total of 153,335 strings of DNA, each string is comprising 117-nt.[13]

Fig. 5: Image file of European bioinformatics institutes.

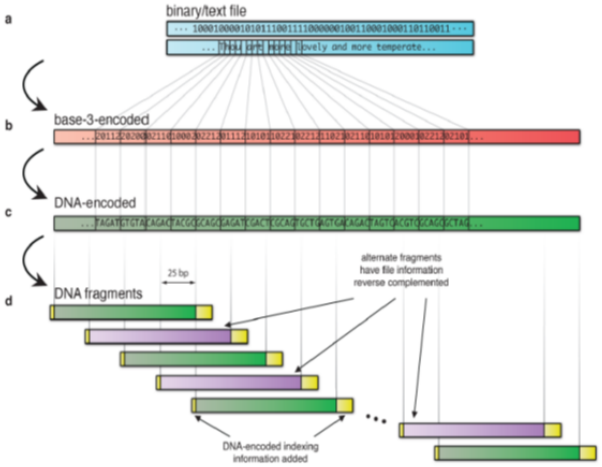


Fig. 6: Digital information encoding in DNA

Digital information (a, in blue), here binary digits holding the ASCII codes for part of Shakespeare's sonnet 18, was converted to base-3 (b, red) using a Huffman code that replaces each byte with five or six base-3 digits (trits).[14] This in turn was converted in silico to our DNA code (c, green) by replacement of each digit with one of the three nucleotides different from the previous one used, so, there is ensure that no homopolymers were generate at this basis formation of large number of overlapping segments of length of 100 with 75 base pair is occur, creating fourfold redundancy (d, green and, with alternate segments reverse complemented for added data security, violet). Indexing DNA codes were added (yellow), also encoded as non-repeating DNA nucleotides. An additional advantage of their encoding scheme that the fragment length is perfect and uniform and absence of Homopolymers. So, obviously the synthesized DNA does not have a natural

(biological) origin and the presence of aimful design encoded information.[15] They designed DNA strings using an updated version of Agilent technologies OLS (Oligo Library Synthesis) process.[16] They created a large number 2.5*106 of copies of each DNA string, with low error (1 error per 500 bases). Then they supply a lyophilized to synthesized DNA for excellent long-term preservation characteristics[17,18] and then this synthesized DNA was shipped (at ambient temperature, specified packaging) from the USA to Germany via the U.K and then they performed resuspension, amplification and purification a sample of resulting library product. Then it was sequenced in paired end mode on an Illumina HIseq 2000 and it was transferred to multiple aliquots and re-lyophilized for long term storage. the full length (117-nt) DNA strings were reconstructed in silico from the read pairs, with those containing uncertainties due to synthesis or sequencing error being discard by using reverse procedure of encoding. This discard string has information is recovered with more sophisticated decoding. So, they prove that DNA storage a potential as a practical solution to the digital archiving problem and may become a cost-effective solution for recovery assessed archives.

Fig. 7: Illumina HIseq 2000

1.7. XOR encoding

While Goldman encoding provide high reliability. It suffers significant overhead: each block in the input string is repeated four types. They propose a simple new encoding that provides similar levels of redundancy to prior work, but with reduced overhead.Encoding shown in figure They tookA⊕B as payload A and B had two strands Which produce new payload. the high bit of the address is used to indicate whether a strand is an original payload or an exclusive or strand. These provided redundancy. Any of this payload is sufficient for recover third.

1.8. James B et al. experiments:

They perform the experiment on random access capability of DNA storage. they encoded four image files using two

Fig. 8: PCR thermo cycler

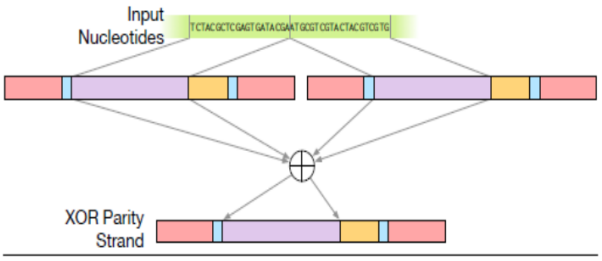


Fig. 9: Our proposed encoding incorporates redundancy by taking the exclusive or of two payloads to form a third. Recovering ant two of the three stands is sufficient to recover the third.

different encoding method.

They performed experiment that they took various in files size from 5Kb to 84 Kb. They synthesize and sequencing these files and resulting DNA to recover the files. They used four images' files for input to the DNA based storage system for each image file x. jpg, they generated DNA sequence related to the output of images (x. jpg. . .). They perform the experiments using two methods.

1. Goldman encoding
2. XOR encoding

They performed this experiment on four images, they used Goldman encoding method for three images and other one is encoded by XOR encoding system. (the Sydney. jpg image). Combine the 8 practical produce 45652 sequences of length 120 nucleotides Represents 151 Kb of datato demonstrate that DNA based storage system allows to effective random access. They synthesized sequence were prepared for sequencing by amplification via the PCR method. The product was sequenced using on Illumina Mises platform. The selected get operation total 16,994 sequences and 42 Kb produced 20.8 M reads of sequences in the pool.[8] They inspected the result and observed no.

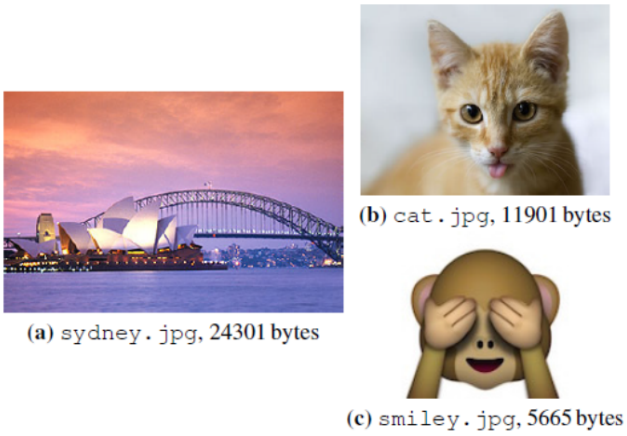


Fig. 10: Three images files we synthesized and sequenced for our experiments.

of reads of sequence that were not selected. So, random access was effective in amplifying only target files. They successfully recovered all four files from sequenced DNA. They conclude that the sequencing depth is reduced so it will give batter results.[8]

1.9. Church et al.:

They developed strategy to encoded arbitatory digital information by using a novel encoding scheme that uses next generation DNA synthesis and sequencing technology. They converted HTML coded draft of a book that included 53,426 words of jpg images and one java script program into 5.27 mega bite bitstream.[4] Then they encoded these bits onto 54,898.[14] 159-nt oligonucleotides, each encoding 96-bit data block (96-nt), A 19-bit address specifying the location of the data block in the bits stream (19-nt) and flanking of 22-nt common sequences for amplification and sequencing. This DNA library pool is synthesized by ink-jet printer, highly fidelity DNA microchips.[14,16] To read this encoded DNA it is necessary to be amplified the DNA library by limited cycle of PCR in thermo cycler and then sequenced on single lane of an Illumina Hi seq. Then they joined overlapping paired end 100-nt. Reads to reduce the effect of sequencing error.[19]

Then only expected 115-nt length and perfect barcode sequences the generated consequence at each base of each data block at an Avg of ~3000-fold coverage. (Fig). All data blocks were recovered with a total of 10 bits error out of 5.27 million (fig). Their method has at least 5 adv. over past DNA storage approaches. They encoded one bit per base instead of two (A or C for 0, G or T for 1). So, they can encode message many ways in order to avoid sequences that are difficult to read or write. By divided the bitstream into address data blocks, they eliminate the need for long DNA constructs. That are difficult to assemble at this scale.

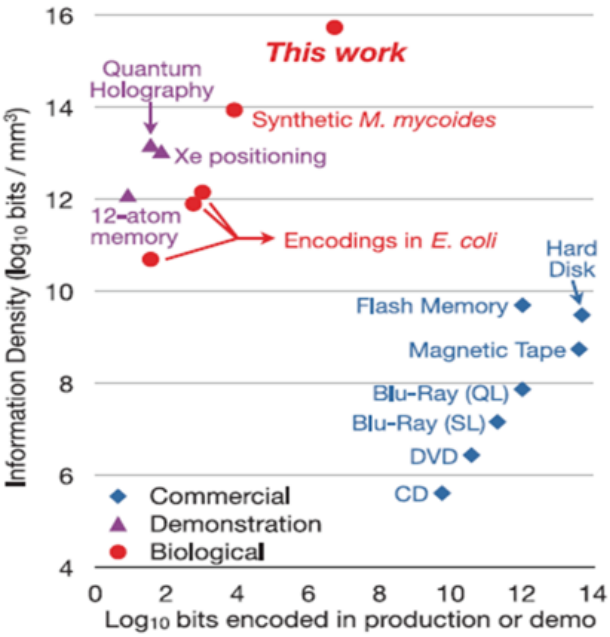


Fig. 11: Comparison to other technologies. We plotted information density (log10 of bits/mm3) versus Current scalability as measured by the log10 of bits encoded in the report or commercial unit.

They synthesized, store and sequenced many copies of each individual oligo. They use purely in-vitro approaches that avoids cloning and stability issue of in-vivo approaches.

So, by this experiment they concluded next generation technologies in both DNA synthesize and sequencing to allow for encoding and decoding of large amount of information for 1, 00,000-fold less cost than first generation encoding.

1.10. Leon anavy et al.:

Oligonucleotide multiplicity, which is an important inherent property of current DNA synthesis and sequencing technologies is not exploited by the aforementioned work. They introduced c DNA letters that constructs and utilize this multiplicity and so, they able to increase the information capacity per synthesized portion. a composite DNA letter is a representation of a position in a sequence that constitutes a mixture of all four standard DNA nucleotides in a specified predetermine ratio. They describe that a composite DNA letters from the basis to a DNA synthesis approach that trades sequence multiplicity for increased the complexity of synthesized DNA effectively and it has higher data capacity per synthesized position. In the early days of DNA sequencings by hybridization, degenerate and semi-degenerate bases were proposed as wildcards for increasing the fidelity of the system.[20–22] Next generation DNA sequence have higher quality and capacity when using degenerate base addition together with error

correction approaches.[23] They demonstrated practical on implementation of a complete large-scale composite DNA storage system by demonstration commercially available DNA synthesis and sequencesing techniques. Their method is superior to the previous method. They improved capacity of system implements an error correction scheme that combines an adaption of the previously repeated fountain code.[24] They used composite DNA coding system to repeat the original DNA fountain experiment and increased 24% capacity per synthesized position. They stared a composite file contain an HTML version of the bible in both Hebrew and English taken from the mamre institute.[25]

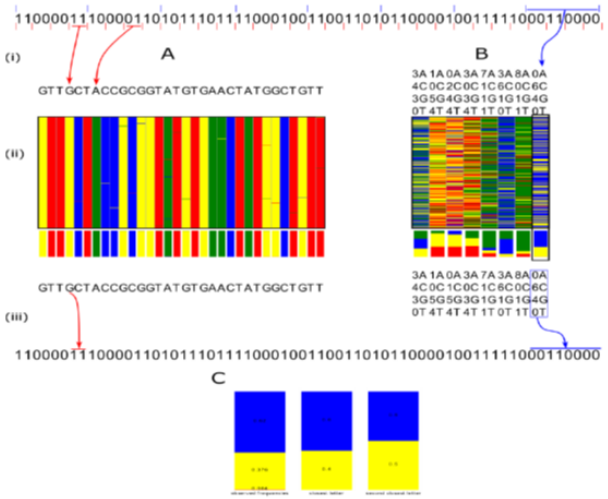


Fig. 12: Encodinga binary message using standard and composite DNA.

1.10.1. Encoding a binary message using standard and composite DNA

A binary message, depicted on top, is encoded into DNA. A. Standard DNA based storage scheme9. The binary message is being encoded to DNA by mapping every 2 bits (depicted by the short red separating lines) to a DNA base or synthesized position (is), the designed DNA sequence will then be synthesized and sequenced by a noisy procedure that introduces some errors (ii). The sequencing output is then used to infer the DNA composition at every position (iii). Decoding of the original message is done assuming the use of an error correcting code over the binary message. B. The same message is encoded using a composite DNA alphabet of resolution k=10 by mapping every 8 bits (depicted by the blue separating lines) of the binary message to a single composite DNA position. Using sufficiently deep sequencing allows to correctly identify the original composite letters (the right most position, in a black frame, is exemplified in C) and to decode the message. The decoding also uses an error correction mechanism (Reed-Solomon over the appropriate finite field, in our implementation), over the composite alphabet. C. An example of the inference step at a single synthesized position. The observed frequencies are used to infer the source, σ= (0, 6, 4, 0), as the closest composite letter, using KL divergence (see text and Online Methods). Note that the inference at any fixed position is affected by the sequencing depth obtained there as well as by sequencing and synthesis error.[25]

1.10.2. The experiment

They used this equation to calculate capacity for storage information:

$$Capacity = \frac{Length\ of\ binary\ message\ i\ after\ huffman\ coding}{Length\ of\ composite\ message\ i\ after\ huffman\ coding}$$

They analysed all performance of function of process to better evaluate challenges and improvements of composite DNA based data storage. They performed a large-scale molecular implementation of a six-letter composite alphabet storage system. This encoded DNA is successfully retrieving the same 2.12 MB data file from erlich et al.[24] That DNA pool consisted of 58,000 six letter composite oligo of length 152-nt, compared to 72,000 oligoes of same length required using standard DNA and then they increased 24% information carrying capacity per synthesized position and they make decoding pipeline that is allowing the correction systematics synthesis biases. We understand this pipeline below.

A compressed input file is being processed by the fountain code to produce binary droplets. A composite DNA encoding flow is then applied on each droplet consisting of the following steps (See Online Methods for details): (is) the binary message is translated into a composite DNA sequence. The seed sequence is translated to standard DNA sequence, which will serve as a barcode for the decoding process. The payload is translated to a six-letter composite DNA alphabet (Σ6) in 5-bit chunks. (ii) Error correction nucleotides are added to the DNA sequence by using a systematic Reed-Solomon (RS) encoding. The barcode is encoded using RS over $GF(13)$ and the payload is padded and encoded using RS over GF. (iii) Each encoded message is then filtered to verify that the RS redundancy letters are all from Σ6. (iv) Experiment identifier and amplification template sequences are appended to each valid sequence. They also examined the minimal sequencing depth required to decode the message correctly for each one of the four-composite alphabet for higher resolution required deeper sequencing. They proved the concept of composite DNA, the properties of current DNA synthesize and sequencing process, to potentially attain higher density DNA based storage system. They proved and improved and implement in other approaches to increase capacity and fidelity of DNA based storage system, such as orthogonal base pair system,[26] efficient coding[13,24,27,28] andrandom-access approaches[13,29–32] incorporating composite DNA

based storage system will require further investment in future.

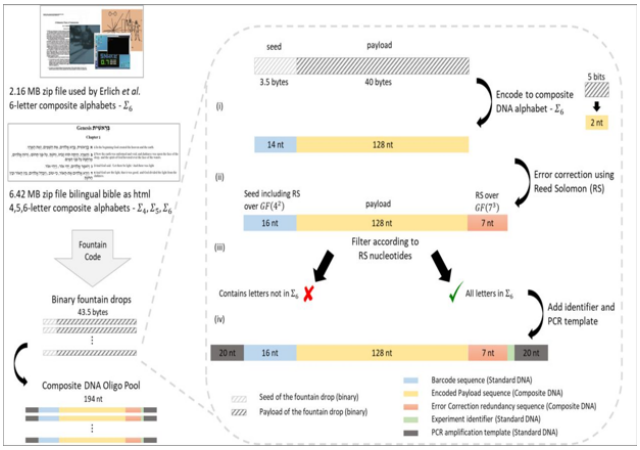
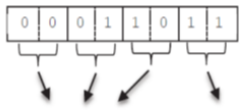


Fig. 13: Encoding pipeline of a large-scale composite DNA based data storage

1.11. M. blawat et al.:

M. blawat and his team also worked on a storage capacity boost and scheme for error correction method. They reported that strong capacity boost strength of storing digital data in synthetic DNA. They also developed an efficient and robust forward error correcting scheme adapted. To the DNA channel they used designed DNA channel model on data from a proof of concept conducted 2012 by a team from the Harvard medical school.[6] They introduce their own method or scheme which is eliminate the all type of error of today's DNA synthesis, amplification and sequencing process ex. Insertion, deletion and swap error by use of their method or scheme.[33] They able to store and retrieve error free 22 Mbyte of digital data in synthetic DNA recently.(34) They also proves that the practically uses of synthetic DNA as long -term Digital data storage system. They analysis of the experiment data of church and his team gathered and produced a new designed forward error correction (FEC) scheme.[6] Ex. Insertion and deletion and swap. They observed one type of Swap error occurred in an oligo, if a nucleotide had been replaced with incorrect one, at that time oligo length stays unchanged. so, an insertion or deletion error occurred in oligo, if their an addition nucleotide has been inserted or removed, the predominant error type the affect oligo corresponds lengthened or shorted. In the experiment data of church and his team. They found that the swap error rate lies between 6.0*10-4& 1.4*10-3, while insertion and deletation error rates are 1.0*10-3, and 5.0*10-3respectively (Blawat). They also fined some sequencing, which is not fined in read sequencing is called "missing oligoes". And they also prove that the DNA storage system is not a memoryless data channel.



Value	Nucleotide
00	A
01	C
10	G
11	T

Table 1: Mapping of first 3 bit-tupel to 3 nucleotides

Value	Option 1	Option 2	Option 3	Option 4
00	AA	CC	GG	TT
01	AC	CG	GT	TA
10	AG	CT	GA	TC
11	AT	CA	GC	TG

Table 2: Mapping of last 2 bits to a pair of nucleotides

Fig. 14: Experimentassumption of Blawat M. et. al.

They synthesized 900 000 230 nt oligonucleotides on Agilent's oligo library synthesis (OLS) microarray platform, divided into four libraries with 225000 oligonucleotides in 100 μL TE. Illumina specific sequencing adaptors were introduced into the synthesized OLC pool in a two-stage serial PCR amplification using the syBR FAST MASTER mix. Reaction was performed using the following protocol on an Eppendorf master cycler realplex 4 real time PCR machine by monitoring the syBR green channel signal.[33] Each reaction was harvested after and cycle of amplification to avoid PCR bias in the resulting library the resulting PCR products after each stage were purified using agencourt amppureXP beads according to manufacturer's instruction.

Fig. 15: Illumina HIseq. 2500 next generation sequencer.

They sequenced the amplified library by loading 1mL of 16mL library on 2 LANs of a rapid sequence 300 cycle SBS kit an Illumina HIseq. 2500 next generation sequencer. They obtained 144,475,005 paired reads with 83.78 % of the reads

Fig. 16: ThesyBR FAST MASTER Mix.

scoring >=Q30. They are successfully work done on storage capacity boost and sequencing of DNA.[33]

We discussed about various DNA data storage approaches which has important role into future of digital data storage into DNA they work at all learn problem and get their solution and give some more efficient, and secure technique is provide to us for our better future.

1.12. Application of digital data storage DNA fountain

1. Biggest application of DNA fountain is to store digital data in more amount.
2. Store a sensitive and secrete data more efficiently with more safety and more security.
3. Huge amount of data is store in only 1 gm of DNA sample.
4. For storage of more information it required more space but by use of this technique we storage of 1 PB data in 1 gm of DNA.
5. We store a data for long time such as billions of years by storage DNA into -180ºC.
6. For long time there is no destroy of our data.

2. Conclusion

We concluded that above review article the DNA storage: DNA fountain is very good option to solve future data storage, security of data problem. In this article we seen step by step evolution into this field. Many of scientist is give big contribution in this field. They performed many series of experiments and result they developed a huge data storage technique for future storage of data. First the concept was introduced by church and his team. They developed strategy to encoded this binary digit or digital information into synthesized DNA by next generation DNA sequencesing. They work on HTML coded draft of book. After this church Leon anvy and his team is work on a capacity of data storage system. They derived new method of encoding of data storage is “encoding pipeline” for increasing the information storage capacity. They derived one “equation of capacity” and they increase the binary message length and decrease the composite message length so; they use this eq. increase the capacity of storing information. And then Goldman and his team are work on capacity as well as maintenance of storage data and they developed high capacity and low maintenance storage of digital information in synthesized DNA. After Goldman M. Blawat and his team is more work on a Goldman work means capacity of storage. M. Blawat developed new method for boosting of capacity and error correction. They mainly work on error which is generated during encoding of data into synthesized DNA. They minimize the error by their method. and last Bronholt and his team is contribution on encoding, storage capacity, error correction, and retrieving of encoded data securely and without any mistake or error. They successfully retrieved all data they are encoded into synthesized DNA. So, that are scientist and their contribution to this field. Recently more research work on this topic for our bright future of digital storage world.

3. Conflict of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

4. Source of Funding

None.

References

1. Shrivastava S, Badlani R. Data storage in DNA. *Int J Electr Energy*. 2014;2:119–24.
2. Castillo M. From hard drives to flash drives to DNA drives. *Am J Neuroradiol*. 2014;35(1):1–2.
3. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci*. 1748;279(1748):4724–33. doi:10.1098/rspb.2012.1745.
4. ExtremeTech. New optical laser can increase DVD storage up to one petabyte; 2020. Available from: http://www.extremetech.com/computing/159245-new-optical-laser-canincrease-dvd-storage-up-to-one-petabyte2013.
5. Clelland CT, Risca V, Bancroft C. Hiding messages in DNA microdots. *Nature*. 1999;399:533–4.
6. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. *Science*. 2012;337(6102):1628.

7. Goldman N, Bertone P, Chen S, Dessimoz C, Leproust EM, Sipos B, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*. 2013;494(7435):77–80.
8. Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G, Strauss K. A DNA-based archival storage system. Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. 2016;.
9. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. *Analytical Chem*. 2011;83(12):4327–41.
10. Huffman DA. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*. 1952;40(9):1098–101.
11. Mackay DJ, Kay M, J D. Cambridge university press. 2003.
12. Watson JD, Crick FH. Molecular structure of nucleic acids. *Nature*. 1953;171(4356):737–8.
13. Yazdi SHT, Yuan Y, Ma J, Zhao H, Milenkovic O. A rewritable, random-access DNA-based storage system. *Scientific Rep*. 2015;5(1):1–10.
14. Nikolados EM. Storing data at the tip of a pencil. In: medium, editor. medium articles. online: medium; 2019. Available from: https://medium.com/@evangelosmariosnikolados/storing-data-at-the-tip-of-a-pencil-a0dd155042be.
15. Cox JP. Long-term data storage in DNA. *TRENDS Biotechnol*. 2001;19(7):247–50.
16. Leproust EM, Peck BJ, Spirin K, Mccuen HB, Moore B, Namsaraev E, et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res*. 2010;38(8):2522–40.
17. Anchordoquy TJ, Molina MC. Preservation of DNA. *Cell Preservation Tech*. 2007;5:180–8.
18. Bonnet J, Colotte M, Coudy D, Couallier V, Portier J, Morin B. Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucleic Acids Res*. 2010;38(5):1531–46. doi:10.1093/nar/gkp1060.
19. J. St. John, SeqPrep (2011), https://github.com/jstjohn/ SeqPrep.
20. Bains W. Hybridization methods for DNA sequencing. *Genomics*. 1991;11(2):294–301.
21. Pevzner PA. Rearrangements of DNA sequences and SBH. *Comput Chem*. 1994;18:221–3.
22. Preparata FP, Oliver JS. DNA sequencing by hybridization using semi-degenerate bases. *J Comput Biol*. 2004;11(4):753–65.
23. Chen Z, Zhou W, Qiao S, Kang L, Duan H, Xie XS, et al. Highly accurate fluorogenic DNA sequencing with information theory-based error correction. *Nat Biotechnol*. 2017;35(12):1170–8.
24. Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. *Science*. 2017;355(6328):950–4.
25. Anavy L, Vaknin I, Atar O, Amit R, Yakhini Z. Improved DNA based storage capacity and fidelity using composite DNA letters. *bioRxiv*. 2018;doi:10.1101/433524.
26. Jiménez-Sánchez A. DNA computer code based on expanded genetic alphabet. *Eur J Comput Sci Inf Technol*. 2014;2:8–20.
27. Yazdi SHT, Kiah HM, Garcia-Ruiz E, Ma J, Zhao H, Milenkovic O, et al. DNA-based storage: Trends and methods. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*. 2015;1(3):230–48.
28. Gabrys R, Kiah HM, Milenkovic O. Asymmetric Lee distance codes for DNA-based storage. *IEEE Trans Inf Theory*. 2017;63(8):4982–95.
29. Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G, Strauss K, et al. Toward a DNA-based archival storage system. *Ieee Micro*. 2017;37(3):98–104.
30. Yazdi ST, Kiah HM, Gabrys R, Milenkovic O. Mutually uncorrelated primers for DNA-based data storage. *IEEE Transactions on Information Theory*. 2018;64(9):6283–96.
31. Raviv N, Schwartz M, Yaakobi E. Rank modulation codes for DNA storage. *2017 IEEE International Symposium on Information Theory (ISIT)*. 2017;p. 3125–9.
32. Levy M, Yaakobi E. Mutually uncorrelated codes for DNA storage. *IEEE Trans Inf Theory*. 2018;65(6):3671–91.
33. Blawat M, Gaedke K, Huetter I, Chen XM, Turczyk B, Inverso S, et al. Forward error correction for DNA data storage. *Procedia Computer Sci*. 2016;80:1011–22. doi:10.1016/j.procs.2016.05.398.

Author biography

Darshan Patel, Student

Sohan Patel, Assistant Professor

Cite this article: Patel D, Patel S. DNA fountain: As storage system. *IP Int J Comprehensive Adv Pharmacol* 2021;6(3):126-135.