Extended nucleic acid memory as the future of data storage technology

Saptarshi Biswas*

Department of Computer Science and Engineering,
Meghnad Saha Institute of Technology,
Kolkata 700150, India
Email: saptarshi.biswas9@gmail.com
*Corresponding author

Subhrapratim Nath and Jamuna Kanta Sing

Department of Computer Science and Engineering,
Jadavpur University,
Kolkata 700032, India
Email: suvro.n@gmail.com
Email: jksing@ieee.org

Subir Kumar Sarkar

Department of Electronics and Telecommunication Engineering,
Jadavpur University,
Kolkata 700032, India
Email: su_sircir@yahoo.co.in

Abstract: The amount of operational data being generated at an exponential rate in various spheres of computing, in turn, has culminated in a pressure on the available silicon memory-constrained by its limited capacity. In recent times, research has been undertaken on DNA computing for memory technology where nucleic acid memory (NAM) was formulated and found to be an efficient alternative for storing a large amount of digital data in the molecular space. This work presents a new encoding scheme which efficiently maps the binary data into a hybrid system of standard as well as non-standard genetic nucleotides to achieve a higher data capacity. Comparative studies have been done with existing encoding schemes, moreover, this work demonstrates the use of unnatural base pairs like Ds-Px and Im-Na which exhibit high stability and high selectivity in a DNA molecule.

Keywords: memory technology; nucleic acid memory; NAM; unnatural base pair; non-standard nucleotide.

Reference to this paper should be made as follows: Biswas, S., Nath, S., Sing, J.K. and Sarkar, S.K. (2020) 'Extended nucleic acid memory as the future of data storage technology', *Int. J. Nano and Biomaterials*, Vol. 9, Nos. 1/2, pp.2–17.

Biographical notes: Saptarshi Biswas is a 4th-year (final year) BTech (Bachelor of Technology) student and is studying Computer Science and Engineering at Meghnad Saha Institute of Technology, Kolkata, India. He is associated with some of the research projects like VLSI Global Routing along with Jadavpur University, Kolkata in collaboration with his parent institution Meghnad Saha Institute of Technology. His current research interests include artificial intelligence, computational biology, memory technology, etc.

Subhrapratim Nath is currently a Head and Assistant Professor of Computer Science and Engineering Department, Meghnad Saha Institute of Technology, (Techno India Group), India. He received his BTech in Electronics and Telecommunication Engineering and MTech in Software Engineering. He is currently pursuing his PhD degree in the Department of Computer Science and Engineering at Jadavpur University, Kolkata. He has published more than 25 technical research papers in international/national journals and peer-reviewed conferences. His research focuses in the area of VLSI physical design, MANET, network security, computational biology and algorithms using soft computing tools. He is also a member of IEEE, IE (India) and CSI.

Jamuna Kanta Sing is currently a Professor at the Department of Computer Science and Engineering, Jadavpur University, India. He received his PhD (Tech.) degree from Jadavpur University, Kolkata, India. He has completed five R&D projects sponsored by different Government of India funding agencies. He has more than 100 Conference and Journal papers and has published six books. His works includes face recognition, video processing, medical image processing and computational intelligence. He is also a senior member of IEEE.

Subir Kumar Sarkar is currently a Professor of Electronics and Telecommunication Engineering Department, Jadavpur University, India. received his PhD (Tech.) in Microelectronics from the University of Calcutta, Kolkata, India. He has completed 18 R&D projects sponsored by different Government of India funding agencies and published more than 590 technical research papers in international/national journals and peer-reviewed conferences. His research interest includes nano-devices and low power VLSI circuits, computer networks, digital watermarking and RFID. He is also a senior member of IEEE, IEEE EDS distinguished lecturer, a life fellow of IE (India) and IETE, life member of ISTE and life member of IACS.

This paper is a revised and expanded version of a paper entitled ‘Storing digital data in nucleic acid memory with extended genetic alphabet’, presented at the 3rd International Conference on 2019 Devices for Integrated Circuit (DevIC 2019) Kalyani Government Engineering College, Kalyani, India, 23–24 March 2019.

1 Introduction

A cell is the atomic unit of every living organism. It was estimated through a study that the approximate number of cells contained in a human body is 3.72×10^{13} (Bianconi et al., 2013). The behaviour and characteristics of a living cell are defined by the genetic content of the cell. A cell is composed of various organelles. Generally, there are two organelles, namely the nucleus and the mitochondria, which contain the genetic material.

Every living cell has nucleic acids as the elementary genetic material. It is a biochemical compound that has the capability to store any form of biological information of a living organism. There are two basic types of nucleic acids. One is the deoxyribonucleic acid (DNA) and the other one is the ribonucleic acid (RNA). With respect to the structure, the nucleic acids can be further classified into a single-stranded and double-stranded nucleic acid. In the single-stranded configuration, the molecule exists in the form of a single-stranded structure whereas, in double-stranded configuration, two single-stranded molecules remain conjugated with each other through intermolecular hydrogen bonds (H-bond) and remain in the form of a double-stranded helical structure. The intermolecular H-bond stabilises the molecule which in turn causes the double-stranded nucleic acids to be more stable chemically than the single-stranded nucleic acid.

There has been an exponential surge in the amount of data produced globally within the last decade and it is also expected that the rate of data generation will continue the same way in the future. It is expected that the silicon memory will get depleted completely in the near future. The nucleic acid memory is chosen as the best alternative due to its numerous advantages. These include high scalability, very high retentivity and data density, very low latency and high robustness (Zhirnov et al., 2016). The cost of sequencing of DNA molecule has also promisingly reduced to economic feasibility. As per the reported results, the cost of sequencing of DNA has reduced to 9×10^{16} bits/USD (Zhirnov et al., 2016). This has led to the inclination towards nucleic acid memory. A comparative analysis of nucleic acid memory with other kinds of storage technologies is presented in Table 1 (Zhirnov et al., 2016). All the data presented in Table 1 are in their approximate values.

Table 1 Comparison between various memory technologies with NAM

Metrics	*Hard disk*	*Flash memory*	*DRAM*	*NAM*
Read/write latency	3–5 ms per bit	100 μs per bit	< 10 ns per bit	< 100 μs per bit
Retention	> 10 years	10 years	64 ms	> 100 years
ON power	0.04 W/GB	0.01–0.04 W/GB	0.4 W/ GB	$< 10^{-10}$ W/GB
Volumetric density	10^{13} bit cm^{-3}	10^{16} bit cm^{-3}	10^{13} bit cm^{-3}	10^{19} bit cm^{-3}

There occur four standard nucleotides in a standard RNA or DNA. Later numerous unnatural nucleotides were designed artificially in the laboratory conditions. It was observed and reported that these unnatural nucleotides behave in the same way as the natural nucleotides do in a living cell (Hirao et al., 2006a). It was also reported that the unnatural nucleotides not only respond to the DNA amplification process PCR (Polymerase Chain Reaction) (Hirao et al., 2006a, 2006b 2007) but also has the capacity to take part in the transcription process and exist within the RNA (Hirao et al., 2006a) just like the standard nucleotides. A recent study reported that it was made possible to design synthetic DNA of *Escherichia Coli* bacteria by injecting an unnatural base pair within the standard DNA molecule (Malyshev et al., 2014). It was also reported that the unnatural base pair not only sustained within the bacterial cell but also participated in the replication process under the natural intracellular conditions.

This proposed paper has taken into consideration the application of both standards as well as non-standard nucleotides within a single system. Each nucleotide is assigned with a 3-bit binary number. This has enhanced the encoding to represent two 3N binary numbers with N number of nucleotides.

2 Fundamentals on DNA and its Molecular Structure

The DNA serves as the genetic material of every living organism. By observing the real world scenario one can find innumerable common properties between successive generations of a living organism. The phenomenon by which an organism passes some common features to its descendent is known as inheritance. This phenomenon is made possible through the DNA of the parent organisms. It can be visualised in a manner that the DNA of the parent organism stores some information that is passed to the next generation. Due to this reason, the DNA is sometimes termed as the genetic memory of an organism. The information about the appearance of an organism and its natural behaviour is also stored in the genetic memory.

The combinations of DNA base pairs encode the information of all the proteins that are needed for the proper functionality of a living cell and to carry out its life process. As a result, it forms the basis of all the metabolic and physiological activities of a living cell. The DNA also plays a vital role in maintaining biological rhythms of a living organism, for, e.g., the sleep and wake cycle of human beings, etc., and is known as the genetic clock (Badiu, 2003). It holds all the information of the proteins needed to sustain the rhythmic metabolic activities within a living being.

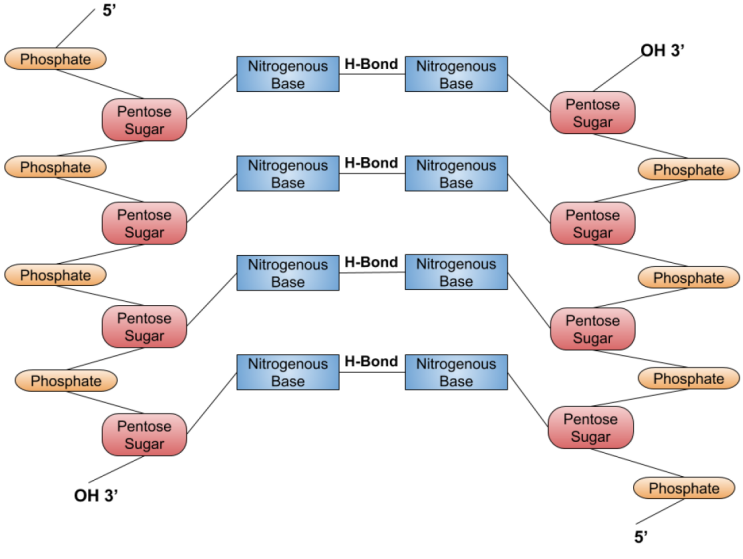
DNA is also involved in the phenomena of ‘evolution’. Evolution is the steady process of occurrence of any change in the metabolic or physical characteristics of a living organism over a long period of time. This effect occurs due to the change in the combination of base pairs in the DNA sequence in the cells of a living organism which in turn reflects the change in the genetic data held by the DNA sequences. The difference in the behaviour of different cells within the same organism is also an effect of different genetic memory that is held by those cells.

According to the model of nucleic acid as proposed by J.D. Watson and F.H. Crick (1953) a DNA is a complex molecule which exists in a helical geometry. A standard DNA molecule is composed of nitrogenous-bases, namely adenine (A), thymine (T), cytosine (C) and guanine (G), and a phosphate-sugar backbone. Naturally, the DNA molecule exists in the living cell in the form of a double-stranded helix. The nucleotide bases remain bonded with a pentose sugar molecule, namely the deoxyribose sugar, which in turn binds with a phosphate molecule. Each phosphate molecule remains bonded with two consecutive sugar molecules forming the phosphate-sugar backbone. The DNA molecule exists in the nucleus of a cell in the form of a highly condensed body known as the chromosome.

Structurally the two strands remain together in an anti-parallel manner, i.e., one strand is oriented in the 5’ to 3’ direction and the other runs in the 3’ to 5’ direction, and they are bonded together with the help of an intermolecular Hydrogen bond (H-bond). The 5’ and 3’ represents the Carbon atom positioned at 5th and 3rd position of the sugar molecule respectively. The 5’ carbon atom gets associated with the phosphate whereas

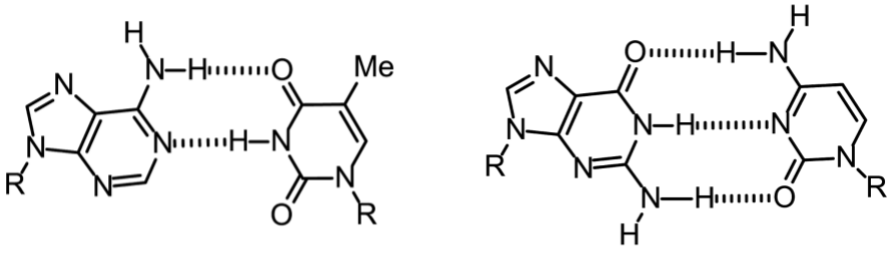
the 3' carbon atom consists of a hydroxyl (–OH) group. This asymmetric geometry gives the DNA molecule a physical direction. The 5'–3' strand is known as the leading strand whereas the 3'–5' strand is known as the lagging strand. Both of these strands are complementary to each other. The fundamental architecture of a DNA molecule is shown in Figure 1.

Figure 1 Basic structure of a DNA molecule (see online version for colours)



Nucleotide bases can be categorised into two types, namely the purines and pyrimidines. Purines are classified into adenine and guanine and pyrimidines are classified into cytosine and thymine. Adenine has the capability to bind with Thymine with a double Hydrogen bond (A = T) whereas cytosine and guanine remain conjugated with each other through a triple hydrogen bond (C ≡ G) in a DNA strand. The molecular structures and their respective conjugated forms are shown in Figure 2(a) and Figure 2(b).

Figure 2 (a) Base pairing of adenine and thymine (b) Base pairing of cytosine and guanine



(a) (b)

3 Non-standard nucleotides

The concept of non-standard nucleotide gained its popularity over the last few years. The non-standard nucleotides were chemically designed in such a manner that the complementary base pairs are highly specific and selective in nature (Malyshev et al., 2014; Saito-Tarashima et al., 2018).

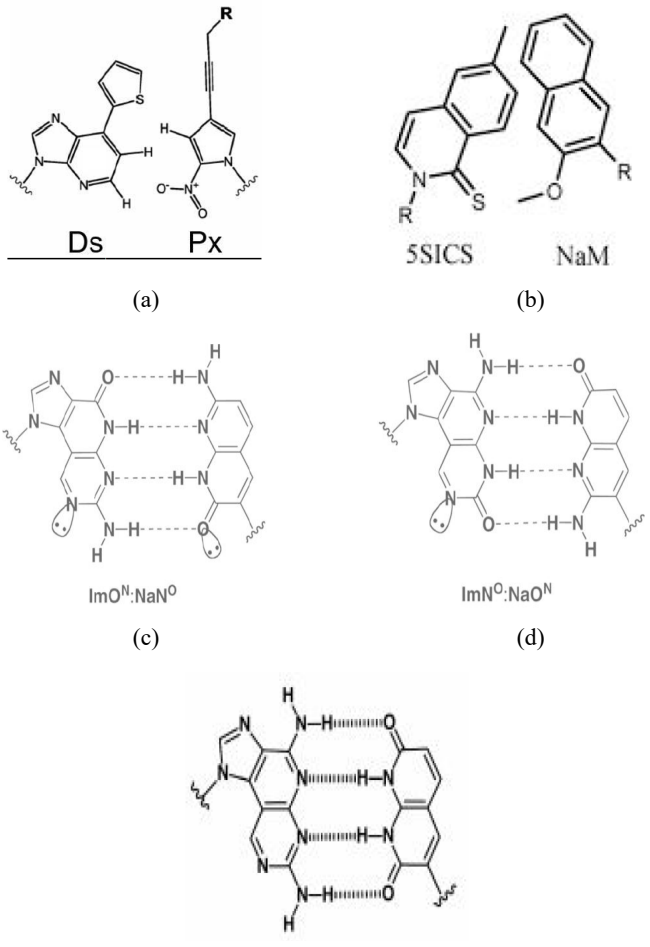
As defined earlier, the standard nucleotide base pairs form intermolecular H-bond. The H-bonding not only makes the nucleotides highly stable but also maintains a high degree of specificity. On the other hand, the stability and specificity of the non-standard nucleotides primarily depend upon three main factors. These factors are shape complementarity base stacking and intermolecular H-bonds (Jahiruddin and Datta, 2015; Lee and Berdis, 2010). Shape complementarity is the steric factor related to the molecular structure. It represents the structural or shape compatibility between the complementary non-standard base pairs. On the other hand, base stacking is the phenomenon by which a non-standard nucleotide gets stacked between neighbouring nucleotides and remains embedded through the phosphate backbone.

Numerous non-standard nucleotides have been discovered in the past (Hirao and Kimoto, 2012). The non-standard base pairs 7-(2-thienyl) imidazo [4, 5-b] pyridine (Ds) and pyrrole-2-carbaldehyde (Pa) have been identified to perform well in the transcription and replication process (Hirao et al., 2006a, 2006b). Later 2-nitropyrrole (Pn) was used in place of Pa which boosted the accuracy of PCR to a greater extent (Hirao et al., 2007). The Pn molecule also reduced the chances of mispairing during PCR amplification. Recently 2-nitro-4-propynylpyrrole (Px) was used with Ds in PCR amplification (Hirao et al., 2007; Hirao and Kimoto, 2012). The Ds-Px pair showed high fidelity and efficiency in Polymerase Chain Reaction amplification (Okamoto et al., 2016). It was noticed that the selectivity of Ds-Px pair in DNA replication was greater than 99.9%. The retention of Ds-Px base pairs in the DNA strand was more than 97% when amplified over 100 PCR cycles (Yamashige et al., 2012). The molecular structure of the base pair Ds and Px is shown in Figure 3(a).

Another hydrophobic base pair, namely 5SICS and NaM, was used within a DNA molecule. The 5SICS-NaM base pair showed high fidelity when PCR was applied on a DNA molecule containing them. It was also reported that the behaviour of 5SICS and NaM was equivalent to any standard nucleotide base pairs in the PCR or PCR based processes (Malyshev et al., 2012). The molecular structure of 5SICS and NaM is shown in Figure 3(b).

It needs to be also observed that 5SICS and NaM are stabilised by steric compatibility and base stacking. Though this base pair has high fidelity and selectivity yet it lacks the ability to form H-bonds which can lead to mispairing of nucleotides. As an alternate base pair, imidazo [5', 4': 4.5] pyrido [2, 3-d] pyrimidines (Im) and 1,8-naphthyridines (Na) is chosen for this work. Besides the steric compatibility and base stacking property, the Im-Na base pair also has the ability to form four intermolecular H-bonds which serves as an added benefit over 5SICS and NaM base pair as it prevents the mispairing of nucleotide base pairs (Saito-Tarashima et al., 2018). There exist multiple molecular formations of Im-Na base pair. The second generation Im-Na base pair, i.e., ImN^{N} and NaO^{O} are used in this work. The molecular structure of Im and Na is shown in Figures 3(c), 3(d) and 3(e).

Figure 3 (a) Molecular structure of Ds and Px base pair (b) Molecular structure of 5SICS and NaM base pair (c) Molecular structure of ImO^N and NaN^O (d) Molecular structure of ImN^O and NaO^N (e) Molecular structure of ImN^N and NaO^O



(a) (b) (c) (d) (e)

Source: Yamashige et al. (2012), Malyshev et al. (2012) and Saito-Tarashima et al. (2018)

4 Feasibility of DNA sequencing and relating digital signals with DNA

At the beginning of the emergence of the concept behind nucleic acid memory the initial strategy that was taken to encode digital data into standard nucleotides was by mapping DNA codons directly to the ASCII characters. The 26 English alphabets along with the numbers from 0 to 9 and some punctuation symbols were related with a triplet DNA codon (Clelland et al., 1999).

It was also made successful in using the standard double-stranded DNA to act as a rewritable molecular memory (Chandrasekaran et al., 2017). The work uses a 5-bit

system which was capable of implementing three basic operations of a memory system namely, write, erase and rewrite. The memory system was demonstrated by creating DNA nanoswitches during the encoding process and the decoding mechanism was carried out through gel electrophoresis. It was also made possible to utilise the DNA nanoswitch system to implement the logic gates of bio-molecular nature. The AND and OR logics were successfully implemented as stated in the literature. The feasibility of interrelating the biological DNA molecule with the digital system was efficiently preserved and proved by the work.

Another method for using DNA as a storage system was proposed in the literature (Yazdi et al., 2015). The system explained the process of designing a DNA based storage system which had the capability of synthesising, sequencing and rewriting the data into the DNA memory with reduced error rate. The work also mentioned about a strategy of DNA memory storage technique which consisted of the several procedures namely, inputting the digital information, encoding into DNA codes, synthesise DNA sequence, storage of the synthesised DNA, editing and reading through high throughput sequencing (HTS) which finally retrieves the digital data from the stored DNA sequence. This effectively proves the feasibility of relating the nucleic acid system with the digital system. It also proves the workability of sequencing, editing, erasing and reading mechanisms using modern sequencers. This has provided a strong momentum towards the elevation of the notion behind the DNA based memory systems.

5 Outline of the proposed encoding scheme

The proposed methodology uses a double-stranded DNA as the memory unit since a double-stranded DNA is more robust and less prone to mutation than a single-stranded DNA. Every nucleotide is assigned with a 3-bit binary number. This encoding map is shown in Table 2. It is to be noted that we are denoting ImN N and NaO O as Im and Na respectively (Saito-Tarashima et al., 2018).

Table 2 Encoding map us

Nucleotide	*Binary equivalent*	*Nucleotide*	*Binary equivalent*
A	000	T	111
C	001	G	110
Ds	010	Px	101
Im	011	Na	100

Using the information theory it can be stated that the information capacity (I) of each nucleotide can be calculated using the following equation:

$$I(x) = \log_2(1/p) \tag{1}$$

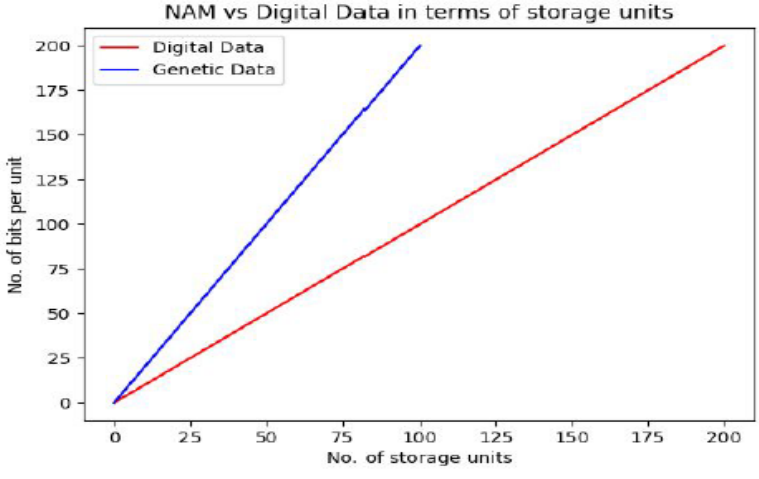
where p is the probability of occurrence of a single base and $x \in X$, where X = {A, C, T, G, Ds, Px, Im, Na}

The probability (p) of the occurrence of each element is 1/8. As a result

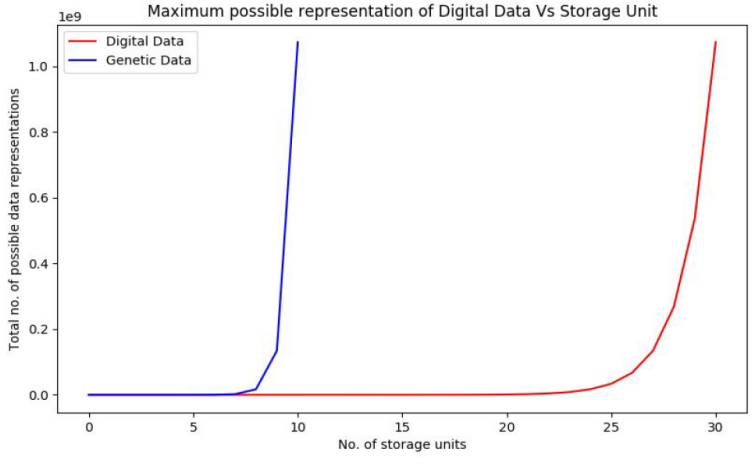
$$I(x) = 3 \quad \forall x \in X \tag{2}$$

It clearly depicts that each nucleotide has the capacity to express data of length 3-bits. As a result, through this mapping, it is made possible to represent eight binary numbers with the help of a single length nucleotide, i.e., the binary numbers from 000 to 111, as the system is linking a 3-bit data to eight different nucleotides. However, the number of unique data that can be represented through N nucleotides is 8^N, i.e., 2 3^N. A comparison between standard silicon memory and nucleic acid memory in an idle system is shown in Figure 4. Figure 4(a) shows a comparison in terms of the storage unit and Figure 4(b) shows a comparison in terms of maximum data representation that can be stored by the storage units.

Figure 4 (a) Comparison with respect to storage units (b) Comparison in terms of maximum data representation (see online version for colours)



(a)



(b)

6 The procedure followed to encode digital data like ASCII characters

1 Detect the ASCII value of an input character.

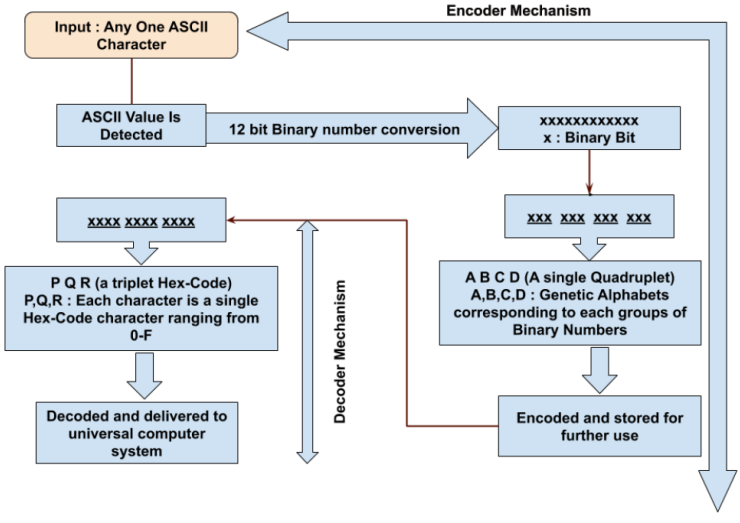
2 Convert it into its corresponding 12-bit binary number by pre-pending the required number of 0s to the original binary ASCII value.

3 Start encoding consecutive three bits at a time into a single nucleotide.

4 Generate a quadruplet consisting of four nucleotides representing a 12-bit binary ASCII value (4 nucleotide × 3 bit/nucleotide).

The whole mechanism of encoding and decoding process is shown in Figure 5.

Figure 5 Encoding and decoding mechanism (see online version for colours)



It can be seen from Figure 5 that the whole process is divided into two subparts, namely the encoding and decoding mechanism. The encoding mechanism deals with the conversion of digital input data into genetic data and its storage for future use. On the other hand, the decoding mechanism deals with the reading of the stored genetic data and reconverting it into the digital data for its use in the computing systems.

In the encoding mechanism, an ASCII character is entered as input. The ASCII value of the character is identified. It is then converted into a 12-bit binary number by pre-pending the required number of 0 to the original binary value. The 'x' in Figure 5 represents binary digits. The 12 binary bits are then grouped into 4 clusters by taking three consecutive bits sequentially. Each of the clusters is then encoded into its respective nucleotides with respect to the encoding scheme defined in Table 2. Finally, the encoded data is stored and preserved for future use.

In the decoding mechanism, the stored quadruplet nucleotide sequence is scanned and is directly reconverted into a 12-bit binary number. This 12-bit binary number is again grouped into three clusters by taking four consecutive binary digits sequentially. Each of

the clusters is converted sequentially into a single hex digit and is integrated to form a triplet hex code which is then delivered to the computer systems.

The proposed encoding scheme described above is used to encode the set of 128 ASCII characters. The results obtained following the above-mentioned procedure is provided in Table 1 in the supplementary material.

7 Result and discussion

The result of this work is compared with the previously reported results. The fields of comparison along with their reported results with respect to the individual work are tabulated in Table 3. The fields of comparison involves theoretical data capacity in bits per nucleotide, types of encoding which are used in the individual works, experimental data capacity as reported in the literature, average theoretical data capacity of each encoding scheme without any constraints and the applicability of the encoding scheme on the types of data which needs to be encoded were examined with the previously reported results.

This work has additionally taken into consideration the presence of phosphate backbone of the DNA molecule. It is done to reach a closer approximation of the data storage capacity with respect to the natural feasibility of a DNA system. From Table 3, it can be analysed that the encoding scheme not only preserves a high data density but it also represents the encoded data through both binaries as well as hexadecimal numbering system. It was also made possible to assimilate 3-bit data into a single genetic alphabet which is better than the previously reported results. Furthermore, binary numbers are mapped with the genetic alphabet such that the complementary bases hold data which are the complement of each other. The average theoretical data density of the proposed methodology was found to be 0.631 ZB/g (Zettabyte per gram).

The standard, as well as non-standard nucleotide base pairs, are highly selective in nature. Moreover, the encoding scheme also has the provision for the self-correction mechanism due to the fact that the complementary base pairs hold the complementary data in the leading and the lagging strand. As a result, the designed model has the inherent ability to enhance the efficiency of the memory storage system. The encoding approach as proposed in this paper is not only explicit and easy to implement but also universal in nature, i.e., the proposed scheme is applicable for encoding any form of digital data into nucleic acid memory. However, the encoding scheme does not require any external use of memory rather the whole mechanism is direct in nature. This can notably enhance the speed of the encoding and decoding process.

This work has been analysed along with two of the pre-existing works through another set of comparison which is provided in Table 4. The calculations related to the respective fields under study are provided in the supplementary material. It provides a comparison with respect to the following fields:

1 Error rate – It is the average mutation frequency which defines the error that can occur due to the mutation or wrong sequencing of nucleotide base pairs.

2 Average redundancy – It defines the average degree of data repetition per unique sequence of nucleotide bases.

Table 3 Comparative analysis of the results with the previously reported results

Field	*Church et al. (2012)*	*Jiménez-Sánchez (2013)*	*Goldman et al. (2013)*	*Jiménez-Sánchez (2014)*	*Grass et al. (2015)*	*Bornholt et al. (2016)*	*Blawat et al. (2016)*	*Sayehira et al. (2017)*	*Erlich and Zielinski (2017)*	*Song et al. (2018)*	*Proposed encoding scheme*
Theoretical capacity (bits/nucleotide)	1	2	1.58	3	1.78	1.58	1.6	1.33	*1.98*	1.9	3
Type of encoding used for digital data	Binary	Binary + quaternary	Binary	Octal	Binary	Binary	Binary	Hexadecimal	Binary	Quaternary	Binary + hexadecimal
Experimental capacity (bits/nucleotide)	0.83	NIL	0.33	NIL	1.44	0.88	0.92	NIL	1.57	NIL	> 2
Average theoretical data density without constraints (bytes/g)	0.227×10^{21}	0.455×10^{21}	0.359×10^{21}	NIL	0.405×10^{21}	0.359×10^{21}	0.364×10^{21}	0.303×10^{21}	0.450×10^{21}	0.432×10^{21}	0.631×10^{21}
Applicability on the type of data to be encoded	Universal	Universal	Universal	Character encoding	Universal	Universal	Universal	Universal	Universal	Universal	Universal

3 Types of encoding used for digital data – This field denotes the type of numbering system namely, binary, hexadecimal etc. that is used for the inter-conversion of genetic data to digital data and vice versa.

4 Applicability on the type of data to be encoded – It expresses the type of data on which the respective encoding schemes can be applied.

The AMF is calculated using the following formula:

$$\text{AMF} = \text{Mutation frequency} \times \text{Frequency of English alphabets} \times (8 - \text{R}) \div 7 \quad (3)$$

where R signifies the redundancies that are associated with the individual encoding schemes presented by the respective literature. It can be noticed that the redundancy of the data obtained from the proposed method of this work is found to be 1 as each nucleotide uniquely maps with the 3-bit binary data, i.e., it follows a one-to-one mapping. As a result, the equation of AMF reduces to:

$$\text{AMF} = \text{Mutation frequency} \times \text{Frequency of English alphabets} \quad (4)$$

From the work of Church et al. (2012), it can be analysed that the error rate or the mutation frequency was 1.8975×10^{-6} errors/bit. The average mutation frequency as reported by the proposal of A. Jiménez-Sánchez (2014) was identified to be 1.2697×10^{-6} errors/bit which was much less than that of the work proposed by Church et al. In this work, the error rate or the average mutation frequency was found to be 1.8973×10^{-6} errors/bit which is near to that of the work of Church et al. The calculations regarding the error rate are provided in the supplementary material.

Table 4 Comparative analysis of results with pre-existing works

Field	*Church et al. (2012)*	*Jiménez-Sánchez (2014)*	*Suyehira et al. (2017)*	*Proposed encoding scheme*
Error rate (errors/bit)	1.8975×10^{-6}	1.2697×10^{-6}	NIL	1.8973×10^{-6}
Average redundancy	2	2.46	2.44	1
Type of encoding used for digital data	Binary	Octal	Hexadecimal	Binary + hexadecimal
Applicability on the type of data to be encoded	All forms of digital data can be encoded	Only 26 English alphabets can be encoded	All forms of digital data can be encoded	All forms of digital data can be encoded

Though the AMF of this proposed methodology is higher than the previously reported result yet it has the potential to eliminate the constraints of the octal system based encoding scheme. Furthermore, it is also not restricted to the encoding of the 26 English alphabets. It was made possible to expand the relevance of the proposed scheme over any existing form of digital data. The modern computers, microprocessors and microcontroller systems are designed to operate on a hexadecimal numbering system for its functionality. The complete system becomes complex and inconsistent if the memory system works in the octal system whereas the data and memory processing sectors work in hexadecimal systems. However, this inconsistency can be removed by implementing

additional inter-conversion algorithms at the cost of greater encoding complexity and speed.

Furthermore, from Table 4 it can be also perceived that the average redundancy of this work is much less than those of the previously proposed methods. This ensures the better utilisation of the mapping between the nucleotide and the digital data. It also increases the productivity of the encoding by exploiting every element of the permutation set of the nucleotide sequences of a specific length.

8 Conclusions

The introduction of non-standard nucleotides and unnatural base pairs besides opening a new dimension in the biological science, it has also given a new momentum in the field of data storage technology like nucleic acid memory. The encoding scheme presented in this paper uses a system of four standard and four non-standard nucleotides. As a result, it has led to an increase in the information capacity of each nucleotide to 3 bits per nucleotide.

Various non-standard nucleotides had been developed in the past. These nucleotides had been studied and are found to behave in the same manner as the standard nucleotide in a DNA molecule. They are also capable of replication through the natural as well as artificial DNA replication processes. It was also studied that these nucleotides are stabilised by three basic factors namely, base stacking, steric compatibility and intermolecular hydrogen bonds.

Besides being simple and universal, the proposed encoding scheme also has the capability to maintain a high data capacity per gram of DNA. The theoretical data capacity was found to be 0.631 ZB/gram of DNA. However, the encoding scheme also integrates an added benefit of reducing the average redundancy due to the implementation of one-to-one mapping of digital data with the genetic data. This ensures the effective utilisation of the permutation set of the genetic alphabet. The average redundancy of this work was found to be 1.

Further modifications can be made by integrating error correction mechanisms along with the encoding scheme to reduce the occurrence of errors during the encoding and decoding process respectively. Other areas include the development of portable and more efficient electronic DNA sequencer which can further reduce the sequencing cost per nucleotide.

References

Badiu, C. (2003) ‘Genetic clock of biologic rhythms’, *Journal of Cellular and Molecular Medicine*, Vol. 7, No. 4, pp.408–416.

Bianconi, E. et al. (2013) ‘An estimation of the number of cells in the human body’, *Annals of Human Biology*, Vol. 40, No. 6, pp.463–471.

Blawat, M., Gaedke, K., Hütter, I., Chen, X-M., Turczyk, B., Inverso, S., Pruitt, B.W. and Church, G.M. (2016) ‘Forward error correction for DNA data storage’, Paper presented in *The International Conference on Computational Science*, *Procedia Computer Science*, Vol. 80, pp.1011–1022.

Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G. and Strauss, K. (2016) 'A DNA-based archival storage system', *Proceedings of the 21st International Conference on Architectural Support for Programming Languages and Operating Systems*, pp.637–649.

Chandrasekaran, A.R., Levchenko, O., Patel, D.S., Maclsaac, M. and Halvorsen, K. (2017) 'Addressable configurations of DNA nanostructures for rewritable memory', *Nucleic Acids Research*, Vol. 45, No. 19, pp.11459–11465.

Church, G.M., Gao, Y. and Kosuri, S. (2012) 'Next-generation digital information storage in DNA', *Science*, Vol. 337, No. 6102, p.1628.

Clelland, C.T., Risca, V. and Bancroft, C. (1999) 'Hiding messages in DNA microdots', *Nature*, Vol. 399, No. 6736, pp.533–534.

Erlich, Y. and Zielinski, D. (2017) 'DNA fountain enables a robust and efficient storage architecture', *Science*, Vol. 355, No. 6328, pp.950–954.

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*, Vol. 494, No. 7435, pp.77–80.

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 Edition*, Vol. 54, No. 8, pp.2552–2555.

Hirao, I. and Kimoto, M. (2012) 'Unnatural base pair systems toward the expansion of the genetic alphabet in the central dogma', *Proceedings of the Japan Academy, Series B*, Vol. 88, No. 7, pp.345–367.

Hirao, I., Kimoto, M., Mitsui, T., Fujiwara, T., Kawai, R., Sato, A., Harada, Y. and Yokoyama, S. (2006a) 'An unnatural base pair system for in vitro replication and transcription', *Nucleic Acids Symposium Series*, Vol. 50, No. 1, pp.33–34.

Hirao, I., Kimoto, M., Mitsui, T., Fujiwara, T., Kawai, R., Sato, A., Harada, Y. and Yokoyama, S. (2006b) 'An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA', *Nature Methods*, Vol. 3, No. 9, pp.729–735.

Hirao, I., Mitsui, T., Kimoto M. and Yokoyama, S. (2007) 'An efficient unnatural base pair for PCR amplification', *Journal of the American Chemical Society*, Vol. 129, No. 50, pp.15549–15555.

Jahiruddin, S. and Datta, A. (2015) 'What sustains the unnatural base pairs (UBPs) with no hydrogen bonds', *The Journal of Physical Chemistry B*, Vol. 199, No. 18, pp.5839–5845.

Jiménez-Sánchez, A. (2013) 'A proposal for a DNA-based computer code', *International Inventions Journal Biochemistry and Bioinformatics*, Vol. 1, No. 1, pp.1–4.

Jiménez-Sánchez, A. (2014) 'DNA computer code based on expanded genetic alphabet', *Journal of Computer Science and Technology*, Vol. 2, No. 4, pp.8–20.

Lee, I. and Berdis, A.J. (2010) 'Non-natural nucleotides as probes for the mechanism and fidelity of DNA polymerases', *Biochimica et Biophysica Acta (BBA) – Proteins and Proteomics*, Vol. 1804, No. 5, pp.1064–1080.

Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J.M., Corrêa Jr., I.R. and Romesberg, F.E. (2014) 'A semi-synthetic organism with an expanded genetic alphabet', *Nature*, Vol. 509, No. 7500, pp.385–388.

Malyshev, D.A., Dhami, K., Quach, H.T., Lavergne, T., Ordoukhanian, P., Torkamani A. and Romesberg, F.E. (2012) 'Efficient and sequence independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet', *Proceedings of the National Academy of Science*, Vol. 109, No. 30, pp.12005–12010.

Okamoto, I., Miyatake, Y., Kimoto, M. and Hirao, I. (2016) 'High fidelity, efficiency and functionalization of Ds-Px unnatural base pairs in PCR amplification for a genetic alphabet expansion system', *ACS Synthetic Biology*, Vol. 5, No. 11, pp.1220–1230.

Saito-Tarashima, N. and Minakawa, N. (2018) 'Unnatural base pairs for synthetic biology', *Chemical and Pharmaceutical Bulletin*, Vol. 66, No. 2, pp.132–138.

Song, W., Cai, K., Zhang, M. and Yuen, C. (2018) 'Codes with run-length and GC-content constraints for DNA-based data storage', *IEEE Communications Letters*, Vol. 22, No. 10, pp.2004–2007.

Suyehira, K., Llewellyn, S., Zadegan, R.M., Hughes, W.L. and Andersen, T. (2017) 'A coding scheme for nucleic acid memory (NAM)', *IEEE Workshop on Microelectronics and Electron Devices (WMED)*.

Watson, J.D. and Crick, F.H.C. (1953) 'Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid', *Nature*, Vol. 171, No. 4356, pp.737–738.

Yamashige, R., Kimoto, M., Takezawa, Y., Sato, A., Mitsui, T., Yokoyama, S. and Hirao, I. (2012) 'Highly specific unnatural base pair systems as a third base pair for PCR amplification', *Nucleic Acids Research*, Vol. 40, No. 6, pp.2793–2806.

Yazdi, S.M.H.T., Kiah, H.M., Garcia, E.R., Ma, J., Zhao, H. and Milenkovic, O. (2015) 'DNA-based storage: trends and methods', *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, Vol. 1, No. 3, pp.230–248.

Zhirnov, V., Zadegan, R.M., Sandhu, G.S., Church, G.M. and Hughes, W.L. (2016) 'Nucleic acid memory', *Nature Materials*, Vol. 15, No. 4, pp.366–370.