Molecular Biology Reports (2018) 45:1479–1490
https://doi.org/10.1007/s11033-018-4280-y

REVIEW

Trends to store digital data in DNA: an overview

Fatima Akram[1] · Ikram ul Haq[1] · Haider Ali[1] · Aiman Tahir Laghari[1]

Received: 8 February 2018 / Accepted: 23 July 2018 / Published online: 2 August 2018

Abstract
There has been an ascending growth in the capacity of information being generated. The increased production of data in turn has put forward other challenges as well thus, and there is the need to store this information and not only to store it but also to retain it for a prolonged time period. The reliance on DNA as a dense storage medium with high storage capacity and its ability to withstand extreme environmental conditions has increased over the past few years. There have been developments in reading and writing different forms of data on DNA, codes for encrypting data and using DNA as a way of secret writing leading towards new styles like stenography and cryptography. The article outlines different methods adopted for storing digital data on DNA with pros and cons of each method that has been applied plus the advantages and limitations of using DNA as a storage medium.

Keywords Alignment · Cryptography · Digital data · Oligonucleotides · Sequencing

Introduction

The excursion of data dates back to the ancient time when it was written, stored and collected on leaves, bones, rocks and paper using different signs, symbols and pictorgrams. The journey then made its way to the use of floppy discs, gramophones, punched cards, magnetic tapes and as technology boomed optical devices including CD's, DVD's, blue-ray discs and flash drives took over. All these devices are subjected to decay and hence their efficacy is lost with time resultantly, the data is not only vulnerable to environmental factors but it perished as well. In addition, these devices are non-biodegradable which can further put a stress on our environment when burnt and cause pollution [1].

Every year the storage necessity is increasing by 50% [2] and there is a dire need to switch towards a more active and reliable method to store the increasing amount of the digital data that is being produced. Memory cards and chips were opted at first; however they only last for 5 years. Hard drives can store up to 100 GB data but are prone to damage by high temperatures, moistures, exposed to magnetic fields and mechanical failures. Therefore, researchers shifted their attention towards exploiting deoxyribonucleic acid (DNA) as a storage medium [3]. DNA can store an unbelievable amount of enormous data just as Castillo has reported that the whole data present on internet can be recorded in a device that would be lesser than a cubic inch [3]. The amount of data that can be stored on DNA is 1 EB (EB or 10^9 GB) per cubic millimeter, it means that eighth order of the amount stored on tapes having a half-life of 500 years in harsh conditions [4].

Data is stored as binary digits of 0's and 1's on a computer therefore, DNA comprising of adenine, guanine, cytosine and thymine that are already paired in the A–T and G–C fashion are utilized as binary codes for storing the information. A single nucleotide represents a 2 bits information; as a result, about 455 EB of data can be stored in just 1 g of single stranded DNA [3]. The data produced by the entire world over a year can be stored only in a mere 4 g of DNA [1]. The 3-D structure of DNA allows it to be versatile and highly dense rewriteable storage medium. We all are aware about the classical base pairing fashion of A–T and G–C in DNA but the advance technology focus on encoding data using new pair systems like A–C and G–T where A–C is employed to 0 and G–T to 1 binary digit [5].

Microvenusproject was started by Joe Davis with the sole purpose to store an image in DNA with an aim to store

✉ Ikram ul Haq
director_dric@gcu.edu.pk; ikmhaq@yahoo.com

Fatima Akram
fatimaakram@gcu.edu.pk

[1] Institute of Industrial Biotechnology, GC University, Lahore 54000, Pakistan

the abiotic data. Encoding was based on the molecular size of bases C-1, T-2, A-3 and G-4. Further, each nucleotide was designated with a phase structure, C-X, T-XX, A-XXX and G-XXXX. Encoding was done by assigning a nucleotide for each repeated position of 0 and 1 bits for example 100101 = CTCCT and 10101 = CCCC. However, at the time of decoding, C could be decoded as both either 0 or 1 which would create errors, for instance CTCCT could be decoded as both 01101 or 100101 that's why this scheme was regarded as inaccurate since decoding was not consistent [6, 7].

Genesis project was introduced by Eduardo Kac. He created an artist's gene by converting a sentence from Genesis, a bibliographical book into Morse code and later into DNA base pairs [8]. Hyphen and full stop were represented by T and C while word space and letter space were replaced by A and G respectively [6]. The genes were fused in the bacterium and were subjected to mutation by ultraviolet radiation. The original sentence was changed when the procedure of decoding was performed [8]. Even though, the above two methods were pioneer in encoding data on DNA but were not applied much due to the complications of inconsistent decoding and alteration in the original content by mutations that produce errors.

Designing DNA codons for data storage

Due to unavailability of standard structures and designs for generating code-words further different encoding models were presented which are explained as follows:

Template map strategy

In this strategy, a solid structure is maintained while allowing a great deal of flexibility which helps in virtual representation of any segment you want to operate using your own unique *strategy*. In this method, constrains on code words are divided into two codes or binary codes by the Codon's group. Template group specify the GC content map which shows mismatch between pair of words. The results of these codes are quaternary codes. Arita [9] expanded this mapping strategy for longer code words by using the Hadmard code. Problem linked to this approach is varying of melting temperature due to GC contents and Comma freeness. Mishybridization was another drawback which can create enlarged multiple words.

De Bruijn construction

To deal with problems like varying melting temperatures, the oligonucleotides were selected; they can also overcome the problem of mishybridization by placing matched pairs consecutively. In this approach, small number of mismatches between words and comma freeness is the major disadvantage.

Stochastic method

In the past, this method was used to find code words with similar melting temperatures by using genetic algorithms. Due to complexity of genetic algorithms, now this method can be applied only to the 25 code words limit and to avoid this problem template map strategy was used [9].

Codes for encrypting data in DNA

Throughout the course of history, scientists have employed three types of codes for the sake of storing data in DNA. These codes had a general consideration that the language encoded in DNA is alphabetic. By alphabetic, some researchers interpreted it as English language while it can also be employed as shorthand which is the writing style for phonetics. Certain conditions must be fulfilled by a code to be proper which are as under:

- After the data have been encoded, it must have the capability to rebuild the message.
- It should be able to use nucleotides in a thrifty manner because the synthesis of extended oligonucleotides is an expensive process.

If there is a mechanism for the detection of error and protection present in the coding system, then it would be of humongous advantage however, it's not a necessary condition. This characteristic on the other hand is not considered indispensably significant, because there are other processes for coping up with this problem such as consuming several copies of DNA. The written language integrally contains a self-correcting mechanisms. It renders this character of error recognition and rectification which is again not fundamentally important. Following are some codes for encrypting data in DNA [10].

Huffman coding

It is a lossless data firmness algorithm. The main idea and principle used by this type of coding is the variation in the length of the symbols which are used to represent a character. The character which appears most recurrently in the text is assigned the lowest number of symbols while the character which repeated less is awarded the highest number of symbols. Using this principle, we can develop a very discrete code. In Huffman coding, 2.2 is the code length average which by far is the least length average attained. Once the initial point is exposed, there is only

one method to read the encoded message which solves the problem of uncertainty of the code [10].

The main drawback of the Huffman coding is its inability to properly code and supply numbers and symbols. This is primarily because of the occurrence of displaying these symbols have exceedingly reliant on the text which analyses the fact that they are incapable to be encompassed in articulating the Huffman code. Secondly, it is not appropriate for long period storage due to the fact that when dissimilar length codons are gathered together, it might not expose a proper pattern. Therefore, the forthcoming generations might not be capable to perceive the worthiness of the pattern [10].

Thus, a scheme is employed to overcome the drawbacks of Huffman coding. In this scheme each tactic used for information storing in DNA varies according to the cost-effective use of nucleotides. Here, Ailenberg and Rotstein employ the principles of Huffman coding to state DNA codes for the whole keyboard, for unambiguous information coding. This immobilizes the drawback of the Huffman code, being restricted only to the letters of the alphabet. This is founded on a creation of a plasmid library with particularly designed primers fixed along with the message for quick retrieval. Index plasmid comprises only details about the structure of the information library. An economical encoding scheme should have inexpensive use of nucleotide per character which is almost 3.5 for this method [10].

The comma-code

A comma code is actually a type of prefix-free code. In this system a comma, a certain symbol or sequence of symbols are present at the end of a code word which otherwise don't exist at the end. A single base (G) is considered as a comma in this approach. Employing this base G the codons having length of 5-bases can be separated from each other. The other three bases namely A, T and C comprise the % base codon. It is further restricted to the two G–C base pairs and one A–T base pair. In the second G–C base pair, base C is always positioned in the upper strand. This scheme of base pairing has an advantage of having an isothermal melting temperature. This code has a reading frame of six codons including G, which is its central feature. Other codes lack this feature. This G as its central feature aids in recognizing a clear reading frame without having the need to cite a starting point. This code provides protection from mutations like insertion and deletion which renders other codes complex. It is an expensive process due to the fact that it reiterates the comma base G for the purpose of creating an automatic reading frame [11].

The alternating code

This scheme is comprised of the six base codons. It contains both purines and pyrimidines, 64 in total number. The message of DNA is designed in a completely synthetic fashion which is the main distinct feature of this coding type. This system has overcome the shortcoming of Huffman coding scheme because it can create an artificial DNA which is appropriate for long term storage. Moreover, it bid the perks of being isothermal and can detect errors but even then it is still inferior to comma code scheme. The primary shortcoming of this coding scheme is that it entails recurrent features which render it costly (non-economical). This has triggered an urge in researchers to develop a scheme with non-repetitive features [11].

Comma-free code

This is also known as prefix-free code or self-synchronizing block codes because it doesn't need any synchronization to find the beginning of a code word. It encompasses base frames, without commas, which have fixed length for the purpose of separating the frames. It employs an automatic frame detection scheme. Comma-free code does not contain four matching base pairs which is the only method of obstructing from natural DNA sequences. These codons are likely can be read merely in one way and errors can be detected by back error detection devices as well [12]. Although, comma-free code is robust and the error rectification works to fix damage at a small-scale which includes point mutations in DNA; however, it does not have the capability to mend shattered data when very large DNA segment is erased from the data encoded DNA region [12].

Encoding schemes

Encoding DNA methods were majorly focused from 2003 to 2009 and data was encoded in form of base triplets till 2009. After application of different Linguistics, Mathematics and Combinatorics, there was a dire need of robust and representative technique that serves as communication link between DNA bases and other languages in which present data may be stored and can have the scope of expanding further data formats. Momentarily, an organized and proper coding strategy was needed at that time. After the breakthroughs in DNA cryptography technique, different schemes for encoding were tested and implemented to store data. Alternating code, Huffman code and Comma codes are interesting examples of these schemes [13].

One of the fine DNA coding scheme prerequisites is the use of per character usage of nucleotide bases. It was proven mathematically that bases may harm safety of messages encodes, therefore, the scientists created

20 bases long complementary strands hence, restriction enzymes were used to insert encoded fragments and thus resultantly they were able to clone strands in plasmids. PCR was used to extract data whenever it was needed and stop codons played role of 'sentinels' for protection of messages. In this way, 57–99 base pairs of information were encoded in bacteria from 7 fragments of DNA that were chemically synthesized. This research was important because it saved data from harsh environmental conditions like radiations, nucleases etc and laid the basis for an idea that DNA can be used for science archival which is a suitable host. They are shown in Fig. 1 that depicts ratio between three coding systems and therefore, researchers prefer Huffman coding. It stores data in form of three bases representation but Ailenberg and Rotstein modified this scheme because DNA needs data to be stored in four bases [13].

Both Ailenberg and Rotstein proposed modification by which we can now store digital data like pictures, text and audio features in DNA. By utilizing Huffman code principles they defined DNA codons on keyboard to finish any uncertainty during coding. They designed primers and synthesized plasmid library for storing data that was exactly same as that of exon and intron structures. It facilitated efficient, speedy and reliable information retrieval. There were also other methods that had high ratios however, they were uneconomical but were still opted because they represented English alphabets and would worked smoothly [14].

In 2010, one of projects from Craig Venter Institute encoded 7920-bits in genome of *Mycoplasma mycoides*. Artificial DNA synthesis was a prerequisite for researchers so they produced synthetic cells which differentiated former from natural cells. Even though, digital data storage was involved in this project but it was a large project to encode huge amount of data storage in DNA. It was also a big achievement for the very first time that synthetic cell was produced, which was a successful breakthroughs in Computing DNA as the project moved from bits to Mb's of information [14].

DNA decryption

Decryption of data is the reversal of encryption in which encrypted text or other form of data is taken and is converted back into the text which anyone (computer) can read easily. Therefore, decryption can be used as a term to unencrypt the data. When the data is stored and encrypted in DNA, it becomes difficult to steal the data however, in order to view it and understand the stored data, it must be decoded. Special computational decryption algorithms are used to decrypt the DNA data [15]. Even the smallest size of DNA can store an immense amount of data thus, encryption and decryption is needed according to the type and purpose of information stored on DNA.

A pseudo DNA cryptography method is one such novel and promising tool in the field of cryptography research. DNA can be used in cryptography in order to store, encrypt and decrypt the data as well [16]. It is a key tool using DNA based molecular cryptographic systems [17]. The main procedure includes in this is: (i) a substitution method, which utilizes distinct pad libraries [17], each of these pads defines a randomly generated, specific and pairwise mapping. (ii) Indexed random key strings and an XOR scheme which utilizes molecular computational methods [17].

These aforementioned approaches are used to store and encrypted both natural and artificial data [17]. Though, combined with this strategy, there is another novel microarray DNA-based chip method technology for a 2D data input and output (decrypted). However, these methods require high tech laboratory requirements and computational limits [16].

Polymerase chain reaction based encoding schemes

In this method the sequence of data is converted into a DNA sequence based on the rule of encryption key or codon. By placing the encoded DNA between the two template regions corresponding to forward and reverse primers, dsDNA is designed and inserted into a genomic DNA. The DNA is amplified by PCR and is decoded by DNA sequences for the readout process [18]. Microdot—a microphotograph is utilized to store data. The method of using microdot was feasible due to the small size of dots and data was secured

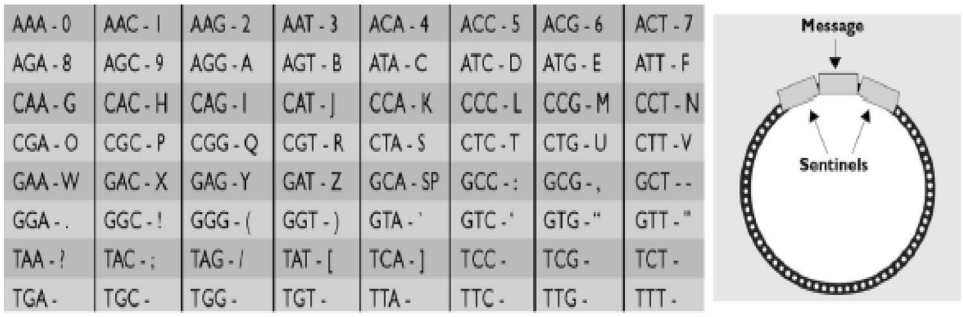

Fig. 1 Showed depicts ratio between three coding systems [48]

in a way that without knowing the sequence of primers it could not be decoded [7]. Though, the amount of data stored was scarce as only 136 bits was stored owing to limited size of microdots [6].This limitation was overcome by modifying this method. Kac proposed the information DNA model, iDNA which consisted of one Polyprimer key (PPK) and have both forward and reverse primers of 5–6 common base pairs which indicated that information is stored in it. Encoding was done by mapping data to three bases (A, C, T) with primer sequences containing four bases consisting of G as the 4th base to prevent mispriming [19]. The decoding was performed by reading the PPK and information was retrieved. The data in the iDNA is thus highly secured because the recipient must have the sequence of encryption key and primer to decode the data. Disadvantages of PCR method includes the requirement to know about the primers, their construction, practical problems and insertion of errors in the template region. The breakage and damage of data due to human errors involved in coding and decoding the information is another drawback [18].

Alignment based encoding models

The alignment based model is independent of PCR. This approach introduced such method of data storage and retrieval which was founded on DNA sequence alignment. This method has a unique specialty that it can carry out retrieval without any parity checks of DNA template and error correcting algorithms [9]. Yachie et al. [20] proposed a method for achieving enormous data inheritance and flexibility of storage. In this method the data was copied and pasted in an organism in order to copy and paste data in an organism's sequence and this step made this method highly suitable to use DNA as valuable transmissible media and also as trademarks for living modified organisms (LMOs). This method requires only the sequence of complete genome thus it can be used to retrieve the data from living cells without using template DNA. Annealing sites of the DNA have importance because the data retrieved by PCR amplification is at high risk in breakage of DNA annealing sites but these sites also have significant importance in reading the even parts of encoded data [21].

The advantage of this method is the low cost of synthetic DNA and fast speed of reading the DNA data. Data inheritance and greater durability are ensured by the presence of multiple copies of data and the capability of each copy to detect and correct any errors in other copy. To prevent the loss of data during evolution, Yachie et al. [20] used such different nucleotide sequences that encoded for the same data through multiple paths of data compression. Multiplications of cassettes lead to unnecessary volumes, which is a big disadvantage of this method. Parity effects cost a certain volume of data sequence. Recovery rate of data is weak and low as compared to data breakage which arises due to long range DNA deletions. The alignment results can be used to identify the positions of the data breakages even if they cannot be recovered.

The other major drawback of this approach is the limited size of the cassette oligonucleotides that were used for encoding the message. If the size is increased to a particular limit, then there comes a possibility of it appearing in host genome by chance. In order to retrieve that data, the entire genome requires sequencing. Improved Huffman coding method was proposed by Alienberg in which specific primers were used efficiently for different file types. Improved Huffman coding describes DNA codes for the entire keyboard to give precise information coding. Improved Huffman coding is based upon plasmid library construction accompanied by specially designed primers embedded with the message for the sake of rapid retrieval. A fine encoding scheme must have economical use of nucleotides per character [20, 21].

Goldman encoding

The encoding used in DNA nucleotides breaks them into segments that overlapped and provided each segment redundancy up to four folds (Fig. 2) [22]. The strands in output encoding were same as that of each four segments of window. Using this encoding the scientists recovered 739 KB message successfully. At that time it was an effective DNA data storage method to our knowledge, so we used this encoding. Its advantage is that it provides redundancy level that is tunable, "by lowering segment widths & repeating them again and again in strands of similar length (for instance, in Fig. 2 if overlapped segments were half long then rather repeating them in 4 strands they should be repeated in eight strands)" [23].

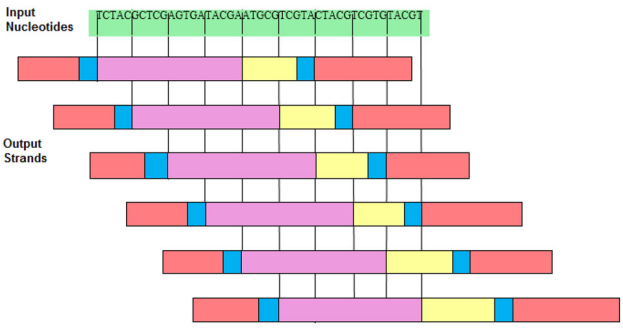

Fig. 2 Goldman et al. [22] proposed DNA encoding. Payloads of each strands of input are overlapping segments in such a way that stream appears as 4 distinct strands

XOR encoding

High reliability and high costs lie in Goldman encoding; as each input string block is four times repeated. A new encoding method which gives result incorporating reduced overheads without compromising the redundancy to prior work. The redundancy of this encoding is same as that of RAID 5; which means that if we have two A, B, of three strands, to recover third strand two strands $A \oplus B$ are enough. This new encoding method has similar reliability as that of Goldman encoding. In wet lab experimentation researchers had reported that objects were recovered with high reliability from both encoding schemes. However, there was an advantage that the density of new encoding method was higher than that of Goldman's method-where repeats of nucleotides were 4-times and our's encoding had average nucleotide repeats of 1.5-time. Practically, there was a little difference between both encoding schemes as overhead primers used were same in both encodings [23]. Previous section shows that our new encoding method practically gives equal reliability both in sequencing as well as in synthesis of DNA and had double the density as that of Goldman encoding scheme.

Existing encodings

Using simple ternary encoding method Bancroft et al. [19] was able to translate text format into DNA form that was later on readable. The 26 letters of English along with an element of "space" was given to three nucleotides; A, T and C so a total of 27 characters were given to these nucleotides. Afterwards scientists successfully recovered 106 letters message encoded on DNA but it had a drawback that recovery of these characters needed a handsome investment and also the results were not so reliable [24].

Factors in encoding design

Mostly, errors are non-uniform in synthesis and sequencing: they are varied in different locations of strands and such varied sequences of nucleotide are relatively under high susceptibility of reactions. As the strand's length increases there are more chances of errors occurrence during synthesis of nucleotides towards the end. Due to truncations, some of these errors are hidden easily and are overlooked. Other subtle errors also lie in synthesis and sequencing with common error of substitution mutation. If our encoding scheme can be improved then these substitution errors can be tolerated by not directly aligning the strands. Let us take an example for better understanding, instead of computing $A \oplus B$, we compute $A \oplus B'$, where B' is inverse of B. Generally, it is seen that at the end of strands the sequence of nucleotides contain errors, thus if one strand is reversed then it ensures an average of constant quality through strand. Every single nucleotide has information regarding redundancy stored in it at position of high reliability [23].

Encodings for reliable storage

In the earlier sections it is discussed that how can we organize DNA based storage systems and which methods are efficient to arrange data according to the specified portions of DNA Strands. The method of encoding implemented on the other hand is quite simple than the approaches which are used for data organization and arrangement. While, decoding data is programmed very carefully so that smallest unit of data (bit) is correctly fed in single position in daughter strands of DNA. It then relies on durability and long term storage of DNA. If a robust design is developed then it would results in data redundancy at stages when data is fed. This part focuses on the work done on redundancy encodings which demonstrate trade-off between the reliability and density implied in designing encoding method, and describes a new encoding method ensuring higher density than existing work without compromising the reliability of results [25].

DNA secret writing

Secret writing is employed in order to avoid illicit access of data by unsanctioned parties. Steganography and Cryptography are the two methods which are being used for this purpose. In cryptography, the information is manipulated for misdirection and in steganography the existence of data is hidden. Conventional cryptography and stenography practices have been recognized as trailing influence and hence, are also becoming fragile because they are fashioned on tough mathematical problems which are developed both in theory and comprehension [26]. Therefore, researchers are exploring developing hybrid cryptosystems relating to DNA procedures into cryptography and stenography. DNA cryptography is defined as hiding data in sequences of DNA. Stenography is defined as data embedding in DNA [27]. This has been made possible via: (i) Writing in between sequences of DNA using insertion and deletion, (ii) Reading DNA known as sequencing.

The core objectives are: (i) Authentication, (ii) Data integrity and (iii) Data confidentiality.

Problems associated with DNA cryptography

Non-availability of an imaginary foundation and absence of knowledge linked with DNA cryptographic approaches are major problems. Likewise, high price and trouble in comprehending is also a drawback in addition to the inaptness which is going to be employed by general community due to the biological tests and trials which in turn have to be achieved in extremely technology fortified laboratories [28, 29].

Data storage styles

This method is used to store DNA words in a medium. DNA words can be stored in two media namely solid media and liquid media. Word "design" has significance in data storage styles. DNA chips played an important role in designing the data storage styles. Both surface and soluble approaches are important to avoid the mishybridization and arrangement data in a systematic way.

Surface based approaches

DNA words in this method are arrayed on a solid surface known as solid phase surface. This approach allows the separation of complement strands from the double helix structure of DNA easily without any laborious work. This reduces the risk of undetected codon combinations and by thus, labeling these codons with fluorophores, they also provide the effective reading information.

Soluble approaches

Advantages of this approach are the information processing by introducing different DNA codons into microbes. Limitation associated with this approach is the access to information stored in molecules which is difficult due to soluble media presence.

Rewritable and random access based DNA storage system

It is the first DNA-based storage architecture that allows random access to DNA data blocks by promoting rewriting capability and nonlinear access of information into random locations; this is considered as the main features of this design. The drawback of this existing method is to read the entire file in order to read a single fragment of data. The other existing methods are read-only methods while this approach is rewritable. By prohibiting the redundancy of information, unwanted cross-hybridization troubles are eliminated in this method. Wikipedia pages of six universities were encoded by Yazdi et al. [21] with this method. Parts of the stored data were carefully chosen and edited text related to three universities was written onto the DNA. The following drawbacks are faced when shifting to rewritable methods from the current read-only methods:

1. The entire content must be rewritten when editing a compressive domain.
2. The rewriting process is made much complex and difficult due to the four-fold coverage used for ensuring the reliability of information. Base modifications of 4 locations are needed just to rewrite one base.
3. Addressing method is utilized only to read the position of a read but it does not perform selective reads.

Access of random data sections and storing of frequently updated data can be effectively carried out by this method which needs to memorize the editing history [13]. In order to access random information, DNA sequences having special strings of addresses were used by Blaum et al. [30]. Error-correction mechanisms have been incorporated into these DNA sequences. Mutually non-interrelated addresses are designed while they satisfied error-control running digital sum constraint [30]. Addresses accurately concluded as prefixes, threaded together which is known as encoding. The method used in accomplishment of rewriting is OE PCR while in decoding of data Sanger sequencing is used. This method is much more efficient than others but it needs long and expensive primers for sequencing OE-PCR method [31]. The other limitation to this method is its high cost that's why the next generation sequencing methods are exploited for being economical [32].

Next generation digital information system

The innovations in digital world are being carried out at high pace with the emerging technology trends. The information which is present in digital form needs to be stored for a long time span with high density. DNA data storage has gained much more attention for storage due to its versatility, data storage capability for long time and its ability to keep data intact and unchanged. DNA is not restricting itself like other storage mediums; it can be read after degradation in the non-ideal conditions over thousands of year. Its natural reading, writing and rewriting function in biological processes made it the current efficient storage media [33]. The messages stored in DNA was initially reported in 1988 and about the 7920 bits encoded data is the largest project till date. The difficulty in reading and writing of perfect long DNA sequences due to small work done on this technique has led to the development of new strategies using next generation DNA synthesis and sequencing technologies to encode random digital information.

In this new technique 11 JPG images, 1 JavaScript program are converted into 5.27 megabit stream and an html coded draft of book that include 53,426 words. Later, all of these bits were encoded onto 54,898159nt oligonucleotides where each of them encoding 96 bits of data and 19 bits of other's data specified locations of data blocks in flanking 22 nucleotides and data blocks are needed for DNA amplification and DNA sequencing. Oligo libraries are synthesized by DNA microchips which are highly reliable for this purpose. The reading of library is another problem and to overcome this problem the libraries are amplified by limited cycle PCR [33].

Among the former DNA strategies, this method has five main merits. Firstly, 1 bit of data is encoded per base instead of using 2 bits, this feature allows the encoding of messages in several ways which permit us to avoid reading sequencing that are difficult to understand or rewriting them in form of repeats, secondary structures and GC contents. The needs of long DNA constructs have been eliminated in this method which caused difficulty in assembly by splitting bit stream into addressed data blocks. The mass, compactness and efficiency are other advantages of this DNA storage system. The portable DNA sequencer for single molecule becomes easily available and we can thus easily simplify reading the information which is encoded in DNA. This new approach of DNA library synthesis, data block address and consensus sequencing will help us in future for DNA synthesis and DNA sequencing. Then again, the use of DNA at large scale for storage of information can be a game changer in making new DNA sequencing and synthesis technologies. In future, we can also opt for redundant encodings, error correction, parity checks and compression for improvement of error rate reduction, density and safety. Other modifications on DNA may include the consideration of maximizing writing, reading and storage abilities [33].

DNA computation

It utilizes DNA, molecular biology and biochemistry hardware systems instead of computer technology. Researches in this area are related to experiments, applications and theory of DNA computing. The term 'molectronics' is used to define this field but now this terminology is mainly used for small scale electronic technology [34].

It was developed for first time in 1994 by Leonard Adleman who explained the use of DNA in computation and solved Hamiltonian pathway problem. Initial experiments by Adleman and their advancement result were considered constructible. The machines used for DNA computation are termed as 'Turing machines'. The initial interest in field was to tackle *NP-Hard* issues but later on this approach proved to be undesirable that's why new proposals were given for "Killer application" approach. A computer scientist in 1997 named Mitsunori Ogihara suggested an application of Boolean circuits in DNA computation [35, 36]. Researchers from *Weizmann institute of science* in Israel in 2002 designed a programmable machine based on computation of DNA and enzymes instead of silicon chips. In April 2004, researchers from the same institute published a journal in which they reported the approach that enabled them to determine cell's cancerous activity. In 2013, researchers stored a JPEG picture and an audio of Martin Luther on DNA [37, 38].

The complexity and organization of living organisms is based on the coding system of 4 nitrogenous bases found in DNA. A rough estimate shows that 6 g of DNA can store up to 3072 EB. Furthermore, the speed of data transfer by this process is very fast. Shelf life of data stored on DNA is also greater than the shelf life of hard discs and flash drives [39].

Capabilities

DNA computing is done in parallel form in a way that it takes various molecules of DNA and makes many possibilities within no time. DNA computers are smaller and faster than any other computer constructed. Jian-Jun Shu and his colleagues designed a DNA GPS system and conducted experiment showing that transport of charges chiefly through DNA is faster in magnetic fields. The application of Strassen matrix multiplication algorithm on DNA computer is provided by Aran Nayebi, though there exist various problems with scaling. Researchers in Caltech designed a circuit which is made of 130 unique strands of DNA and were able to find square root up to 15 numbers [40, 41]. DNA computation does not provide new capabilities of computability theory and problems can be solved computationally. For instance, as compared to size of problem if space for solution of problem increases with the double rate like EXPSPACE problems, the problem level will increase at exponential rate on machines of DNA [42, 43].

Methodology

These are multiple DNA computation devices with every machine having its own pros and cons. Most of these are associated with basic logic gates (AND, OR, NOT) to digital logics from DNA. The methods include are explained below.

DNAzymes

When a matching oligonucleotide enters the machine, the catalytic DNA enzymes chop the DNA and catalyze the reaction according to given inputs. These DNAzymes help in building logic gate which are analogues of digital logics on silicon chips. DNAzymes are limited to 1, 2, 3-input gates and there is no implementation of evaluation of the series of statements. When oligonucleotide in machine binds to nucleotide with matching sequence, it changes its structure and fluorescent substrate is set free. Fluorescent substrate is used because it is easily detectable. The signal strength emitted from it shows the efficiency of reaction. The DNAzyme used in a reaction cannot initiate reaction once again hence, media is to be added continuously in reactions occurring in the stirred fermenter in order to get product continuously [44, 45].

The two important DNAzymes are E6 and 8–17 and because they cleave substrate in arbitrary location of DNA that's why they are being commonly used. MacDonald used

E6 enzyme in constructing MAYA I and MAYA II machines. He also determined basic gates utilizing 8–17 DNAzyme. Both of these DNAzymes are helpful in construction of logic gates but are not useful in vivo because of their requirement of cofactors like Zn^{2+} and Mn^{2+} that's why a new design that has loop on one end is called "Stem loop" is also being used. It opens and closes when DNA binds to loop and is therefore, found helpful in design of logic gates [41].

Enzymes

Turing machines are useful while we use enzymes for DNA computation. Here enzyme is hardware while DNA is software. Benenson et al. [37] used Fokl enzyme to elaborate DNA computation by showing automata that can diagnose and make reaction with prostate cancer in which PIM1 and HPN are over expressed while PPAP2B and GSTP1 were under expressed genes. This automata elaborated the expression of these genes and positive diagnosis was observed for DNA antisense for MDM2 using it. MDM2 is a tumor suppressor and for negative diagnosis only suppressor for positive diagnosis was added. A limitation to this process is that 2 automata are required for inducing different drugs separately while on the other hand it is as efficient as the reaction take only 1 h to complete [46].

Toehold exchange

Toehold exchange is also a concept in which DNA computers can be developed. In this system, the DNA added to computer binds to toehold or sticky end of other DNA molecule. This process allow the generation of basic logic gates which are linked to other large computers. These computers don't have requirement of enzymes and is based in binding of 2 DNA molecules [47].

Self-assembly of algorithms

In this approach, DNA nanotechnology is used for DNA computation. Tiles of DNA are designed that contain a range of sticky ends. An array is designed that demonstrate assembly encoding XOR operation. It allows DNA array to implement cellular automaton that produces a fractal named *Sierpinski gasket* [48].

Pros and cons

The processing speed of DNA computers is slow as compared to silicon computers. DNA computers perform their work in minutes, hours or maximum within a span of few days rather than performing their activity in milliseconds, this drawback however, is compensated by using many computations at once. It allows system to take similar amount of time for a complex calculation as for a simple one. It is achieved when million and billions of molecules interact at once. Again, there is a difficulty in analyzing results which are difficult to read because they are in enormous amounts [35].

Computation method

On 16th August, 2011 Gao, Church and Kosuri published a landmarked paper in which they elaborated the methods of storing information on DNA. They converted 11 JPG images, a Java script and a html draft of 53,000 words book into 5.27 MB stream of 0's and 1's. They assigned 0 as C or A (Cytosine or Adenine) and 1as G or T (Guanine or Thymine). Then, they converted 0's and 1's into oligonucleotides of 96 bit data and all of these were encoded onto 54,898,159 nucleotides. It was too lengthy so sequence was broken down into sets of 96 nucleotides. Among them, 19-nts (bits) depicted text; 22-nts were amplified by PCR. The results of recovery were 100% with a little error of 10 bits per 5.27 million nucleotides. This error arose due to coverage of single sequence. However, it was difficult to synthesize long stretch of DNA therefore, Church and colleagues devised a strategy to break DNA into sequence of nucleotides which were 96-bases long. Each of them had bar code of 19-bits to indicate where these chunks belong to.

The synthesized DNA was inkjet printed on DNA chip or microarray and then dried forming 50 ng smaller clumps. In this method, rather than encoding 2 bits per nucleotides, Church and colleagues added information 1 bit per nucleotide and it showed distinct advantages like efficient reading. The information stored was in form of palindromes which means that strands had same information in forward and backward direction. This entire process was done ex vivo and no living organism was involved. A distinct aspect of this work was scale and amount of information which was much better than any other storage techniques, principle of this work has shown in Fig. 3.

After this, another milestone on DNA data storage was published on 23rd January, 2013 by Nick Goldman and his colleagues in EBI (European Bioinformatics Institute). Nick's main aim was to make a system which had large capacity, require less maintenance, cost effective and is feasible for storing the data. The major breakthrough was the design of an appropriate medium for science archival in which the synthesized and sequenced DNA with proper reconstruction methods an encoded 740 KB of digital data into it. The data they stored comprised of Watson and Crick's 1954 paper about DNA, Shakespeare's Sonnets, Huffman code in ASCII scheme and an audio speech of Martin Luther King in MP3 format with size of 757,051 bytes which gave them 88% efficiency. Both these techniques are the concrete evidence that DNA based storage can be scaled

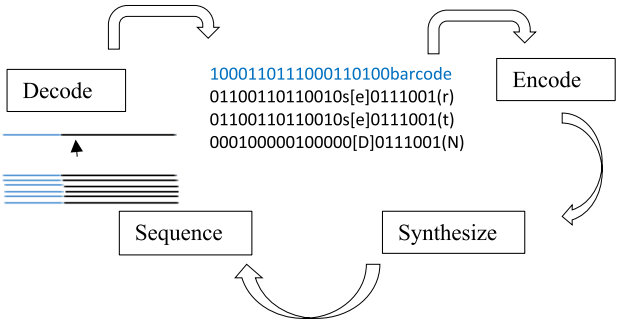

Fig. 3 Overall process of decoding self-referential DNA that encodes the notes

beyond present world volumes of data for large-scale practical mediums, long-term and sequentially accessed digital archiving [14].

DNA editing

DNA can store desired information by synthesizing it in the form of heteroduplexes and therefore, can be written again by means of DNA editing. DNA editing means the addition or deletion of one or more nucleotides to specific sites. For this purpose one should synthesize DNA fragments of short and medium length. This can be done by means of two techniques namely gBlocks Gene Fragments which are known to be building tool for deletion and insertion of DNA edit and Overlap-Extension PCR (OEPCR). Both are used for adding blocks which are result of mutation [49, 50]. gBlocks have wide range of applications due to their characteristics of controlled and double stranded building block in qPCR, protein engineering, CRISPR-mediated genome editing and PCR. They can be constructed by using short DNA strings containing 18 consecutive nucleotides at very low fraction. The gene library products are tested carefully such as length is verified by capillary electrophoresis, composition of sequence through mass spectrometry. Protocols are being followed to avoid errors. The 80% of generated result should match with the DNA strings. This % is reduced in case of DNA strings containing secondary structure. Therefore, these structures should be controlled to minimize the errors [51].

DNA substring editing is done by means of specified PCR reactions. DNA rewriting is done by means of a process known as OEPCR. IN OEPCR two primers can be used to flank with DNA site to be edited. These flanking primers function as zippers to join with the segment which is to be spliced. The primers which are present at the end are designed so they have overhanging part which is complementary to the other primer overhanging part. After completion of extension sequence primers are elongated by means of PCR amplification and overlapping sequences are fused with each other. Restriction sites or enzymes are not required for this process. OEPCR is used for insertion of oligonucleotides having length of more than 100 nucleotides. Two strands can be modified by OEPCR having mutation at opposite site. Different hybridization products can be achieved by means of denaturation. Only one of all the products will allow polymerase extension by introducing a primer without overlapping the heterodimer at the 5′ end. This duplex is again denatured and is created by the polymerase and DNA strand is created by means of hybridizing another primer. The result of DNA replication in the formation of sequence containing the desired insert [51].

Tunable redundancy

Recent experiments show that for each data structure there is no need of most precise storage [23, 52]. For instance, in JPEG file the data included in headers is important for reliable decoding because little imprecision lies in every decoding method so cons are acceptable only if they are in minute quantity, otherwise, there is no use of encoding data into DNA if afterwards data obtained contains high amount of errors. A major significant advantage of encoding schemes is per-block there is tunable level of redundancy. If we have critical data we can pair this data block with many other blocks to obtain a high redundancy level; it means that if A is critical than it generates $A \oplus B$, $A \oplus C$, etc. While on other hand, we can further reduce the redundancy of encoding when data blocks have low critical value rather than including 2 blocks having n or excluding n in such a way that $n - 1$ in any of the blocks are enough to recover last string. This would be having average density $1/n$. This encoding has the advantages of significant tunable redundancy as well as improved density. With large amount of data, the DNA sequencing and synthesis are error prone and slow methods. Another problem is that error rate does not grow in size linearly during DNA synthesis or sequencing. Therefore, it is recommended to make DNA pools that are small in size with an accurate method rather than synthesizing long strands which contain errors. An optimized balance is maintained by tunable redundancy between efficiency and reliability in storage techniques [24].

Discussion and future work

This section highlights the explanation of experiments performed and improvements done on them so far. The results described above indicate that error rates are high enough to necessitate redundant in a complete storage pipeline. In this section, the causes of these errors have been discussed. The identification of causes will contribute to reduce these errors. A small proportion of strands were synthesized using

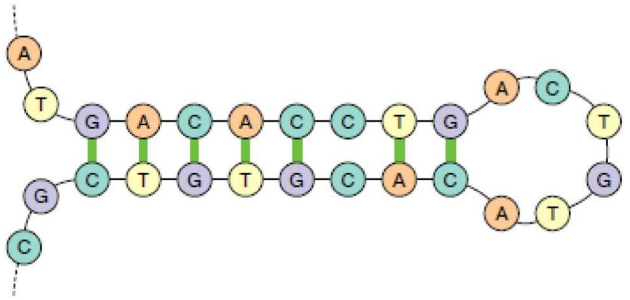

Fig. 4 Formation of a hairpin where some portions of strand are self-complementary. These hair pins make sequencing and DNA amplification error prone [52]

high fidelity synthesis process. The drawback of this process is that, it requires synthesizing each sequence individually giving high costs but low error rate and also low throughput with non-compatibility with applications of information stored. If we move to method of microarray technique, results produced from microarray were not reliable and contain errors but it has an advantage of rapid generation of lowest cost alternative of DNA synthesis. It is clear that sequencing error is most common for all loci and error rate is of higher magnitude if taken on the average. Microarray and other sequencing methods demand that there is a need of improvement in future sequencing techniques. Microarray synthesis sometimes also generates *truncated* strands as well as inappropriate strands. These shortened strands are missed in standard sequencing processes and are useless for recovering data from it, so these strands are rendered waste. Gel electrophoresis is used for determining the length of strands synthesized to highlight losses that may be possibly produced as a result of truncation [52].

If length of target strand contains 120 nucleotides, lower than 5% of strand is truncated earlier or rarely longer. It shows that if we work on smaller fragments then error rate may be reduced and thus synthesis efficiency is enhanced. Let us take an example of *hairpin* formation (Fig. 4). In this a sequence of DNA attaches to itself and the result is folding of both ends together. This binding of two ends results as they were opposite to each other. "This binding prevents the easy amplification and sequencing of strand". When representation is robust it means it avoids generating self-complementary sequences which reduce self-hybridization chances. Certainly, controlling "self-complementarity" also lowers density of representation, thus it results in trade off of density and reliability.

Likewise, the chances of self-hybridization due to binding increases if different strands are partly complementary, encoding with more robustness would try to alleviate this self-hybridization As an alternative method, per strand basis can be used by selecting the encoding that produces the minimal self-complementary and partly complementary strands [52]. The objectives in this current time to discover the dimensions for future work minimize errors in DNA sequencing and synthesizing, improving present techniques to minimize error rates etc. These mentioned things have little effect on experiments thus far can be improved to produce reliable results.

Conclusion

This review highlights various methods used for data storage in DNA. Using diverse codes the data is stored into DNA and numerous methods used to store the data are discussed. Furthermore, pros and cons of these different methods have also been mentioned. We have particularly focused on the challenges and limitations in these methods that allowed large data storage in DNA.

Acknowledgements This work is carried out with the help of prestigious material of the libraries and special thanks to Institute of Industrial Biotechnology, Government College University, Lahore.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Disclosure The authors assure the integrity and quality of our research work. It is also stated that there is no plagiarism in this work and all points taken from other authors are well cited in the text. This study is completely independent and impartial.

Research involving human participants and/or animals This article does not contain any studies conducted on human or animal subjects.

References

1. Shrivastava S, Badlani R (2014) Data storage in DNA. Int J Electr Energy 2:119–124
2. Hakami HA, Chaczko Z, Kale A (2015) Review of big data storage based on DNA computing. In: Proceedings of the Asia-Pacific Conference on Computer-Aided System Engineering (APCASE'15), Quito Ecuador, pp 113–117
3. Castillo M (2014) From hard drives to flash drives to DNA drives. Am J Neuroradiol 35:1–2
4. Allentoft ME, Scofield RP, Oskam CL, Hale ML, Holdaway RN, Bunce M (2012) A molecular characterization of a newly discovered megafaunal fossil site in North Canterbury, South Island, New Zealand. J R Soc N Z 42:241–256
5. Borda M, Tornea O (2010) DNA secret writing techniques. In: Proceedings of the 8th International Conference on Communications (COMM'10). Bucharest, Romania, pp 451–456
6. Davis J (1996) Microvenus. Art J 55:70–74
7. DeSilva PY, Ganegoda GU (2016) New trends of digital data storage in DNA. Biomed Res Int 8072463:14
8. Kac E (1999) "Genesis-art of DNA," http://www.ekac.org/genin fo

9. Arita M (2004) Writing information into DNA. Asp Mol Comput 2950:23–35
10. Smith GC, Fiddes CC, Hawkins JP, Cox JPL (2003) Some possible codes for encrypting data in DNA. Biotech Lett 25:1125–1130
11. Yatchie N, Ohashi Y, Tomita M (2008) Stabilizing synthetic data in the DNA of living organisms. Syst Synth Biol 2:19–25
12. Doig AJ (1997) Improving the efficiency of the genetic code by varying the codon length—the perfect genetic code. J Theor Biol 188:355–360
13. Ailenberg M, Rotstein OD (2009) An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques 47:747–754
14. Sanger F, Nicklen S, Coulson AR (1997) DNA sequencing with chainterminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467
15. Cui G, Li C, Li H, Li X (2009) dna computing and its application to information security field. In: Proceedings of the 5th International Conference of Natural Computation, Tianjian, China; IEEE, pp 14–16
16. Ning K (2009) A pseudo DNA Cryptography method. http://arxiv.org/abs/0903.269
17. Gehani A, LaBean T, Reif J (2003) DNA-based cryptography. In aspects of molecular computing, pp 167–188. Springer, Berlin
18. Yachie N, Ohashi Y, Tomita M (2008) Stabilizing synthetic datain the DNA of living organisms. Syst Synth Biol 2:19–25
19. Bancroft C, Bowler T, Bloom B, Clelland CT (2001) Long term storage of information in DNA. Science 293:1763–1765
20. Yachie N, Sekiyama K, Sugahara J, Ohashi Y, Tomita M (2007) Alignment-based approach for durable data storage into living organisms. Biotechnol Prog 23:501–505
21. Yazdi SMHT, Yuan Y, Ma J, Zhao H, Milenkovic O (2015) A rewritable, random-access DNA-based storage system. Sci Rep 5:14138
22. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, Birney E (2013) Towards practical, high-capacity, low maintenance information storage in synthesized DNA. Nature 494:77–80
23. Chan CY, Ioannidis YE (1999) An efficient bitmap encoding scheme for selection queries. ACM SIGMOD Record ACM 28(2):215–226
24. Cosemans S, Dehaene W, Catthoor F (2008) A 3.6 pJ/access 480 MHz, 128Kbit on-Chip SRAM with 850 MHz boost mode in 90 nm CMOS with tunable sense amplifiers to cope with variability. In Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European IEEE, pp 278–281
25. Cruz RPG, Withers JB, Li Y (2004) Dinucleotide junction cleavage versatility of 817 deoxyribozyme. Chem Biol 11:5767. https://doi.org/10.1016/j.chembiol.2003.12.012
26. Sangwan N (2012) Text encryption with huffman compression. Int J Comput Appl 54:29–32
27. Zhang Y, Bochen Fu LH (2012) Research on DNA cryptography. In: Sen J (ed) Applied cryptography and network security. pp 357–376, InTech, Rijeka, Croatia, http://www.intechopen.com/books/applied-cryptography-and-networksecurity/ research-on-dna-cryptography
28. Borda M (2011) Fundamentals in information theory and coding. Springer, Berlin
29. Borda ME, Tornea O, Hodorogea T (2009) Secret writing by DNA hybridization. Acta Technica Napocensis Electron Telecommun 50:21–24
30. Blaum M, Litsyn S, Buskens V, Tilborg HC (1993) Error correcting codes with bounded running digital sum. IEEE Trans Inf Theory 39:216–227
31. Bryksin AV, Matsumura I (2010) Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. Biotechniques 48:463–465
32. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nature 5:16–18
33. Church GM, Gao Y, Kosuri S (2012) Next-generation digital information storage in DNA. Science 337:1628
34. Ogihara M, Ray A (1999) Simulating Boolean circuits on a DNA computer. Algorithmica 25:239–250
35. Boneh D, Dunworth C, Lipton RJ, Sgall JÍ (1996) On the computational power of DNA. Discret Appl Math 71:79–94. https://doi.org/10.1016/S0166-218X(96)00058-3. (Describes a solution for the boolean satisfy ability problem)
36. Kari L, Gloor G, Yu S (2000) Using DNA to solve the bounded post correspondence problem. Theor Comput Sci 231:192–203. https://doi.org/10.1016/s0304-3975(99)00100-0. (Describes a solution for the bounded Post correspondence problem, a hard-on-average NP-complete problem)
37. Benenson Y, Gil B, Ben-Dor U, Adar R, Shapiro E (2004) An autonomous molecular computer for logical control of gene expression. Nature 429:423–429
38. Jerome B, Yin P, Monica EO, Subsoontorn P, Endy D (2013) Amplifying genetic logic gates. Science 340:599–603
39. Amos M et al (2002) Topics in the theory of DNA computing. Theor Comput Sci 287:3–38. https://doi.org/10.1016/s0304-3975(02)00134-2
40. Ravinderjit SB (2001) Solution of a satisfiability problem on a gel-based DNA computer. DNA computing. Springer, Berlin, pp 27–42
41. Macdonald J, Stefanovic D, Stojanovic M (2009) Des assemblages d'ADN rompus au jeu et au travail, Pour la Science, pp 68–75
42. Nayebi A (2009) Fast matrix multiplication techniques based on the Adleman-Lipton model, arXiv: 0912.0750
43. Wong JR, Lee KJ, Jian-Jun S, Shao F (2015) Magnetic fields facilitate DNA-mediated charge transport. Biochemistry 54:33923399. https://doi.org/10.1021/acs.biochem.5b00295
44. Santoro SW, Joyce GF (1994) A general purpose RNA-cleaving DNA enzyme. Proc Natl Acad Sci 94:4262–4266. https://doi.org/10.1073/pnas.94.9.4262
45. Stojanovic MN, Stefanovic D (2003) A deoxyribozyme-based molecular automaton. Nat Biotechnol 21:10691074. https://doi.org/10.1038/nbt862
46. Seelig G, Soloveichik D, Zhang DY, Winfree E (2006) Enzyme-free nucleic acid logic circuits. Science 314:1585–1588
47. Rothemund PWK, Papadakis N, Winfree E (2004) Algorithmic self-assembly of DNA Sierpinski triangles. PLoS Biol 2:e424. https://doi.org/10.1371/journal.pbio.0020424
48. Huffman DA (1953) A method for the construction of minimum-redundancy codes. Proc IRE 40:1098–1101
49. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
50. Milenkovic O, Kashyap N (2006) On the design of codes for DNA computing. In coding and cryptography. Springer, New York, pp 100–119
51. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
52. Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G, Strauss K (2016) A DNA-based archival storage system. ASPLOS, ACM, New York. https://doi.org/10.1145/2872362.2872397