International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Storing Digital Data in DNA: Advances Toward Practical Implementation

Sri Venkata Aravindbabu Malempati

Abstract: The exponential growth of global digital data, projected to exceed 175 zettabytes by 2025, has exposed fundamental limitations in conventional storage infrastructure, including finite media lifespan and unsustainable energy consumption. This review examines synthetic deoxyribonucleic acid (DNA) as a next-generation archival storage medium, systematically analyzing the encoding architectures, synthesis methodologies, error correction strategies, and retrieval mechanisms that define the DNA data storage pipeline. DNA offers volumetric storage densities of approximately 500 GB/mm³ and molecular stability, enabling data preservation across centuries. The four-nucleotide alphabet theoretically permits 2 bits per nucleotide, though biochemical constraints reduce practical efficiency to 1.5–1.98 bits per nucleotide. Error correction has advanced substantially, with fountain code implementations achieving near-theoretical efficiency at sequencing coverage depths below 11x. However, significant barriers impede deployment: synthesis costs of $0.05–$0.10 per base, write throughput limited to kilobytes per second, and absent universal standards. Recent advances, including CRISPR-based overwriting, microfluidic platforms achieving 97.67% synthesis yields, and enzymatic methods producing oligonucleotides exceeding 1,000 nucleotides, offer promising pathways toward economic viability. The review concludes that DNA storage is approaching feasibility for hybrid archival hierarchies targeting rarely accessed data, with initial commercial deployment anticipated within 10–20 years.

Keywords: *DNA Data Storage, Archival Storage Systems, Error Correction Codes, Enzymatic DNA Synthesis, Data Encoding Architectures*

1. Introduction

Digital data production in the world has gone to an extent that it essentially threatens the ability of current storage systems. It has been estimated that around 2.5 quintillion bytes of data are produced on a daily basis, and estimates have shown that the world datasphere will exceed 175 zettabytes by 2025. This blistering growth results from the increasing artificial intelligence workloads, Internet of Things (IoT) sensor networks, genomic sequencing pipelines, and multimedia content generation [1].

According to Sensintaffar et al. [1], this trend is described as a mismatch of the rate of data creation and the nature of traditional storage technologies. Current technologies, such as Solid State Drives (SSDs) with quad-level or penta-level cells, Shingled Magnetic Recording (SMR) drives, and Linear Tape-Open (LTO) cartridges, are becoming less suitable to address the long-term storage needs of the current data infrastructure. As an example, the life span of a durable disk drive is usually 3-5 years, and even the most resilient magnetic taping types need physical data transfer within intervals of 7-10 years to avoid data loss and degradation. This cycle of replacement and migration contributes greatly to the cost of operations in hyperscale.

The physical footprint and energy consumption of traditional storage infrastructure further exacerbate these problems. The current hyperscale data centers require millions of square feet of real estate, and today they are consuming about 1-2% of the global electricity. This is estimated to increase to 1,929 TWh per year in 2030 as the volume of global data keeps growing at compound growth rates of more than 40 percent every year [2]. Economic and environmental damages related to the creation, operation and cooling of storage facilities of ever-

California State University, Los Angeles, USA

increasing size are increasingly untenable, especially when it comes to cold archival data—information that needs to be stored but is accessed a very limited number of times. The preservation of such data of the infrastructure that is energy intensive creates little operational value as compared to the preservation costs [1].

It is against this background that synthetic Deoxyribonucleic Acid (DNA) has been one of the most promising mediums of storing next-generation archival data. Dimopoulou and Antonini [2] reveal that DNA has the potential to attain a volumetric storage density of around 500 GB/mm³, and this is about 1000 times higher than the traditional hard disk drives. Besides, DNA molecules are able to last hundreds or even thousands of years when stored under the right physicochemical environment; that is, they have a long shelf life when compared to the current digital storage medium by several orders of magnitude.

The information density of DNA is theoretically limited only by the number of bits per position of a nucleotide:

$$I_{DNA} = log_2(4) * N = 2N\ bits$$

Where N is the total number of nucleotide positions in a DNA strand that is synthesized. Since a given position has 4 possible nucleotides {A, T, G, C}, then it can theoretically convey 2 bits of information.

At an average oligonucleotide length of 200 bases, the result would be a theoretical payload of 400 bits before considering encoding overhead, redundancy to correct errors, and address indexing. These other requirements in real-world applications attenuate the efficient storage density by a factor of about 30-50 [1][2].

In spite of these strong benefits, there are a number of technical and economic hurdles that still prevent the mass use of DNA based storage systems. The modern cost of DNA synthesis is between about 0.01 and 0.10 per base, and write speed is also relatively slow, commonly in bytes to kilobytes per second. Moreover, chemical synthesis processes are associated with error rates of about 1-100 bases per 100-1000 bases, and this necessitates effective encoding and error-correction methods in maintaining data integrity [1].

This dissertation is a systematic review that evaluates the encoding systems, synthesis processes, corrective strategies, retrieval processes, and new commercialization directions that are bringing the DNA-based data storage close to the viable large-scale implementation.

2. Related Work

Evolution of DNA-based data storage has been achieved as a progression of several historic studies that have increased the theoretical knowledge as well as the practical viability of the subject. The original assumption is that it developed from computational biology research, which identified DNA as an extremely dense and stable carrier of molecular information. With the fast-growing information technology in the world today, with an estimated compound growth of about 5 percent per year, the amount of digital data is projected to grow to 175 zettabytes by the year 2025. This accelerated growth caused scientists to explore the possibility of using the biochemical properties of synthetic DNA in order to resolve the long-term data storage crisis that was being faced [3].

Over the past few years, there has been a rise in the institutional interest in DNA-based forms of storage. One of the significant milestones in this transition was the creation of the DNA Data Storage Alliance that now has almost 50 member organizations operating in common to create standards and interoperable frameworks for the DNA-based storage technologies. The partnership represents an even larger change from the solitary scholarly study to uniform industrial development and commercialization initiatives [3].

The initial experimental studies were more concentrated on proving the presence of proof-of-concept encoding and retrieval systems. The first experiments were able to encode digital information close to the 1 MB scale, and thus the practicality of mapping binary data to nucleotide sequences was established. Later experiments greatly extended this capacity, showing that at least 200 MB files could be encoded and accurately recovered by synthesizing them into oligonucleotide pools using DNA sequencing technologies [3].

These encoding schemes are based on the four-nucleotide alphabet of DNA, adenine (A), thymine (T), guanine (G) and cytosine (C), which in theory allow each nucleotide position to carry two bits of information. The optimal maximum encoding rate can be written as

$$E_{max} = log_2(4) * L = 2L\ bits$$

where

L is the length of the generated number of nucleotides in the DNA strand.

In real-life applications, efficiency of the encoding is lower because of biochemical limitations. Encoding codes should not contain homopolymer runs (repeats of the same long nucleotide) and should have a balanced GC-content to have trustworthy production and sequencing. Consequently, encoding schemes in the real world normally reach 1.5-1.8 bits per nucleotide, as opposed to the theoretical 2 bits per nucleotide [3].

The DNA storage also has remarkable density and energy efficiency despite these limitations. Volumetric storage density of DNA is more than six orders of magnitude higher than the magnetic storage technologies, and the passive energy storage of DNA long-term storage is some eight orders of magnitude lower than in the traditional media like magnetic tape or hard disk drives [3]. These measures have become standard yardsticks for measuring new encoding structures and system designs in the domain.

Besides the demonstrations provided in experiments, survey-based research has been significant in the synthesis of further progress and defining critical challenges in the DNA storage pipeline. To gain an in-depth investigative study of the DNA data storage paradigm, Akhtar and Rawol reviewed the advances in the encoding, synthesis, sequencing, retrieval, and error handling layers of the system architecture [4]. Their work emphasizes the idea of considering DNA storage not as a biochemical innovation but as a data management paradigm that is interdisciplinary and involves the combination of the fields of molecular biology, computer science and distributed computing systems [4].

Such a view is specifically applicable to the context of research on digital systems, as it reevaluates the concept of DNA storage in the greater context of data engineering and design of computational infrastructures. Massive implementation will demand a concerted effort to optimize algorithmic encoding methods, biochemical synthesis methods, and high-throughput sequencing technologies.

Together, the literature that is available creates three broad themes that characterize the present state of research.

The DNA density and longevity merits are first experimentally confirmed. According to theoretical estimates, in up to a single year, the total volume of the digital data generated in the world could be stored in about four grams of DNA [3].

Second, throughput asymmetry is also a major technical constraint. The contemporary DNA synthesis technologies have top write speeds documented in the kilobytes/second range, and this is many-fold fewer than the gigabytes/second throughput levels needed in large-scale digital storage systems [3].

Third, the cost of DNA synthesis should be lowered in order to make DNA storage economically viable. According to Goldman et al., DNA is at an estimated ten-fold lower cost reduction per base of DNA synthesis, so DNA storage would become cost-effective when compared to magnetic tape when used in long-term archival horizons of 50 years and 500 years [3].

Based on these earlier efforts, the current paper will discuss the role of algorithmic encoding schemes, system-level pipeline optimization, and error correction algorithms in closing the gap between the hypothetical storage capacity of DNA and its actual implementation as a scalable archival storage medium.

3. DNA Data Storage: System Architecture and Pipeline

A DNA-based data storage system end-to-end architecture takes the form of a six-stage pipeline with the encoding stage, synthesis (writing) stage, physical storage stage, random access stage, sequencing (reading) stage, and finally decoding stage of this pipeline. This architecture is similar to

the logical process of the traditional digital storage systems, except that it implements computations in both computational and biochemical domains, forming a hybrid digital-molecular system that needs both co-designed algorithms and laboratory protocols [5]. As the world storage need is expected to grow to about 1.75 x 1014 GB by 2025, whereas traditional storage media have a maximum physical density of about 103 GB/mm3, the architectural efficiency of this pipeline is one of the critical factors to evaluate DNA storage as a potentially useful long-term archival storage level in the changing data hierarchy [5].

3.1 Encoding

The encoding phase codes a binary input file, using a sequence of the four-character DNA alphabet of the set of nucleotide symbols, A,C,G,T. This transformation is determined by two major encoding paradigms, which are constrained coding and unconstrained coding.

The constrained coding implements biochemical constraints required to obtain trustworthy DNA synthesis and sequencing. Such limitations are not allowing homopolymer runs (e.g., long sequences like AAAA) and balancing of GC-content in order to stabilize molecular structure. Even though these rules enhance the reliability of synthesis, they diminish the possible storage density [6].

The unconstrained coding, on the other hand, eliminates such barriers by using data randomizing measures, which allow the theoretical maximum density of information in DNA storage. Such a theoretical ability may be presented as follows:

$$D = log_2(|\Sigma|) = log_2(4) = 2\ bits\ per\ nucleotide$$

where
$|\Sigma|=4$ represents the cardinality of the nucleotide alphabet.

Sharma et al. showed that unconstrained coding together with outer Reed-Solomon error-correcting codes could successfully handle synthesis and sequencing errors and be more efficient in terms of overall coding efficiency. Their findings indicate that the dense storage of unconstrained coding schemes is more stable across feasible error-rate regimes than constrained schemes [6].

Practical applications are based on the subdivision of encoded files into short segments of oligonucleotides (oligos), which are usually 60-200 nucleotides long. The syntax errors are rapidly added to longer strands, and they may severely impair the reconstruction results in the decoding phase [5].

3.2 DNA Synthesis (Writing)

The writing step alters an encoded nucleotide sequence into actual DNA molecules in phosphoramidite synthesis chemistry. It involves a four-step cyclic reaction whereby one nucleotide is added to a growing chain of oligonucleotides immobilized to a solid support. Conventional column-based synthesis systems have the capability to produce 96-384 distinct sequences at a time in parallel systems [5].

The subsequent progress led to array-based synthesis technologies, which were invented in the 1990s, greatly lowering the cost of synthesis. The price per nucleotide in array-based systems could be lowered to about 10^{-4} USD compared to 0.05-0.15 USD per nucleotide in the traditional column-based ones.

Though such improvements may be made, the difference in price between the use of DNA storage and the traditional media is still significant. With an assumed conservative encoding efficiency of 1 bit per nucleotide, encoding 1 terabyte of data into DNA incurs an estimated price of approximately 800 million dollars now, compared with an estimated cost of 16 pennies per terabyte in magnetic tapes. This variation is a difference of costs of about 7-8 orders of magnitude [5].

To cope with these challenges in systems architecture terms, Sharma et al. suggested a modular open-source codec and simulation framework, which decomposes both stages of the pipeline into replaceable modules that are independent of each other. The design gives researchers the ability to compare new encoding algorithms, synthesis models, and error-correction strategies to physical laboratory experiments, which speeds up the system-level innovation in the field [6].

3.3 DNA Sequencing (Reading) and Decoding

Data stored in the online system can be retrieved by means of the reading stage, where the DNA

molecules can be sequenced. There are two popular sequencing technologies, namely:

- Illumina sequencing-by-synthesis (SBS), which offers high precision with about 0.5 percent per-base error rates.
- Oxford Nanopore Technology (ONT) has longer read lengths but generally has about 10 percent error rates in single reads [5].

Nanopore sequencing usually uses longer fragments of DNA than 1 kilobase in order to make efficient use of the pore. Consequently, overlap-extension PCR is commonly used before sequencing in order to construct shorter oligos into longer ones.

After the sequencing process, the decoding pipeline recreates the initial data in a series of computational steps. Clustering is initially done by comparing similarity with the Levenshtein distance (edit distance) as the most popular metric. The minimum path through which one can change a sequence into another using insertion, deletion or substitution operations is the Levenshtein distance.

Once the clusters are distinct, consensus reconstruction algorithms, usually known as trace reconstruction methods, determine the most likely original DNA strand in each cluster of noisy reads [6]. After the reconstruction, the strands are rearranged with the help of the address of indexes, and the final encoded nucleotide sequences are then decoded to restore the initial binary data file.

This pipeline division resulting in a modular assembly of the DNA storage pipeline is necessary to allow systematic assessment and optimization. There are powerful interdependencies between the encoding component, synthesis component, sequencing component and decoding component that may otherwise obscure improvement in performance. The isolation of such stages allows researchers to more reliably attribute the increase in the reliability, efficiency, and storage density to particular algorithmic or biochemical advances [6].

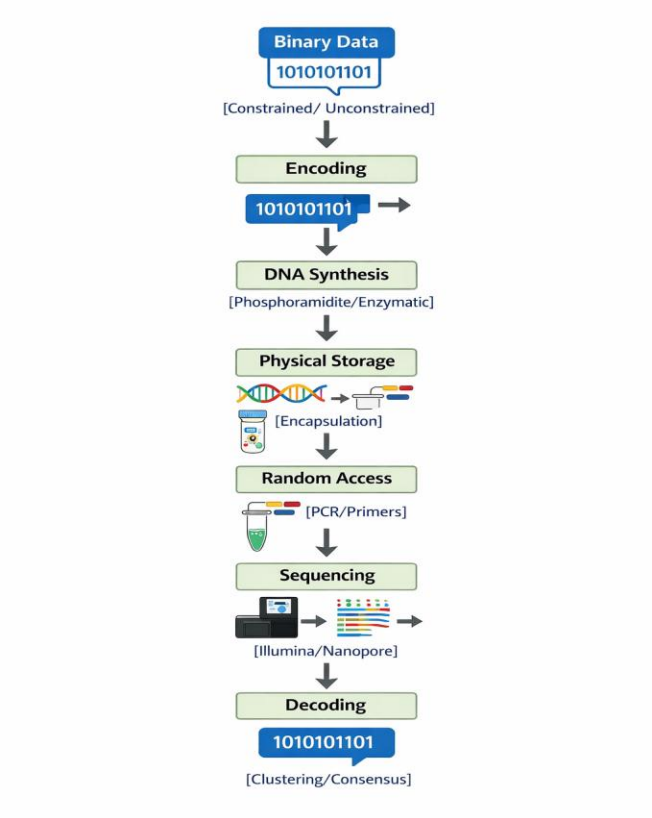

Figure 1: DNA Data Storage Pipeline

4. Error Correction and Data Integrity

Error correction is one of the most computationally intensive components of any data storage system of DNA. DNA synthesis, storage, polymerase chain reaction (PCR) amplification, and sequencing processes are all biochemical reactions that introduce various forms of additional errors, which are substitution, insertion, and deletion. These errors have a probability of occurrence that renders raw stored DNA sequences to be unreliable, unless the algorithm has an ample degree of redundancy and error-checking.

The DNA storing channel is fundamentally different from the traditional communication channel, like the binary symmetric channel or the Gaussian noise channel. Rather, DNA storage is based on a four-base quaternary nucleotide alphabet of the four bases {A, C, G, T} and has sequence-dependent error properties. Further, DNA encoding should be able to meet biochemical requirements, including having balanced GC-content (usually 45-55%) and no longer

than three or four consecutive bases of the same type. The traditional error-correction methods designed in telecommunications are not as useful as when applied directly in the DNA storage systems without any modification due to these needs [7].

These limits present information density penalties that are impossible to avoid. Though the theoretical maximum storage capacity of DNA is 2 bits per nucleotide, the real coding schemes with GC-content limitation (45-55%) and homopolymer limitation (maximum length of 3) have an effective coding potential of about 1.98 bits per nucleotide with a sequence length of about 150 nucleotides [7].

4.1 Error-Correction Architectures

There are three main architectures that can be used to define the existing error-correction methods of DNA storage systems:

1. Direct-mapping schemes with minimal redundancy
2. Inner–outer coding architectures
3. Fountain code-based approaches

The first direct-map systems proved the possibility of storing the data in DNA, but their reliability was not so high. As an example, the encoding scheme suggested by Church et al. (2012) transformed binary data into a direct sequence of the bases of DNA where A/C was mapped as 0 and T/G was mapped as 1. In this method, a dataset of about 5.2 MB was coded in 159-nucleotide sequences, comprising 96 nucleotides of payload information and 22-nucleotide PCR primer flanking sequences. Nonetheless, the system also had quite high raw error rates and needed a large sequencing coverage to guarantee proper reconstruction [8].

Later, engineering provided more advanced inner-outer code frameworks. In the system suggested by Grass et al., there were two levels where Reed-Solomon error-correcting codes were used:

- Inner codes operating at the individual DNA strand level
- Outer codes operating at the data-block level

This strategy enhanced trustworthiness. The system was able to reassemble an 83 KB text file using an average sequencing coverage of just 372x, 158-nucleotide sequences of data payload, and hence was able to reassemble the text file.

Based on this architecture, in 2018 Microsoft and the University of Washington announced random-access addressing primers of about 20 nucleotides. The primers were created to meet both GC-content and minimum Hamming distance criteria, which allows the retrieval of specific files in large DNA pools to be selective. This innovation enabled the recovery of 35 different files at an average of 5x sequencing coverage, which was very efficient in terms of accessibility of large-scale DNA storage systems [8].

4.2 Fountain Code-Based Approaches

Fountain codes are the densest known information error-correction paradigm that has been implemented to date on the storage of DNA data. Such probabilistic erasure codes can come close to storage efficiencies of 80 percent of the theoretical maximum information density [8].

The most influential of them has become the DNA Fountain scheme, which was suggested by Erlich and Zielinski (2017). The data, which was encoded into about 72,000 DNA oligonucleotides of length 152 oligonucleotides, each with a length of 152 nucleotides, in this system comprised 2.15 MB of data, which included text, images, video, and a computer operating system. It was found that data recovery occurred at an average depth of sequencing of 10.5x, which exhibited high density as well as high reliability in recovery.

A better variant of the compound-alphabet encoding system subsequently brought the number of oligonucleotides needed down to about 58,000, with about 24 parts of better logical storage density but with the same success rate of decoding [8].

The efficiency of fountain codes is commonly described by the code rate RRR, defined as

$$R = \frac{k}{n + \epsilon}$$

where:

- k represents the number of original source symbols
- n denotes the number of encoded packets, and
- ϵ represents the decoding overhead required for successful reconstruction.

A study conducted by Schwarz and Freisleben had indicated that the Raptor codes, a sophisticated version of the conventional Luby Transform (LT) codes, are much more efficient in decoding. Raptor codes can be easily put into practice with an overhead of only 1-2 extra packets to attain a decoding success probability of 99.99, in contrast to the robust-soliton distribution employed in standard DNA Fountain algorithms, which demand a significantly larger redundancy [7].

Even their optimized implementations of fountain codes reach the limits of coding of about 1.84 bits per nucleotide even with index sequences of about 13 nucleotides per 150-nucleotide strand. Moreover, the schemes are resistant to up to 38 percent DNA strand dropout rates, which is more resilient than previous inner-outer coding architectures and has a higher overall storage efficiency [7].

A key lesson of this is that the degree distribution functions applied in fountain coding must be customized to the quaternary DNA storage channel instead of being directly inherited as a result of binary communication systems. These distributions optimized to the specific error patterns in DNA result in reduced decoding decisions, improved yield rates of valid packets, and reduced costs to synthesize each stored bit, which makes the entire theory of DNA based archival storage systems more feasible in practice [7].

5. Current Challenges and Limitations

Despite substantial technical progress establishing DNA data storage feasibility, significant challenges impede practical implementation at scales relevant to addressing global data storage requirements. Cost barriers, speed limitations, scalability constraints, and standardization gaps collectively define the current developmental frontier requiring resolution before DNA storage transitions from laboratory demonstrations to operational infrastructure [9]. Understanding these limitations in quantitative terms enables realistic assessment of developmental timelines and identification of priority research directions.

Synthesis costs remain the most formidable barrier. Current phosphoramidite synthesis, the dominant method used across demonstrated systems from Church et al. (650 KB) to Blawat et al. (22 MB), prices oligonucleotide production at approximately $0.05–$0.10 per synthesized base, placing the cost of encoding even a single gigabyte of data in the range of millions of dollars [9]. While enzymatic synthesis approaches are emerging, as demonstrated by Lee et al., who encoded just 18 bytes using column-based enzymatic methods, throughput remains far too low for commercial deployment. Additional cost reductions of three to four orders of magnitude are necessary before DNA storage can approach cost competitiveness with magnetic tape archival systems, which store data at roughly $0.002 per gigabyte.

Sequencing costs for data retrieval have declined sharply due to genomics investments, yet remain prohibitive for frequent-access scenarios. Systems such as Organick et al.'s 200.2 MB demonstration required sequencing coverage depths of 5× to 36× depending on the platform (Illumina NextSeq vs. ONT Nanopore), and earlier demonstrations like Church et al.'s demanded 3,000× coverage to achieve reliable reconstruction [9]. These oversampling requirements multiply sequencing costs substantially, restricting practical DNA storage to write-once-read-rarely archival workflows where retrieval costs can be amortized over years or decades.

Operational write and read speeds fall dramatically below conventional storage performance. The leading information density achieved in demonstrated systems, 1.95 bits per nucleotide by Ping et al. and 1.71 bits per nucleotide by Yazdi et al., confirms that encoding efficiency is approaching theoretical limits, yet synthesis throughput constrains effective data writing to bytes or kilobytes per second, compared to the gigabytes per second achievable with modern solid-state drives [9]. On the retrieval side, Bar-Lev et al. demonstrated that their Deep Neural Network-based pipeline achieved up to a 3,200× speed improvement over prior leading solutions in processing 3.1 MB of sequenced data, yet even this breakthrough still leaves DNA read latencies orders of magnitude above conventional storage benchmarks [10].

Scalability challenges extend beyond individual instrument throughput to encompass the complete infrastructure needed for petabyte- or exabyte-scale

systems. No operational DNA storage facility currently approaches conventional data center capacities, and the largest demonstration, Organick et al.'s 200.2 MB archival, still required purpose-built workflows and months of preparation [9]. Transitioning to exabyte-class infrastructure would require tens of thousands of synthesis instruments, automated liquid-handling robotics, and robust quality assurance pipelines, none of which yet exist in integrated form.

Standardization and error management remain critical unresolved gaps. Across the 11 systems compared by Cao et al., encoding densities ranged from 0.19 bits/nt (Goldman et al.) to 1.95 bits/nt (Ping et al.), and error correction strategies spanned simple repetition codes to concatenated Reed-Solomon and fountain codes, with no universal standard in place [9]. Bar-Lev et al. addressed this fragmentation by combining Tensor-Product based error-correcting codes with deep learning to achieve a 40% improvement in reconstruction accuracy at a code rate of 1.6 bits per base under high-noise sequencing conditions, demonstrating that algorithmic convergence is possible even absent platform standardization [10]. Until encoding protocols, biosafety disposal frameworks, and quality benchmarks are harmonized, interoperability between vendors and regulatory compliance will remain barriers to commercial DNA storage deployment.

6. Recent Advances and Emerging Directions

The DNA data storage field continues experiencing rapid advancement across multiple technological fronts, with innovations in synthesis methodologies, storage architectures, automation systems, and computational approaches collectively accelerating progress toward practical implementation. A defining concern driving this innovation is the environmental cost of conventional de novo chemical synthesis: Sadremomtaz et al. calculated that storing just 5 minutes of a 1080p YouTube video in commercially synthesized DNA currently costs over $7 million, consumes more than 100 kWh of energy, takes over 4 days, and generates over 15 liters of toxic waste [11]. Extrapolating this to the projected ~6×10^{23} bytes of global data by 2030 would produce nearly 85 petabytes of hazardous chemical by-products, surpassing the volume of water the Mississippi River discharges into the Gulf of Mexico over 40 years [11]. These calculations establish the urgent necessity of synthesis alternatives.

CRISPR-based genome editing has emerged as a transformative approach, replacing synthesis-dependent writing with targeted molecular overwriting. Sadremomtaz et al. introduced the "DNA Mutational Overwriting Storage" (DMOS) system, which uses combinatorial, addressable, in vitro CRISPR base-editing reactions to write data onto pre-made "blank" DNA tapes replicated in bacteria, entirely eliminating the need for repeated de novo synthesis [11]. Each DMOS register consists of 16 distinct 23 bp domain "bits," each paired with a 40 bp index sequence, and a block of 48 registers encodes 768 bits per codeword using a Protograph LDPC code with 25% redundancy. The system demonstrated writing and recovery of 1,250 bits, including a school logo bitmap across 32 registers, with 100% accuracy after 100,000 nanopore sequencing reads, and an initial bit recovery plateau of 91.37 ± 1.7% after just 10,000 reads [11]. Critically, once the blank tapes are synthesized and replicated in bacteria, the DNA can be reproduced indefinitely, with writing driven solely by commercially available enzymes and decoded using a Bayesian classifier operating on nanopore sequencing output.

Microfluidic integration represents another decisive advance, enabling end-to-end DNA storage within miniaturized, automated platforms. Li et al. developed DNA-DISK, a tabletop device measuring only 33.8 × 33.0 × 22.6 cm^3 and weighing 8.7 kg, which integrates enzymatic single-nucleotide synthesis, agarose-based on-chip encapsulation, and pyrosequencing on a 96-electrode digital microfluidic (DMF) chip [12]. Using an engineered terminal deoxynucleotidyl transferase (E-ZaTdT) with reversible 3′-ONH_2 terminators, the platform achieved an average stepwise synthesis yield of 97.67% across 18-mer oligonucleotides, with a full-length yield of 92.76% and average accuracy of 99.35% per base [12]. Deletion errors, the dominant error type at 1.85% for 18-mers with biocapping, were substantially reduced from 6.34% without capping. The system successfully encoded and retrieved 228 bits of digital music, the first eight

measures of "MoLiHua," with a write-to-read latency of just 4.4 minutes per bit across 8-plex parallelized operation [12]. DNA encapsulation using 0.5% agarose on-chip achieved a DNA recovery rate of 98.27%, and DNA protected by agarose retained 90.94% recovery after 15 days at room temperature, compared to only 74.37% for unprotected DNA [12].

Enzymatic synthesis commercialization is also accelerating independently of microfluidic integration. Industry progress already includes synthesis lengths of 280 nucleotides demonstrated by DNA Script, 300 nucleotides by Camena, and 1,005 nucleotides by Ansa Biotechnologies, lengths unachievable via phosphoramidite chemistry, with enzymatic methods projected to reduce costs by several orders of magnitude compared to current chemical synthesis [12]. The theoretical storage density achievable on a standard-sized DMF chip already exceeds 50 TB cm^{-3}, representing a 6,144-fold improvement over LTO-9 magnetic tape density (8.2 GB cm^{-3}), establishing a compelling long-term density argument even at current throughput [12].

Encoding diversity is expanding in parallel. De Silva and Ganegoda document that practical coding schemes range from simple 2-bit-per-base direct mapping to sophisticated approaches achieving 1.917 bits/nt using quaternary constraints, 1.98 bits/nt with GC- and HP-constrained fountain-style codes, and as high as 2.14 bits/nt through compression-integrated image storage strategies [13]. Across more than 20 documented CODEC implementations, the field now offers encoding densities from 1.33 to 3.9 bits/nt depending on constraint sets and degenerate base approaches, demonstrating that information-theoretic headroom remains substantial for future improvement [13].

Platform	Write Mechanism	Accuracy/ Yield	Demonstrated Capacity	Storage Density
DMOS	CRISPR base editing (dCas9 + APOBEC3A)	100% data recovery; 91.4% at 10,000 reads	1,250 bits	768 bits/codeword (25% LDPC)
DNA-DISK	Enzymatic synthesis (E-ZaTdT) on DMF chip	99.35%/base; 97.67% stepwise yield	228 bits (music sheet)	50 TB cm^{-3} (theoretical)
Enzymatic synthesis (commercial)	TdT-based polymerization	Toxic-waste-free; cost reduction of several orders of magnitude projected	Up to 1,005 nt strands	6,144× LTO-9 tape density
CODEC advances	Fountain, Huffman, degenerate bases	GC ~50%, HP-constrained	KB–MB scale demos	1.33–3.9 bits/nt

Table 1: Key Emerging DNA Storage Platforms

7. Future Prospects and Impact on Digital Systems

The trajectory of DNA data storage technology suggests transformative impacts on digital information infrastructure as technical barriers progressively diminish and economic competitiveness improves. Gervasio et al. frame the developmental horizon across four timeframes: a near future of 10–20 years requiring standardized metadata and biocybersecurity protocols; a medium future of 20–100 years demanding universal CODEC standards; an archaeological future of 100–1,000 years calling for fully standardized sample handling and sequencing methodologies; and a paleontological future exceeding 1,000 years requiring biological watermarks distinguishing synthetic from natural DNA [14]. These timeframes reflect a fundamental advantage: DNA can retain information for millions of years, as demonstrated by the recovery of environmental DNA dating back 2 million years, vastly surpassing the 15–30 year rated lifespan of LTO magnetic tape [14].

Near-term deployment is most likely to target "cold" and "glacial" storage, data accessed fewer than once every 50 years, where DNA's density and longevity advantages justify current cost premiums. Illumina's 2023 NovaSeq X Plus platform announcement

demonstrated sequencing of 16 trillion bases within 48 hours, yielding approximately 92 million bases per second or roughly 23 MB/s, a read rate increasingly competitive with industrial storage standards [14]. On the write side, Gervasio et al. report that the phosphoramidite synthesis route currently operates at 1 base every 4–6 minutes, yet parallelized chip platforms have achieved densities of 25 million oligonucleotides per cm^2, and Custom Array's Miniature Semiconductor Technology is projected to write up to 200 billion oligonucleotides per chip [14]. Writing platforms are expected to achieve massive parallel synthesis of gigabytes per chip within a few years and terabytes per chip within the current decade [14].

Hybrid storage hierarchies represent the most realistic near-term architectural scenario. DNA storage would occupy a new "glacial" tier below magnetic tape in existing hierarchies, with automated tiering software routing infrequently accessed data to DNA based on access frequency, retention policy, and cost targets. The energy economics are compelling: up to 49% of all data center operating costs arise from temperature control alone, and DNA stored in sealed, inert-atmosphere encapsulation requires negligible active cooling, potentially operating indefinitely at room temperature [14]. Healthcare, regulatory compliance archives, scientific data repositories, and entertainment content preservation represent the likeliest initial commercial adoption segments, as demonstrated by Netflix and Twist Bioscience's proof-of-concept encoding of the Biohackers series in synthetic DNA [14].

Biocybersecurity presents a concrete challenge requiring resolution before widespread deployment. Synthetic malware encoded within DNA sequences can be designed to attack sequencing pipelines and bioinformatics software during the readout process, an attack vector documented as early as 2017 [14]. Gervasio et al. argue that robust detection of malicious DNA sequences must become a mandatory pipeline step, with deep learning models proposed as promising automated screening tools [14]. DMOS-style systems that eliminate de novo synthesis by writing on prereplicated blank tapes offer a partial structural safeguard against this class of attack, since no new sequence pool is synthesized on demand [11].

Standardization is identified by multiple sources as the single most critical systemic requirement. The DNA Data Storage Alliance, whose membership includes IBM, Microsoft, Seagate, Lenovo, Dell, and Western Digital, has initiated working groups to address unified CODEC specifications, yet no universal standard currently defines how binary data maps to nucleotides or what error correction rates qualify as production-grade, or how biosafety disposal of stored DNA should be regulated [14]. As Gervasio et al. note, without a single standardized CODEC, data stored by one company today may become inaccessible within decades if the originating organization ceases operations [14]. The imperative for a universal standard mirrors historical parallels: analog storage formats from the 1960s–80s are already partially unreadable due to software and hardware obsolescence.

Looking further ahead, the convergence of DNA storage with synthetic biology opens prospects for self-replicating, biologically maintained archives. Sadremomtaz et al. demonstrate that DMOS blank tapes can be replicated indefinitely in bacteria, providing a biological copy-propagation mechanism with no electronic infrastructure [11]. Li et al. project that active-matrix DMF platforms scaling to several thousand addressable electrodes could enable DNA Hard Disk Drive architectures storing terabytes per chip footprint, fundamentally redefining storage density norms [12]. The field has shifted from a speculative research domain to one with measurable industrial momentum, with the core question no longer being whether DNA storage is feasible, but how rapidly the cost-per-bit curve will converge with magnetic tape's current $0.002 per gigabyte threshold [14].

Conclusion

This review has systematically examined the technical foundations, current capabilities, and developmental trajectory of DNA-based data storage as a response to the escalating global data preservation crisis. The analysis demonstrates that DNA storage has transitioned from a speculative concept to an experimentally validated technology, with encoding capacities approaching 2 bits per nucleotide, successful retrieval of datasets exceeding

200 MB, and error correction architectures achieving near-theoretical efficiency at sequencing coverage depths below 11×. These achievements establish DNA as the only known storage medium capable of simultaneously satisfying the density, longevity, and energy efficiency requirements necessary for sustainable archival infrastructure. However, the seven-to-eight order-of-magnitude cost disparity between DNA synthesis and magnetic tape storage remains the most formidable barrier to commercialization. The technological diversification documented here, encompassing CRISPR-based mutational overwriting, microfluidic enzymatic platforms, and massively parallelized array synthesis, suggests multiple viable pathways toward cost reduction, with the DMOS approach representing a particularly significant innovation that simultaneously addresses cost, throughput, and environmental sustainability concerns.

The practical deployment scenario emerging from this analysis positions DNA storage not as a replacement for existing media but as a "glacial" tier within hybrid storage hierarchies, receiving data accessed fewer than once per decade and retained for periods spanning decades to centuries. Realization of this potential requires coordinated progress across three domains: synthesis cost reduction toward the $0.002 per gigabyte threshold, universal standardization of encoding formats and biosafety protocols through initiatives such as the DNA Data Storage Alliance, and development of biocybersecurity frameworks addressing synthetic sequence vulnerabilities. The question confronting the field is no longer whether DNA can store digital data, but how rapidly the remaining engineering and economic barriers can be overcome to enable deployment at the scale demanded by an inexorably expanding global datasphere.

References

[1] Alex Sensintaffar et al., "Advancing Archival Data Storage: The Promises and Challenges of DNA Storage System," ACM, 2025. https://dl.acm.org/doi/pdf/10.1145/3723166

[2] Melpomeni Dimopoulou and Marc Antonini, "Data and image storage on synthetic DNA: existing solutions and challenges," Springer, 2022, no. 23, 2022. https://link.springer.com/content/pdf/10.1186/s13640-022-00600-x.pdf

[3] Linda C. Meiser et al., "Synthetic DNA applications in information technology," Nature, 2022. https://www.nature.com/articles/s41467-021-27846-9

[4] Zarif Bin AKHTAR and Ahmed TAJBIUL RAWOL, "Unlocking the Future for the New Data Paradigm of DNA Data Storage : An Investigative Analysis of Advancements, Challenges, Future Directions," JIS, 2024. https://revues.imist.ma/index.php/JIS/article/view/47102

[5] Andrea Doricchi et al., "Emerging Approaches to DNA Data Storage: Challenges and Prospects," ACS, 2022. https://pubs.acs.org/doi/10.1021/acsnano.2c06748

[6] Puru Sharma et al., "DNA Storage Toolkit: A Modular End-to-End DNA Data Storage Codec and Simulator." https://prongs1996.github.io/assets/pdf/DNA_Storage_Toolkit.pdf

[7] Peter Michael Schwarz and Bernd Freisleben, "Optimizing fountain codes for DNA data storage," ScienceDirect, 2024. https://www.sciencedirect.com/science/article/pii/S2001037024003581

[8] Chenyang Wang et al., "DNA Storage Toolkit: A Modular End-to-End DNA Data Storage Codec and Simulator," Springer, 2022. https://link.springer.com/article/10.1007/s42514-022-00094-z

[9] Ben Cao et al., "Efficient data reconstruction: The bottleneck of large-scale application of DNA storage," ScienceDirect, 2024. https://www.sciencedirect.com/science/article/pii/S2211124724000275

[10] Daniella Bar-Lev et al., "Deep DNA Storage: Scalable and Robust DNA-based Storage via Coding Theory and Deep Learning," arXiv:2109.00031, 2024. https://arxiv.org/abs/2109.00031

[11] Afsaneh Sadremomtaz et al., "Digital data storage on DNA tape using CRISPR base editors,"

Nature, 2023. https://www.nature.com/articles/s41467-023-42223-4

[12] Kunjie Li et al., "DNA-DISK: Automated end-to-end data storage via enzymatic single-nucleotide DNA synthesis and sequencing on digital microfluidics." PNAS, 2024. https://www.pnas.org/doi/10.1073/pnas.2410164121

[13] Pavani Yashodha De Silva and Gamage Upeksha Ganegoda, "New Trends of Digital Data Storage in DNA," Wiley, 2016. https://onlinelibrary.wiley.com/doi/full/10.1155/2016/8072463

[14] Joao Henrique Diniz Brandao Gervasio et al., "How close are we to storing data in DNA?" ScienceDirect, 2024. https://www.sciencedirect.com/science/article/pii/S0167779923002354