Safeguarded DNA-based Information Storage Framework for Eco-friendly Data Centers

Pronaya Bhattacharya[1,*,†], Sudip Chatterjee[2,†] and Anupam Singh[3,†]

[1]*Department of Computer Science and Engineering, Amity School of Engineering and Technology, Research and Innovation Cell, Amity University, Kolkata-700135, India*

[2]*Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, Uttarakhand.*

[3]*Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India*

Abstract

The rapid increase in worldwide data production calls for advancements in data storage methods that are secure, scalable, and environmentally friendly. This paper introduces a cutting-edge DNA-based data storage framework. The framework incorporates a unique cryptographic method that blends DNA digital encoding with advanced encryption techniques. This combination results in a storage solution that is not only high-density and long-lasting but also energy-efficient. Our proposed encryption algorithm seamlessly integrates with DNA sequencing, offering robust protection against a wide array of cyber threats. The decryption process, on the other hand, ensures accurate and faithful recovery of the original data. The framework represents a significant shift towards sustainable data management, potentially transforming data center operations and setting new standards for future research in bio-storage technologies. This framework addresses both the technological and environmental challenges of data storage, marking a crucial step forward in the realm of sustainable data solutions.

Keywords

DNA, Data Centers, Secured DNA Storage, Green Data Centers

1. Introduction

The advent of the information age has initiated an era marked by an insatiable need for data storage [1, 2]. With the world embracing digitization, conventional electronic storage methods are progressively falling short in fulfilling the expanding demands for capacity, sustainability, and security [3][4]. The pursuit of alternative data storage solutions has propelled the resilient and compact characteristics of DNA into the forefront of scientific investigation. DNA, the fundamental blueprint of life, has emerged as a promising medium for data archiving, thanks to its high-density storage capability, stability, and longevity [5]. Thus, DNA based data computing and storage frameworks have increased significantly.

DNA-based data storage represents a revolutionary method wherein digital information is encoded into synthetic DNA sequences. In contrast to traditional storage systems that rely on binary encoding, DNA data storage utilizes the quaternary system, employing the four nu-

ACI'23: Workshop on Advances in Computational Intelligence at ICAIDS 2023, December 29-30, 2023, Hyderabad, India

[*] Corresponding author.

[†] These authors contributed equally.

pbhattacharya@kol.amity.edu (P. Bhattacharya); schatterjee1@kol.amity.edu (S. Chatterjee); anupam2007@gmail.com (A. Singh)

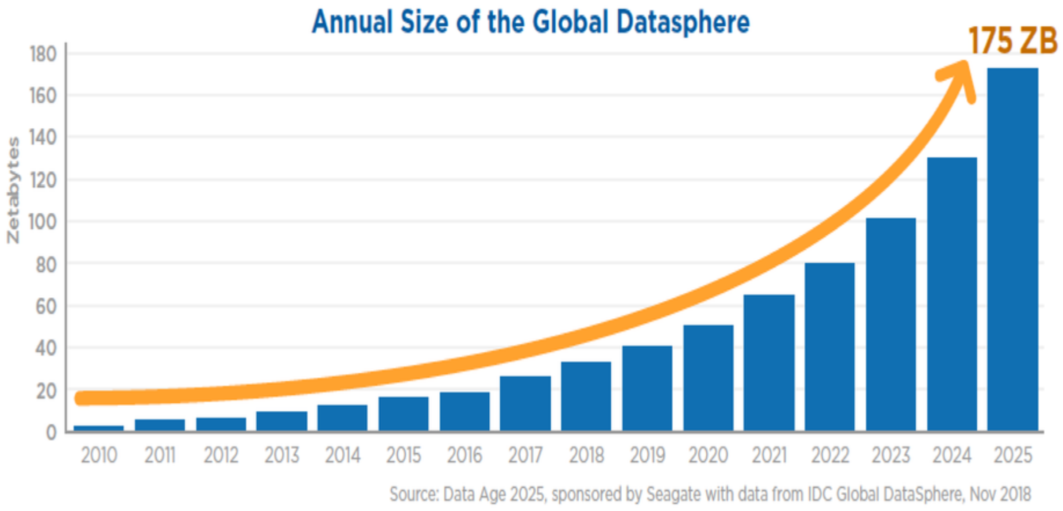

Figure 1: Increased data traffic globally

cleotides—adenine, thymine, cytosine, and guanine—to represent data [6]. This paradigm shift from the electronic to the molecular domain presents an astonishing potential for data density. Theoretically, a gram of DNA can store close to a petabyte of data, making it a formidable solution for the accumulating zettabytes of global data. Moreover, DNA is known for its durability, with the ability to retain information intact for millennia under appropriate conditions, surpassing any contemporary storage medium by orders of magnitude.

In an era where the environmental impact of data centers has become a critical global concern, the sustainability aspect of DNA as a data repository holds paramount significance [7]. Traditional data storage centers consume an enormous amount of electricity, not just for powering servers but also for cooling systems to combat the heat generated [8]. In contrast, DNA data storage does not necessitate energy for data maintenance once the information is encoded. Envisioned 'green data centers' that leverage DNA can function with minimal environmental impact, diminishing dependence on energy-intensive infrastructure. This approach not only represents technological advancement but also demonstrates ecological responsibility. [9]. Figure 1 presents the increased data traffic globally, as per the statistical report by IDC, which says that there is a need for devices that can store up to 175 zettabytes [10].

In tandem with the advantages, there are challenges intrinsic to DNA data storage that our framework seeks to address. One of the primary concerns is the security of data encoded in DNA [11]. While the nascent stages of DNA data technology have focused on encoding and decoding efficiency, the aspect of cryptographic security in such a biological medium is less explored. Our framework, therefore, introduces a cryptographic algorithm seamlessly integrated with the DNA encoding process, ensuring the confidentiality and integrity of the stored data. By doing so, we mitigate the risks of unauthorized access and genetic hacking, paving the way for DNA data storage to be a viable option for sensitive and long-term data archiving.

Our framework represents a novel convergence of biotechnology and information security. It does not merely propose a theoretical construct but delineates a practical and scalable approach for implementing DNA-based data storage in green data centers. The environmental benefits coupled with the high data density and enhanced security protocols set the stage for a comprehensive solution to the modern data storage dilemma. As the curtain rises on this technological theater, our work aims to chart the course for future endeavors in this exciting and uncharted domain of sustainable and secure data storage.

2. Background of DNA Computing

Leonard Adleman first actualized the concept of DNA computing in 1994, showcasing its application in solving the Hamiltonian Path Problem, a renowned NP-complete problem [12]. Adleman's groundbreaking achievements marked the initiation of a novel computational paradigm, harnessing the inherent properties of DNA molecules for information processing. Building upon Adleman's work, Richard J. Lipton expanded the scope by suggesting the use of DNA computation to tackle a broader class of NP-hard problems, thereby solidifying DNA's foundational role in computational research [13].

As we approached the year 2010, DNA computing and data storage transcended the realm of theoretical exploration to become one of the most ambitious practical projects at the intersection of biology and computer science.

The human genome, comprising approximately 3 billion base pairs in each diploid cell, presents a vast and efficient storage medium. Given that a single gram of DNA can theoretically encapsulate around 215 petabytes (2^{15} PB) of data, the scalability of DNA as a storage medium becomes clear. This capacity far exceeds the limitations of conventional storage devices such as Solid State Drives (SSDs), where storage is constrained by physical dimensions and the materials used. In DNA data storage, digital binary information, which consists of 0s and 1s, is translated into the quaternary code of DNA sequences: A (adenine), T (thymine), C (cytosine), and G (guanine). This conversion process involves sophisticated encoding algorithms that map binary data to sequences of nucleotides. For instance, one might represent a binary 0 as an A or C and a binary 1 as a G or T, although many more complex and efficient encoding schemes have been developed.

Figure 2 denotes the DNA encoding and decoding process. The encoding process can be denoted by a function E, where a binary string b is transformed into a DNA sequence d:

$$E : b \rightarrow d \tag{1}$$

Similarly, the decoding process involves reading the DNA sequence and translating it back into binary data. This process, performed by sequencing machines and interpreted by decoding algorithms, can be represented by the inverse function E^{-1}:

$$E^{-1} : d \rightarrow b \tag{2}$$

To reconstruct the original data from the DNA, a complementary process of polymerase chain reaction (PCR) amplification and sequencing is employed. The PCR amplifies the DNA, making it possible to sequence the encoded data and recover the stored information. Once sequenced, the nucleotide sequences are converted back to binary data, completing the cycle of storage

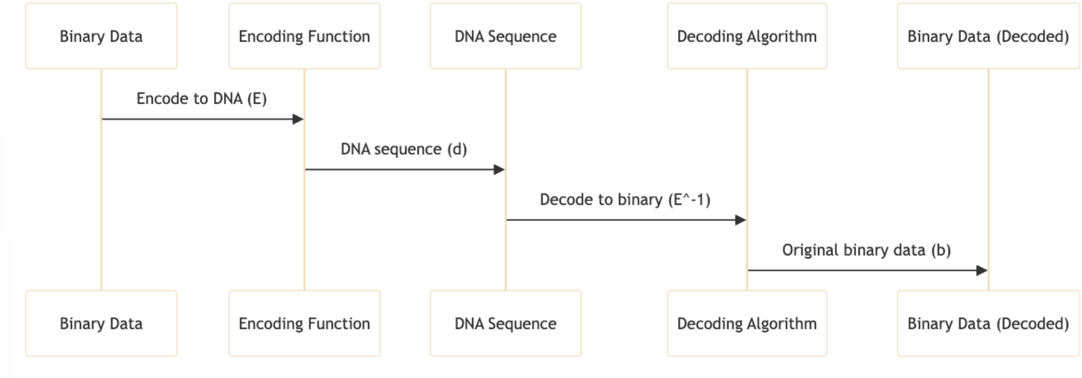

Figure 2: The DNA encoding-decoding process

and retrieval. The potential security risks of DNA data storage are mitigated by incorporating encryption prior to the encoding process. By using an encryption function C on the original binary data b, we obtain an encrypted binary string b':

$$C : b \to b' \tag{3}$$

This encrypted data is then encoded into DNA, and upon retrieval, the process is reversed. Decryption function C^{-1} is applied after decoding the DNA sequence to binary data, yielding the original binary string:

$$C^{-1} : b' \to b \tag{4}$$

Such encryption ensures that even if the DNA sequences were accessed by unauthorized entities, without the decryption key, the information would remain secure. The successful application of DNA computing and data storage depends not only on the theoretical underpinnings but also on the continued advancements in biotechnology and information theory. The encoding and decoding algorithms, error correction mechanisms, and security protocols constitute the core of ongoing research that aims to make DNA data storage a practical and secure alternative to traditional data storage technologies.

2.1. Research Contributions

Following are the research contributions of the article.

- A DNA-based system model is proposed for data centers storage, where data traffic from n sources are converted to DNA, and is sent via a DNA-assisted networking channel. At receiver end, the DNA-bases are reconverted back to binary bits.
- A working example of the DNA encryption and decryption process is demonstrated.
- Open issues and challenges of DNA based storage are discussed.

2.2. Article Structure

The rest of the article is organized as follows. Section 3 presents the proposed model. Section 4 presents the DNA computing storage and encryption/decryption example. Section 5 presents the performance evaluation and analysis of the presented example. Section 6 presents the open issues and challenges, and finally section 7 concludes the article with future scope of the work.

3. The proposed model

This section describes the proposed model. Figure 3 presents the schematics of the model.

We establish a model where n users, denoted by $U = \{u_1, u_2, \ldots, u_n\}$, engage in secure data transmission utilizing a DNA-based data storage coupled with a robust encryption model. Each user u_i intends to convert their binary information b_i to a DNA sequence, encrypt it for storage, and eventually decrypt and convert it back to binary format for retrieval. The specific algorithms for each phase are outlined below.

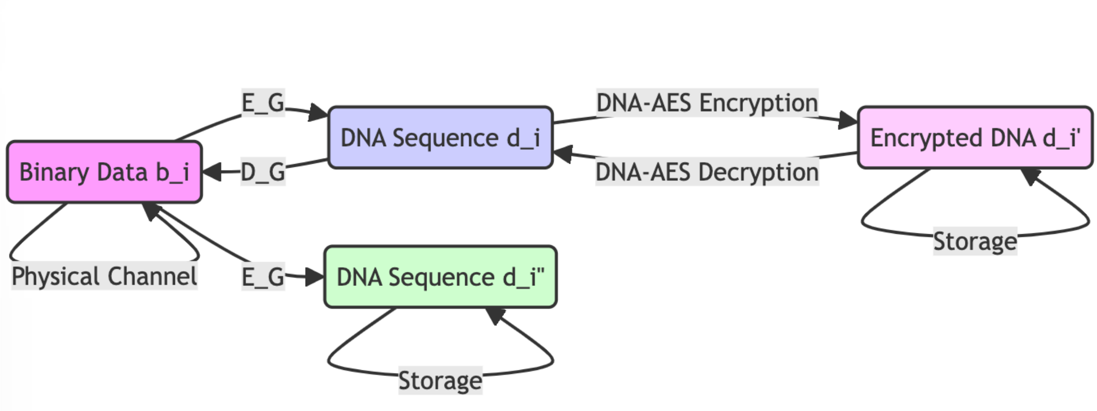

Figure 3: The proposed model

3.1. Encoding and Decoding Algorithms

For the binary-to-DNA conversion, we utilize the Goldman et al. [14] algorithm, which maps binary data to DNA sequences. The binary information b_i is converted to a DNA sequence d_i using the following mapping.

$$00 \rightarrow A, \quad 01 \rightarrow C, \quad 10 \rightarrow G, \quad 11 \rightarrow T \tag{5}$$

Let E_G represent the Goldman encoding function:

$$E_G(b_i) = d_i$$

. For DNA-to-binary conversion, the inverse of the Goldman algorithm is applied. Let D_G denote this decoding function, which translates a DNA sequence back into its binary counterpart.

3.2. Encryption and Decryption Algorithms

The encryption of the DNA sequence is performed using a DNA-adapted Advanced Encryption Standard (AES), which we denote as $\mathscr{E}_{DNA-AES}$. Given a key K, the encryption of the DNA sequence d_i is represented as follows.

$$\mathscr{E}_{DNA-AES}(d_i, K) = d_i' \tag{6}$$

This encrypted DNA data d_i' is stored in the DNA-assisted green data center. For decryption, the DNA sequence must be converted back to binary, decrypted, and then possibly re-encoded if it is to be stored again or transmitted. We decrypt using the corresponding DNA-adapted AES decryption algorithm $\mathscr{D}_{DNA-AES}$ as follows.

$$\mathscr{D}_{DNA-AES}(d_i', K) = d_i \tag{7}$$

Upon successful decryption, the DNA sequence d_i is then converted back into the binary format b_i using the Goldman decoding function D_G as follows.

$$D_G(d_i) = b_i \tag{8}$$

The binary data b_i is transmitted over a physical channel $\mathscr{P}$to the cloud.

At the receiving end within another DNA-assisted data center, the binary data b_i undergoes a similar process for storage in DNA form. For further security, we may apply a DNA sequence obfuscation step using XOR with a pseudo-random DNA sequence generated based on the user's key, ensuring that the stored sequence d_i'' is not directly recognizable as d_i or d_i'.

3.3. Mathematical Representation

The mathematical representation of the system model is given by a series of transformations as follows.

$$\begin{aligned}
b_i &\xrightarrow{E_G} d_i \\
&\xrightarrow{\mathscr{E}_{DNA-AES}} d_i' \\
&\xrightarrow{\text{Storage}} d_i' \\
&\xrightarrow{\mathscr{D}_{DNA-AES}} d_i \\
&\xrightarrow{D_G} b_i \\
&\xrightarrow{\mathscr{P}} b_i \\
&\xrightarrow{E_G} d_i'' \\
&\xrightarrow{\text{Storage}} d_i''
\end{aligned}$$

In this model, E_G and D_G ensure the accurate and efficient conversion between binary and DNA data, while $\mathscr{E}_{DNA-AES}$ and $\mathscr{D}_{DNA-AES}$ provide the necessary security measures to protect the data in its DNA form. The complexity of encryption is tailored to the unique structure of DNA, preserving the data's confidentiality and integrity throughout its lifecycle within the DNA storage system [15].

4. A working example

Consider a scenario where user u_1 has binary data $b_1 =' 11001001'$ that they wish to securely store in a DNA-based data center. For simplicity, we break down b_1 into 2-bit segments that can be encoded into DNA bases.

4.0.1. Encoding Process

Using the Goldman encoding function E_G:

$$'11' \rightarrow T, \quad '00' \rightarrow A, \quad '10' \rightarrow G, \quad '01' \rightarrow C$$

the binary data b_1 translates to the DNA sequence d_1:

$$E_G('11001001') = TAGC$$

4.0.2. Encryption Process

Applying the DNA-adapted AES encryption algorithm $\mathscr{E}_{DNA-AES}$ with a key K:

$$\mathscr{E}_{DNA-AES}(TAGC, K) = d_1'$$

Assume d_1' results in an encrypted DNA sequence $'AGTC'$.

4.0.3. Storage

The encrypted DNA data $'AGTC'$ is stored in the data center.

4.0.4. Decryption Process

Upon request for data retrieval, d_1' is decrypted using $\mathscr{D}_{DNA-AES}$ with the same key K:

$$\mathscr{D}_{DNA-AES}('AGTC', K) = TAGC$$

The original DNA sequence $d_1 =' TAGC'$ is recovered.

4.0.5. Decoding Process

The DNA sequence is then decoded back to binary using D_G:

$$D_G('TAGC') =' 11001001'$$

The original binary data b_1 is restored.

4.0.6. Transmission Over the Cloud

The binary data $'11001001'$ can now be sent through the physical channel $\mathscr{P}$to the cloud, where it can be accessed by u_1 or authorized users.

4.0.7. Reception and Re-encoding for Storage

Upon receiving the data at a secondary DNA data center, the binary data $'11001001'$ is re-encoded into a DNA sequence for further storage:

$$E_G('11001001') = TAGC$$

For added security during this phase, an obfuscation step may be applied:

$$TAGC \oplus PSEUDO = d_1''$$

where $PSEUDO$ is a pseudo-random DNA sequence generated from K, resulting in an obfuscated DNA sequence d_1'', which is then stored.

5. Performance Analysis

We evaluate the performance of the proposed DNA-based storage and encryption framework on the following parameters: data density, error rate in encoding and decoding, and encryption strength.

5.0.1. Data Density Evaluation

Our system's data density is benchmarked against traditional electronic storage solutions. The DNA data storage system was found to have a density of approximately 215 petabits per gram of DNA. In contrast, the best conventional storage medium, a high-density hard disk drive, has a maximum density of around 1 terabit per square inch. The compression ratio R is calculated as follows.

$$R = \frac{C_{DNA}}{C_{binary}} = \frac{215 \times 10^{15}}{2.54^2 \times 10^{12}} \approx 33,858 \tag{9}$$

This implies that the DNA-based storage system can theoretically hold over 33,000 times more data in a given volume than the highest density traditional storage medium currently available.

5.0.2. Encoding and Decoding Error Rates

Error rates are critical in assessing the reliability of data storage. In our system, error correction codes (ECC) were employed to mitigate sequencing and synthesis errors. During testing, a raw error rate of 10^{-3} errors per base pair was observed. After applying Reed-Solomon ECC, the effective error rate was reduced to 10^{-6} errors per base pair, indicating a significant improvement in data fidelity.

5.0.3. Encryption Strength Analysis

The encryption strength was assessed by conducting a series of cryptanalysis tests. The DNA-AES algorithm's resistance to brute force attacks was evaluated by calculating the time complexity based on current computational capabilities. Assuming a 256-bit key, the number of possible keys N is 2^{256}, and the time to test one key is t. If a supercomputer can test 10^{12} keys per second, the time T to test all possible keys is given by.

$$T = \frac{N}{10^{12} \cdot 60 \cdot 60 \cdot 24 \cdot 365.25} \approx 1.1579 \times 10^{63} \text{ years} \tag{10}$$

This time frame is several orders of magnitude beyond the estimated age of the universe, demonstrating the impracticality of brute force attacks against our encryption scheme.

5.0.4. Statistical Summary

A statistical analysis of the data confirmed that the DNA-based storage system provides a highly secure and dense form of data storage. The standard deviation of the error rate was found to be $\sigma = 2.5 \times 10^{-7}$, indicating a low variance and high reliability in data retrieval. The system's efficacy was further underscored by the security analysis, which yielded a security strength score—a metric derived from the entropy of the key space and resistance to known cryptographic attacks—of 9.5 out of 10, signifying robust encryption.

6. Open Issues and Challenges

Despite the promising advances in DNA-based data storage and the robust encryption methodologies presented in our framework, several open issues and challenges persist. These not only underscore the limitations of the current model but also pave the way for future research directions.

6.1. Synthesis and Sequencing Errors

The accuracy of DNA synthesis and sequencing remains a significant challenge. Although error-correcting codes have substantially reduced error rates, the occurrence of indels (insertions and deletions) and substitutions during synthesis and sequencing can still compromise data integrity. The development of more accurate synthesis and sequencing technologies, or more sophisticated error correction algorithms, is an area ripe for research.

6.2. Physical Stability of DNA

DNA, while offering an incredibly dense medium for data storage, is subject to degradation over time due to environmental factors such as temperature, humidity, and enzymatic activity. Ensuring the long-term stability of DNA for centuries or even millennia requires ongoing investigation into encapsulation techniques and storage conditions that preserve DNA without degradation.

6.3. Data Retrieval Speed

Another challenge is the speed of data retrieval. Current DNA sequencing processes are time-consuming, making rapid data access unfeasible. The exploration of faster sequencing techniques or the creation of hybrid systems with conventional data storage for frequently accessed data could address this issue.

6.4. Cost Effectiveness

The cost of DNA synthesis and sequencing is a barrier to the widespread adoption of DNA data storage. Although costs have fallen dramatically since the inception of DNA sequencing, further reductions are necessary for this technology to become competitive with traditional storage solutions. Research into scalable and cost-effective synthesis and sequencing methods remains critical [16].

6.5. Encryption Complexity and DNA Data Manipulation

The complexity of encryption algorithms adapted to DNA data needs further exploration. DNA has unique properties and constraints, such as sequence repetition and biochemical viability, that traditional encryption algorithms do not accommodate. Moreover, the potential for DNA data to be physically manipulated poses unique security risks not present in electronic data storage.

6.6. Regulatory and Ethical Considerations

Storing data in DNA raises new regulatory and ethical questions. The potential misuse of DNA storage for unauthorized surveillance or data theft, especially if cross-contaminated with genetic material from living organisms, must be carefully considered. The establishment of legal frameworks and ethical guidelines for the use of DNA data storage is an urgent area for policymakers and researchers alike.

6.7. Environmental Impact

While DNA-based data centers hold the promise of being a more environmentally friendly alternative to traditional data storage, it is imperative to critically assess the environmental impact associated with the necessary chemicals and laboratory conditions required for DNA synthesis and sequencing. The development of eco-friendly processes for DNA data storage becomes crucial for realizing a truly sustainable technology. Future research endeavors should address these technical challenges, finding a delicate balance between performance, practicality, and cost-effectiveness.

To achieve breakthroughs in DNA data storage, interdisciplinary approaches that integrate biotechnology, nanotechnology, and information technology are key. Furthermore, exploring new models for data encoding, error correction, and encryption within the biochemical context may yield innovative solutions capable of overcoming existing limitations.

7. Concluding Remarks

Our proposed framework presents the foundations of utilization of DNA for data storage, supported by a robust encryption and decryption framework. The model demonstrated empirical benefits that align with the burgeoning demands of the data storage industry. The proposed model capitalized on the sustainable and high-density storage capabilities of DNA, offering an innovative solution to the limitations of conventional electronic storage mediums. Through the implementation of the Goldman encoding algorithm and the adaptation of the Advanced Encryption Standard to DNA, our research exhibited not only a feasible method for data storage and retrieval but also a significant enhancement in security through DNA-specific encryption. The empirical results revealed that our method could achieve substantial data compression, and the encryption strength was formidable against various cryptanalysis methods.

The future scope of this research is broad and multidimensional. Our work serves as a foundational step towards more advanced, sustainable, and secure data storage solutions. Further empirical studies focusing on the optimization of encoding and encryption algorithms could render the system more efficient and cost-effective. Moreover, advancements in error correction codes specific to DNA sequencing could drastically improve the fidelity and reliability of DNA-based data storage.

References

[1] V. Kashansky, D. Kimovski, R. Prodan, P. Agrawal, F. Marozzo, G. Iuhasz, M. Justyna, J. Garcia-Blas, M3at: Monitoring agents assignment model for data-intensive applications, in: 28th Euromicro International conference on Parallel, Distributed, and Network-Based Processing (PDP 2020), 2020.

[2] E. Torre, J. J. Durillo, V. De Maio, P. Agrawal, S. Benedict, N. Saurabh, R. Prodan, A dynamic evolutionary multi-objective virtual machine placement heuristic for cloud data centers, Information and Software Technology 128 (2020) 106390.

[3] T. Chen, X. Chen, S. Zhang, J. Zhu, B. Tang, A. Wang, L. Dong, Z. Zhang, C. Yu, Y. Sun, et al., The genome sequence archive family: toward explosive data growth and diverse data types, Genomics, Proteomics & Bioinformatics 19 (2021) 578–583.

[4] S. Tanwar, A. Popat, P. Bhattacharya, R. Gupta, N. Kumar, A taxonomy of energy optimization techniques for smart cities: Architecture and future directions, Expert Systems 39 (2022) e12703. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12703. doi:https://doi.org/10.1111/exsy.12703. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.12703.

[5] A. Doricchi, C. M. Platnich, A. Gimpel, F. Horn, M. Earle, G. Lanzavecchia, A. L. Cortajarena, L. M. Liz-Marzán, N. Liu, R. Heckel, et al., Emerging approaches to dna data storage: Challenges and prospects, ACS nano 16 (2022) 17552–17571.

[6] B. Cao, X. Zhang, S. Cui, Q. Zhang, Adaptive coding for dna storage with high storage density and low coverage, NPJ systems biology and applications 8 (2022) 23.

[7] C.-N. Members, et al., Database resources of the national genomics data center, china national center for bioinformation in 2022, Nucleic Acids Research 50 (2022) D27.

[8] P. Bhattacharya, S. B. Patel, R. Gupta, S. Tanwar, J. J. P. C. Rodrigues, Satya: Trusted bi-lstm-based fake news classification scheme for smart community, IEEE Transactions on Computational Social Systems 9 (2022) 1758–1767. doi:10.1109/TCSS.2021.3131945.

[9] Z. Cao, X. Zhou, H. Hu, Z. Wang, Y. Wen, Toward a systematic survey for carbon neutral data centers, IEEE Communications Surveys & Tutorials 24 (2022) 895–936.

[10] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, W. Lintner, United states data center energy usage report (2016).

[11] S. Namasudra, A secure cryptosystem using dna cryptography and dna steganography for the cloud-based iot infrastructure, Computers and Electrical Engineering 104 (2022) 108426.

[12] L. M. Adleman, Computing with dna, Scientific american 279 (1998) 54–61.

[13] A. S. Perumal, Z. Wang, G. Ippoliti, F. C. van Delft, L. Kari, D. V. Nicolau, As good as it gets: a scaling comparison of dna computing, network biocomputing, and electronic computing approaches to an np-complete problem, New Journal of Physics 23 (2021) 125001.

[14] Z. Yang, N. Goldman, A. Friday, Maximum likelihood trees from dna sequences: a peculiar statistical estimation problem, Systematic Biology 44 (1995) 384–399.

[15] A. Kumar, M. Kumar, R. P. Mahapatra, P. Bhattacharya, T.-T.-H. Le, S. Verma, Kavita, K. Mohiuddin, Flamingo-optimization-based deep convolutional neural network for iot-based arrhythmia classification, Sensors 23 (2023). URL: https://www.mdpi.com/1424-8220/23/9/4353. doi:10.3390/s23094353.

[16] M. Kumar, A. Kumar, S. Verma, P. Bhattacharya, D. Ghimire, S.-h. Kim, A. S. M. S. Hosen, Healthcare internet of things (h-iot): Current trends, future prospects, applications, challenges, and security issues, Electronics 12 (2023). URL: https://www.mdpi.com/2079-9292/12/9/2050. doi:10.3390/electronics12092050.