INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 1 - V11I1-1285)
Available online at: https://www.ijariit.com

Encoding Digital Information in DNA: Advances, Techniques, and Applications

Ananya Chandra
ananyac2324@gmail.com
National Post Graduate College, Lucknow, Uttar Pradesh

Mahesh Tiwari
maheshyogi26@gmail.com
National Post Graduate College, Lucknow, Uttar Pradesh

ABSTRACT

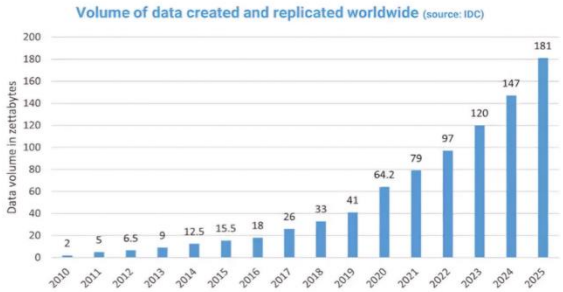
In 2020 approximately 64 zettabytes of Data were generated and it was predicted that by 2025 this number will be greater than the twice of it. This prediction is proving itself as everyday approximately 402.74 million Terabytes of data is created and as of 2024 the number has risen up to 147 Zettabytes already and it's assumed that this amount will be 181 Zettabytes by the end of 2025.
This data primarily includes IoT data that is the fastest growing segment of data which is then followed by social media.

The existing storage technologies cannot cater the needs of the Zettabyte Era, as they have considerable issues like limited durability, high power consumption and environmental impact they cause.

DNA is the nature's best alternative to these problems and can store such high amounts of data for a longer period of time without or very less decay. One gram of DNA can store up to 215 Petabytes of data. It's longevity of thousands of years and enormous information density without harming the environment by generating lesser e-waste makes it a promising archival storage medium.

KEYWORDS: *DNA Storage, Data Growth, Archival Storage, Synthetic Biology, Long-Term data Preservation, Digital Data Explosion, Next-Gen Data Storage*

INTRODUCTION



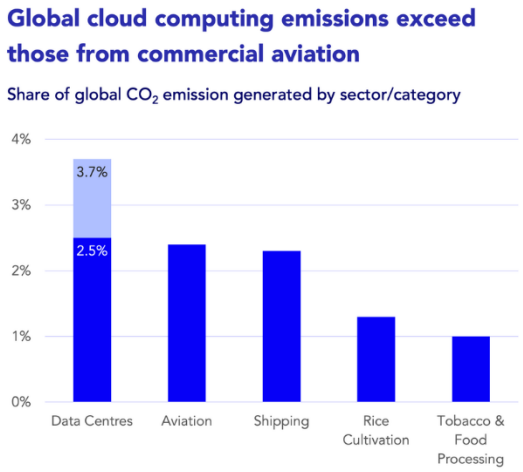
- **Data Growth**:
  - **Post-COVID Digital Surge**: Post Covid-19 digital content consumption and creation has an unprecedented increase. Short format videos like Instagram reels, TikTok and YouTube shorts produce Terabytes or even Zettabytes of data daily.
  - **Data Production Metrics**: The average usage of an Instagram or TikTok user is 21.5 hours and daily billions of uploads are done. Facebook alone generates 4 petabytes of data daily while Instagram deals with approximately 95 million photo and video uploads every day. WhatsApp users are sending over 100 billion messages daily and

Google processing of 20 petabytes is done each day. It may seem a large number but the average usage after covid has accelerated and is still increasing daily which has been testing our data Storage infrastructure.

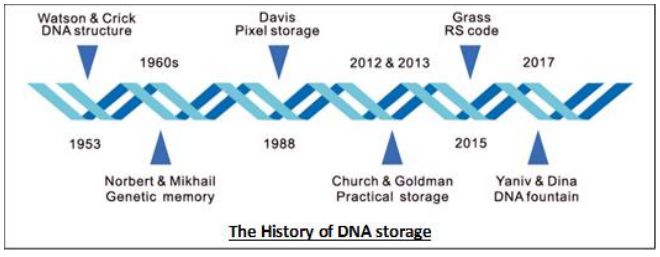
- **Traditional Storage Comparison**: Traditional storage devices like SSDs and HDDs have a really low lifespan and have many storage limitations which we can overcome easily by using DNA as it's high density and long-term stability capabilities can manage such high data storage requirements easily.

- **Energy costs of traditional storage:**
  - Every year data centres produce excessive CO_2 emissions and have a high energy consumption. This is because traditional storage methods have a high energy requirement due to which companies like OpenAI, Google, Amazon and Microsoft have started investing in nuclear plants for energy to be able to handle the Ai processing systems and large datasets. All of this can be minimised by using DNA to store the data that would use minimal energy and help stabilize the environment.

- **Need for DNA:**
  - The high density and archival capabilities of DNA is ideal for storing the rapidly growing data volume. 1 gram of DNA can store 215 Petabytes of data, that is more than the storage in data centres and in a very less physical space comparatively while also reducing the energy usage drastically.
  - Microsoft and University of Washington conducted an experiment and successfully stored and retrieved 200 MB data in DNA which establishes its practical applicability and potential for an innovating data storage solution.



1. History and Development

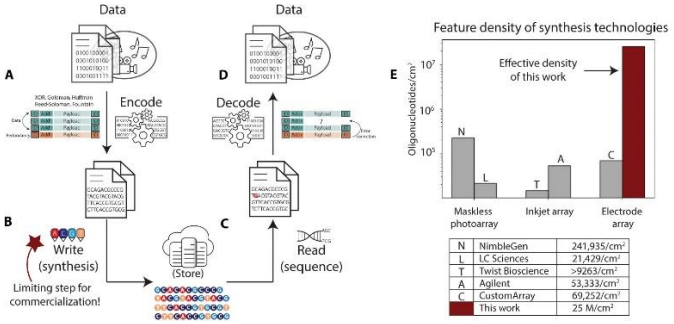


- **Initial Concepts**:
  - Using DNA as a storage concept was introduced in 1960s when scientists drew comparison between biological information storage and digital data storage data in parallel.
  - The natural data encoding and retrieving capabilities of DNA are inspired from genetic transcription and translation. These capabilities helped scientists visualize the potential of DNA as a digital storage medium.

- **Pioneering Research and Experiments**:
  - In 2012, Harvard scientists encoded a 53,000-word book in DNA and retrieved it successfully. It was stored at a density of 5.5 petabits/mm³. This established the base of DNA data storage for future.

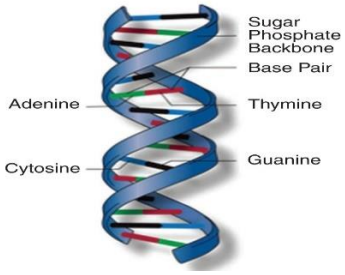
- **Microsoft Experiment**: Microsoft and Twist Bioscience collaborated and developed DNA storage systems that are scalable and cost efficient. This was done to meet the technological advancements and higher data demands, while also implementing error correction and efficient encoding methods.

- **Current Technologies and Major Contributors**:
  - Institutions like Twist Bioscience, ETH Zurich and Harvard are working towards affordable DNA storage solutions and faster sequencing processes. These institutions are focusing on making DNA synthesis and advance retrieval techniques more accessible and efficient.



3. DNA Data Storage Mechanism

- **DNA Structure**:
  - DNA is organized in a double helix structure consisting of the nucleotides Adenine(A), Thymine(T), Cytosine(C) and Guanine(G) that are used to encode the binary data.
  - **Binary Mapping**: Binary data is converted into nucleotide sequences (e.g., 00 = A, 01 = C, 10 = G, and 11 = T). The equivalent of the binary code "1101 0101" has "T G C C" as it's data encoding representation.



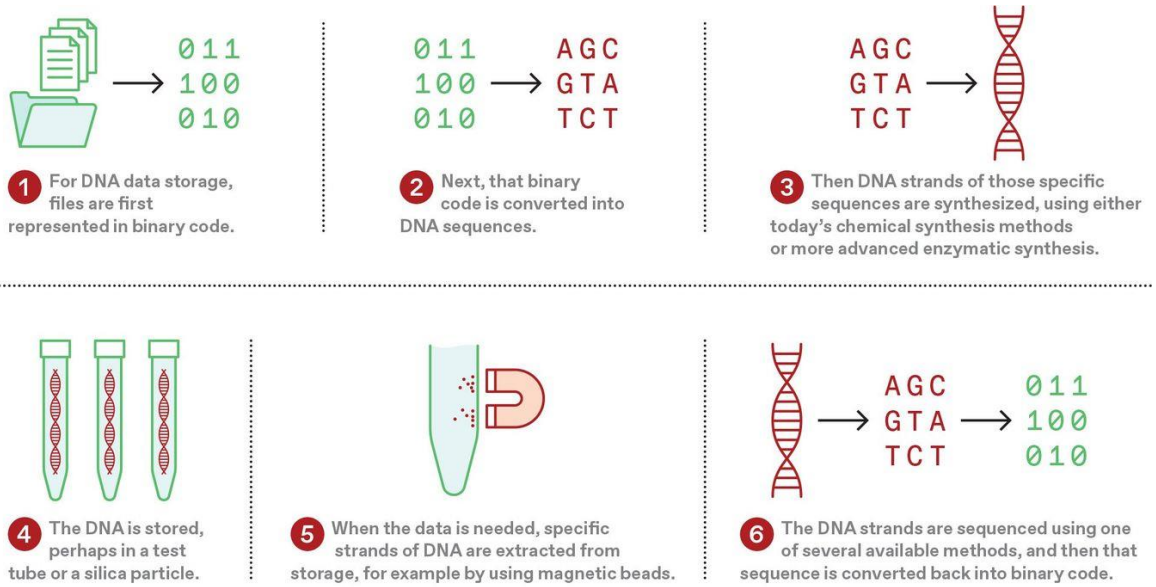
- **Encoding Process for Social Media Data**:
  - High density encoding techniques can help us to store high resolution data like short format videos and images into DNA strands efficiently.
  - **Encoding Example**: YouTube generates approximately 440,000 TB of data everyday while platforms like X and TikTok daily create 12TB and 7.35TB respectively. Converting trillions of bytes of data from these applications into DNA and compressing them would be an ideal solution where trillions of video files can be stored in a compact and durable storage format.

- **Synthesis**:
  - DNA synthesis machines daily produce DNA strands which are used for high precision and error free data storage.
  - **Example**: The automated DNA synthesizer of Twist Bioscience produces millions of oligonucleotides which is ideal for storing data from huge datasets of social media and IoT.

- **Storage and Encapsulation**:
  - DNA strands are stored in dry and dark conditions which are necessary to provide stability and longevity to them.
  - **Enhanced Longevity Example**: Storing DNA in glass or silicon capsules can keep it safe for many centuries which provides a cost efficient and durable data archival solution.

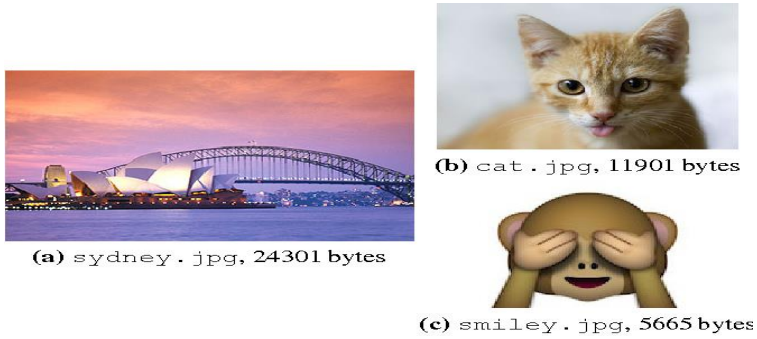
- **Data Retrieval and Decoding**:

- To retrieve the data, nucleotide sequences are decoded back to their binary form using DNA sequencing. **Error correction techniques** like Reed-Solomon fix the retrieval error and ensure the data accuracy.



4. Goldman Encoding

➢ **Goldman's Encoding System Experiment**



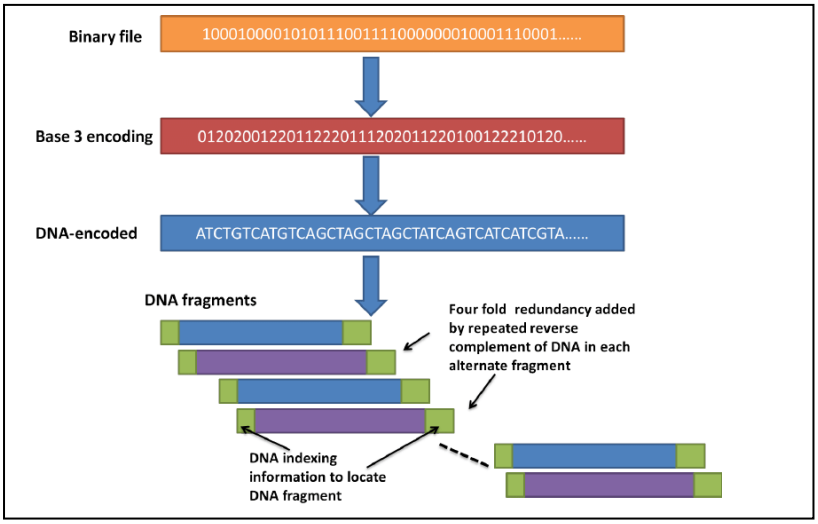
- Image Files Stored Using Goldman's XOR Encoding:
  Scientists used Goldman encoding system, specifically XOR encoding to store 3 image files into DNA.
  - Their results:
    1. **Sydney.jpg** - 24,301 bytes (Stored and retrieved successfully)
    2. **Cat.jpg** - 11,901 bytes (Stored and retrieved successfully)
    3. **Smiley.jpg** - 5,665 bytes (1-byte error due to bugged sequencing)
  - In this experiment, 2 files were successfully encoded and retrieved, while the 3rd file had an error due to a bug in the sequencing process.
- Goldman's Individual Achievements:
  Goldman successfully stored various types of data, including:
  - 154 sonnets of Shakespeare (ASCII format)
  - One PDF file
  - One colour photograph
  - A MP3 file

➢ **Advantages and Disadvantages of Goldman Encoding**

- **Advantages:**
  - DNA is a highly durable and stable storage medium.
  - DNA can easily be synthesized and it can store data for thousands of years.
  - This technology is very reliable since only 1 gram of DNA can store 2.2 Petabytes of data.
- **Disadvantages:**
  - The process of copying data and its retrieval is slow and needs improvement.
  - Maintenance of DNA replicators is costly and challenging.
  - Technology is not fully developed yet, and identifying and correcting errors in sequencing process is difficult.

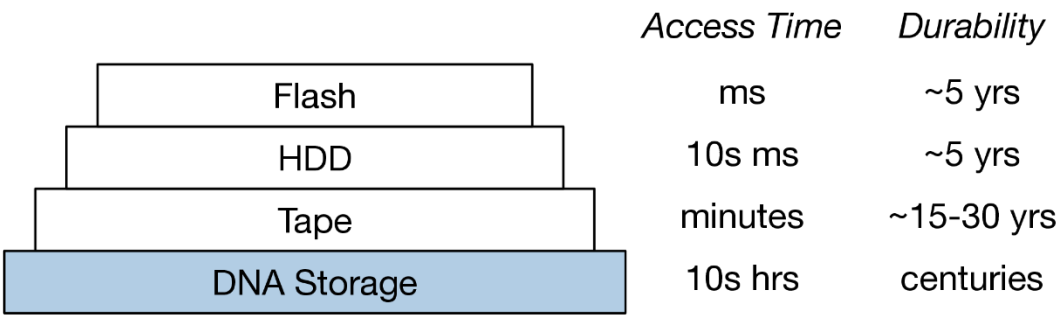
Goldman encoding experiments showcase immense potential of DNA data storage, but also highlights its current limitations. Improvements in retrieval speeds, error correction, and cost reduction can make this technology more accessible and efficient.

Stepwise encoding of data into DNA using Goldmans approach:



5. Advantages of DNA Data Storage

- **High Density Storage for Increasing Data**:
  - Due to its high storage density one gram of DNA can store up to 215 Petabytes of data. Storing data in DNA could replace data centres managing Google's 20 Petabytes and YouTube's ~440,000 Terabytes daily in while potentially saving large physical space and energy. This space saving and high-density feature makes DNA uniquely capable for long term storage needs.
  - **Real-World Example**: DNA can potentially support the archiving of exponentially growing data daily from 500 million tweets on X, 2 billion calls on WhatsApp, and 95 million Instagram posts. By storing them in DNA in a highly compact, energy efficient and more stable form can help secure digital record for future generations.
- **Longevity**:
  - DNA is naturally durable and can maintain stability and readability for centuries if preserved in suitable conditions.
  - **Historical Insight**: Scientists have successfully sequenced thousands of years old DNA of Wooly Mammoths and Mummies which is a proof of its long-term preservation potential.
- **Environmental Impact**:
  - The adoption of DNA storage can drastically reduce generation of e-waste and emissions of CO2.
  - **Environmental Case Study**: If daily social media data is archived into DNA, it would result in reducing gigatons of CO_2 emissions and tons of e-waste annually.



6. Challenges and Limitations

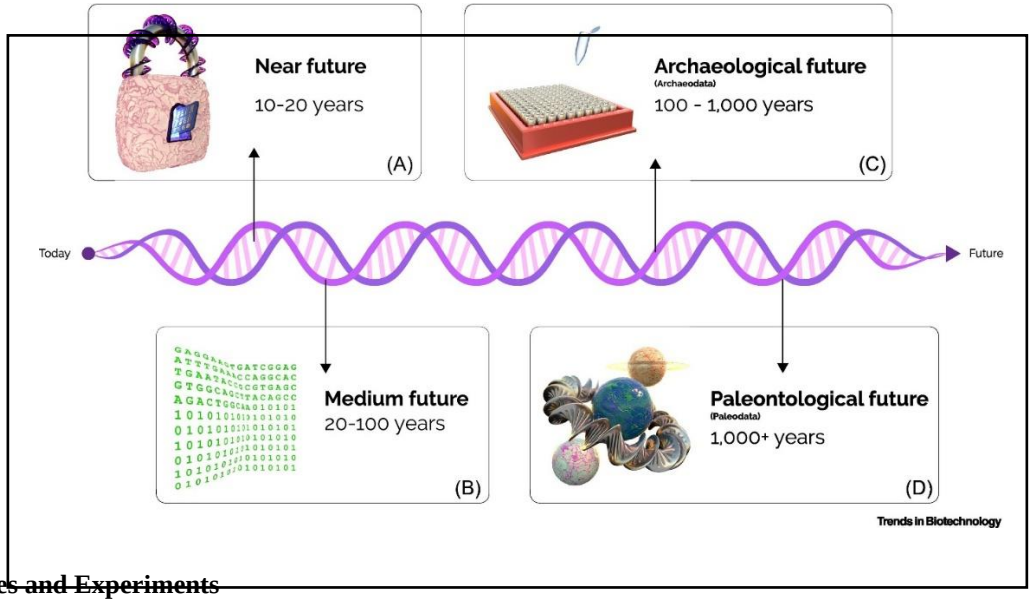
- **High Cost**:
  - In the current scenario of 2024 DNA synthesis and sequencing are quite expensive, which is not feasible for large scale adoption. Storing one megabyte of data in DNA costs approximately $3000. Storing data like Facebook's 4 Petabytes currently faces high costs in synthesis and sequencing, which impacts DNA's commercial viability. This reinforces the need for feasible solutions for DNA storage in order for it to be adopted in sectors like social media where data production is high.
  - **Cost Comparison**: Traditional storage methods cost a few cents per GB while DNA storage costs are in dollars which is economically challenging for large scale data archival.
- **Retrieval Speed Limitations**:
  - As of now, the retrieval process of data from DNA is slow, which makes it difficult to meet the needs of social media platforms for high-speed access of data.
  - **Efficiency Comparison**: 200MB of data retrieval from traditional storage takes a few seconds while it can take hours to retrieve the same amount of data from DNA.
- **Error Management and Data Integrity**:
  - During DNA synthesis and sequencing processes, errors like insertions, deletions and substitutions impact data accuracy.
  - **Solution Example**: Advanced error correction codes like Hamming and Reed-Solomon codes are used to detect and rectify these errors.

7. Applications and Future Potential

- **Archiving Social Media and Streaming Data**:
  - For storing historical archives and daily uploads from social media that generate 7.35 terabytes of data daily, DNA storage can be used to store them for long periods of time. DNA can be used for permanently archiving old videos and posts from TikTok and Instagram.
  - **Use Case Example**: Converting daily generated data from TikTok and Instagram into DNA storage can reduce load on physical data centres and make long-term archiving of data more feasible.
- **Cloud Storage and DNA Hybrid Models**:
  - We can create a hybrid model by integrating DNA storage with cloud storage. This hybrid system can store frequently accessed data in traditional cloud storage while the archival data can be shifted to DNA storage. Such hybrid models could be a cost-effective solution for managing immense data flow from companies like Google and Facebook.
  - **Benefit Example**: Hybrid models can help enterprises minimize the risk of high cost archival storage and risks of data loss. For example, research organisations that generate massive datasets can move their infrequently accessed data into DNA, reducing cloud storage costs and minimizing the environmental footprint.
- **Medical and Genomic Data Storage**:

- DNA storage can have a huge scope in medical and genomic research, where long-term preservation of genetic records and patient data is necessary. In sectors like healthcare, the growth of data science market has reached up to $63.97 billion, utilising minimal space DNA could store patient records and genetic data for decades.
- **Example**: Genomics companies like 23andMe and AncestryDNA can store their vast datasets in DNA in order to preserve that data for generations and making it accessible for further researches.

- **Digital Legacy Preservation**:
  - DNA archival storage is ideal for preserving social media data, digital photographs and family records. This data is ever growing and the longevity advantage of DNA storage helps in securing this data and keeping it accessible for future generations.
  - **Application Example**: Companies like Facebook and Google can permanently archive their users' legacies by storing them in DNA that will protect family history and digital memorabilia.



8. Case Studies and Experiments

➢ **Microsoft's Experiment and Real-World Feasibility**

- **DNA Data Storage Research Initiatives**:
  - Microsoft successfully entered the field of DNA data storage by conducting it's research and experiments.
  - They purchased 10 million DNA strands (Oligonucleotides) from Twist Bioscience which were used to encode digital data into the DNA.
  - They stored 200 MB data, which included literary and other articles, and successfully retrieved it with 100% accuracy. This was a major milestone in the tech field.
  - Based on the research it was estimated that 1 cubic millimeter of DNA can store 1 Exabyte (1 billion gigabytes) of data, which is more compact and efficient compared to traditional methods.
- **Addressing Massive Data Storage Needs**:
  - Microsoft saw DNA data storage as an efficient solution to the growing amount of data load from social media platforms like Facebook that generate Petabytes of data daily.
- **Replacing Data centres with DNA Technology**:
  - Microsoft plans that the data centres which spread across thousands of acres of land that are based on drives and servers to be replaced with DNA-based storage systems.
  - In the first step, an Azure data centre is planned to be converted into a DNA-based data centre.
  - They also plan on integrating DNA data storage in Azure cloud services, which will be a hybrid data storage model.

➢ **Ongoing Research on Data Retrieval**

- **Improvements in Retrieval Speeds**:
  - Earlier, the data retrieval speed was 400 bytes/second which was very slow. Due to recent researches and advancements this speed has reached upto 100 MB/second.
  - This is important for platforms that demand access to real-time data, like Google that processes 20 PB data everyday.

- **Collaborations and Future Plans**:
  - Microsoft is collaborating with University of Washington to develop new algorithm and efficient encoding-decoding techniques.
  - Ongoing experiments show that DNA data storage is feasible along with being a revolutionary solution for the future.

➢ **Columbia University's Living Bacteria Experiment**

- **Using Living Bacteria for DNA Storage**:
  - Researchers of Columbia University conducted a unique experiment in which they engineered DNA of living bacteria (E. Coli) to store data.
  - Electrical signals and CRISPR biology were used to encode binary data in the bacteria's DNA. The presence of electrical signal represented '1' and the absence represented '0'.
  - Researchers used this method to encode 72 bits of data which formed the message "Hello world!".
- **Durability and Data Recovery**:
  - Researchers mixed Encoded E. Coli into the soil to successfully recover their message. The DNA stored in the bacteria was protected from degradation.
  - Mutations occurred during the replication of bacteria, due to which the data was stable for only upto 60 generations.
- **Future Potential**:
  - This method explores new possibilities, in which data can be hidden inside natural microbial communities. If the mutation and stability issues are solved, then this can be a secure and innovative medium of storing data.
  - Dr. Harris Wang and Dr. Sung Sun's research demonstrates that engineered bacteria can be a cost effective and sustainable alternative to current data storage technologies.

An image that contrasts the idea of traditional data centers with the concept of DNA data storage in a serene, green environment. [DALL-E]

9. CONCLUSION

➢ **Summary of DNA Storage Benefits:**

- **Unmatched Storage Capacity:**
  - The biggest advantage of DNA storage is it's massive storage capacity.
  - The estimate of 0.495 ZB of data to be generated by 2025 could possibly only be efficiently handled by an advanced data storage medium like DNA.
- **Compact and Durable Solution:**
  - DNA storage provides a compact and durable medium.
  - It can protect the data from degradation for thousands of years.
  - The density and lifespan of DNA, both are superior than that of traditional storage methods comparatively.
- **Energy and Space Efficiency:**
  - Data centers of today consume a large area of land and energy.

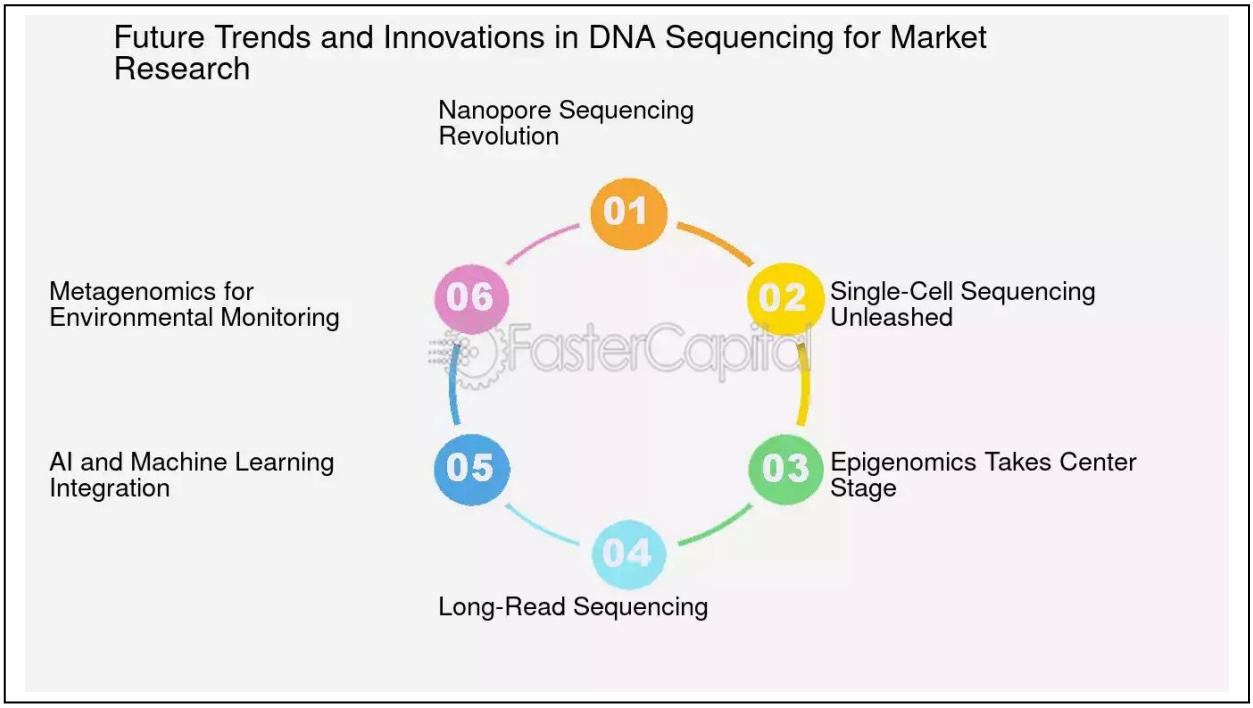
- DNA-based storage systems use significantly less resources.
- This is not only an eco-friendly solution but also very impact ful in space and energy savings.

➢ **Future Scope for Research and Development:**

- **Advanced Research and Innovation:**
  - Due to the exponential growth of digital content, development of DNA storage technology is very crucial.
  - It can be a viable solution for social media platforms, search engines and daily data generating services like Facebook and Google.
- **Improving Speed and Cost:**
  - The DNA encoding and retrieval processes are relatively slow and expensive.
  - The future research should focus on optimizing these two aspects so that this technology becomes more accessible and practical.
- **Error Correction Mechanisms:**
  - Sequencing Robust error correction techniques shall be developed to resolving sequencing bugs and retrieval errors.
  - Without accurate and reliable data retrieval, this technology cannot reach full scale deployment.
- **Integration with Existing Systems:**
  - Combining DNA storage and cloud storage, hybrid systems can be implemented in future.
  - These systems will provide both flexibility and scalability.
- **Security Enhancements:**
  - Advanced encryption and security measures need to be adopted to prevent unauthorised access to DNA data storage and protect it from hackers.

➢ **A Vision for the Future:**

- DNA storage is a revolutionary technology that can solve the storage challenges of today as well as of the upcoming digital era.
- For its full-scale adoption continuous R&D and practical implementation is needed.
- If we efficiently address the challenges like speed, cost and accuracy, then this technology can set a new standard in the global storage systems.



REFERENCES

[1] *Raza, M. H., Desai, S., Aravamudhan, S., & Zadegan, R.* (2023). An outlook on the current challenges and opportunities in DNA data storage. *Journal of Biotechnology, 365*, 1-19. Elsevier.

[2] *Meiser, L. C., Antkowiak, P. L., Koch, J., Chen, W. D., Kohll, A. X., Stark, W. J., Heckel, R., & Grass, R. N.* (2019). *Reading and writing digital data in DNA*. *Nature Protocols, 14*, 3104–3129.

[3] *Anavy, L., Vaknin, I., Atar, O., Amit, R., & Yakhini, Z.* (2019). Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nature Biotechnology, 37*(10), 1229–1236.

[4] *Li, B., Song, N. Y., Ou, L., & Du, D. H. C.* (2020). Can we store the whole world's data in DNA storage? *Proceedings of the 20th International Conference on DNA Computing and Molecular Programming (DNA 2020)*, 1-15.

[5] *Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., & Strauss, K.* (2016). *A DNA-Based Archival Storage System*. In *Proceedings of the 21st International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '16)* (pp. 637-649). ACM.

[6] *Sharma, D. K., Kumar, S., & Kumar, A.* (2018). DNA-based storage: Introduction, characteristics, applications, and challenges. *International Journal of Machine Learning and Networked Collaborative Engineering, 2*(4), 163–169.

[7] *Ramanamurthy, S. V., Kiranmayi, P. L., Sarayu, S. D., & Iftekharuddin, M.* (2018). *DNA Data Storage*. *Journal for Research, 3*(11), 32-37. ISSN: 2395-7549.
Panimalar, A. S., Henry, A., Balu, T. A., & Nishanth, R. (2018). *DNA Digital Data Storage*. *International Research Journal of Engineering and Technology (IRJET), 5*(2), 636-640. ISSN: 2395-0056.

[8] *Twist Bioscience.* (2017). *DNA-Based Digital Storage: White Paper*.

[9] *Toppo, A., & Jacob, E.* (2017). DNA: The future supernatural storage. *National Conference on Contemporary Research and Innovations in Computer Science (NCCRICS)*. *International Journal of Engineering and Technology*, ISSN: 2395-1303.

[10] *Atos.* (2021). Could DNA be the next big thing in data storage? *Atos White Paper on Thought Leadership*.

[11] *Brush, K.* (2016). Review of "Could the molecule known for storing genetic information also store the world's data?". *Nature, 537*.

[12] *Lakshmi, N. R. R.* (n.d.). DNA as an information storage device. *Vasireddy Venkatadri Institute of Technology, Dept. of Electronics and Communication*.

[13] *Robbins, R. J.* (1995). DNA as a mass-storage device. *Johns Hopkins University*.