Table of content

[1. Abstract 1](#_Toc20814)

[2. Introduction 2](#_Toc28187)

[3. Literature Review 2](#_Toc26651)

[4. Problem observed 3](#_Toc19800)

[5. Objectives 3](#_Toc5786)

[6. Comparison with Existing Projects: 3](#_Toc14296)

[7. Unique features 4](#_Toc25920)

[8. Software used 4](#_Toc20750)

[9. Methodology 4](#_Toc31551)

[10. Expected results and Potential impact 5](#_Toc5833)

[11. Timeline: 5](#_Toc3336)

[12. Budget and Resources 5](#_Toc18982)

[13. Potential and Ethical Considerations 6](#_Toc1000)

[14. References 6](#_Toc20802)

[15. Contact Information: 6](#_Toc20355)

Title: DNA vault

1. Abstract

This project will build a secure, fast and adaptive DNA based data storage system. It will store digital data (text and images) in DNA by converting binary data into DNA sequences using compression, encryption and adaptive encoding. It will use machine learning to improve error correction during data retrieval and will have a web based interface for real-time data storage and retrieval. This will solve modern data storage problems such as high volume storage needs, environmental sustainability and secure long term archival. By using DNA as a storage medium this will be a big step towards next gen data storage.

# Introduction

With the data production growing exponentially globally there is an immediate need for scalable, energy efficient and durable storage systems. Traditional data storage methods like hard drives and cloud based systems consume a lot of energy and have limited lifespan. DNA as a medium offers a promising alternative, to store huge amount of data in very small volume with long lifespan. DNA’s durability and density is the foundation for long term data storage solutions. This project will enhance DNA based data storage by adaptive encoding, real-time processing and improved security features to make it more accessible and practical.

# Literature Review

DNA as a storage medium has been explored by many.

* Church et al. (2012) showed DNA can store digital information.
* Goldman et al. (2013) encoded text and images into DNA. But these systems had errors and real-time retrieval issues.
* Erlich & Zielinski (2017) introduced DNA Fountain, a system that maximized data density using fountain codes and achieved big gains in DNA data storage.
* Grass et al. (2015) encoded image data into DNA, focused on long term archival.

Despite these advances, current systems still have cost, real-time data access, security and adaptive encoding for different file types issues. This project will address these limitations by combining data compression, encryption and machine learning based error correction into a real-time DNA storage system.

# Problem observed

Data storage is growing faster than traditional storage methods can keep up with. They are energy hungry, have a limited lifespan and take up physical space. DNA storage is eco-friendly and scalable but existing solutions are limited by cost, scalability, error rates and no real-time data access. This project solves these problems by building a secure, adaptive and efficient DNA storage system with real-time and error correction.

# Objectives

**General objective**: Creating adaptive, secure and efficient DNA data storage with real-time processing and error correction.

**Specific Goals:**

* Compression and encryption for DNA data storage.
* Adaptive encoding for text and image data.
* Web interface for real-time data storage and retrieval.
* Machine learning for error correction during data retrieval.
* Test the system for storage efficiency, retrieval accuracy and scalability.

# Comparison with Existing Projects:

* DNA Fountain (Erlich & Zielinski, 2017): Maximizes data density using fountain codes and corrects errors, but no security and real-time access.
* Church et al. (2012): Early text in DNA, but small data size and error handling.
* Goldman et al. (2013): Successfully stored multimedia data in DNA, but scalability and cost issues.
* Grass et al. (2015): Encoded images in DNA but didn’t explore real-time data retrieval or adaptive encoding for different data types.

# Unique features

### **Adaptive Encoding**: Text and image encoding separately for maximum efficiency.

### **Compression:** Less data stored for more efficiency.

### **Encryption**: AES encryption for data stored in DNA.

### **Real-Time:** Web interface to upload, encodetrieve data in real-time.

### **Machine Learning for Error Correction:** Machine learning to predict and correct errors during data retrieval so you don’t get incomplete or corrupted data.

# Software used

### **Language**: Python because of its web development, encryption, data processing and machine learning ecosystem.

### **Web Framework:** Flask/Django to create the web interface to upload and retrieve files in real-time.

### **Compression**: Zlib/Gzip libraries to reduce the data size before encoding into DNA sequences.

### Encryption: AES encryption using PyCryptodome to encrypt data stored in DNA.

### Machine Learning: Scikit-learn to build models to predict and correct errors during data retrieval.

### Storage System: Binary data will be converted to DNA sequences using standard base-pair encoding (A, T, C, G).

# Methodology

**Data Preprocessing:**

* Convert uploaded text and image files into binary.
* Compress and encrypt binary.

**Adaptive Encoding**:

* Use different encoding for text and images based on file type.
* Convert binary to DNA sequences using a nucleotide map.

**Web Interface**:

* Build a Flask or Django web based interface for file upload, real-time encoding and retrieval.

**Machine Learning**:

* Train machine learning models to detect and correct errors in DNA sequences during retrieval.

**Testing and Evaluation**:

* Test on multiple datasets, measure efficiency, security and accuracy.

# Expected results and Potential impact

**Expected results:** A working DNA storage system to compress, encrypt and encode text and image files. Accurate and error free data retrieval. A user friendly web interface for real time file handling.

Potential impact: No need for physical storage devices, environmentally friendly data storage solutions.Secure and efficient long term archiving through DNA storage, potentially changing the game for industries like healthcare, finance and historical archives.

1. Timeline:

# Budget and Resources

|  |  |  |
| --- | --- | --- |
| Task | Duration | Completion date |
| Project Proposal and Setup | 3 days | Day 3 |
| Core Feature Development | 6 days | Day 9 |
| Web Interface and Real-Time Features | 5 days | Day 14 |
| Machine Learning and Error Correction | 5 days | Day 19 |
| Final Testing and Documentation | 2 days | Day 21 |

|  |  |
| --- | --- |
| Resource | Cost |
| Python Libraries (Open-source) |  |
| Cloud Hosting for Web Interface |  |
| Data Storage Costs |  |
| Machine Learning Model Training |  |
| Miscellaneous |  |
| Total |  |

# Potential and Ethical Considerations

**Data Privacy**: AES encryption will keep your data safe during storage and retrieval. Users must be aware of data privacy measures.

**Ethical use of DNA**: Since we’re using simulated DNA sequences digitally, any ethical concerns around data in DNA must be addressed.

# References

* Erlich, Y., & Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. Science, 355(6328), 950-954.
* Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in DNA. Science, 337(6102), 1628-1628.
* Goldman, N., et al. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 494(7435), 77-80.

# Contact Information:

* **Name**: Lidiya Bokona
* **Email**: mystery.written@gmail.com
* **Phone**: 0951248018
* **GitHub**: