Blockchain Implementation for Secure Data Management with Genetic Storing Integration

Abdelrhman Elnamaki[1]
[1] Digital Career Institute, e23p07 , abdoelnamaki@gmail.com

August 12, 2024

Abstract

Blockchain technology provides a robust framework for secure data management, ensuring integrity, transparency, and immutability. This paper explores the implementation of a blockchain system in Python tailored for integrating an alternative chain to securely store genetic sequences extracted from plant disease reports, scientific papers, and datasets in each block. The system employs SHA-256 hashing for cryptographic security and DNA sequence storage to ensure long-term data preservation.

This project is a pivotal component of fulfilling graduation requirements from Digital Career Institute (DCI), demonstrating the practical application of blockchain technology and AI agents in the management and analysis of complex datasets. AI agents operating over the blockchain ecosystem enhance data management capabilities through automated analysis, validation, experimental contracts, and decision-making processes.

The applications of this integrated system span across healthcare, scientific research, and other data-intensive domains, showcasing blockchain's potential to improve data integrity, foster collaboration, and streamline information management. This study highlights how blockchain and AI technologies address contemporary challenges in data security and management, emphasizing their transformative impact on data-driven industries.

1 Introduction

Advancements in data-intensive fields such as Plants diseases and scientific research have underscored the critical need for secure and authentic data management systems. Traditional approaches often face challenges related to data integrity, transparency, and centralized control. Blockchain technology offers a promising solution by providing a decentralized and immutable ledger that ensures data integrity and security through cryptographic principles.

This project focuses on leveraging blockchain for secure data management, particularly in handling disease reports, scientific papers, datasets, and genetic

sequences . The integration of genetic data adds a unique dimension, requiring specialized methods for data storage and authentication. The system utilizes SHA-256 hashing for robust cryptographic security and incorporates DNA sequence storage to facilitate future advancements in genetic research and personalized medicine.

The blockchain implementation presented in this paper comprises two core components: the `Blockchain` and `GenChain`. The `Blockchain` facilitates continuous block generation, secure data validation using RSA cryptography, and management of diverse data types. Meanwhile, the `GenChain` specializes in storing and authenticating genetic sequences , ensuring data integrity through specialized hashing and DNA sequence conversion.

In this paper, we discuss the mathematical concepts and functions underlying RSA cryptography and SHA-256 hashing, providing a comprehensive understanding of their role in securing our blockchain operations and data storage. We also introduce the `Block` and `GenChainBlock` detailing their attributes and functionalities within the blockchain framework.

By demonstrating the implementation details and applications of blockchain technology in data management, this paper aims to showcase its potential in enhancing security, integrity, and efficiency across plant science field , and beyond.

2 RSA Encryption and Signing

2.0.1 Key Pair Generation

To generate an RSA key pair, the following steps are performed:

- **Generate large primes p and q**: Two distinct large prime numbers are generated.
- **Compute the modulus**:

$$n = p \times q \tag{1}$$

- **Compute the totient**:

$$\phi(n) = (p-1) \times (q-1) \tag{2}$$

- **Select the public exponent**:

$$e = 65537 \tag{3}$$

- **Compute the private exponent d**:

$$d \times e \equiv 1 \pmod{\phi(n)} \tag{4}$$

The public key consists of (e, n) and the private key consists of (d, n).

2.0.2 Private Key Serialization

The private key (d, n) is serialized to PEM format using PKCS8 without encryption.

2.0.3 Public Key Serialization

The public key (e, n) is serialized to PEM format using the SubjectPublicKeyInfo format. This standard format ensures compatibility across different systems and applications.

2.0.4 Data Signing

To sign data using the RSA algorithm:

$$\begin{aligned} &\text{hash}(data) \rightarrow \text{Apply a hash function (e.g., SHA-256) to the data} \\ &\text{signature} = \text{hash}(data)^d \mod n \end{aligned}$$

PSS (Probabilistic Signature Scheme) padding is applied to ensure security against certain types of attacks:

$$\text{PSS padding} = \text{PSS}(\text{hash}(data)) \tag{5}$$

2.0.5 Signature Verification

To verify a signature:

$$\begin{aligned} \text{verification} &= \text{signature}^e \mod n \\ &= \text{hash}(data) \quad \text{if the signature is valid} \end{aligned}$$

PSS padding is used during verification to ensure the integrity of the hash function and the signed data:

$$\text{PSS padding} = \text{PSS}(\text{hash}(data)) \tag{6}$$

3 SHA-256 Hashing Algorithm

3.1 Message Padding and Parsing

SHA-256 processes the padded message M in 512-bit blocks:

$$M_{\text{padded}} = \text{pad}(M) \tag{7}$$

The padded message M_{padded} is divided into 512-bit blocks for processing.

3.2 Message Schedule

Each 512-bit block is divided into 16 32-bit words, which are then extended to 64 words:

$$W_t^i = \begin{cases} M_t^i & \text{for } 0 \leq i < 16 \\ \sigma_1(W_{t-2}^i) + W_{t-7}^i + \sigma_0(W_{t-15}^i) + W_{t-16}^i & \text{for } 16 \leq i < 64 \end{cases} \tag{8}$$

3.3 Compression Function

For each 512-bit block M_t:

$$\text{Initialize } H_0^t, H_1^t, \ldots, H_7^t \tag{9}$$
$$\text{For } t = 0 \text{ to } 63: \tag{10}$$
$$a = H_0^t, b = H_1^t, c = H_2^t, d = H_3^t, e = H_4^t, f = H_5^t, g = H_6^t, h = H_7^t \tag{11}$$
$$\text{Compute } T_1 = h + \Sigma_1(e) + \text{Ch}(e, f, g) + K_t + W_t \tag{12}$$
$$\text{Compute } T_2 = \Sigma_0(a) + \text{Maj}(a, b, c) \tag{13}$$
$$h = g \tag{14}$$
$$g = f \tag{15}$$
$$f = e \tag{16}$$
$$e = d + T_1 \tag{17}$$
$$d = c \tag{18}$$
$$c = b \tag{19}$$
$$b = a \tag{20}$$
$$a = T_1 + T_2 \tag{21}$$

Where:

$$\Sigma_0(x) = \text{ROTR}^2(x) \oplus \text{ROTR}^{13}(x) \oplus \text{ROTR}^{22}(x) \tag{22}$$
$$\Sigma_1(x) = \text{ROTR}^6(x) \oplus \text{ROTR}^{11}(x) \oplus \text{ROTR}^{25}(x) \tag{23}$$
$$\text{Ch}(x, y, z) = (x \wedge y) \oplus (\neg x \wedge z) \tag{24}$$
$$\text{Maj}(x, y, z) = (x \wedge y) \oplus (x \wedge z) \oplus (y \wedge z) \tag{25}$$

3.4 Final Hash Value

Concatenate the outputs of each compression function to obtain the final hash value:

$$\text{Final Hash Value} = H_0^n \parallel H_1^n \parallel H_2^n \parallel H_3^n \parallel H_4^n \parallel H_5^n \parallel H_6^n \parallel H_7^n \tag{26}$$

4 Blockchain Components

4.1 Blockchain

The `Blockchain` serves as the core framework for securely managing data operations within the system.such as plant diseases, scientific research, and beneficial datasets. The chain is meticulously designed to user contributions, enabling the reporting of plant diseases, submission of plant-related research, and sharing datasets crucial for biologists and plant scientists . This ensures a robust platform for collaborative knowledge exchange and comprehensive plant science advancements.

4.1.1 Functionality

- **Chain Operations**: The `Blockchain` facilitates seamless block management operations, including block creation, validation, and chain traversal. Each block in the chain is linked sequentially, with each block containing a cryptographic hash of the previous block, ensuring the integrity of the entire chain.

- **Data Storage**: It provides robust mechanisms for storing diverse data types such as disease reports, scientific papers, datasets, and translated genetic information. Each block within the chain can encapsulate multiple data entries, allowing for efficient and organized data management.

- **Cryptographic Validation**: Utilizing cryptographic methods such as SHA-256 hashing and RSA encryption, the `Blockchain` class ensures that each block's contents are securely hashed and validated. This cryptographic validation guarantees the authenticity and immutability of stored data, preventing tampering and unauthorized modifications.

- **Continuous Block Generation**: Through continuous block generation, the `Blockchain` class supports the dynamic addition of new data entries to the chain. This feature enables real-time data updates and enhances the chain's utility in applications requiring up-to-date information, such as disease surveillance and research collaborations.

- **B2DNA**: The system includes a unique feature serves as alternative chain called the GenChain, which tokenizes binary data into DNA sequences. This process, we call it B2DNA, facilitates the storage of potentially billions of data entries in a highly compact and efficient manner on the GenChain. By converting binary data into DNA sequences, the GenChain enhances data storage capabilities, ensuring that vast amounts of information can be preserved and accessed securely.

5 Components

- **Block (`Block`)**: Each block within the `Blockchain` is represented by the `Block` class, encapsulating essential attributes such as index, previous hash, timestamp, data entries, block type, public keys, hash value, creator identifier, description, and additional metadata. These attributes collectively ensure comprehensive tracking and verification of data transactions within the chain.

index	:	Integer index of the block within the blockchain
previous_hash	:	Hash value of the preceding block in the chain
timestamp	:	Unix timestamp indicating the block's creation time
data	:	List of data entries (e.g., DiseaseReport, SciencePaper, Dataset)
block_type	:	Type of the block (e.g., 'genesis', 'data')
public_keys	:	List of public keys associated with data in the block
hash_value	:	Hash value computed for the entire block
creator	:	Identifier of the block's creator or miner
description	:	Optional description providing additional context
additional_info	:	Dictionary for storing supplementary metadata

- GenChain

  The `GenChain` class specializes in storing and authenticating genetic data within the blockchain, utilizing SHA-256 hashing and DNA sequence conversion.

- GenChainBlock

  The `GenChainBlock` class manages genetic data entries within the `GenChain`, storing DNA sequences and ensuring data integrity.

index	:	Index of the block within the GenChain sequence
previous_hash	:	Hash of the preceding block in the GenChain
timestamp	:	Timestamp indicating the block's creation time
dna_sequence	:	Sequence of DNA derived from binary data
hash	:	Hash value computed for the current block

- **GenChain Integration**: The `Blockchain` optionally integrates with `GenChain`, a component that tokenizes binary data into DNA sequences. This B2DNA (Binary to DNA) process allows for highly compact and efficient data storage. By converting binary data into DNA sequences, the `GenChain` can potentially store billions of data entries securely. The

`GenChain` continuously mines new blocks and adds them to its chain, enhancing the overall data storage capacity and ensuring the preservation of vast amounts of information.

6 Applications

The `Blockchain` class finds diverse applications across industries, including:

- **Plant Disease Management**: Facilitating secure and traceable storage of plant disease reports, research data, and diagnostic results. Blockchain ensures data integrity, transparency, and collaboration among agricultural researchers, farmers, and biologists. This enhances the tracking and management of plant disease outbreaks and supports the development of effective treatments and preventive measures.
- **Scientific Research**: Enhancing transparency and reproducibility in plant science studies by securely storing and sharing research findings, datasets, and peer-reviewed publications. Blockchain's immutable ledger enables verifiable citations and intellectual property protection, fostering trust and collaboration within the scientific community.
- **Supply Chain in Agriculture**: Providing transparency and traceability in agricultural supply chain management by recording transactional data, product provenance, and compliance certificates. Blockchain mitigates fraud, counterfeit products, and enhances supply chain efficiency and accountability, ensuring the quality and authenticity of agricultural products.
- **Genetic Data Storage**: Specialized applications include the storage and authentication of genetic data within the `GenChain` framework. The B2DNA (Binary to DNA) process tokenizes binary data into DNA sequences, allowing efficient and compact storage of potentially billions of data entries over the time . Blockchain ensures the integrity of genetic sequences, supports genetic storage collaborations.

7 Data Management in the Blockchain

Our blockchain framework securely manages diverse data types critical for agricultural and scientific research. This includes detailed disease reports, scientific papers, and datasets, each structured to ensure integrity and accessibility.

7.1 Disease Reports

Disease reports provide comprehensive insights into plant health, detailing disease identifiers, plant types, symptoms, diagnosis, treatments, incident dates,

geographic coordinates, submitters, notes, severity, and environmental conditions. These reports facilitate effective disease monitoring and response strategies.

7.2 Scientific Papers

Scientific papers stored in the blockchain include research findings on plant diseases, treatments, and agricultural studies. Each paper features identifiers, titles, authors, abstracts, publication dates, journals, URLs, keywords, citation counts, related topics, DOIs, research fields, methodologies, and additional attributes. This ensures transparent and traceable dissemination of scholarly information.

7.3 Datasets

Datasets contain essential data for agricultural research, including identifiers, names, descriptions, creation dates, URLs, creators, formats, sizes, licenses, tags, versions, data sources, quality metrics, and hash values. They support evidence-based research and analysis, ensuring data integrity and reproducibility.

Storing these data types on our blockchain enhances transparency, security, and collaboration in agricultural and scientific communities, fostering innovation and informed decision-making.

8 Conclusion

This paper has presented a comprehensive exploration of blockchain technology's application in secure data management, with a specific focus on integrating genetic data storage. By leveraging Python and employing SHA-256 hashing for cryptographic security, our framework ensures data integrity, transparency, and immutability across diverse applications such as healthcare, scientific research, and data-intensive domains.

The implementation of RSA cryptography for data validation underscores our commitment to secure data handling practices, enhancing trust and reliability within the blockchain structure. Furthermore, the introduction of the B2DNA approach facilitates efficient storage of genetic sequences, showcasing our innovative approach to long-term data preservation.

Applications highlighted in this paper, including plant disease management, scientific research transparency, and genetic data storage, illustrate the versatility and profound impact of blockchain technology in safeguarding sensitive information and fostering collaboration across industries.

In conclusion, this study underscores the transformative potential of blockchain in revolutionizing data management practices, ensuring robust security measures and facilitating seamless data exchange in critical fields of study and industry applications.

9 References

1. Digital Career Institute. (n.d.). Digital Career Institute. Retrieved from https://www.digitalcareerinstitute.org

2. Mukherjee, A., Dutta, A., Bhaumik, C. (2019). Blockchain technology for secure data management: A survey. IEEE Communications Surveys Tutorials, 21(4), 3230-3243. https://ieeexplore.ieee.org/document/9076545

3. National Institute of Standards and Technology. (2015, August 5). Secure Hash Standard (SHS). Retrieved from https://www.nist.gov/publications/secure-hash-standard

4. Menezes, A., van Oorschot, P. C., Vanstone, S. A. (1996). Handbook of applied cryptography. CRC press.

5. Amin, M. R., Zhang, G., Sun, Z., Khan, F. H. (2021). Blockchain for Agri-Food Supply Chain Management: A Use Case Based Review of the Literature. Sustainability, 12(1), 40. https://www.mdpi.com/2077-0472/12/1/40

6. Zhang, Y., Wang, Z., Xu, X., Li, M. (2020). Blockchain Applications in Improving Scientific Data Management and Transparency. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8409015/

7. DNA storage: research landscape and future prospects. https://academic.oup.com/nsr/article/7/6/1092/5711038

8. Yan, Y., Yang, J., Li, J., Li, S., Liu, Z. (2022). A Survey on DNA-Based Information Storage. arXiv preprint arXiv:2205.05488. https://arxiv.org/abs/2205.05488

9. Li, Z., Liu, J., Sun, Y. (2017). Blockchain for provenance tracking: A survey from the perspectives of applications, techniques, and future research directions. Proceedings of the IEEE, 106(5), 977-1007. https://ieeexplore.ieee.org/iel7/6287639/9668973/09936616.pdf

10. Yusuf, A., Khan, F. I., Imran, M., Xia, L. (2021). Blockchain technology in agriculture: A systematic review of applications and challenges. Sustainable Agriculture Research, 10(5), 545-565. https://www.sciencedirect.com/science/article/pii/B9780128214701000033

11. Blockchain analytics and artificial intelligence. (2018). IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/abstract/document/8645631

12. Artificial Intelligence and Blockchain Integration in Business. (2022). Information Systems Frontiers, 1-18. Retrieved from https://link.springer.com/article/10.1007/s10796-022-10279-0