Microservices Architecture for Blockchain-Based Proteomic Data Storage System Integrated with PDB and UniProt

Victor Ikenna Kanu, Josiah Isong, Simeon Okechukwu Ajakwe, Jae Min Lee, Dong-Seong Kim
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
(kanuxavier, isongjosiah, simeonajlove)@gmail.com, (ljmpaul, dskim)@kumoh.ac.kr

***Abstract*—Proteomic data storage faces challenges in security, consistency, and integration with unconventional bioinformatics systems. This paper proposes a microservices-based blockchain architecture to address these issues, focusing on amino acid sequences and 3D protein structures. Using HyperLedger and AES-256 encryption ensures secure access and data consistency. Integration with UniProt and the Protein Data Bank (PDB) enables seamless retrieval. The results of the performance tests show scalability, 197 TPS peak throughput, and stable latency under higher loads.**

***Index Terms*—blockchain, bioinformatics, data sharing, data storage, microservices, proteomics, security.**

I. Introduction

Proteomic data integration faces challenges due to the rapid growth and diversity of datasets. Centralized repositories like PDB and ProteomicsDB offer large-scale management but suffer redundancy, limited interoperability, and security issues, hindering collaboration [1]. Efforts like the ProteomeXchange consortium improve metadata consistency but struggle with synchronization and data heterogeneity, underscoring the need for unified frameworks for seamless and efficient management [2].

Various frameworks address bioinformatics challenges in data integration, interoperability, and security, focusing on accessibility and reliability. “Precision omics” improve data standardization but lack scalability, while FAIR-compliant frameworks enhance reproducibility yet face storage and privacy challenges [3]. Semantic Web technologies enable large-scale integration but struggle with heterogeneity, and service-oriented architectures face metadata synchronization issues. These limitations highlight the need for scalable, secure, and user-friendly solutions [4]. Blockchain technology addresses these limitations by providing decentralized, secure storage, enhancing data interoperability, and mitigating fragmentation risks [5]. Integrating AES-256 encryption and role-based access control ensures sensitive data protection and secure permissions, enabling seamless collaboration [6] and robust data management in proteomics.

This study introduces a microservices-based architecture for blockchain-based proteomic data storage, addressing integration and interoperability challenges in bioinformatics. The framework leverages the blockchain’s decentralized nature to ensure secure data ingestion and integrity while integrating databases like PDB and UniProt using standard protocols such as REST APIs and compatible formats like FASTA. Advanced security measures, including OAuth authentication and HTTPS encryption, safeguard sensitive data and enable secure access.

II. Methodology

The proposed microservices architecture for a blockchain-based proteomic data storage system integrates decentralized blockchain networks with bioinformatics repositories, specifically PDB and UniProt. This system utilizes a modular design, as illustrated in the architecture diagram in Fig 1. The framework operates through several interconnected components.

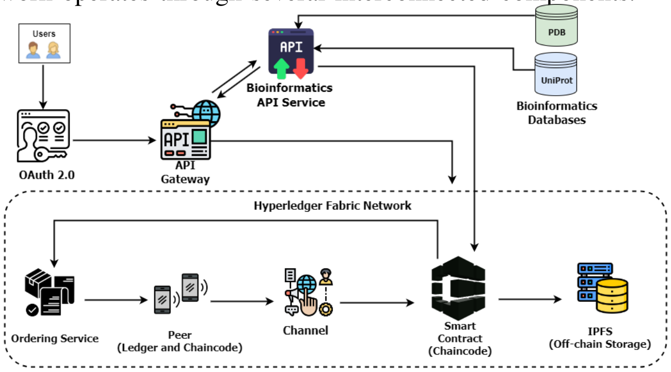

Fig. 1. Overview of the Proposed Microservices-based Architectural Design

The framework begins with ingesting proteomic data from PDB or UniProt into the hyperledger fabric network, facilitating seamless integration between it and the bioinformatics platforms. Advanced Encryption Standard (AES) with a 256-bit key is employed for data confidentiality. Data $(\mathtt{D})$ is encrypted into ciphertext $\mathtt{C}$ using Equation (1):

$$\mathtt{C} = \mathtt{E}_{\mathtt{K}}(\mathtt{D}) \tag{1}$$

where $\mathtt{E}_{\mathtt{K}}$ represents the encryption function with key $\mathtt{K}$. The decryption process, ensuring authorized access, is expressed as Equation (2):

$$\mathtt{P} = \mathtt{D}_{\mathtt{K}}(\mathtt{C}) \tag{2}$$

Here, $\mathtt{D}_{\mathtt{K}}$ is the decryption function using the same key $\mathtt{K}$. This ensures end-to-end security during data transmission and storage. Using cryptographic hashing, each proteomic dataset $\mathtt{D}$ is uniquely represented as $\mathtt{H}(\mathtt{D})$, calculated using a secure hash algorithm, as shown in Equation (3):

$$\mathtt{H}(\mathtt{D}) = \mathtt{SHA256}(\mathtt{D}) \tag{3}$$

Here, $\mathtt{H}(\mathtt{D})$ is the cryptographic hash of the dataset $\mathtt{D}$, ensuring its tamper-proof integrity. The hashed representation is stored on the blockchain, while the encrypted raw dataset $\mathtt{C}$ resides securely off-chain in an InterPlanetary File System (IPFS). A reference $\mathtt{R}$ is obtained and associated with $\mathtt{H}(\mathtt{D})$.

Role-Based Access Control (RBAC) mechanisms provide granular permission enforcement. The permissions $\mathtt{R}(\mathtt{u})$ assigned to a user $\mathtt{u}$ are evaluated against the requested dataset $\mathtt{D}$, as shown in Equation (4):

$$\mathtt{R}(\mathtt{u}) = \{\mathtt{p}_1, \mathtt{p}_2, \ldots, \mathtt{p}_n\} \tag{4}$$

Access is granted if the user's role satisfies the required permission, $\mathtt{T}(\mathtt{u}, \mathtt{D}) = \mathtt{Valid}$, expressed in Equation (5):

$$\mathtt{Access}(\mathtt{u}, \mathtt{D}) \equiv \mathtt{p}_\mathtt{i} \in \mathtt{R}(\mathtt{u}) \land \mathtt{T}(\mathtt{u}, \mathtt{D}) = \text{Valid} \tag{5}$$

where the logical AND ($\land$) ensures both conditions are met, and the equivalence ($\equiv$) guarantees access is tightly controlled.

This secure and decentralized system also ensures seamless interoperability through smart contracts. These automate data validation and synchronization processes across repositories like PDB and UniProt. Synchronization is achieved by periodically verifying on-chain and off-chain data consistency, with the synchronization condition defined as $\mathtt{H}(\mathtt{D})_{\text{external}} \neq \mathtt{H}(\mathtt{D})_{\text{on-chain}}$. Discrepancies trigger updates to ensure real-time alignment.

Performance benchmarking was conducted using Hyperledger Caliper to evaluate blockchain throughput, latency, and consistency. Locust testing validated API performance and schema compliance with UniProt and PDB. OAuth 2.0 authentication and Role-Based Access Control (RBAC) secured access, while hash-based comparisons ensured synchronization between external repositories and the blockchain for seamless integration. The simulation platform for the pytest (v.8.3.4) was a Python environment (v.3.12.8) using macOS (Darwin). While the endorsing node of Hyperledger Fabric for our blockchain experiments was on an Ubuntu Linux environment equipped with 1 vCPU (2.3 GHz) and 2 GB of RAM.

III. RESULT AND DISCUSSION

Table I summarizes the benchmarking results of the Protein Metadata Chaincode ingested from PDB and UniProt, tested using Hyperledger Caliper.

TABLE I

SUMMARY OF BLOCKCHAIN PERFORMANCE METRICS

Parameters	Store	Query
No. of TXs	1000	100
Succ	1000	1000.
Send Rates (TPS)	100, 200, 300, 400, 500	100, 200, 300, 400, 500.
Latency (s)	0.10, 1.68, 2.11, 2.38, 1.96	0.11, 0.81, 1.97, 1.93, 2.20.
Throughput (TPS)	99.8, 160.6, 191.1, 168.7, 197.0	99.6, 185.2, 184.4, 196.3, 193.5.

The blockchain performance metrics for 1000 transactions demonstrate consistent scalability and efficiency across varying send rates for store and query operations. For store operations, throughput ranges from 99.8 TPS to 197 TPS, with latency increasing from 0.10s to 2.38s. Similarly, query operations achieve throughput between 99.6 TPS and 193.5 TPS, with latency ranging from 0.11s to 2.20s. These results highlight the system's robustness in handling bioinformatics transactions under different configurations.

Fig. 2 shows performance metrics under 1000 users: requests peaked at 90 RPS, failures remained low, and 95th-percentile response time spiked to 20s before stabilizing. User load scaled consistently, highlighting the system's handling of high concurrency and stress conditions effectively.

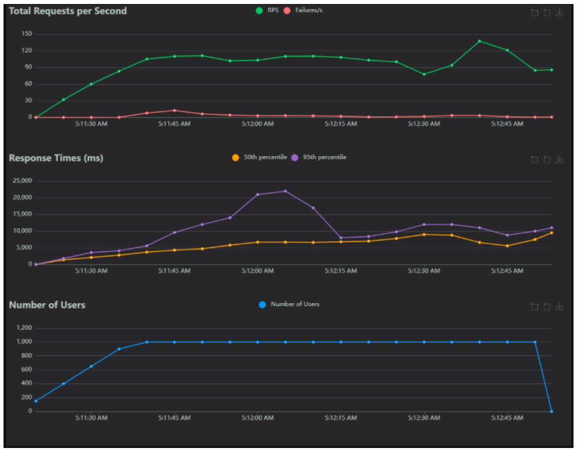

Fig. 2. Performance Metrics of Locust Load Testing

IV. CONCLUSION

This study proposed a blockchain-based proteomic data storage system microservices architecture that ensures secure, interoperable bioinformatics integration. Performance tests validated robust throughput, secure access, and data consistency. Future improvements will focus on enhancing scalability and optimizing API performance for broader applications in proteomic data management.

ACKNOWLEDGMENT

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korea government(MSIT) (IITP-2025-RS-2020-II201612, 33%) and by Priority Research Centers Program through the NRF funded by the MEST(2018R1A6A1A03024003, 33%) and by the MSIT, Korea, under the ITRC support program(IITP-2025-RS-2024-00438430, 34%).

REFERENCES

[1] P. Samaras, T. Schmidt, M. Frejno, S. Gessulat, M. Reinecke, A. Jarzab, J. Zecha, J. Mergner, P. Giansanti, H.-C. Ehrlich *et al.*, "Proteomicsdb: a multi-omics and multi-organism resource for life science research," *Nucleic acids research*, vol. 48, no. D1, pp. D1153–D1163, 2020.

[2] J. Ma, T. Chen, S. Wu, C. Yang, M. Bai, K. Shu, K. Li, G. Zhang, Z. Jin, F. He *et al.*, "iprox: an integrated proteome resource," *Nucleic acids research*, vol. 47, no. D1, pp. D1211–D1217, 2019.

[3] Z. Wang and Y. He, "Precision omics data integration and analysis with interoperable ontologies and their application for covid-19 research," *Briefings in Functional Genomics*, vol. 20, no. 4, pp. 235–248, 2021.

[4] M. R. Kamdar, J. D. Fernández, A. Polleres, T. Tudorache, and M. A. Musen, "Enabling web-scale data integration in biomedicine through linked open data," *NPJ digital medicine*, vol. 2, no. 1, p. 90, 2019.

[5] S. O. Ajakwe, I. I. Saviour, J.-H. Kim, D.-S. Kim, and J. M. Lee, "Banda: A novel blockchain-assisted network for drone authentication," in *2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2023, pp. 120–125.

[6] S. O. Ajakwe, I. I. Saviour, V. U. Ihekoronye, O. U. Nwankwo, M. A. Dini, I. U. Uchechi, D.-S. Kim, and J. M. Lee, "Medical iot record security and blockchain: Systematic review of milieu, milestones, and momentum," *Big Data and Cognitive Computing*, vol. 8, no. 9, p. 121, 2024.