Amazon Biobank: a Blockchain-based Genomic Database for Bioeconomy

Leonardo T. Kimura[1], **Marcos A. Simplício Jr.**[1]

[1]Escola Politécnica - Universidade de São Paulo (USP)

{lkimura, mjunior}@larc.usp.br

Abstract. *The bioeconomy, an industrial production model based on biological resources and sustainable development, can be considered an emerging opportunity for biodiversity-abundant regions, such as the Amazon rainforest. However, existing genomic repositories lack data traceability and economic benefit-sharing mechanisms, resulting in limited motivation for data providers to contribute. To address this challenge, we present Amazon Biobank, a community-driven genetic database. By leveraging blockchain and peer-to-peer (P2P) technologies, we enable distributed and transparent data sharing; meanwhile, by using smart contracts directly registered in the system, we enforce fair benefit-sharing among all system participants. Moreover, Amazon Biobank is designed to be auditable by any user, reducing the need for trusted system managers. To validate our approach, we implemented a prototype using Hyperledger Fabric and BitTorrent and evaluated its performance. Our results show that the prototype can support at least 400 transactions per second in a small network and that it can be further improved by adding new nodes or allocating additional computational resources. We expect that Amazon Biobank will serve as a vital tool for collaborative biotechnology research, fostering sustainable development in high-biodiversity regions.*

1. Introduction and Motivation

High-biodiversity regions have great potential to develop economic activities in a sustainable manner [Nobre and Nobre 2019, Nobre et al. 2016]. The Amazon rainforest alone, for example, provides critical ecological services whose annual value is estimated to be worth trillion dollars [Strand et al. 2018]. Moreover, the biodiversity in these regions, built over millions of years of evolution [Hoorn et al. 2010], has the potential to stimulate biotechnological development in several fields. Examples include biomimetic engineering, synthetic biology, and the development of new materials, chemical compounds, and biofuels [Nobre and Nobre 2019, Rech 2011].

However, there are still significant challenges to fostering bioeconomic development in high-biodiversity regions. One is the scale of effort necessary to survey millions of species across extensive forest areas – e.g., the Amazon rainforest exceeds 5 million square kilometers. Thus, the collaboration of local residents is of immense value, especially in critical activities such as identifying species with specific properties (e.g., medicinal). Nevertheless, there is not much incentive for such collaboration. After all, it is not always obvious how to ensure fair compensation for these residents. In addition, this collaboration is hindered by the practice of biopiracy in deprived areas [Mgbeoji 2007], as had occurred in the field of genomic research [Li 2021]. Partly for this reason, the

principle of benefit-sharing is one of the main objectives of the Convention on Biological Diversity (CDB), a global agreement that aims at the conservation and the sustainable use of biodiversity [Glowka et al. 1994]. This principle can thus be seen as a strong requirement to better promote data sharing and, ultimately, promote biotechnology initiatives.

One proposal in that direction is to build a collaborative and highly scalable genomic database. This database could be populated by any resident of areas of interest, who would retain data ownership and receive appropriate compensation for their contribution. This approach contrasts with (and also complements) the many genomic repositories that currently support biotechnological research. For example, the US National Center for Biotechnology Information (NCBI) maintains a genomic database, which is also done by the European Bioinformatics Institute (EBI) in Europe, and the DNA Databank of Japan (DDBJ). However, these repositories make data publicly available without any kind of usage tracing. This model may facilitate data re-usage, but it does not contribute to an adequate sharing of economic benefits. For example, even if a highly profitable medicine is developed using genomic data from these repositories, their profits are usually not distributed to the corresponding data provider. Consequently, even people with easier access to genomic data (e.g., residents of high-biodiversity regions) are not encouraged to contribute. This results in less data variety and less development in the local bioeconomy.

To address these issues, recent works in the literature have suggested the deployment of collaborative technologies as an integral part of genomic repositories. For example, many studies discuss the potential benefits of blockchain for healthcare genomics [Ozercan et al. 2018, Alghazwi et al. 2022, Beyene et al. 2022]. In this scenario, blockchain would not simply be an overhyped technology but could be used as a transparent and verifiable record of transactions involving digital assets (e.g., DNA data). Moreover, with the development of a special-purpose currency, blockchain could contribute to fair benefit-sharing among all players of the system. Some of the opportunities created by the technology include data integrity, data ownership (i.e., the owner controls the use of the data), and decentralization (to avoid a single point of failure or to enable distributed data processing).

These characteristics of blockchains motivated us to develop the Amazon Biobank, a community-based genetic database designed to better support biodiversity research in high-biodiversity regions. This would result in (1) larger data variety, since data providers would be compensated for their contribution; and (2) cost reduction, since it is potentially cheaper to purchase genomic data from a database than to organize an expedition to the remote places where that information is found. Amazon Biobank uses blockchain to transparently trace biotechnology products and research to genomic data in the repository. It also uses smart contracts to appropriately share the benefits among all the participants that collect, insert, process, store, and validate genomic data. In addition, it uses other peer-to-peer (P2P) technologies like BitTorrent [Cohen 2003] to build a highly scalable and collaborative computing environment, in which users can contribute (and be paid for) genomic data, computational, and bandwidth resources. Finally, it provides auditability not only for internal system managers but also for any external users. Thus, the correct operation of the system does not critically depend on system administrators, making the architecture adherent to the zero-trust principle [Gilman and Barth 2017].

2. Goals

The primary goal of the Amazon Biobank is to facilitate collaborative biotechnology research in regions with ecologically rich ecosystems. This involves improving traceability by linking each research project to the DNA data used and associating uploaded DNA data with the identity of the uploader. The system must also be auditable, allowing for independent verification of its correct operation without critically trusting administrators.

Thus, this research seeks to answer the following research question: ***"How to build a genetic database with transparency, scalability, and benefit-sharing properties?"***

Since Amazon Biobank improves transparency, by providing easier access to genomic data, while keeping auditability and benefit-sharing, it contributes to the "Grand Research Challenges in Information Systems in Brazil 2016–2026", specifically in the theme of "Information Systems and the Open World Challenges" [Boscarioli et al. 2017].

3. Research Methodology

Our research methodology consists of the following steps

1. **Brainstorming**: we discussed with several stakeholders, including Biologists, layers, economists, NGOs, and engineers. We defined the research problems and produced a brief draft of the solution.
2. **Requirements definition**: we defined the Amazon Biobank requirements, based on previous discussions.
3. **Informal review of related works**: we investigated the state-of-the-art of blockchain-based genomic databases, examining surveys and analyzing solutions from genomic-related companies.
4. **Design**: We defined the Amazon Biobank operation, including roles, operation flows, and how benefit-sharing would work.
5. **Prototype implementation**: We build a prototype with the main operations of Amazon Biobank, including inserting, processing, distributing, and purchasing genomic data.
6. **Performance evaluation**: We evaluate the prototype performance, assessing its viability in real-world deployment.

4. Related Works

Many blockchain-based genomic repositories aim to remove brokers and increase user control over their data. Table 2 lists relevant examples, comparing some of their features that are relevant to biodiversity research: support for data validation, distributed processing, sequence search, benefit sharing, and an association between the data and its owner (see Table 1).

One example of a proposal is a genomic marketplace such as Encryptgen [Encrypgen 2017], a platform in which users can provide their genetic data in exchange for cryptocurrency tokens. Similarly to the Amazon Biobank, it stores DNA data in encrypted form and registers metadata and transactions in a blockchain. However, the platform has little support for collaboration among end-users. For example, collectors cannot request other users to contribute to tasks such as genomic sequencing and assembling.

Tabela 1. Some biobank properties relevant to biotechnology research

Property	Description
Data validation	Check the correctness of the data or metadata entered
Distributed processing	Outsource the genetical data assembling and sequencing to other users
Sequence search	Query genetic data based on similarity to a sequence of interest
Benefit sharing	Distribute the economic benefits to all involved players
Owner association	Associate each uploaded DNA sequence with its uploader

Tabela 2. Comparison between blockchain-related genetic projects

	Data Validation	Distributed processing	Sequence Search	Benefit Sharing	Owner Association
Encryptgen	●	-	-	-	-
Zenome	●	●	-	-	-
Nebula Genomics	●	●	●	-	-
Genesy	●	●	●	-	-
Global ABS Tracker	-	-	-	●	●
Amazon Biobank	●	●	●	●	●

● = provides property; - = does not provide property;

Additionally, buyers cannot perform any genomics-relevant operations directly on the uploaded data, without downloading it. Consequently, it does not support sequence-based searches, i.e., finding relevant data based on a given DNA sequence.

The Zenome platform [Kulemin et al. 2017], in turn, allows end-users to take on the role of a computational node, providing storage and CPU time in exchange for ZNA tokens. Therefore, like Amazon Biobank, the costs of processing and sharing data can be distributed among participants. In addition, Zenome employs a data rating system, in which high-quality data are identified, while less valuable data are penalized. Like Encryptgen, though, the Zenome platform does not support sequence search operations.

Nebula Genomics [Grishin et al. 2018] allows computation over encrypted DNA data by using privacy-preserving techniques, such as partially homomorphic encryption and secure computation over Intel Secure Guard Extension (SGX). Hence, it can better reconcile data privacy with distributed data processing operations. However, Nebula's business model tends to decrease user control over genomic data: once the data is registered in the system, users have little visibility or oversight over how it is handled. In contrast, Amazon Biobank allows users to better control their genomic data through configurable smart contracts, defining the price and the conditions for its use.

Genesy [Carlini et al. 2019] is another genomic marketplace that provides sequencing services, such as access to genomic data and the distributed sharing of DNA sequences. Similar to the Amazon Biobank, the authors of Genesy argue that permissioned blockchains (e.g., Hyperledger Fabric) better meet the complex requirements of genetic data storage, such as user identification. It supports the purchase of access to genomic data in both fiat and cryptocurrency transfers, using third-party APIs for this purpose. In addition, Genesy aims to eventually aggregate more organizations into its consortium, strengthening its open governance model and encouraging collaboration and fairness.

We note that Nebula Genomic, Genesy, and other platforms prioritize human genetics, rather than focusing on biodiversity as an asset. Consequently, they provide limited support for intellectual property protection or benefit sharing. In addition, to safeguard user privacy, these platforms often restrict the identification of data owners to the company or federation only. While this is appropriate in the context of human genetics, it is not necessarily desirable in the case of the Amazon Biobank. In particular, the anonymization of data owners possibly hinders the preservation of their intellectual property rights and restricts their fair compensation.

In the context of non-human genetic data, in 2021, the United Nations Development Programme (UNDP) conducted a blockchain-based project to improve the traceability of genomic resources and benefit-sharing [UNDP 2021]. With the major goal of implementing the Nagoya Protocol [Buck and Hamilton 2011], the Global ABS Tracker project is currently in the early stages, with a pilot prototype launched. The project, nonetheless, tries to handle all kinds of natural products, such as plants or natural substances, and does not focus solely on genetic data. Hence, the system does not support collaborative and private storage of genomic data, nor the analysis, validation, and search of DNA sequences. Also, one of the challenges of the project is that it requires global coordination among countries, something that is still a work in progress.

5. Amazon Biobank

In this section, we discuss the overall design of Amazon Biobank

5.1. System Requirements

Based on Amazon Biobank's goals, we consider some functionalities as essential requirements:

- **Data insertion:** The ability to collect DNA sequences in different forms (raw, assemblies, or annotated) and upload them into the system together with any relevant metadata (common name, scientific name, where it was collected, information about its common usages, etc.)
- **Owner association:** Association of the uploaded DNA sequences with the identity of the uploader, to preserve the latter's rights and to ensure fair compensation. This procedure must be performed in a verifiable manner, i.e., it must not depend critically on the trust deposited in the system entities.
- **Data validation:** The ability to validate the correctness of inserted data (e.g., that processed DNA sequence corresponds to some previously registered raw DNA data), or at least giving confidence of its correctness (e.g., by means of a reputation system).
- **Sequence search:** The ability to search for specific data among the entries inserted into the system.
- **Benefit-sharing:** The possibility of purchasing access rights to data of interest and then downloading it. All actors who helped in making that data available (e.g., by collecting, processing, validating, and/or distributing it) should then be properly remunerated.

A few non-functional requirements are similarly relevant:

Tabela 3. Main players of Amazon Biobank.

Player	Operation	Description
Collector	Data insertion	Usually a resident of target high-biodiversity region
Processor	Sequence and Assembly	Any user with computational power to spare
Distributor	P2P distribution	Any user with enough storage and bandwidth capabilities
Buyers	Purchase access	Any user interested in genomic data (e.g., researchers)
Curator	Verify data and metadata	Users with a background in genomics (e.g., biologists)
Validator	Verify processing	Processors who validate the work of their peers
Federation	Maintain system operation	System stakeholders, such as Universities and (non-)governmental organizations

- **Traceability**: the system must provide some level of traceability for biotechnology developments resulting from DNA data stored in the Biobank (e.g., scientific discoveries or intellectual property). This feature promotes the reproducibility of results, which can be reliably traced to Biobank entries. This feature is useful both for academic purposes and to support claims about the prior existence of some data in the Biobank when handling disputes involving data misuse.
- **Scalability**: the system must be able to handle many users uploading and accessing data stored by the Biobank One challenge for this is that DNA files are usually large (e.g., many gigabytes) and operations on them (e.g., sequencing raw data, or searching for specific sequences) can be very time-consuming. We do not require that a large number of entities participate in the Federation though, as only some authorized entities would maintain the system.

5.2. Overview

The main players of the Amazon Biobank [Kimura et al. 2023, Kimura et al. 2021] and their respective roles in the system are as follows (see Table 3).

Collectors are responsible for one of the main operations of the system, the registration of raw DNA data. Typically, residents in regions of high-biodiversity extract raw DNA data using a portable sequencing device and enter them into the Amazon Biobank application. The application encrypts the genomic data and creates a ".torrent" file with the corresponding magnet link. This magnet link is recorded in the blockchain with any relevant metadata entered by the Collector (e.g., common name, place of extraction). Distributors can then use the magnet link to download (and later upload) the encrypted DNA data via BitTorrent. Moreover, Collectors may allow Curators to access plaintext data and metadata to assess and endorse its correctness, adding value to the corresponding records.

Raw DNA data typically require computational processing (e.g., assembling and sequencing) to be more usable in biotechnology research. Collectors can therefore outsource this task to Processors, players who offer their computing power in exchange for a reward. Processors contact Distributors to download raw DNA data, and then submit the processing results to the system as processed DNA. To avoid any misconduct (e.g., recording data that does not correspond to the raw DNA), these results are verified by the Validators, which are other Processors who have worked simultaneously or subsequently on the same DNA sequences. Any malicious behavior is punished accordingly, either by suspension of rewards, loss of reputation, or even eviction from the system.

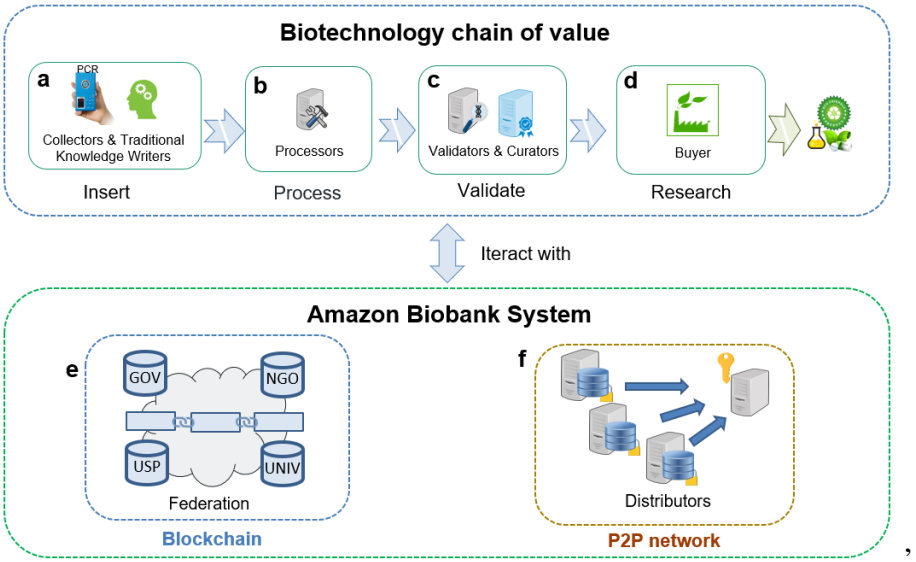


Figura 1. Amazon Biobank: overview and main operations

Finally, Buyers gain access to DNA data by paying some biocoins, the Biobank's internal cryptocurrency. This purchase is registered in the blockchain, and the paid biocoins are distributed via smart contracts to all entities involved in the acquisition and processing of DNA data (Collectors, Processors, Validators, and Curators). To help find DNA data of interest, Amazon Biobank supports some searching procedures, based on keywords matching the corresponding metadata, or on genomic sequences matching the DNA data content.

5.3. Architecture

The Amazon Biobank architecture is organized into three layers (see Figure 2): infrastructure, access management, and application. The infrastructure layer comprises the core components of Amazon Biobank, namely the blockchain network and the BitTorrent network. The access management layer manages access permissions to the Amazon Biobank data access. Finally, the application layer provides functionalities that users can directly interact with, comprising the interface applications, the auditing service, and the data search mechanism.

Blockchain: It is responsible for storing the magnetic links and all the information related to the transactions of the genetic data. The blockchain network is maintained by the Federation, which must register, order, and validate all Amazon Biobank transactions. For this purpose, each organization from the Federation deploys at least one Endorsing node with the necessary ledger and smart contracts; nevertheless, some more engaged organizations may also run an additional Orderer node. The smart contracts deployed in these Endorsing peers contain all the logic necessary for the operation of the Amazon Biobank. These include (1) tasks related to genetic data (insertion, processing, validation, and purchase); (2) auxiliary functions (e.g., reputation system procedures); (3) and cryptocurrency-related functions (e.g., biocoin minting, transferring, and payments for data access).

BitTorrent: This module enables Distributors to store and share encrypted DNA

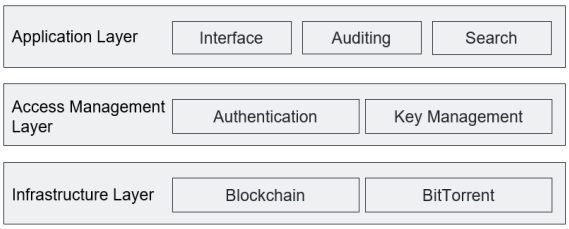


Figura 2. Main modules of Amazon Biobank

sequences in a collaborative and highly scalable manner. At its core, this module relies on Federation nodes that store backup copies of all genomic data. However, to increase data redundancy and facilitate distribution without overloading the Federation nodes, Distributors can also download and seed genomic data. This balance between Federation and independent nodes is achieved with the help of a BitTorrent tracker, which monitors the number of Distributors seeding a given torrent. If this number is lower than a configurable threshold, Federation nodes are expected to play the role of seeders; in this case, registered Distributors may be allowed to download the corresponding genomic data from Federation nodes without paying any fee. Conversely, when many seeders are available for a torrent, Federation nodes may refrain from participating in its distribution, keeping the corresponding data in cold storage only.

Authentication: In Amazon Biobank, user creation is managed by each organization via Certificate Authorities (CAs). Thus, each organization can validate users in the most convenient identification methods (e.g., password-based authentication, in-person ID verification, or a corporate email integrated with OAuth). User authentication is done through certificates and private keys. The user inserts their certificates and private keys in the local Biobank interface to sign all transactions sent to the blockchain.

Key Management: The key management module, or Keyguard, is responsible for storing and protecting the encryption keys of DNA data from unauthorized access. It (1) receives all secret key requests, (2) verifies if the requesting user is registered as the data owner in the blockchain; (3) if approved Keyguard gets a secret key from the system database and returns it. That way, only Buyers who acquire the rights to the key can decrypt the corresponding DNA data. Note that, to protect the data encryption keys within the Federation, Keyguard can use a secure storage device such as a Hardware Security Module (HSM).

The Application Layer: The application layer comprises the Amazon Biobank modules through which users can interact with the system. These modules are: (1) the user interface, from which the user can enter, download, and purchase access to genetic data; (2) the audit module, which allows users to verify the correct operation of the Biobank; and (3) the search tool, which is responsible for allowing users to search for data of interest (e.g., traditional keyword search, encrypted keyword search, or sequence-based search).

6. Prototype implementation

To illustrate the Amazon Biobank operation, we built a prototype that combines the technologies and layers described in Section 5.3. Our prototype supports the main operations on genetic data, including data registration, purchase, and download (see Fi-

gure 3). The prototype and the corresponding documentation are publicly available at `https://github.com/amazon-biobank/biobank`.

For the blockchain layer, we used Hyperledger Fabric to build a basic network with three endorsing nodes and one orderer. On each endorsing node, we developed and deployed smart contracts that implement the core functionalities of Amazon Biobank. These contracts ensure, for example, that each magnet link to DNA data is uniquely registered on the blockchain, preventing duplication.

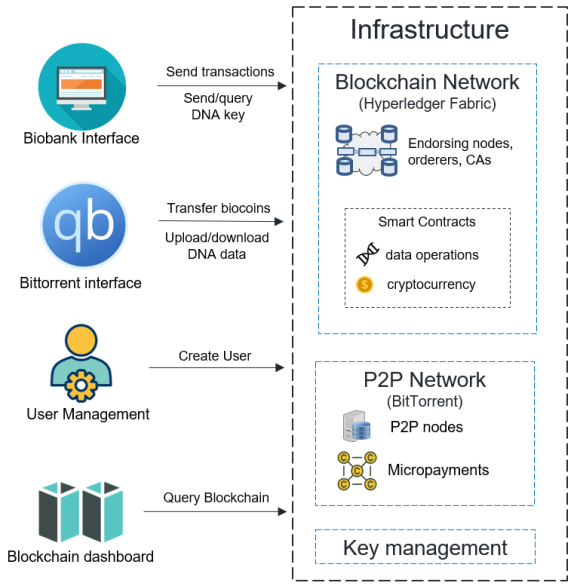


Figura 3. Main modules of the Amazon Biobank prototype

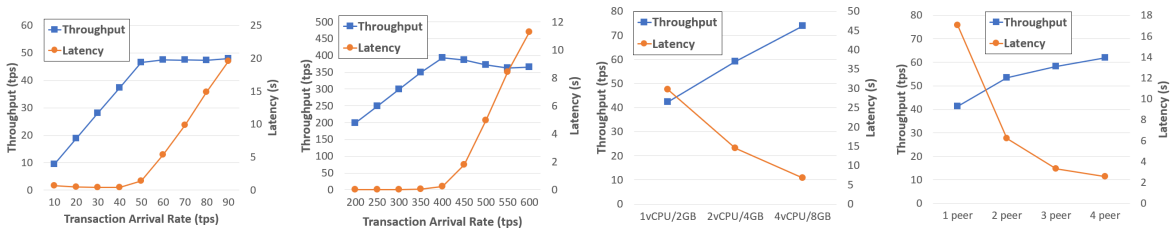
To interact with the blockchain layer, we built the `Biobank Interface`, an express.js web application. Users interact with this interface to register new genetic data, buy decryption keys for them, or perform any other operations in the system. To enable these functions, the interface interacts with the blockchain layer by sending signed transactions.

For the BitTorrent layer, we used `Torrente` [Shiraishi et al. 2021], a dedicated client application based on qBittorrent[1]. By installing this application, users can turn their machines into BitTorrent nodes, contributing to the genetic data distribution via P2P. `Torrente` implements a micropayment mechanism, which requires a small biocoin-based payment for each piece of data transmitted. These payments require communication with the blockchain layers so that the biocoins for a given account can be redeemed.

Furthermore, we have built a preliminary version of the modules related to access management in Amazon Biobank. Our `Key Management` module is responsible for storing the DNA decryption keys and was built using express.js and MySQL. We also built a `User Management` system to simulate the controls that an organization participating in the Federation would implement.

Finally, we deployed a `Blockchain Visualizer` tool based on the Hyperledger Explorer. This visualizer facilitates auditing procedures as it allows direct access

[1]`https://www.qbittorrent.org`



(a) Data write **(b)** Data read **(c)** Resources **(d)** Peers

Figura 4. Blockchain prototype performance in: (a,b) base configuration; (c) improving computing resources; (d) increasing the number of peers in each organization.

to blockchain data, including the transactions sent to the system, the blocks that were approved, and the number of peers and orderers operating in the network.

7. Performance Evaluation

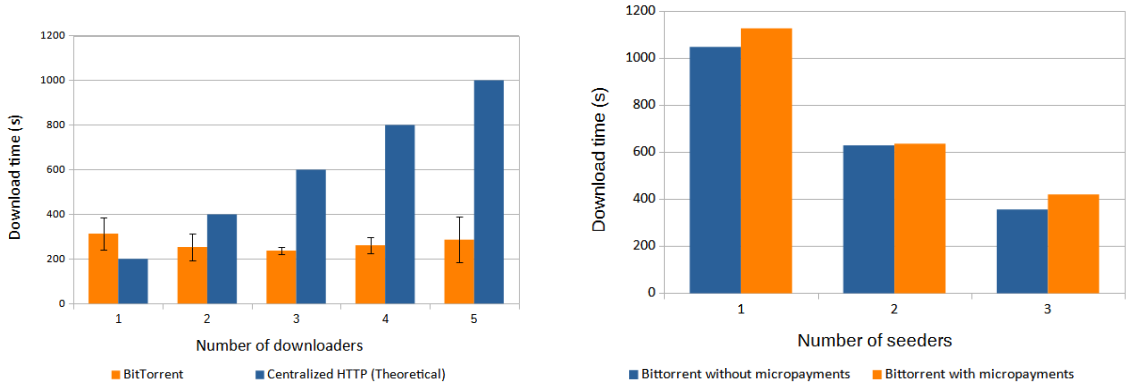
In this section, we evaluate the performance of the Amazon Biobank. We started by testing several blockchain configurations, varying the number of nodes, organizations, and computing resources. We then measured the download time of DNA data using BitTorrent, while also evaluating the overhead introduced by the Torrente micropayment protocol and comparing these results with a scenario that relies on centralized servers.

7.1. Blockchain performance

We then measured the latency and the throughput of two types of transactions. The first type is a data write transaction, representing the registration of new DNA data. This transaction inserts a new `rawData` object, with the corresponding metadata and magnet link, in the blockchain. The second type is a data read transaction, representing a user who inspects the details of a given entry. It consists of querying a `rawData` in the Hyperledger Fabric ledger using the corresponding identifier. To generate the transaction workload, we used Hyperledger Caliper, a performance benchmarking tool specialized in Hyperledger Fabric.

We first tested the prototype in a configuration that simulated the initial deployment of Amazon Biobank. Thus, we deployed 3 organizations with 1 endorsing peer each, representing a Federation made up of a few universities or NGO institutions. With this setup, our prototype was able to support at least 50 transactions per second (tps) for data write operations, and 400 tps for data reads (see Figure 4a and 4b). As it is reasonable to expect limited use of the system in the early phases of an actual deployment, this performance can be considered satisfactory.

On the other hand, the Amazon Biobank may require support for more operations as it gains popularity. Therefore, we have also investigated how some alternative settings could improve its overall performance. One possible strategy is to upgrade the computing resources available on each node (vertical scaling). For example, our experiments showed that an upgrade from 1 vCPU to 2VCPU can increase the transaction rate by 50% and reduce the latency time by less than half (see Figure 4c). Another option is to increase the number of Endorsing peers in each organization (horizontal scaling). In our experiments,



(a) Centralized servers **(b)** Micropayments

Figura 5. Experimental results for BitTorrent, comparing: (a) BitTorrent and centralized server; (b) download with and without micropayments

the simple strategy of adding an extra peer per organization is enough to reduce latency time by less than half and increase throughput by almost 50% (see Figure 4d). These results indicate that the performance of the Amazon Biobank responds well to both vertical and horizontal scaling strategies. We also evaluate the effect of adding more contribution organizations to the Federation, concluding that the blockchain performance is slightly improved in data writes, and linearly improved in data queries.

7.2. Bittorrent performance

For our experiments with DNA data exchanges, BitTorrent nodes were deployed in personal computers with Intel core i5/i7 and 8GB RAM or more. We configured the upload rate of each node to 1MiB/s, and we divided a large file containing random data into 16 KiB pieces, following the BitTorrent protocol.

For this setting, our first experiment considers a Collector inserting new DNA data into the system. The Collector thus acts, thus, as a data seeder, and the Distributors become leechers while acquiring the corresponding data pieces. To test this scenario, we set up 1 seeder with a 200MB file and 1-5 simultaneous downloaders. Figure 5a shows the resulting download times. This figure shows that, with BitTorrent, the download time is quite similar even when we increase the number of downloaders. This behavior is expected, since BitTorrent relies on the upload capabilities of all nodes during file exchange, and is a suitable technology for handling large files (which is typical of genetic data) and many simultaneous downloaders. In contrast, the download time for a centralized HTTP server, whose bandwidth has been also limited to 1 MiB/s, only increases with the number of downloaders. This scenario illustrates how a centralized architecture would quickly lead to bottlenecks when distributing DNA data among interested users.

We also evaluated the performance of the Torrente micropayments mechanism. Specifically, we measured the overhead introduced by this mechanism by comparing the download time when micropayments are enabled or disabled in the system. Our experiments show that the micropayment increases the download time by an average of 10% (Figure 5b). We also observed that this already low overhead can be lowered even further when extra seeders are available (each extra seed leads to almost a 30% decrease in the

download time). Hence, we can argue that the benefits of financially incentivizing new seeders can be considered enough to compensate for these overheads.

8. Discussion

The Amazon Biobank promotes the collaborative development of biodiversity-based research in regions with rich ecosystems. By combining blockchain, smart contract, and P2P technologies, the system fulfills all requirements described in Section 5.1:

- **Data Insertion:** Collectors can upload raw DNA data and the corresponding metadata, and Processors can then download and process those raw sequences, uploading annotated sequences. Distributors reinforce the system's data storage and bandwidth. Those contributing players are rewarded with biocoins, the system's internal currency.
- **Owner association:** The system registers all operations related to a DNA sequence. That way, in case of legal disputes, users can use the transparent log provided by the blockchain as proof of such events. This approach is expected to facilitate real-world actions regarding intellectual property protection.
- **Data validation:** Curators and Validators help to promote data and metadata quality, increasing the confidence in the accuracy of the corresponding entries. If misconduct is detected, the culprit's reputation is penalized; in case of repeated misbehavior, the user may be evicted from the system, losing access to existing funds and to future profit opportunities.
- **Sequence search:** The system is organized in such a manner that Federation nodes can perform searches in local data, as well as collaborate with other nodes (e.g., Collectors, Processors, and Buyers) for that purpose.
- **Benefit-sharing:** To legitimately access some DNA sequence, Buyers must invest some biocoins. Except for eventual system fees, the total amount paid by Buyers is then shared among all players responsible for the availability of that data entry (not only Collectors but also Processors, Distributors, Validators, and Curators). If some profit is made thanks to that data, a certain amount of royalties is also expected to be reverted to those players, according to the rules established in smart contracts configured by the data stakeholders.

Similarly, the system also fulfills its non-functional requirements:

- **Traceability**: all data is converted into torrent files for integrity protection, and operations are recorded in the blockchain to create a temporal and transparent log of events. Hence, any research or product developed from the registered data can be securely traced back to the corresponding entry in the Biobank. For example, this allows Buyers to prove the relationship between their products and Amazonia's biodiversity. Once some data entry is registered in the system, attempts to modify its contents or place in time can be detected by auditors. Anyone can audit the blockchain and verify the correctness of operations thereby registered, so the system's transparency is independent of the amount of trust bestowed upon the Federation itself.
- **Scalability**: the system leverages collaborative distributed technologies, like BitTorrent, to avoid many of the scalability issues typically found in centralized platforms (e.g., bandwidth and storage limitations). In addition, the burden of executing computationally intensive tasks, like DNA processing and sequence-search

operations, can be shared among system players. Also, the blockchain stores only a small reference to DNA data, employing a permissioned consensus that allows high-throughput and low-latency transactions. The resulting collaborative architecture is, thus, expected to be capable of handling a large number of users and data.

8.1. Remaining Challenges

Although Amazon Biobank addresses many of the concerns faced by collaborative genetic repositories, some challenges remain. These issues are similar to those discussed in other blockchain-based management systems, and the solutions have not yet been definitively established [Ito and O'Dair 2019].

One significant challenge is how to prevent the insertion of false or forged DNA data. Although automatic detection mechanisms exist [Seppey et al. 2019], it can be difficult to identify data that has been specifically forged to bypass these strategies. To mitigate this challenge, Amazon Biobank uses a data validation process that involves Curators, Processors, and Buyers checking genomic data after accessing its plaintext. Collectors are only rewarded once their data are deemed valid (or, at least, receive no credible complaints), and misbehaving users can face punishments, such as loss of reputation or suspension from the system.

A second challenge is that Amazon Biobank cannot prevent DNA data distribution outside the system. For example, a user with access to the decryption key may decide to share it freely with others. However, we argue that the typical consumers of DNA sequences have stronger incentives to use data from the Amazon Biobank than from rogue sources, due to the traceability and auditability benefits the system provides. For instance, companies with environmental, social, and corporate governance (ESG) policies usually need to produce evidence of their efforts, and researchers need to provide reproducible results for their publications. For them, sharing data outside the system only harms themselves, by benefiting competitors without bringing any obvious gain.

In addition to these technical challenges, Amazon Biobank also needs to consider local and international regulations. In many countries, there are quite strict laws that regulate biodiversity exploration, varying from notification to the authorities to royalty payments. In addition, the monetary value of biocoins may require some adjustments to enforce financial compliance. Hence, collaboration with multiple stakeholders, such as government agencies, cryptocurrency exchanges, and potential data users, is fundamental for the success of a real Amazon Biobank deployment.

9. Conclusion

In this work, we present the Amazon Biobank, a community-based genetic database that implements monetary incentives for users who collaborate with data, knowledge, and computational resources. The resulting system provides strong traceability and auditability features, making it easier to link biotechnology assets to registered data and to verify compliance with data usage and benefit-sharing agreements. In addition, by leveraging collaborative technologies like BitTorrent and blockchain, the proposed architecture becomes highly scalable and less dependent on the trust deposited in any particular system player.

Our system serves as an alternative to several existing databases that register biodiversity genetic data, such as NCBI and EBI. Despite the relevance of those repositories, they lack adequate sharing of economic benefits resulting from exploring genomes. In our solution, people with easy access to high-biodiversity areas, such as local community members, are encouraged to insert genetic data. This will increase the variability of DNA data cataloged, especially in challenging and extensive areas such as the Amazon Rainforest.

9.1. Publications

Resulting of the research carried out during this work, we produced the following publications:

- **Journal Article:**
  - Kimura, L. T., Shiraishi, F. K., Andrade, E. R., Carvalho, T. C., & Simplicio, M. A. (2024). Amazon Biobank: Assessing the Implementation of a Blockchain-based Genomic Database. *IEEE Access*.
  - Kimura, L. T., Andrade, E. R., Nobre, I., Nobre, C. A., de Medeiros, B. A., Riaño-Pachón, D. M., ... & Simplicio Jr, M. A. (2023). Amazon Biobank: A collaborative genetic database for bioeconomy development. *Functional & Integrative Genomics*, 23(2), 101.
- **Conference Papers:**
  - Kimura, L. T., Andrade, E. R., Carvalho, T. C., & Junior, M. A. S. (2021, October). Amazon Biobank - A community-based genetic database. In Anais Estendidos do XXI Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (pp. 74-81). SBC.
  - Shiraishi, F. K., Perles, V. H., Yassuda, H. K., Kimura, L. T., Andrade, E. R., & Junior, M. A. S. (2021, October). Torrente, a micropayment-based Bittorrent extension to mitigate free riding. In Anais Estendidos do XXI Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (pp. 82-89). SBC.

10. Future Work

The next steps for Amazon Biobank include deploying the prototype in the Amazon region, first still as a small-scale demonstration. It would include presenting the prototype to target populations (e.g., traditional community members) and populating it with real genetic data to collect feedback. Thus, we intend to make the system as user-friendly as possible, avoiding Collectors having to rely on others to insert the data. This deployment most probably would include collaboration with other universities or non-profit organizations (such as Amazon 4.0 initiative[2]).

After those tests, we plan to execute a larger-scale assessment. This would include establishing an intercommunication channel with existing systems that mediate and lay the legal grounds for the usage of genetic data in the Amazon Forest, like the Brazilian National System for the Management of Genetic Heritage and Associated Traditional Knowledge [SisGen 2021]. Other non-technological aspects also must be emphasized, such as regulations on biodiversity and financial assets. We expect that this deployment will result in a more mature specification that can be used to implement the definitive version of Amazon Biobank.

[2]https://amazonia4.org/

Referências

Alghazwi, M., Turkmen, F., Van Der Velde, J., and Karastoyanova, D. (2022). Blockchain for genomics: A systematic literature review. *Distrib. Ledger Technol.*, 1(2).

Beyene, M., Toussaint, P. A., Thiebes, S., Schlesner, M., Brors, B., and Sunyaev, A. (2022). A scoping review of distributed ledger technology in genomics: thematic analysis and directions for future research. *Journal of the American Medical Informatics Association*, 29(8):1433–1444.

Boscarioli, C., de Araujo, R. M., and Maciel, R. S. P. (2017). I GranDSI-BR Grand Research Challenges in Information Systems in Brazil 2016-2026.

Buck, M. and Hamilton, C. (2011). The Nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. *Review of ECIEL*, 20(1):47–61.

Carlini, R., Carlini, F., Dalla Palma, S., and Pareschi, R. (2019). Genesy: a blockchain-based platform for DNA sequencing. *DLT@ ITASEC*, 2019:68–72.

Cohen, B. (2003). Incentives build robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer systems*, volume 6, pages 68–72.

Encrypgen (2017). The clinical and investment potential in the gene-chain project the unprecedented growth of genomic data.

Gilman, E. and Barth, D. (2017). *Zero Trust Networks*. O'Reilly Media, Incorporated.

Glowka, L., Burhenne-Guilmin, F., Synge, H., McNeely, J. A., and Gündling, L. (1994). A guide to the convention on biological diversity.

Grishin, D., Obbad, K., Estep, P., Cifric, M., Zhao, Y., and Church, G. (2018). Blockchain-enabled genomic data sharing and analysis platform. Technical report, Nebula Genomics.

Hoorn, C., Wesselingh, F. P., ter Steege, H., Bermudez, M. A., Mora, A., Sevink, J., Sanmartín, I., Sanchez-Meseguer, A., Anderson, C. L., Figueiredo, J. P., Jaramillo, C., Riff, D., Negri, F. R., Hooghiemstra, H., Lundberg, J., Stadler, T., Särkinen, T., and Antonelli, A. (2010). Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science*, 330(6006):927–931.

Ito, K. and O'Dair, M. (2019). A critical examination of the application of blockchain technology to intellectual property management. In *Business transformation through blockchain*, pages 317–335. Springer.

Kimura, L., Andrade, E., Carvalho, T., and Junior, M. S. (2021). Amazon biobank - a community-based genetic database. In *Proc. of the XXI SBSeg*, pages 74–81, Porto Alegre/RS, Brazil. SBC.

Kimura, L. T., Andrade, E. R., Nobre, I., Nobre, C. A., de Medeiros, B. A. S., Riaño-Pachón, D. M., Shiraishi, F. K., Carvalho, T. C. M. B., and Simplicio, M. A. (2023). Amazon biobank: a collaborative genetic database for bioeconomy development. *Functional & Integrative Genomics*, 23(2):101.

Kulemin, N., Popov, S., and Gorbachev, A. (2017). The zenome project: Whitepaper blockchain-based genomic ecosystem. *Zenome. io*, page A.

Li, F.-W. (2021). Decolonizing botanical genomics. *Nature Plants*, 7(12):1542–1543.

Mgbeoji, I. (2007). *Global biopiracy: patents, plants, and indigenous knowledge*. ubc Press.

Nobre, C. A., Sampaio, G., Borma, L. S., Castilla-Rubio, J. C., Silva, J. S., and Cardoso, M. (2016). Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. *Proc. of the National Academy of Sciences*, 113(39):10759–10768.

Nobre, I. and Nobre, C. A. (2019). The Amazonia third way initiative: the role of technology to unveil the potential of a novel tropical biodiversity-based economy. *Land use. Assessing the Past, Envisioning the Future*.

Ozercan, H. I., Ileri, A. M., Ayday, E., and Alkan, C. (2018). Realizing the potential of blockchain technologies in genomics. *Genome research*, 28(9):1255–1263.

Rech, E. (2011). Genomics and synthetic biology as a viable option to intensify sustainable use of biodiversity. *Nature Precedings*.

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). *BUSCO: Assessing Genome Assembly and Annotation Completeness*, pages 227–245. Springer, New York, NY.

Shiraishi, F., Perles, V., Yassuda, H., Kimura, L., Andrade, E., and Simplicio, M. (2021). Torrente, a micropayment based Bittorrent extension to mitigate free riding. In *Proc. of the XXI Brazilian Symposium on Information and Computational Systems Security (SBSeg)*, pages 82–89, Porto Alegre/RS, Brazil. SBC.

SisGen (2021). Sisgen - sistema nacional de gestão de patrimônio genético e do conhecimento tradicional associado. `https://www.mma.gov.br/patrimonio-genetico/conselho-de-gestao-do-patrimonio-genetico/sis-gen`. [Online; accessed 23-February-2021].

Strand, J., Soares-Filho, B., Costa, M. H., Oliveira, U., Ribeiro, S. C., Pires, G. F., Oliveira, A., Rajão, R., May, P., van der Hoff, R., Siikamäki, J., da Motta, R. S., and Toman, M. (2018). Spatially explicit valuation of the Brazilian Amazon forest's ecosystem services. *Nature Sustainability*, 1(11):657–664.

UNDP (2021). A pilot to improve genetic resources traceability through blockchain technology launched by the UNDP GEF Global ABS project. `https://bit.ly/3hnMqEh`. Acessed on 29-06-2021.