International Journal of Advanced Smart Convergence Vol.13 No.3 150-155 (2024)
http://dx.doi.org/10.7236/IJASC.2024.13.3.150

IJASC 24-3-17

Genomic data Analysis System using GenoSync based on SQL in Distributed Environment

Seine Jang*, Seok-Jae Moon**

** The master's course, Graduate School of Smart Convergence, Kwangwoon University, Seoul, Korea*
*** Professor, Graduate School of Smart Convergence, Kwangwoon University, Seoul, Korea*
E-mail : {seine1025, msj8086}@kw.ac.kr

Abstract

Genomic data plays a transformative role in medicine, biology, and forensic science, offering insights that drive advancements in clinical diagnosis, personalized medicine, and crime scene investigation. Despite its potential, the integration and analysis of diverse genomic datasets remain challenging due to compatibility issues and the specialized nature of existing tools. This paper presents the GenomeSync system, designed to overcome these limitations by utilizing the Hadoop framework for large-scale data handling and integration. GenomeSync enhances data accessibility and analysis through SQL-based search capabilities and machine learning techniques, facilitating the identification of genetic traits and the resolution of forensic cases. By pre-processing DNA profiles from crime scenes, the system calculates similarity scores to identify and aggregate related genomic data, enabling accurate prediction models and personalized treatment recommendations. GenomeSync offers greater flexibility and scalability, supporting complex analytical needs across industries. Its robust cloud-based infrastructure ensures data integrity and high performance, positioning GenomeSync as a crucial tool for reliable, data-driven decision-making in the genomic era.

Keywords:, *Big data, Distributed cloud system, Genomic data, Hadoop framework, SQL-based*

1. INTRODUCTION

Genomic data has significant applications in medicine, biology, and forensic science, transforming each field with advanced technology and insights. In the medical field, genomic data is essential for clinical diagnosis and personalized medicine. High-speed sequencing technologies allow for the identification of genetic variations associated with various diseases, enabling accurate diagnosis and personalized treatment plans [1]. For example, in cancer research, genomic data helps understand the genetic basis of tumors and predict patient prognosis, facilitating personalized treatments and improving survival rates [2]. In forensic science, genomic data is crucial for individual identification and crime scene investigation. Techniques such

Manuscript Received: July. 19. 2024 / Revised: July. 25. 2024 / Accepted: July. 31. 2024
Corresponding Author: msj8086@kw.ac.kr
Tel: 02-940-8283, Fax: 02-940-5443
Author's affiliation: Professor, Graduate School of Smart Convergence, Kwangwoon University, Seoul, Korea

as short tandem repeat (STR) analysis and genetic genealogy have revolutionized crime scene investigations, allowing for the resolution of cold cases and the identification of unknown individuals through comparison with genetic databases [3]. Additionally, genomic data is used in disaster victim identification by comparing DNA from remains with personal belongings or relatives' DNA [4].

However, most genomic data analysis tools developed so far are tailored to specific research or data types, which limits their flexibility for data integration and overall analysis [5]. One of the main challenges in the analysis of genomic data is the lack of compatibility between various formats and tools that hinders the integrated use of data. Most tools are optimized for specific research or genomes, making it difficult to analyze diverse data sources and formats [6]. To enhance the effectiveness of genomic data analysis, the development of modular analysis tools, standardization of data formats, sophisticated data preprocessing techniques, efficient management of computational resources, and the introduction of analysis techniques using machine learning and artificial intelligence are necessary [7].

This study proposes a system called 'GeneSync' to enable integrated analysis of genomic data. This system is based on the Hadoop framework, which can handle large-scale data, and integrates various genomic databases [8]. It provides data search and matching functions using SQL along with machine learning techniques. This approach offers a more effective way to utilize genomic data in medicine, biology, and forensic science.

2. PROPOSED SYSTEM

2.1 System Overview

The system proposed in this paper integrates and manages various types of genomic databases using the Hadoop framework. This approach resolves compatibility issues between existing data formats and enhances data accessibility. GenoSync allows users to access the database through SQL queries, providing flexible and intuitive data searching and matching capabilities. This is particularly advantageous for rapid data exploration and analysis in clinical diagnostics and forensic fields. By applying machine learning techniques to the integrated dataset, the system analyzes patterns in genomic data and builds predictive models. This can be used to analyze the genomic data of cancer patients, identify the genetic characteristics of tumors, and suggest personalized treatment methods. Additionally, it aids in identifying suspects and solving cold cases by matching DNA evidence collected at crime scenes with the database. In disaster victim identification processes, it improves the accuracy of identification by utilizing genomic data for verifying identities.

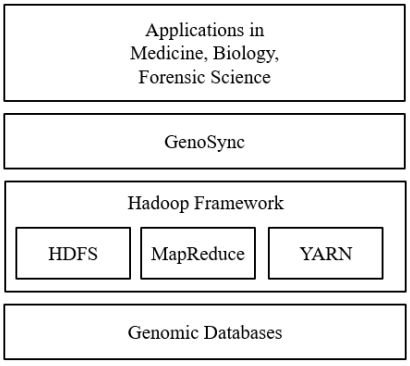

Figure 1. System overview

2.2 System Workflow

The core of this proposal system module is an SQL Cluster, which is composed of a central Master node and multiple subordinate Nodes. The Master node coordinates and distributes tasks within the cluster, assigning tasks to each Node for parallel processing. Each Node contains Executors, the units of execution for data, and each Executor interacts directly with the database to execute SQL queries. This setup allows data to be processed in parallel, maximizing efficiency and processing speed.

The Hadoop Layer plays a central role in data storage and management within this system. The NameNode manages the metadata of HDFS, storing the file system's structure and the location information of each file block. This enhances data access efficiency and allows centralized management of the cluster's data. The Secondary NameNode periodically creates checkpoints to reduce the risk of data loss in the NameNode and ensure system stability.

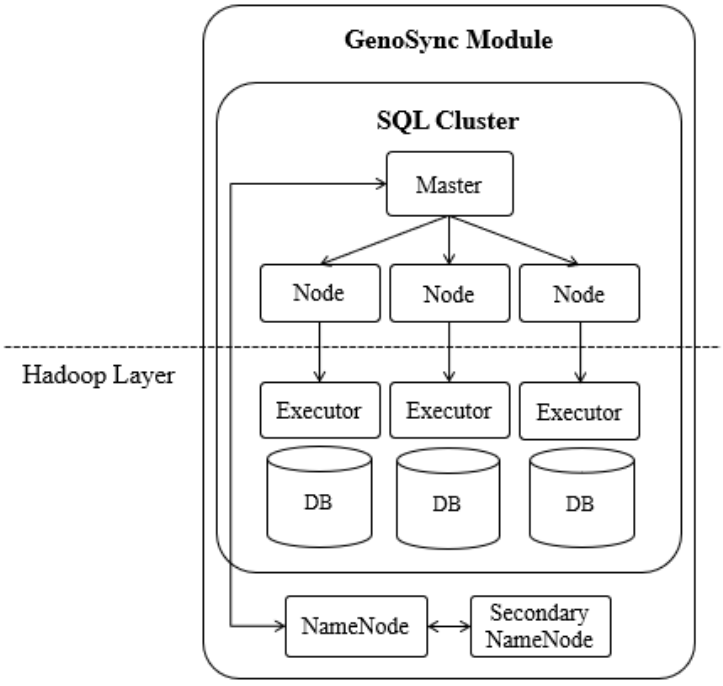

Figure 2. Data Workflow

The overall data processing flow is as follows:

1. The SQL Cluster's Master Node receives the list of file blocks and the location information of DataNodes where these blocks are stored from the HDFS NameNode.
2. Using this information from the NameNode, the Master selects the most accessible DataNodes.
3. The Master assigns tasks to each subordinate Node based on the DataNode information, instructing them to process specific blocks. Each Node utilizes its contained Executors to efficiently process the assigned blocks.
4. Executors execute SQL queries directly on the database. During this process, all Executors within the SQL Cluster work in parallel, enhancing the overall processing speed.

5. Upon completion of the SQL query processing, Executors report the completion and JOIN data back to the Master. The Master then schedules additional tasks if necessary or passes the data to the ML Cluster's Master Node as preparation data for machine learning.

Algorithm 1 describes the preprocessing steps to search for similar genomic data based on the DNA profile found at a crime scene and to predict genetic traits. Initially, the DNA data from the crime scene is loaded. Next, genomic data with similar allele values are searched and similarity scores are calculated. These scores are used to aggregate similar genomic data, filtering out those with a similarity score of 0 or less and sorting the remaining data in descending order of similarity. Finally, genomic data with the highest similarity scores are retrieved, providing preprocessed data suitable for training a machine learning model.

Table 1. SQL searching among raw genomic DB

Algorithm 1: Crime Scene DNA Similarity Scoring Algorithm

```
LOAD CrimeSceneData FROM database

FOR EACH crime_data IN CrimeSceneData DO
  LOAD GenomicData FROM database
    WHERE marker_name = crime_data.marker_name

  FOR EACH genomic_data IN GenomicData DO
    IF (genomic_data.allele_1 = crime_data.allele_1 OR genomic_data.allele_2 = crime_data.allele_2) OR
       (genomic_data.allele_1 = crime_data.allele_2 OR genomic_data.allele_2 = crime_data.allele_1) THEN
      SET match_score = 1
    ELSE
      SET match_score = 0
    END IF

    INSERT INTO SimilarGenomicData
      VALUES (genomic_data.id, genomic_data.marker_name, genomic_data.allele_1, genomic_data.allele_2,
match_score)
  END FOR
END FOR

GROUP BY SimilarGenomicData.id
  CALCULATE total_match_score = SUM(match_score)

FILTER SimilarGenomicData
  WHERE total_match_score > 0
  ORDER BY total_match_score DESC

SELECT id, total_match_score
FROM SimilarGenomicData
LIMIT 100
```

3. COMPARISON OF SYSTEMS

HIrisPlex, Crossbow, and 8-plex systems are specialized tools each designed for distinct analytical purposes in their respective fields. HIrisPlex is a system used in forensic science and anthropology to predict an individual's genotype and physical appearance characteristics. It is based on DNA analysis and infers biological traits such as hair color and skin color. HIrisPlex utilizes genotype markers and DNA sequencing to estimate the phenotype of specific individuals. Crossbow is a tool designed for processing and analyzing large-scale genomic data, particularly specialized in detecting variations in human genomic data. This system integrates genomic analysis tools like Bowtie and SOAPsnp with a Hadoop-based framework, enabling parallel

processing of large datasets. It is effective in aligning genomic data and detecting variations quickly and efficiently. 8-plex is used for protein analysis and employs LC-MS/MS (Liquid Chromatography-Mass Spectrometry) technology to perform identification and quantitative analysis of proteins. This system is specialized in protein identification, offering high sensitivity and specificity to effectively identify individual proteins within complex protein mixtures. The proposed system in this paper supports various data sources and formats, making it adaptable to diverse analytical needs. It can manage database operations and execute SQL queries simultaneously, providing high flexibility in complex data processing environments. Moreover, the cloud-based distributed computing environment efficiently distributes the data processing load, enabling rapid handling of large-scale datasets. These features are designed to address the growing volume and complexity of data, contributing to real-time analysis and decision support across various industries.

Table 2. Comparison of systems

	GenoSync	HIrisPlex [9]	Crossbow [10]	8-plex [11]
Main Purpose	Large-scale genomic data processing and SQL query execution	DNA analysis and ancestry prediction	Large-scale genetic data analysis	Protein analysis and identification
Components	GeneSync, Hadoop	PCR-based, DNA markers	MapReduce, Hadoop	LC-MS/MS
Data Processing Method	Distributed cloud computing and parallel processing	Biological analysis and prediction of ancestry	Distributed computing and parallel genetic analysis	Mass spectrometry-based data processing
Advantages	Scalability, reliability, versatility	Accurate prediction capabilities, biological context	High speed, large-scale data processing capabilities	High sensitivity and specificity
Disadvantages	Complexity of explanation and maintenance	Specific application environment	Complex setup and operation	Requires expensive equipment and data processing expertise
Versatility	Capable of processing various data types	Ancestry prediction	Supports various genetic data analysis	Optimized for protein identification

4. CONCLISION

The GenoSync system proposed in this paper offers flexibility and scalability to meet diverse data processing requirements and presents an effective solution for large-scale data analysis. While existing systems like HIrisPlex, Crossbow, and 8-plex are specialized for specific analytical purposes and possess unique strengths, GenoSync addresses the limitations of these systems by providing the capability to process datasets of various formats in parallel. By leveraging a distributed cloud computing environment, GenoSync maintains data integrity and delivers high performance. It efficiently supports the complex analytical needs across various industries. Therefore, the GenoSync system is positioned to overcome the limitations of current data analysis technologies and provide a more reliable analytical environment, contributing to data-driven decision-making in the future.

ACKNOWLEDGMENT

※ This paper was supported by the KwangWoon University Research Grant of 2024.

REFERENCES

[1] O. A. Montesinos-López et al., "A review of deep learning applications for genomic selection," BMC Genomics, vol. 22, no. 1. Springer Science and Business Media LLC, 06-Jan-2021.
DOI: https://doi.org/10.1186/s12864-020-07319-x

[2] F. S. Collins and H. Varmus, "A New Initiative on Precision Medicine," New England Journal of Medicine, vol. 372, no. 9. Massachusetts Medical Society, pp. 793–795, 26-Feb-2015.
DOI: https://doi.org/10.1056/NEJMp1500523

[3] J. M. Butler, "The future of forensic DNA analysis," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 370, no. 1674. The Royal Society, p. 20140252, 05-Aug-2015.
DOI: https://doi.org/10.1098/rstb.2014.0252

[4] M. Jiang, C. Bu, J. Zeng, Z. Du, and J. Xiao, "Applications and challenges of high performance computing in genomics," CCF Transactions on High Performance Computing, vol. 3, no. 4. Springer Science and Business Media LLC, pp. 344–352, 19-Oct-2021.
DOI: https://doi.org/10.1007/s42514-021-00081-w

[5] S. Das, C. J. McClain, and S. N. Rai, "Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges," Entropy, vol. 22, no. 4. MDPI AG, p. 427, 10-Apr-2020
DOI: https://doi.org/10.3390/e22040427

[6] V. Marx, "The big challenges of big data," Nature, vol. 498, no. 7453. Springer Science and Business Media LLC, pp. 255–260, 12-Jun-2013.
DOI: https://doi.org/10.1038/498255a

[7] Z. D. Stephens et al., "Big Data: Astronomical or Genomical?," PLOS Biology, vol. 13, no. 7. Public Library of Science (PLoS), p. e1002195, 07-Jul-2015.
DOI: https://doi.org/10.1371/journal.pbio.1002195

[8] S. Hedayati, N. Maleki, T. Olsson, F. Ahlgren, M. Seyednezhad, and K. Berahmand, "MapReduce scheduling algorithms in Hadoop: a systematic study," Journal of Cloud Computing, vol. 12, no. 1. Springer Science and Business Media LLC, 10-Oct-2023.
DOI: https://doi.org/10.1186/s13677-023-00520-9

[9] S. Walsh et al., "The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA," Forensic Science International: Genetics, vol. 7, no. 1. Elsevier BV, pp. 98–115, Jan-2013.
DOI: https://doi.org/10.1016/j.fsigen.2012.07.005

[10] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," Genome Biology, vol. 10, no. 11. Springer Science and Business Media LLC, p. R134, 2009.
DOI: https://doi.org/10.1186/gb-2009-10-11-r134

[11] K. L. Hart, S. L. Kimura, V. Mushailov, Z. M. Budimlija, M. Prinz, and E. Wurmbach, "Improved eye- and skin-color prediction based on 8 SNPs," Croatian Medical Journal, vol. 54, no. 3. Croatian Medical Journals, pp. 248–256, Jun-2013.
DOI: https://doi.org/10.3325/cmj.2013.54.248