Academic Journal of Science and Technology
Vol. 2 (2019)
https://academicpinnacle.com/index.php/ajst

Anomaly Detection Systems for Protecting Genomic Databases from Cyber Attacks

Aravind Kumar Kalusivalingam
Northeastern University, Boston, USA
Corresponding: karavindkumar1993@gmail.com

Abstract

Anomaly Detection Systems play a pivotal role in safeguarding genomic databases from cyber-attacks by identifying irregular patterns and potential threats within vast amounts of genomic data. These systems leverage advanced algorithms to detect deviations from normal genomic behavior, enabling early detection of unauthorized access, data breaches, or malicious activities. By analyzing diverse genomic data sets, including DNA sequences, gene expressions, and variations, anomaly detection systems can identify anomalies that might indicate cyber threats such as data tampering, injection attacks, or unauthorized data access. Integrating machine learning and statistical techniques, these systems continually adapt to evolving attack strategies, providing proactive defense measures to protect sensitive genomic information from exploitation and ensuring the integrity and confidentiality of genetic data for research and medical purposes.

Keywords: Anomaly Detection Systems, Genomic Databases, Cyber Attacks, Genetic Data, DNA sequences

1. Introduction

Genomic databases play a crucial role in modern biological research and medical advancements, containing vast amounts of genetic information critical for understanding diseases, developing treatments, and personalized healthcare. However, the increasing reliance on digital platforms to store and share genomic data has also exposed these databases to various cyber threats. Cyber-attacks targeting genomic databases pose significant risks, including data breaches, unauthorized access, and potential misuse of sensitive genetic information [1]. Thus, there is an urgent need for robust security measures to protect genomic databases from such threats. Anomaly Detection Systems have emerged as a key component in the defense against cyber-attacks on genomic databases. These systems employ advanced algorithms and techniques to

monitor genomic data for irregular patterns or deviations from expected behavior. By analyzing diverse genomic datasets, including DNA sequences, gene expressions, and variations, anomaly detection systems can detect anomalies that may indicate potential cyber threats, such as data tampering, injection attacks, or unauthorized access [2]. Through proactive detection and response, these systems help mitigate risks and ensure the integrity, confidentiality, and availability of genetic information stored in databases. This paper provides an in-depth exploration of the role of anomaly detection systems in protecting genomic databases from cyber-attacks. We will discuss various anomaly detection techniques, including machine learning-based approaches and statistical methods, tailored to the unique characteristics of genomic data[3]. Additionally, real-world case studies and examples will be examined to illustrate the effectiveness of anomaly detection systems in mitigating cyber threats to genomic databases. Furthermore, challenges, evaluation metrics, and future directions in this field will be addressed to provide insights into advancing genomic data security. Genomic databases serve as repositories of genetic information crucial for various fields, including biomedical research, personalized medicine, and genetic diagnostics. These databases store vast amounts of data ranging from DNA sequences to gene expressions and variations, facilitating insights into human health, disease susceptibility, and treatment responses. However, the digitization of genomic data and the increasing interconnectedness of systems have heightened concerns about cybersecurity risks. Genomic databases are vulnerable to cyber-attacks due to their sensitive nature and the potential consequences of unauthorized access or manipulation. The theft, tampering, or exposure of genetic information can lead to privacy breaches, identity theft, and even discrimination based on genetic predispositions [4]. Moreover, the interconnectedness of genomic databases with healthcare systems and research networks amplifies the risk, making them lucrative targets for cybercriminals.

Protecting genomic data is of paramount importance to safeguard individual privacy, maintain data integrity, and uphold ethical standards in genomic research and healthcare [5]. Ensuring the confidentiality of genetic information is crucial for maintaining trust between patients, researchers, and healthcare providers. Additionally, protecting genomic databases is essential for preserving the integrity of research findings, preventing data tampering, and ensuring that genetic data is used ethically and responsibly. Furthermore, the potential misuse of genetic information for discriminatory purposes underscores the urgent need for robust cybersecurity measures to safeguard genomic databases from cyber threats [6]. Anomaly detection systems play a vital role in enhancing

the cybersecurity posture of genomic databases by identifying and mitigating potential threats and anomalies. These systems employ sophisticated algorithms to monitor genomic data for abnormal patterns, deviations, or suspicious activities that may indicate cyber-attacks or unauthorized access. By continuously analyzing genomic data and detecting anomalies in real-time, anomaly detection systems provide early warning signs of potential security breaches, enabling prompt responses and mitigating potential damage. Through proactive detection and response mechanisms, anomaly detection systems contribute to enhancing the overall security and resilience of genomic databases against evolving cyber threats [7]. Genomic databases encompass a wide array of repositories that store genetic information critical for biomedical research, clinical applications, and personalized medicine. These databases can be categorized into different types based on their content and purpose. Public databases, such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), host vast collections of genetic data from various organisms, freely accessible to researchers worldwide. Private databases, often maintained by research institutions or pharmaceutical companies, contain proprietary genetic information used for specific research or commercial purposes. Clinical databases store patient genomic data, aiding in diagnosis, treatment selection, and disease monitoring. The structure of genomic databases varies but typically includes DNA sequences, gene annotations, genetic variations, and associated metadata. Applications of genomic databases range from basic research on gene functions, evolutionary studies, and pharmacogenomics, to clinical applications like disease diagnosis, drug development, and personalized medicine [8].

Despite their critical importance, genomic databases face various cyber threats that endanger the confidentiality, integrity, and availability of genetic data. Common cyber threats targeting genomic databases include unauthorized access, where malicious actors gain entry to databases to steal sensitive genetic information. Data breaches, either through external hacking or internal leaks, can lead to the exposure of confidential genetic data, compromising patient privacy and confidentiality. Injection attacks, such as SQL injection or command injection, can manipulate database queries or commands to extract or alter genetic data unlawfully. Additionally, ransomware attacks can encrypt genomic databases, rendering them inaccessible until a ransom is paid. Furthermore, malicious tampering of genetic data can lead to incorrect research findings, affecting scientific integrity and potentially impacting clinical decisions [9]. Unauthorized access and data breaches pose significant risks to

genomic databases and the individuals whose genetic information they contain. Beyond the immediate privacy concerns, unauthorized access can lead to identity theft, discrimination based on genetic predispositions, or misuse of genetic data for illicit purposes. Furthermore, compromised genetic data can erode public trust in genomic research and healthcare systems. Protecting against these risks requires robust security measures, including encryption, access controls, regular security audits, and, importantly, advanced anomaly detection systems to detect and mitigate cyber threats in real time.

2. Anomaly Detection Techniques

Anomaly detection systems are critical components of cybersecurity strategies aimed at protecting genomic databases from cyber threats. These systems are designed to identify unusual patterns, deviations, or outliers in genomic data that may indicate potential cyber-attacks, data breaches, or unauthorized access. Anomaly detection relies on the principle that normal behavior in genomic data follows certain patterns, and deviations from these patterns may signal malicious activities. By continuously monitoring genomic databases, anomaly detection systems can provide early warnings of potential security breaches, allowing for timely responses to mitigate risks and protect sensitive genetic information[10]. Anomalies in genomic data can manifest in various forms, each posing unique challenges to anomaly detection systems. Common types of anomalies include **Structural Anomalies:** Deviations in the structure of genomic data, such as missing values, duplications, or inconsistencies in DNA sequences or gene annotations.

Temporal Anomalies: Changes in gene expressions or variations over time that deviate from expected patterns, indicating potential biological changes or experimental errors.

Statistical Anomalies: Outliers or extreme values in genomic data that do not conform to the statistical distribution of normal data points.

Semantic Anomalies: Instances where genetic data contains unexpected patterns or combinations that do not align with known biological principles. Understanding these different types of anomalies is crucial for developing effective anomaly detection systems tailored to the unique characteristics of genomic data [11].

As shown in Figure 1, the process begins by inferring the sex from the sex chromosomes. Next, metadata is used to determine the person's state and age. The person's surname is then recovered using publicly available genetic-genealogy databases such as Ysearch. Public record search engines like

PeopleFinders.com generate a list of potential individuals. Finally, social engineering or pedigree analysis is employed to triangulate the person's identity. This pioneering study underscores the importance of protecting the genetic privacy of data originators. The figure illustrates a possible route for identity tracing using anonymous genomic data. Initially, DNA is extracted from an anonymized sample and sequenced to obtain a genetic profile [12]. This profile is then compared to public and private genetic databases to find potential relatives. By identifying genetic matches, investigators can construct partial family trees. Next, metadata, such as age and location, is integrated to narrow down the search. Public records and genealogy databases are utilized to infer potential surnames and identify individuals. Social engineering techniques and pedigree analysis further refine the list of candidates. This comprehensive approach allows for the triangulation of an individual's identity. The figure highlights the importance of genetic privacy and the potential risks associated with the misuse of genomic data [13].

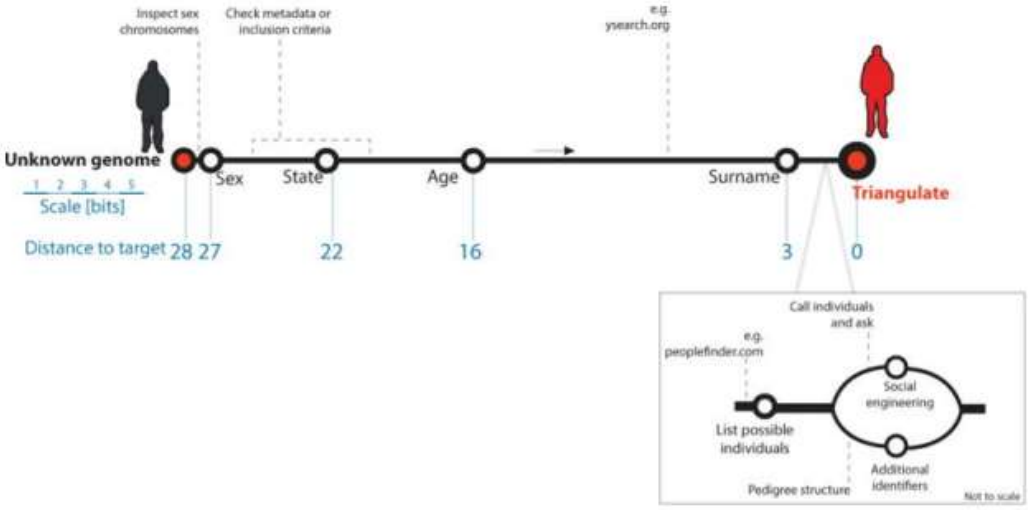

Figure 1: A possible route for identity tracing using anonymous genomic data.

Machine learning techniques have shown promising results in anomaly detection for genomic databases due to their ability to handle complex data patterns and adapt to evolving threats. Supervised learning algorithms, such as Support Vector Machines (SVM) or Random Forests, can be trained on labeled genomic data to classify normal and anomalous instances. Unsupervised learning algorithms, including clustering methods like k-means or density-based approaches like DBSCAN, can detect anomalies in genomic data without the need for labeled examples. Deep learning techniques, such as autoencoders, are also increasingly used for anomaly detection in genomic sequences, leveraging neural networks to reconstruct normal patterns and

detect deviations. Statistical methods form the foundation of many anomaly detection systems for genomic databases[14]. These approaches utilize statistical models, hypothesis testing, and probability distributions to identify anomalies in genetic data. Techniques like z-score analysis, where anomalies are detected based on deviations from the mean or median, are commonly used. Bayesian methods, such as Gaussian mixture models, estimate the probability distribution of normal genomic data and flag instances with low probability as anomalies. Additionally, time-series analysis methods can detect temporal anomalies in gene expression data by modeling trends and seasonality. Integrating statistical approaches with machine learning techniques can enhance the accuracy and efficiency of anomaly detection systems for protecting genomic databases [15].

3. Implementation and Future Directions

Implementing anomaly detection systems for genomic databases faces several challenges, stemming from the complexity and scale of genomic data: Data Complexity: Genomic data is highly complex, comprising sequences of nucleotides, gene expressions, and variations. Detecting anomalies amidst this complexity requires sophisticated algorithms capable of capturing subtle deviations. Scalability: Genomic databases are vast and continuously growing, posing scalability challenges for anomaly detection systems. These systems must efficiently process and analyze large volumes of data generated by high-throughput sequencing technologies. Data Quality: Genomic data may contain errors, artifacts, or noise introduced during sequencing or data processing steps. Ensuring data quality is crucial for accurate anomaly detection, requiring preprocessing and quality control measures.

Real-time Detection: Timely detection of anomalies is essential to prevent data breaches or cyber-attacks. Implementing real-time anomaly detection systems capable of monitoring genomic data streams continuously is challenging but necessary for effective cybersecurity. Interpretability: Anomaly detection algorithms should provide interpretable results to understand why certain instances are flagged as anomalies. In genomic research, transparency and reproducibility are crucial, necessitating interpretable anomaly detection methods. Ethical Considerations and Privacy Concerns: Protecting genomic data raises significant ethical and privacy concerns, necessitating careful consideration and safeguards: Data Ownership: Clarifying data ownership and control is important, especially in collaborative research or commercial settings. Establishing clear policies on data ownership and usage rights can prevent unauthorized access or exploitation. Data De-identification:

Implementing robust de-identification techniques to remove personally identifiable information while preserving data utility is critical for privacy protection. Techniques like anonymization or differential privacy can help mitigate privacy risks. Risk of Discrimination: Genetic information may be used for discriminatory purposes in employment, insurance, or social contexts. Safeguarding against such risks and implementing antidiscrimination policies is imperative to protect individuals' rights and prevent harm.

Privacy-Preserving Techniques: Develop advanced cryptographic techniques like homomorphic encryption or secure multiparty computation to perform analysis on encrypted genomic data without compromising privacy. Blockchain Technology: Exploring the use of blockchain for secure and decentralized storage of genomic data, ensuring data integrity, traceability, and auditability. Explainable AI: Advancing explainable AI methods to improve transparency and interpretability of anomaly detection algorithms, enabling stakeholders to understand and trust the results. Collaborative Security Frameworks: Establishing collaborative security frameworks and standards to facilitate information sharing, threat intelligence, and best practices for genomic data protection. Addressing these challenges and embracing emerging technologies and research directions will be crucial for ensuring the security, privacy, and ethical use of genomic data in the future.

4. Conclusion

In conclusion, Anomaly Detection Systems represent a critical defense mechanism for protecting genomic databases from cyber threats. By leveraging advanced algorithms and techniques, these systems can effectively monitor genomic data, identify irregular patterns, and detect potential cyber-attacks in real time. Despite challenges such as data complexity, scalability, and ensuring interpretability, anomaly detection systems offer a proactive approach to mitigating risks associated with unauthorized access, data breaches, and tampering. However, addressing ethical considerations and privacy concerns surrounding genomic data is equally crucial to maintaining public trust and ensuring the responsible use of genetic information. Looking ahead, embracing emerging technologies like privacy-preserving techniques, blockchain, and explainable AI, while fostering collaborative security frameworks, will be essential for advancing genomic data security and safeguarding the integrity, confidentiality, and ethical use of genetic information for research and medical purposes.

Reference

[1] J. Caswell *et al.*, "Defending our public biological databases as a global critical infrastructure," *Frontiers in bioengineering and biotechnology,* vol. 7, p. 58, 2019.

[2] P. Wlodarczak, "Cyber Immunity: A bio-inspired cyber defense system," in *Bioinformatics and Biomedical Engineering: 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26–28, 2017, Proceedings, Part II 5*, 2017: Springer, pp. 199-208.

[3] P. A. A. Resende and A. C. Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling," *Security and Privacy,* vol. 1, no. 4, p. e36, 2018.

[4] B. A. Vinatzer, L. S. Heath, H. M. Almohri, M. J. Stulberg, C. Lowe, and S. Li, "Cybersecurity challenges of pathogen genome databases," *Frontiers in bioengineering and biotechnology,* vol. 7, p. 106, 2019.

[5] P. M. Ney, "Securing the future of biotechnology: A study of emerging bio-cyber security threats to DNA-information systems," 2019.

[6] A. B. Carter, "Considerations for genomic data privacy and security when working in the cloud," *The Journal of Molecular Diagnostics,* vol. 21, no. 4, pp. 542-552, 2019.

[7] J. L. Raisaro *et al.*, "Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy," *IEEE/ACM Transactions on computational biology and bioinformatics,* vol. 15, no. 5, pp. 1413-1426, 2018.

[8] D. Grishin, K. Obbad, and G. M. Church, "Data privacy in the age of personal genomics," *Nature Biotechnology,* vol. 37, no. 10, pp. 1115-1117, 2019.

[9] S. Wang *et al.*, "Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States," *Annals of the New York Academy of Sciences,* vol. 1387, no. 1, pp. 73-83, 2017.

[10] R. Pizzolante, A. Castiglione, B. Carpentieri, A. De Santis, F. Palmieri, and A. Castiglione, "On the protection of consumer genomic data in the Internet of Living Things," *Computers & Security,* vol. 74, pp. 384-400, 2018.

[11] R. Ghasemi, M. M. Al Aziz, N. Mohammed, M. H. Dehkordi, and X. Jiang, "Private and efficient query processing on outsourced genomic databases," *IEEE Journal of biomedical and health informatics,* vol. 21, no. 5, pp. 1466-1472, 2016.

[12] F. K. Dankar, A. Ptitsyn, and S. K. Dankar, "The development of large-scale de-identified biomedical databases in the age of genomics—principles and challenges," *Human genomics,* vol. 12, pp. 1-15, 2018.

[13] W. A. Valdivia-Granda, "Big data and artificial intelligence for biodefense: A genomic-based approach for averting technological surprise," *Defense Against Biological Attacks: Volume I,* pp. 317-327, 2019.

[14] D. Deuber *et al.*, "My genome belongs to me: controlling third party computation on genomic data," *Proceedings on Privacy Enhancing Technologies,* 2019.

[15] J. N. Tuazon, "Privacy-Preserving Genomic Disease Susceptibility Testing Using Secure Multiparty Computation," 2016.