MZ Journals

Journal of Engineering and Technology

Secure Multi-Party Computation in Genomics: Protecting Privacy While Enabling Research Collaboration

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

Abstract

Secure Multi-Party Computation (SMPC) in genomics presents a groundbreaking approach to preserving privacy while fostering collaboration in genomic research. By leveraging cryptographic techniques, SMPC allows multiple parties to jointly analyze genomic data without sharing sensitive information. This innovative method ensures that individual genetic data remains encrypted and confidential, protecting participants' privacy rights. Moreover, it enables researchers to collaborate across institutions or countries, pooling data resources for more comprehensive analyses while adhering to strict privacy regulations. Thus, abstract SMPC not only empowers research advancements in genomics but also ensures ethical data handling and privacy protection, crucial for maintaining public trust in scientific endeavors.

keywords: Secure Multi-Party Computation (SMPC), genomics, privacy protection

Introduction

Genomic research collaboration is essential for unlocking the complexities of the human genome and accelerating scientific breakthroughs in healthcare. The vast amount of genomic data available holds the key to understanding diseases, identifying genetic predispositions, and developing targeted therapies. However, genomic research often requires large-scale data analysis involving diverse datasets from multiple sources [1]. Collaborative efforts among researchers, institutions, and even countries are crucial to leveraging this wealth of data effectively. To enable such collaboration, while ensuring data privacy and security, Secure Multi-Party Computation (SMPC) emerges as a revolutionary approach. SMPC allows multiple parties to jointly analyze data without sharing the raw data itself, thus overcoming the privacy concerns associated with traditional data-sharing methods. This cryptographic technique enables researchers to perform computations on encrypted data while preserving the privacy of individual contributors. As genomic data often contains highly sensitive information, SMPC provides a powerful tool to facilitate collaborative research while protecting participants' privacy. Collaboration in genomic research is vital for several reasons. Firstly, diseases

often have complex genetic origins that can only be fully understood through the analysis of large-scale genomic datasets. By pooling together diverse datasets from different populations, researchers can gain insights into genetic variations, disease prevalence, and treatment responses across various demographics. Collaborative efforts allow for more comprehensive analyses and enhance the generalizability of research findings. Furthermore, collaborative genomic research fosters interdisciplinary collaboration, bringing together experts from various fields such as genetics, bioinformatics, medicine, and computer science. This multidisciplinary approach is crucial for tackling complex scientific questions and translating research findings into clinical applications. Moreover, collaboration facilitates data sharing, reduces duplication of efforts, and accelerates the pace of discovery, ultimately leading to improved patient outcomes and healthcare practices. Therefore, fostering collaboration in genomic research is essential for maximizing the potential of genomic data to revolutionize healthcare and personalized medicine [2].

Genomic data is inherently sensitive, containing information about an individual's genetic makeup, predispositions to diseases, and familial relationships. As such, the privacy concerns surrounding genomic data are substantial. Unauthorized access or misuse of this data can lead to breaches of privacy, discrimination, and other ethical concerns. Individuals may be reluctant to participate in research or share their genomic data if they feel their privacy is at risk. Therefore, protecting the privacy of genomic data is paramount to maintaining trust in research endeavors and ensuring ethical data handling practices. Sharing genomic data for collaborative research poses several challenges. Firstly, there are legal and regulatory barriers, including privacy laws such as GDPR and HIPAA, which impose restrictions on the sharing of sensitive personal data. Compliance with these regulations adds complexity and may limit the ability to share data across institutions or jurisdictions. Additionally, technical challenges such as data interoperability, standardization, and data integration arise when combining datasets from different sources. Moreover, concerns about data security, including the risk of data breaches and unauthorized access, further hinder collaborative efforts in genomic research [3]. Traditional methods for sharing genomic data often involve centralized databases or data repositories where researchers upload their data for others to access. While these methods facilitate data sharing to some extent, they come with significant limitations. Firstly, centralized repositories may raise privacy concerns as they require sharing raw genomic data, increasing the risk of re-identification or unauthorized access. Additionally, data-sharing agreements and consent processes can be cumbersome and time-consuming, hindering the efficient exchange of data [4]. Furthermore, centralized approaches may not be suitable for collaborative research involving sensitive or proprietary data due to concerns about data ownership and control. Overall, traditional methods of data sharing often struggle to balance the need for collaboration with the imperative of protecting individual privacy and data security.

Applications of SMPC in Genomics

Secure Multi-Party Computation (SMPC) can be applied in various genomic research scenarios to enable collaborative studies while ensuring data privacy. One key application is in multi-institutional studies where different research groups hold genomic data that, when combined, can provide more comprehensive insights into genetic disorders [5]. For instance, researchers from various hospitals can collaboratively analyze genomic datasets to identify genetic markers for cancer without directly sharing patient data, thus protecting patient privacy. Another scenario involves pharmaceutical companies and academic researchers jointly analyzing genomic data to discover new drug targets. SMPC allows these parties to perform secure computations on combined datasets, ensuring that proprietary information and sensitive patient data remain confidential. Additionally, SMPC can be employed in large-scale genome-wide association studies (GWAS) where data from diverse populations are needed to understand the genetic basis of complex traits and diseases [6]. By using SMPC, researchers can aggregate and analyze data from different cohorts without compromising the privacy of individual participants. A notable case study demonstrating the successful implementation of SMPC in genomic research is the collaboration between Boston University, MIT, and Harvard for analyzing genomic data related to Alzheimer's disease. In this project, SMPC protocols were used to perform joint computations on data from different institutions. The results of the analysis contributed to identifying potential genetic factors associated with Alzheimer's, all while ensuring that the data remained confidential and secure [7].

Another example is the iDASH (integrating Data for Analysis, Anonymization, and SHaring) Secure Genome Analysis Competition, which has promoted the development of SMPC techniques for genomic research. In one of its challenges, participants developed SMPC-based solutions for securely comparing genetic variants across multiple datasets [8]. These solutions demonstrated that SMPC could effectively enable collaborative research without compromising data privacy, setting a precedent for future genomic studies. SMPC stands out among other privacy-preserving techniques in genomics due to its ability to perform computations on encrypted data. Traditional methods such as data anonymization and de-identification aim to remove personally identifiable information (PII) from datasets, but they are often vulnerable to re-identification attacks, especially when combined with other datasets. In contrast, SMPC ensures that raw data remains encrypted and never fully accessible to any party, significantly reducing the risk of privacy breaches. Differential privacy is another technique used to protect individual privacy by adding noise to the data or the results of queries [9]. While effective in many scenarios, differential privacy can introduce inaccuracies and may not be suitable for all types of genomic analyses. SMPC, on the other hand, enables precise computations without altering the data, providing more accurate results for research purposes. Homomorphic encryption is a technique that allows computations on encrypted data without decrypting it. While similar to SMPC in preserving data privacy, homomorphic encryption can be

computationally intensive and less practical for large-scale genomic analyses. SMPC often offers a more feasible approach by distributing the computational workload among multiple parties. In summary, SMPC provides a robust solution for enabling collaborative genomic research while ensuring data privacy. Its ability to facilitate secure computations on encrypted data makes it a superior choice compared to traditional privacy-preserving techniques, paving the way for more.

Privacy Protection Mechanisms in SMPC

Secure Multi-Party Computation (SMPC) employs a variety of cryptographic techniques to ensure the privacy and security of data during collaborative computations [10]. The core principle of SMPC is to enable multiple parties to jointly compute a function over their inputs while keeping those inputs private. This is achieved through the use of cryptographic protocols such as secret sharing, oblivious transfer, and garbled circuits. In secret sharing, data is divided into shares that are distributed to different parties. Each party holds a portion of the data, and only when these shares are combined can the original data be reconstructed. During computation, operations are performed on the shares without revealing the actual data to any party. For instance, in an additive secret-sharing scheme, each input is split into several random shares, and the sum of these shares equals the original input [11]. Computations are then performed on these shares, ensuring that the inputs remain confidential. The oblivious transfer is another mechanism used in SMPC, allowing one party to send multiple pieces of information to another party, where the receiver obtains only one piece without the sender knowing which piece was chosen. This is crucial for ensuring that parties do not learn more than they need to during the computation process. Garbled circuits involve converting the computation into a circuit of gates, where each gate operation is encrypted. Parties can evaluate these encrypted gates without learning the underlying data. This method is particularly useful for complex computations that can be represented as Boolean circuits.

To facilitate secure data sharing and analysis, SMPC combines these cryptographic techniques with secure communication protocols. Data owners first encrypt their data or generate shares before sharing them with other parties involved in the computation [12]. These shares or encrypted data pieces are then processed using SMPC protocols, ensuring that each party only sees the processed results without accessing the original data. One common technique is to use a trusted setup phase where parties agree on cryptographic keys or parameters. Subsequent computations leverage these keys to maintain privacy. Additionally, zero-knowledge proofs can be employed to verify computations' correctness without revealing the actual data, ensuring integrity and trustworthiness in the results.

Figure 1 illustrates the DNA sample lifecycle begins with sample collection, where biological material such as blood or saliva is obtained from individuals. Next, the samples undergo DNA extraction to isolate the genetic material. Once extracted, the DNA is sequenced using high-throughput sequencing technologies to generate raw sequencing

data. Bioinformatics tools come into play during the data preprocessing stage, where sequences are quality-checked, trimmed, and aligned to a reference genome [13]. Following alignment, variant calling tools identify genetic variations such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels). The next step involves variant annotation, where bioinformatics tools annotate identified variants with information about their genomic location, functional impact, and population frequencies. Subsequently, downstream analyses are performed, including pathway analysis, gene expression profiling, and variant prioritization. Throughout the lifecycle, bioinformatics tools play a crucial role in interpreting and analyzing genomic data to extract meaningful insights. They aid in understanding the genetic basis of diseases, identifying biomarkers, and guiding personalized treatment strategies. Finally, the lifecycle concludes with data sharing and publication, contributing to the advancement of genomic research and personalized medicine.

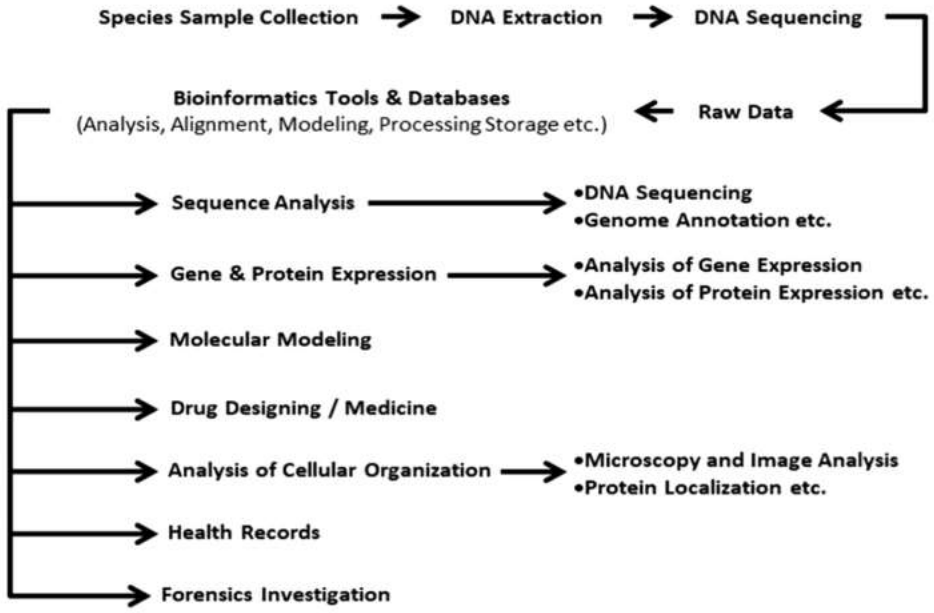

Figure 1: DNA sample lifecycle and the position of bioinformatics tools.

Despite its advantages, SMPC faces several limitations and challenges, particularly in the context of genomic research. One significant challenge is the computational overhead associated with cryptographic operations. SMPC protocols often require extensive computational resources and time, making them less efficient than traditional methods, especially for large-scale genomic datasets. Another limitation is the complexity of protocol implementation. Designing and deploying SMPC protocols requires specialized knowledge in cryptography and secure computation, which may not be readily available in many research institutions. Additionally, the communication overhead can be substantial, as parties need to exchange multiple encrypted messages or shares during the computation process[14]. Scalability is also a concern, as SMPC protocols may struggle to

handle the vast amount of data generated in genomic studies. While advances in cryptographic techniques and computing power are gradually addressing these issues, practical implementation for large-scale projects remains challenging. Moreover, the legal and ethical landscape of genomic data sharing can complicate the use of SMPC. Compliance with regulations such as GDPR and HIPAA adds layers of complexity to implementing secure data-sharing protocols. Ensuring that SMPC implementations adhere to these regulations while maintaining efficiency and usability is an ongoing challenge. In conclusion, while SMPC offers robust privacy protection mechanisms for collaborative genomic research, its practical application is hindered by computational, implementation, and regulatory challenges. Continued research and development are needed to optimize these protocols and make them more accessible and scalable for the broader scientific community.

Implementing Secure Multi-Party Computation (SMPC) in genomics faces several challenges. Firstly, computational complexity remains a significant hurdle, especially for large-scale genomic datasets. The computational overhead of cryptographic operations can be prohibitive, requiring substantial processing power and time. Additionally, ensuring interoperability and standardization across different SMPC implementations is challenging, hindering seamless collaboration. Moreover, navigating regulatory frameworks such as GDPR and HIPAA adds complexity, as compliance requirements must be carefully addressed to ensure legal and ethical data handling. Future advancements in SMPC techniques aim to address these challenges and improve efficiency and usability. One direction involves optimizing cryptographic protocols to reduce computational overhead, making SMPC more practical for large-scale genomic analyses[15]. Additionally, research efforts are focused on developing user-friendly software libraries and tools to streamline SMPC implementation and deployment. Moreover, advancements in hardware acceleration, such as secure enclaves and specialized processors, may significantly improve the performance of SMPC protocols. Continued collaboration between researchers, industry, and policymakers is essential to drive innovation and standardization in SMPC techniques for genomics.

In addition to SMPC, emerging trends and technologies offer promising avenues for enhancing privacy and collaboration in genomic research. Differential privacy techniques are gaining traction for protecting individual privacy while allowing for meaningful analysis of genomic data. Blockchain technology holds the potential for creating decentralized and immutable data-sharing platforms, ensuring transparency and integrity. Furthermore, federated learning approaches enable collaborative model training across distributed datasets without sharing raw data. These emerging technologies, combined with SMPC, offer a diverse toolkit for safeguarding privacy and fostering collaboration in genomic research, paving the way for transformative discoveries in personalized medicine and healthcare.

Conclusion

Secure Multi-Party Computation (SMPC) represents a transformative approach for balancing the dual imperatives of privacy protection and collaborative research in genomics. By enabling secure computations on encrypted data, SMPC ensures that sensitive genomic information remains confidential while allowing researchers to derive valuable insights from combined datasets. This technology overcomes the limitations of traditional data-sharing methods, which often compromise privacy and security. Despite challenges such as computational overhead and regulatory compliance, ongoing advancements in cryptographic techniques and computing capabilities are enhancing the feasibility and efficiency of SMPC. As SMPC continues to evolve, it promises to unlock new potentials in genomic research, fostering greater collaboration and accelerating scientific discoveries while maintaining the highest standards of data privacy.

Reference

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

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

[3] C. Zhao *et al.*, "Secure multi-party computation: theory, practice, and applications," *Information Sciences,* vol. 476, pp. 357-372, 2019.

[4] H. Tang *et al.*, "Protecting genomic data analytics in the cloud: state of the art and opportunities," *BMC Medical Genomics,* vol. 9, pp. 1-9, 2016.

[5] D. Evans, V. Kolesnikov, and M. Rosulek, "A pragmatic introduction to secure multi-party computation," *Foundations and Trends® in Privacy and Security,* vol. 2, no. 2-3, pp. 70-246, 2018.

[6] C.-A. Azencott, "Machine learning and genomics: precision medicine versus patient privacy," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 376, no. 2128, p. 20170350, 2018.

[7] D. Demmler, K. Hamacher, T. Schneider, and S. Stammler, "Privacy-preserving whole-genome variant queries," in *Cryptology and Network Security: 16th International Conference, CANS 2017, Hong Kong, China, November 30—December 2, 2017, Revised Selected Papers 16*, 2018: Springer, pp. 71-92.

[8] O. Tkachenko, C. Weinert, T. Schneider, and K. Hamacher, "Large-scale privacy-preserving statistical computations for distributed genome-wide association studies," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 221-235.

[9] F. Chen *et al.*, "Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions," *Bioinformatics,* vol. 33, no. 6, pp. 871-878, 2017.

[10] M. Blanton and F. Bayatbabolghani, "Efficient server-aided secure two-party function evaluation with applications to genomic computation," *Proceedings on Privacy Enhancing Technologies,* 2016.
[11] J. S. Sousa *et al.*, "Efficient and secure outsourcing of genomic data storage," *BMC Medical Genomics,* vol. 10, pp. 15-28, 2017.
[12] M. M. A. Aziz *et al.*, "Privacy-preserving techniques of genomic data—a survey," *Briefings in bioinformatics,* vol. 20, no. 3, pp. 887-895, 2019.
[13] M. S. R. Mahdi, M. M. Al Aziz, D. Alhadidi, and N. Mohammed, "Secure similar patients query on encrypted genomic data," *IEEE Journal of biomedical and health informatics,* vol. 23, no. 6, pp. 2611-2618, 2018.
[14] Z. Huang, "On Secure Cloud Computing for Genomic Data: From Storage to Analysis," EPFL, 2018.
[15] M. Hosseini, D. Pratas, and A. J. Pinho, "Cryfa: a secure encryption tool for genomic data," *Bioinformatics,* vol. 35, no. 1, pp. 146-148, 2019.