Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230350

A Reliable and Secure Method for Sharing Genomic Data

Dinis. B. CRUZ[a], João R. ALMEIDA[a,b,1], Jorge M. SILVA[a] and José L. OLIVEIRA[a]
[a] *DETI/IEETA, LASI, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*

Abstract. Genomics has significantly impacted the field of medicine, with advances in DNA sequencing leading to personalized medicine and a deeper understanding of the genomic basis of various diseases. The ability to share genomic data is crucial for advancing this field and developing new approaches to understanding the genome. However, the sensitive nature of this data requires secure methods for protecting it during storage and transfer. In this paper, we present a new tool for the secure encryption and decryption of FASTA files without sharing a common secret and with a reduced number of shared keys between the pairs. Our proposal combines symmetric and asymmetric encryption techniques, including the AES (Advanced Encryption Standard) cypher and RSA (Rivest–Shamir–Adleman). The tool is fast, reliable, and secure, outperforming existing tools in terms of security and ease of use. This makes it a valuable solution for the secure sharing and use of sensitive genomic data, representing a significant advancement in the field of genomics.

Keywords. Genomics, DNA sequencing, data security, encryption, decryption, FASTA files.

1. Introduction

The technological advances in DNA sequencing led genomics to the frontlines of research in medicine, agriculture, and environmental sciences. They opened the door to the development of personalized medicine and the understanding of the impacts in climate change and pollution [1]. In this evolutionary path, sharing genomic data is crucial for advancing research and developing new approaches to better understand the genomes. Nonetheless, especially in the medical field, the nature of genomic data requires secure methods for protecting it during storage and transfer [2].

With this in mind, we propose SecureFASTA, a new tool for the secure encryption and decryption of FASTA files, which are commonly used to store DNA sequences and are frequently shared among researchers. This tool uses symmetric and asymmetric encryption techniques to ensure data confidentiality with a reduced number of necessary keys shared between the pairs.

2. Methods

The proposed tool aims to simplify the encryption process of genomic data represented in the Fasta format. It implements a strategy combining symmetric and asymmetric

[1] Corresponding Author: João Rafael Almeida, E-mail: joao.rafael.almeida@ua.pt.

encryption algorithms to optimize the procedure and reduce the number of necessary keys shared between the pairs.

The proposed tool uses AES cypher to encrypt the Fasta files. AES-256 is a widely used and efficient symmetric cypher that is hard to crack using brute force and dictionary attacks when using random keys. The tool generates a random secret key for each file. The secret key is then used to initialize an instance of the AES cypher. The input file is read, encrypted, and a new file is written. The generated key is then encrypted using asymmetric encryption. We used RSA (Rivest–Shamir–Adleman) cypher [3] since it is widely used for secure data transmission.

3. Results and Discussion

We validated the implementation of the proposed tool using synthetic Fasta files containing nucleotide and protein sequences of various sizes. In terms of security, the encrypted files produced by this tool were resistant to tampering. Any attempt to modify the encrypted file resulted in a decryption error, indicating that the tampering had been detected.

Regarding security, our method follows best practices outlined in the GA4GH File Encryption Standard [4]. This standard recommends using authenticated encryption to ensure the confidentiality, integrity, and authenticity of the data and using cryptographic key management systems to store and manage encryption keys securely.

The file's checksum generated by SecureFASTA allows the receiver to confirm its integrity before decrypting it. The algorithm used for this was SHA-256 since it is resistant to collisions. In this context, a collision is when two distinct pieces of data in a hash table share the same hash value.

In this work, we proposed a valuable tool for ciphering Fasta files. Its features include the use of both symmetric and asymmetric encryption and a checksum function that allows the recipient to verify the integrity of each file before deciphering.

Acknowledgments

This work has received funding from the EC under grant agreement 101081813, Genomic Data Infrastructure.

JMS and JRA have been funded by FCT (Foundation for Science and Technology) under the grants SFRH/BD/141851/2018 and SFRH/BD/147837/2019, respectively.

References

[1] Silva JM, Almeida JR. The value of compression for taxonomic identification. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2022. p. 276-81.

[2] Bradley T, Ding X, Tsudik G. Genomic Security (Lest We Forget). IEEE Security & Privacy. 2017;15(5):38-46.

[3] Rivest RL, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM. 1978;21(2):120-6.

[4] Rehm HL, Page AJ, Smith L, Adams JB, Alterovitz G, Babb LJ, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell genomics. 2021;1(2):100029.