CODIS AND PROBABILISTIC GENOTYPING: NAVIGATING SUCCESS AND SECURITY

A Research Paper submitted to the Department of Computer Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Lily Roark

April 27, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Daniel G. Graham, Department of Computer Science

CODIS and Probabilistic Genotyping: Navigating Success and Security

CS 4991 Capstone Report, 2022

Lily Roark
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
lar6jk@virginia.edu

Abstract
The Combined DNA Index System (CODIS), a forensic DNA database developed and used by the U.S. Federal Bureau of Investigation (FBI), needs to be scalable to handle an ever-increasing dataset as well as remain secure in protecting sensitive biological data. Recent developments in probabilistic genotyping software exert additional pressure to increase the efficiency of the system without sacrificing the security of its information. Although DNAxs and other probabilistic genotyping software improve the efficacy of DNA databasing in cases with complex samples, their integration into the laboratory analysis software system introduces opportunities for software bugs and security vulnerabilities.

This state-of-the-art report is a synthesis of secure and scalable design paradigms in the computer science subfields of databases and algorithms as they could be applied to CODIS and probabilistic genotyping software. This will include analyzing the current state of such systems, and where the design is unavailable, recommend specific solutions, such as AES encryption. General architectural observations show that the properties of the data and the processes required for its analysis also contribute to the security of the system.

1 Introduction

DNA forensic databases such as CODIS and their accompanying user software are essential to modern-day criminal justice. As the world becomes increasingly globalized and individuals obtain the ability to move freely across state and national lines, the necessity to have a large, efficient DNA database linking crimes to their perpetrators grows. The demand for fast, successful DNA profile analysis is currently not met; huge backlogs of samples wait to be analyzed and entered into CODIS [1]. These samples not only connect unsolved cases to repeat offenders, but may also constitute case-to-case "hits", grouping crimes across county or state lines [1]. In addition to meeting the needs of individual unsolved cases, database growth offers a statistical advantage. As the database grows, so does its efficacy at generating hits: the likelihood of the database containing a record of a perpetrator's previous crime increases as more samples are added to the database. As such, the issues of scalability, defined as the ability of the database to grow, and efficiency, defined as the rates at which samples are processed and the software executes queries, are intertwined with each other in this case.

2 Motivating Complications

Further complications to the entangled issues of efficiency and scalability are the incorporation of newer technologies in addition to CODIS: namely, RapidDNA automated DNA analysis and probabilistic genotyping software. RapidDNA analysis is a fully-automated instrument to be used by law enforcement booking agencies. RapidDNA utilizes buccal (inner cheek) swabs from arrestees and completes the process of DNA analysis and entry into CODIS within a time span of only two hours [2].

Although this is a great improvement on the efficiency of DNA collection and processing, it is only applicable to a limited number of arrestees, cannot be used for crime scene sample analysis, and further strains the need for scalability in the database by opening up an influx of new records [2]. On the opposite spectrum of DNA sample complexity is DNAxs. DNAxs,

developed by the Netherlands Forensic Institute, is a modular and portable software able to apply probabilistic genotyping in complicated case samples [3]. These complex DNA samples are those that have more donors, greater allele dropout, fewer loci, more common alleles, and among donors a high level of allele sharing [4].

Although DNAxs and other probabilistic genotyping software are worthwhile endeavors to improve the efficacy of DNA databasing in cases with complex samples, their integration into the laboratory analysis software system introduces opportunities for software bugs and security vulnerabilities [3]. Figure 1 details the flow of data through multiple software suites in this system of DNA analysis. DNAxs is the user interface acting as an intermediary between the laboratory information management system (LIMS), the DNA databases and querying software like CODIS and SmartRank, and the calculation engine that performs complicated algorithms to find likelihood ratios associated with DNA samples containing varied numbers of contributors [3].

With the introduction of DNAxs probabilistic genotyping software and the fully-automated RapidDNA processing instrument, CODIS grows in scope, necessitating efficiency and scalability, and it is at a security risk with the expansion. There is a tradeoff between the scalability of a project and its security, which should not be ignored given the extremely sensitive, private nature of the biological data these systems contain [5].

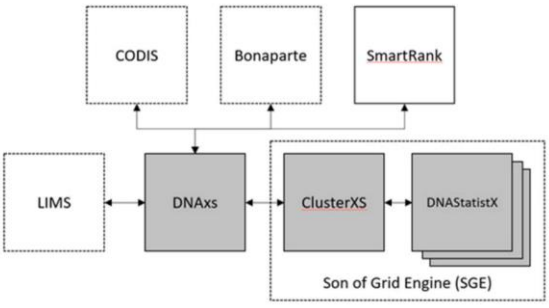

Figure 1: Overview of the software systems connected to DNAxs v2.0: Outlines the integrated software programs connected via the DNAxs software suite; DNAxs interfaces laboratory information management systems (LIMS) with probabilistic genotyping software Bonaparte and SmartRank as well as with CODIS, the DNA database itself [3].

3 Current and Proposed Designs for Security

The data contained by CODIS is extremely secure. The database is physically secured on servers located within FBI premises, and only accessible over a network within local, state and federal levels of government agencies and law enforcement approved by the FBI [6]. At the network level, unspecified firewalls prevent data packets from external sources outside of the network requesting CODIS data [7]. At the application level of database security, data backups and firewall-approved network communications are also encrypted with a non-specified algorithm.

Within the network of authorized personnel who can access CODIS, there are severe ramifications for disclosing private data, such as up to $100,000 in fines. For unauthorized attackers, hacking into CODIS can incur $250,000 in fines and up to a year in jail. Even if hackers managed to get to the data itself, it has trivial value as hostage information. The records within CODIS do not link individuals directly to sample data, and the sample data does not reveal any part of the individual's phenotype.

Figure 2 shows a sample DNA fingerprint record stored in CODIS's relational database. If the 13 core loci of a tested sample match with a record in CODIS, the ID is cross-referenced in a state-owned offline database for the individual's name. The loci stored are 13 pairs of numbers corresponding to locations in the genome where sequences of non-coding nucleobases repeat a variable number of times. There are 26 numbers because the number of repeats differ between the chromosome inherited from the mother and the chromosome inherited from the father. The number of core loci, or pairs of short tandem repeats (STRs) recorded in CODIS has increased to 20 since 2017, but the older, smaller records remain [6].

*Originating Laboratory Identifier*_____**LabXYZ**
*Specimen ID #*_____**0012152**
*13 Core Loci*_____**06,09,11,12,10,10,22,24,9.3,10,08,09, 14,14,15,17,17,22,25,12,12,9,10,09,13**
*Analyst Identifier*_____**DHL**

Figure 2: Example record within CODIS, consisting of only the lab and analyst identifiers, the unique specimen ID number, and the 13 (or more recently, 20) core loci pair values [7].

Furthermore, the samples are stored in the originating laboratory for quality control checks and audits outlined by the Scientific Working Group on DNA Analysis Methods (SWGDAM). Should a hit occur, the matched record will be retested, and if a mistake is found, the lab and analyst identifiers help retrace the origin and magnitude of the error or contamination [7].

So far, this covers all six levels of database security as defined by Prof. Nada Basit in the Databases (CS 4750) course at UVA: database level, application level, operating system level, network level, physical level and human level. However, given the non-specificity of publicly available CODIS design, I propose that the encryption used at application level should be a symmetric algorithm, as assuming the key remains secret within the network, this requires less overhead than an asymmetric public-key algorithm. Of the two prominent symmetric encryption algorithms, the Advanced Encryption Standard (AES) outperforms the older Data Encryption Standard (DES) in time, but not memory [8]. Given the time pressure exerted by testing backlogs and the computational resources of forensic laboratories, the AES encryption best applies to CODIS. AES is also functionally unhackable provided the key remains secure; using current processors, it would take billions of years to find the key based on the encryption, and it is the only public cipher approved by the National Security Administration for top-secret encryption [9]. This detail strongly supports the theory that AES could actually be the algorithm used to encrypt CODIS data.

4 Results

The information in this report is unlikely to be read by a developer working on CODIS, as the FBI has little to gain from an externally sourced, security-focused overview of its own system. However, this report may help demystify the challenge of CODIS security for other students in computer science, forensics, or law. The application of secure software principles to CODIS includes: the separation of identifiers in the records, the minimization of necessary information to perform the matching task, storing tracking information, retesting procedures, maintaining physical server security, disincentivizing unauthorized access, and implementing firewalls and encryption to protect the data at a network level. Especially given the rising threat of personal data leaks and abuse, in addition to the exponential acceleration of DNA forensic technology evidenced by probabilistic genotyping and rapid automated sequencing, it is important to understand how the US government manages to keep large quantities of private biometric data safe. The findings of this investigative research effort should hopefully qualm any fears of data insecurity from the FBI, and encourage other students to engage with the security of real software systems.

5 Conclusion

DNA fingerprinting is an essential tool to modern forensics that relies on an efficient and accurate, yet secure, relational database. In the US, this is the Combined DNA Index System. Within CODIS, records linking perpetrators' DNA to crime scene samples consist of 20 loci pair lengths in the genome, revealing no personal information unless matched by the testing laboratory. Still, it is necessary to protect the raw data contained within CODIS so that the immutable, private biometric fingerprint of citizens is not abused. Layers of database security are applied in addition to non-functionality of the isolated records to ensure that only authorized persons can access and manipulate the data. The security layers range from physical security of keeping the storage servers within a protected space, to the network level of firewalls and AES encryption, and the human level of proactively establishing laws prohibiting misuse of these developing technologies. Only once one has understood the system as it exists can they stay ahead of potential vulnerabilities in it, and fortify its defense against either mistakes or deliberate, criminal abuse.

6 Future Work

The synthesis of scalable, secure design paradigms applied to CODIS could be expanded by a coded implementation of the AES encryption algorithm on CODIS-formatted data points, to further explore the way AES confuses and diffuses the plain text into a cipher. This approach could also be expanded to implement differing encryption algorithms, such as homomorphic encryptions which maintain linearity (the ability to preserve addition and multiplication logic in the cipher). Additionally, this synthesis mentioned but did not explain how firewalls work and which particular ones might be used in CODIS; another technical report could explore this avenue of research in detail.

References

[1] John M. Butler. 2005. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2nd ed.). Academic Press, Cambridge, MA. DOI: https://doi.org/10.1016/C2009-0-01945-X

[2] Federal Bureau of Investigation, 2020. Standards for the operation of Rapid DNA booking systems by law enforcement booking agencies. (September 2020). Retrieved from https://www.fbi.gov/file-repository/standards-for-operation-of-rapid-dna-booking-systems-by-law-enforcement-booking-agencies-eff-090120.pdf/view

[3] Martin Slagter, Dennis Kruise, Larissa van Ommen, Jerry Hoogenboom, Kristy Steensma, Jeroen de Jong, Pauline Hovers, Raymond Parag, Jennifer van der Linden, Alexander L.J. Kneppers, and Corina C.G. Benschop. 2021. The DNAxs software suite: A three-year retrospective study on the development, architecture, testing and implementation in forensic casework. *Forensic Science International: Reports* 3, 100212 (July 2021), 1-12. DOI: https://doi.org/10.1016/j.fsir.2021.100212

[4] Corina C.G. Benschop, Linda van de Merwe, Jeroen de Jong, Vanessa Vanvooren, Morgane Kempenaers, C.P. (Kees) van der Beek, Filippo Barni, Eusebio López Reyes, Léa Moulin, Laurent Pene, Hinda Haned, and Titia Sijen. 2017. Validation of SmartRank: A likelihood ratio software for searching national DNA databases with complex DNA profiles. *Forensic Science International: Genetics*, 29 (July 2017), 145-153. DOI: https://doi.org/10.1016/j.fsigen.2017.04.008

[5] Amit Manjhi, Anastassia Ailamaki, Bruce M. Maggs, Todd C. Mowry, Christopher Olston, and Anthony Tomasic. 2006. Simultaneous scalability and security for data-intensive web applications. In Proceedings of the 2006 ACM SIGMOD international conference on Management of data (SIGMOD '06). Association for Computing Machinery, New York, NY, USA, 241–252. DOI: https://doi.org/10.1145/1142473.1142501

[6] Federal Bureau of Investigation. 2020. CODIS and NDIS fact sheet. (June 2020). Retrieved from https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet

[7] Nevada State Legislature. 2011. Fact sheet on forensic DNA analysis. Retrieved from https://www.leg.state.nv.us/Session/76th2011/Exhibits/Assembly/JUD/AJUD697T.pdf

[8] Shaify Kansal and Meenakshi Mittal. 2014. Performance evaluation of various symmetric encryption algorithms. *2014 International Conference on Parallel, Distributed and Grid Computing*, 105-109. DOI: https://doi.org/10.1109/PDGC.2014.7030724

[9] Lynn Hathaway. 2003. National policy on the use of the Advanced Encryption Standard (AES) to protect national security systems and national security information. (June 2003). Retrieved from https://web.archive.org/web/20101106122007/http://csrc.nist.gov/groups/ST/toolkit/documents/aes/CNSS15FS.pdf