NAVIGATING SUCCESS AND SECURITY OF FORENSIC DNA DATABASES

STANDARDS OF USE OF DNA DATABASES: PRIVACY AND PUBLIC INFORMATION

A Thesis Prospectus
In STS 4500
Presented to
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Lily Roark

November 1, 2021

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.
Signed: Lily Roark Date: November 1, 2021

ADVISORS

Catherine Baritaud, Department of Engineering and Society

Daniel Graham, Rosanne Vrugtman, Department of Computer Science

Deoxyribonucleic acid (DNA) profiling for criminal cases is one of the greatest technological innovations in criminal justice of the past few decades, but uncertainty and controversy still surrounds the standards of DNA collection and analysis. The US employs local, state, and national levels of DNA index systems. All of these systems are supported by the Federal Bureau of Investigation (FBI)-created Combined DNA Index System (CODIS), which includes the DNA database itself and the software suite used to manage it (Butler, 2005, p. 441). Additionally, DNAxs software developed abroad in the recent past builds on CODIS using probabilistic genotyping algorithms (Slagter et al., 2021, p. 1). These technologies have great potential to solve crime and identify victims, but require the cost of collecting tremendous amounts of sensitive biological data from the public. Public privacy is balanced with the need for the system to be accessible, efficient, and accurate.

The technical portion of this thesis, under the advice of Professor Daniel Graham, will examine how the scalability, efficiency and effectiveness of DNA database software affects the security of the system. The STS portion of the thesis, under the advice of Professor Catherine Baritaud, will then examine DNA database software through an actor-network theory (ANT) perspective (Jolivet & Heiskanen, 2010) and investigate how standards of DNA data entry should be set in order to maintain public transparency and legality. These two topics are tightly coupled, both at the surface-level of DNA databases such as CODIS and at the ethical level of protecting the privacy of the public. The majority of the work on these two documents will be research to establish the technical and ethical context surrounding the topics. Although journal articles and court documents will likely serve as the main sources of information into these domains, any relevant open-source code available will be examined. Research should continue through the academic year and be completed by the end of February 2022, followed by the synthesis and analysis of this research into a scholarly article delivered in April.

NAVIGATING SUCCESS AND SECURITY OF FORENSIC DNA DATABASES

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 (Butler, 2005, p. 436). 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 (Butler, 2005, p. 446). 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.

COMPLICATIONS MOTIVATING FURTHER RESEARCH

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 (FBI, September 2020, p. 4). 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 (FBI, September 2020, p. 5). 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 (Slagter et al., 2021, p. 2). 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 (Benschop et al., 2017, p. 147). 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 (Slagter et al., 2021, p. 8). Figure 1 (p. 4) 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 (Slagter et al., 2021, p. 4). With the introduction of DNAxs probabilistic genotyping software and the fully-automated RapidDNA processing instrument, CODIS grows in scope, necessitating efficiency and scalability, and 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 (Manjhi et al., 2006, p. 1).

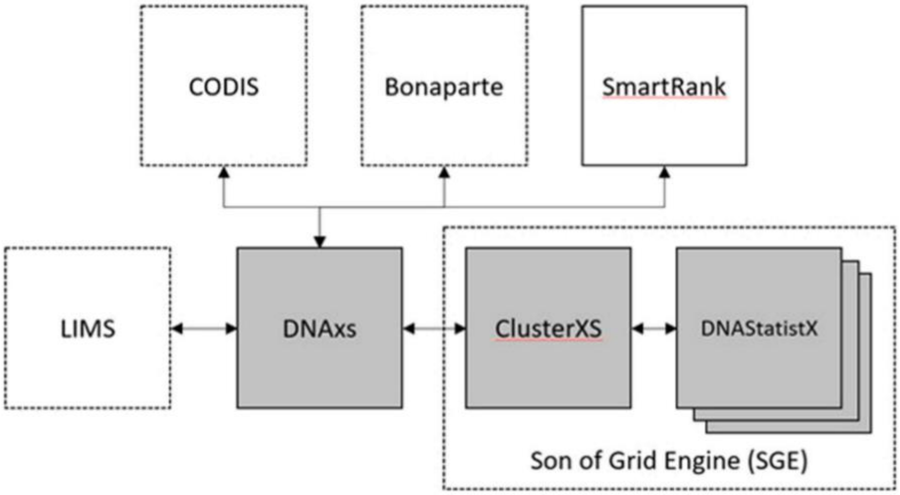

Figure 1: Overview of the software systems connected to DNAxs v2.0: Outlines the integrated software programs connected via the DNAxs software suite; DNAxs interfaces laboratory information management systems (LIMS) with probabilistic genotyping software Bonaparte and SmartRank as well as with CODIS, the DNA database itself (Slagter et al., 2021).

MODES OF RESEARCH AND ANTICIPATED OUTCOMES

Academic articles, available open-source code, and government-issued standards of practice are the primary documents to be used in the research effort. These will serve as evidence in a state-of-the-art technical report that synthesizes and extends the sub-fields of databases and algorithms in computer science. Specifically, this report will examine how the scalability, efficiency and efficacy of DNA databases affects the security necessitated by such sensitive information, and will recommend avenues of improvement for these metrics in the context of integrated probabilistic genotyping applications.

STANDARDS OF USE OF DNA DATABASES: PRIVACY AND PUBLIC INFORMATION

Tightly coupled to the security of DNA database software is the STS thesis topic: the standards of use and management of DNA databases. Again, CODIS is the primary case to study, with relatively well-defined regulations set by the FBI (July 2020, p. 1). In addition to the FBI's quality assurance standards for DNA databasing laboratories, the Scientific Working Group on DNA Analysis Methods (SWGDAM), a meeting community of forensic scientists from the United States and Canada, have produced standards for the efficient DNA processing of Sexual Assault Evidence Kits (SAKs) in a laboratory (SWGDAM, 2016, p. 2).

The FBI standards are clear and concise, demanding trained managerial and technical staff (FBI, July 2020, p. 14), separation of tasks, a documented chain of custody for evidence, and peer review of scientific validation studies (SWGDAM, 2015, p. 5). They also require software review of functionality, reliability, precision, accuracy, sensitivity and specificity (FBI, July 2020, p. 25). In comparison, the standards offered by SWGDAM are more applicable to the victims, nurses, law enforcement and judiciary involved in the collection of sexual assault evidence than the laboratory analysts (SWGDAM, 2016, p. 1). There is an unfortunately high throughput of SAKs to be analyzed, and this requires measures to increase the efficiency of processing (SWGDAM, 2016, p. 3); SAK processing usually takes two to six months by state, but with the *Direct to DNA* and other SWGDAM processing recommendations, this could ideally be cut down to only two to four weeks (SWGDAM, 2016, p. 22). Figure 2 (p. 7) outlines the flow of evidence processed in a lab to increase turnaround. Note should be taken of the *Direct to DNA* approach, as DNA analysis is now often more sensitive and faster than serological (bodily fluid) testing, a manual process that should be performed secondarily if necessary (SWGDAM, 2016, p. 14). Regarding the enforcement of these two standards documents, the FBI quality assurance standards are enforced via an annual audit, whereas the SWGDAM protocol is merely recommended practice on behalf of the National Institute of Justice (NIJ) for the Sexual Assault Forensic Evidence Reporting (SAFER) Act (SWGDAM, 2016, p. 1).

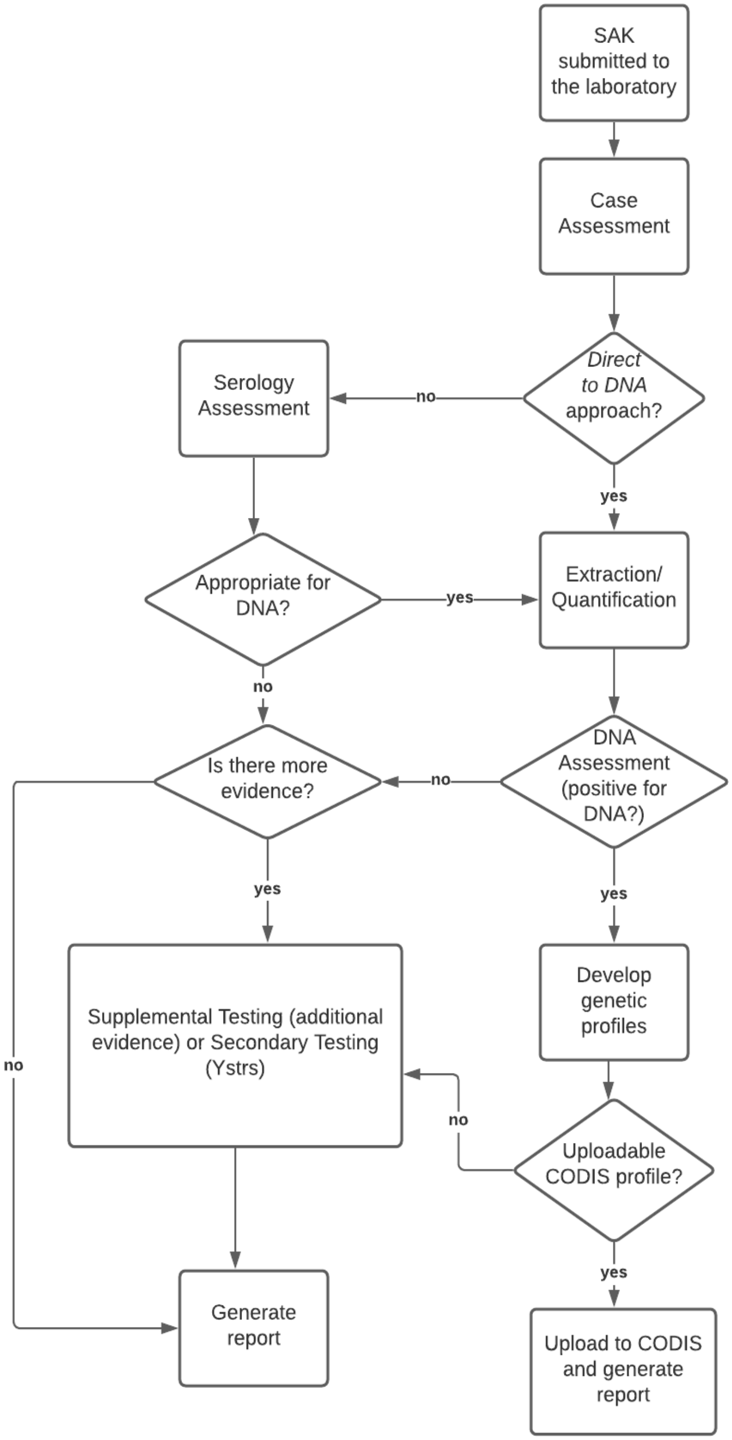
SAK submitted to the laboratory
Case Assessment
Direct to DNA approach?
no
Serology Assessment
yes
Appropriate for DNA?
yes
Extraction/ Quantification
no
Is there more evidence?
no
DNA Assessment (positive for DNA?)
yes
yes
Supplemental Testing (additional evidence) or Secondary Testing (Ystrs)
Develop genetic profiles
no
no
Uploadable CODIS profile?
yes
Generate report
Upload to CODIS and generate report

Figure 2: High throughput process flow for sexual assault evidence kits: Adapted from "SWGDAM recommendations for the efficient DNA processing of sexual assault evidence kits", the flow chart tracks the recommended handling of SAKs by a lab, via the *Direct to DNA* approach (SWGDAM, 2016).

COMPLICATIONS MOTIVATING FURTHER RESEARCH

These existing standards are sufficient in the limited definition of technology; the FBI outlines the setup of labs such that DNA analysis is accurate, and SWGDAM guides the growing efficiency of the practice. However, the practice of a technology is in constant relation with the societal applications of that technology, which expand past the confines of the laboratory. Overlooking this relationship can result in real-world consequences; for example, in the Lifecodes band-shifting debacle of 1989, Lifecodes, a privately contracted DNA testing laboratory, withheld their erroneous methodology from the prosecution, in aid of the defense, resulting in a reversed testimony (Annas, 1990, p. 37). This case would have benefitted from an integrated in-court explanation of the scientific process of DNA analysis, as well as a review of Lifecodes' adherence to standards of practice. It would be a mistake to believe that the developers of database software and the analysts using it are immune to outside pressures to unethically misuse or fabricate the authenticity of their work. In addition to standards of use, there must be independent oversight and external code review, as developers are biased towards their own correctness when testing their own code (Buckleton et al., 2021, p. 6). The uncontrollable growth of computer databases containing a wealth of private information means that even if the individual has little control over how their information is used, they should still be informed of what is collected, how it is generally analyzed, and to what purpose it serves (Weiss, 2004, p. 62). The question is: how should additional standards of forensic DNA analysis be set in order to maintain public transparency and avoid malpractice?

MODES OF RESEARCH AND ANTICIPATED OUTCOMES

Academic articles, official government codes such as the FBI standards discussed above, and case studies in the United States justice system will be the primary sources for context informing how to extend these standards of technology-practice. In addition to examining current standards, research will delve further into the complications that result in consequences detrimental to public privacy, such as conflicts of interest, tradeoffs between automation and the complexity of a given case, and navigation of security and efficiency within the DNA analysis system. Actor-network theory as defined by Jolivet and Heiskanen (2010, p. 1) will be utilized to consider all possible actors that could be affected by the expansion of standards, such as in the preliminary ANT diagram, Figure 3 (p. 10). Mapping the actors in a network helps visually illuminate parties that need to be consulted before moving forward with expanding standards. As seen in the article “Blowing against the wind”, a town within the visual scope of a proposed wind-energy farm in France was not included in the preliminary planning stage, resulting in unexpected consequences of actor “overflow”, effectively derailing the project (Jolivet & Heiskanen, 2010). As shown by the myriad of different actors in Figure 3, DNA databases are at the intersection of biology and computer science, the law, and public interest, so an ANT perspective should help untangle those relationships into a digestible model. Ideally, the STS academic article produced from these research efforts should propose avenues of standards reform that go beyond the laboratory, and suggest discussion between the relevant groups identified by ANT.

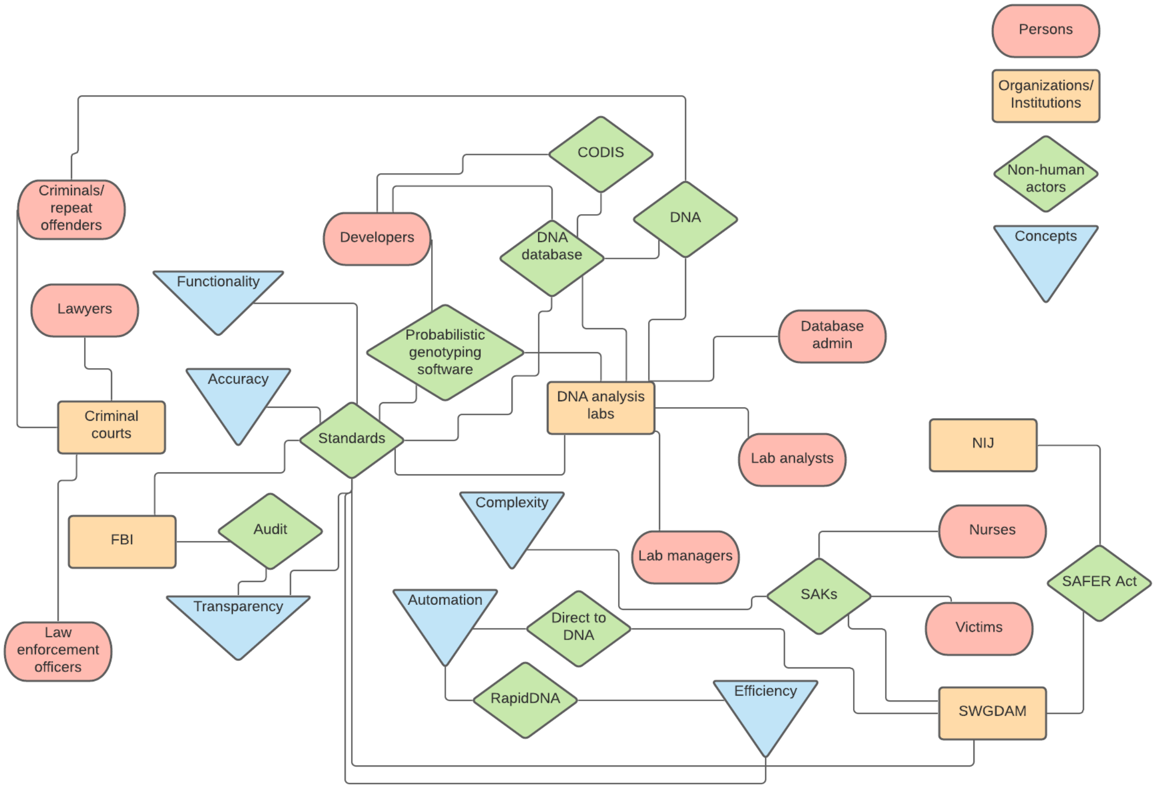

Figure 3: Preliminary DNA database ANT diagram: an overview of key actors influencing the development and standards of practice for forensic DNA databases, featuring probabilistic genotyping software and RapidDNA technologies (Roark, 2021).

MOVING FORWARD

Recent developments in probabilistic genotyping software have created demand for increased scalability, efficiency, and security in DNA databases, primarily CODIS. They also introduce a new layer of controversy to the discourse surrounding how to ensure quality standards and proper use of databases. In a field where the results generated from such software play a hand in the life or death of citizens, the technical components of the software need to be efficient, effective, secure; the state-of-the-art report will analyze the tradeoffs between these metrics and recommend solutions for their optimization. Additionally, the scholarly STS article produced from parallel research efforts will use ANT to address the requisite standards of transparency and reliability to the public via government agencies and contracted laboratories in the face of these growing technologies.

REFERENCES

Annas, G.J. (1990). At law: DNA fingerprinting in the Twilight Zone. *The Hastings Center Report*, 20(2), 35-37. https://www.jstor.org/stable/3562618

Benschop, C.C.G., van de Merwe, L., de Jong, J., Vanvooren, V., Kempenaers, M., van der Beek, C.P., Barni, F., Reyes, E.L., Moulin, L., Pene, L., Haned, H., & Sijen, T. (2017, April). Validation of SmartRank: A likelihood ratio software for searching national DNA databases with complex DNA profiles. *Forensic Science International: Genetics*, 29, 145-153. https://doi.org/10.1016/j.fsigen.2017.04.008

Buckleton, J.S., Curran, J., Taylor, D., Bright, J.A. (2021, September). What can forensic probabilistic genotyping software developers learn from significant non-forensic software failures? *WIREs Forensic Sci*, 3(2), 1-8. https://doi.org/10.1002/wfs2.1398

Butler, J.M. (2005). Combined DNA Index System (CODIS) and the use of DNA databases. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2nd ed., pp. 435-452). Elsevier Science & Technology. https://www.sciencedirect.com/science/article/pii/B9780123749994000126

Federal Bureau of Investigation (2020, July 1). Quality assurance standards for DNA databasing laboratories. https://www.fbi.gov/file-repository/quality-assurance-standards-for-dna-databasing-laboratories.pdf/view

Federal Bureau of Investigation (2020, September 1). Standards for the operation of Rapid DNA booking systems by law enforcement booking agencies. https://www.fbi.gov/file-repository/standards-for-operation-of-rapid-dna-booking-systems-by-law-enforcement-booking-agencies-eff-090120.pdf/view

Jolivet, E. & Heiskanen, E. (2010, November 15). Blowing against the wind-An exploratory application of actor network theory to the analysis of local controversies and participation processes in wind energy. *Energy Policy*, 38(11), 6746-6754. https://doi.org/10.1016/j.enpol.2010.06.044

Manjhi, A., Ailamaki, A., Maggs, B.M., Mowry, T.C., Olston, C., & Tomasic, A. (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, 241–252. https://doi.org/10.1145/1142473.1142501

Roark, L. (2021). *Preliminary DNA database ANT diagram*. [Figure 3]. Prospectus (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Scientific Working Group on DNA Analysis Methods (2015, June 15). SWGDAM guidelines for validation of probabilistic genotyping systems. http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf

Scientific Working Group on DNA Analysis Methods (2016, December 5). SWGDAM recommendations for the efficient DNA processing of sexual assault evidence kits. http://media.wix.com/ugd/4344b0_4daf2bb5512b4e2582f895c4a133a0ed.pdf

Slagter, M., Kruise, D., van Ommen, C., Hoogenboom, J., Steensma, K., de Jong, J., Hovers, P., Parag, R., van der Linden, J., Kneppers, A.L.J., & Benschop, C.C.G. (2021, July). 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), 1-12. https://doi.org/10.1016/j.fsir.2021.100212

Weiss, M.J. (2004, January 1). Beware! Uncle Sam has your DNA: legal fallout from its use and misuse in the U.S. *Ethics and Information Technology*, 6(1), 55 - 54. https://doi.org/10.1023/b:etin.0000036159.90081.cc