Forensic Algorithms

GAO Expert Panel, January 15-16, 2020
Mark Perlin, PhD, MD, PhD, March 11, 2020

GAO Questions

Written answers provided by Cybergenetics panelist Dr. Mark Perlin regarding probabilistic genotyping methods for analyzing complex DNA evidence. The outline follows the PowerPoint slides provided by the GAO for their two-day meeting.

Background Information

Complex DNA evidence

Most DNA evidence is mixture of two or more people. Manual review of such DNA data leads to lost information or inaccurate results. Computer interpretation of the same data using statistical modeling can preserve identification information for accurate results.

Partial-data allele methods

DNA may be present in small amounts. Human review discards low level DNA data. Statistical modeling can preserve this evidence.

Full-data genotype methods

DNA components are present in differing amounts, such as with mixtures that contain more material from one person than from another person. Human review ignores such quantitative data, treating all components as if they had the same amount of DNA. Sophisticated computer interpretation of the same DNA data can use these quantities.

Crime laboratory incentives

Crime laboratories are currently rewarded for how many samples they process. Instead, their incentives should be based on how much identification information they produce.

Session 1: Overview of Forensic Algorithms and their Operational Use

What forensic algorithms do federal law enforcement agencies use to associate evidence with civilian suspects? We are aware of probabilistic genotyping, facial recognition, fingerprinting, gait, voice, handwriting, and iris methods.

a. *Do you agree with this list?*
b. *What algorithms, if any, are missing from this list?*
c. *Which federal law enforcement agencies use these algorithms to associate evidence with suspects?*

I am familiar with forensic genotyping of DNA evidence, sometimes called "probabilistic genotyping" (PG). The "PG" moniker is a misnomer; "forensic genotyping" would be better.

All scientific variables have uncertainty (e.g., arising from a measurement process). This uncertainty is expressed through probability (e.g., credible or confidence intervals). In science or conversation, we do not discuss "probabilistic" speed or "probabilistic" length; all variables are "probabilistic," so no such adjective is needed or used.

What are the key algorithm components or steps used in these methods?

Sophisticated computing uses probability modeling of the short tandem repeat (STR) laboratory process to produce genotypes from observed DNA data. Solving high-dimensional probability equations is often done using statistical sampling, such as Markov chain Monte Carlo (MCMC). The genotype is a summary statistic of the STR data.

Subsequent comparison of two genotypes, relative to a random population, yields a match statistic known as a likelihood ratio (LR). The base ten logarithm, or $\log_{10}(LR)$, is a standard measure of information. Forensic genotyping measures the identification information contained in the DNA evidence.

A positive log(LR) provides statistical support for a person having contributed their DNA to biological evidence. Conversely, a negative log(LR) statistically indicates that a person did not contribute their DNA to the biological evidence.

What common features or steps do these algorithms share, if any?

Accurate genotyping of complex DNA evidence typically involves Bayesian probability modeling, MCMC computer search, and LR match statistic reporting.

Calling these genotyping systems "algorithms" may be misleading. Rather, they are measurement systems that measure DNA identification information. Like most laboratory instrumentation and measurement tools, they use computers.

What standards exist for the data used in forensic algorithms, who issued them, and who certifies the data against these standards?

Standards for the validation of "probabilistic" genotyping are in place from at least (a) the Scientific Working on DNA Analysis Methods (SWGDAM) in 2015, (b) the Federal Bureau of Investigation (FBI)'s Quality Assurance Standards (QAS) in 2011 & 2020, and (c) the American National Standards Institute (ANSI) with American Academy of Forensic Sciences (AAFS) Academy Standards Board (ASB) in 2018.

How mature are these algorithms and how are they being used by federal law enforcement agencies?

Cybergenetics TrueAllele® system was developed twenty years ago [1]. It was used fifteen years ago to reanalyze the World Trade Center disaster DNA data. Court testimony on TrueAllele® results has been given for over ten years. The system has been used in thousands of criminal cases, including federal cases, for both prosecution and defense. TrueAllele® has undergone over three dozen validation studies, including eight studies published in peer-reviewed journals [2, 3, 4, 5, 6, 7, 8, 9]. Going beyond the limitations of manual DNA review, TrueAllele® has helped exonerate ten innocent men.

How does the maturity of each algorithm compare and contrast with each other?

Some genotyping systems are twenty years old, while others have appeared in the last year. Mature systems have more extensive statistical modeling that can handle more complex DNA data, and have undergone more extensive empirical validation.

How is accuracy determined for each algorithm?

a. *Vendors (software companies)*
b. *Users (law enforcement agencies)*
c. *Others (e.g., NIST)*

Accuracy is assessed through empirical validation studies, based on testing the computer systems on complex DNA evidence data. The studies often measure predictable properties of DNA information (e.g., the linear relationship between DNA quantify and identification information). They also measure error rates for different DNA compositions.

All groups can use the same validation methods.

a. Vendors have the most knowledge and experience using the systems. They conduct the most extensive *developmental* validation studies, publishing them in peer-reviewed journals (QAS developmental validation requirement).
b. User laboratories need to understand and operate their systems. They conduct thorough *internal* validation studies, writing them up as reports for auditors (QAS internal validation requirement).
c. NIST has less genotype validation expertise and experience than vendors or labs. They have made conceptual and scientific errors when attempting validation.

Session 2: Characterizing the Accuracy of Forensic Algorithms

What metrics are used to determine accuracy and how is it reported?

Typical metrics include sensitivity, specificity, and reproducibility. These metrics are reported through log(LR) distributions, error rates, and statistical quantities.

What are the requirements for verification and validation of these algorithms?

The mathematical and statistical methods are published, in both reports and peer-reviewed journals. Validation studies are published, in both reports and peer-reviewed journals.

Normative science ensures independent assessment through peer-review. Studies have compared the results of different methods when they are applied to the same data.

Session 3: Strengths and Limitations of Forensic Algorithms

Summarize the known space of what software can and cannot do.

- *What are the strengths of PGS and fingerprint analysis software?*
- *What are the limitations of PGS and fingerprint analysis software?*

Limited PG software merely mimics simple human review methods for deriving forensic statistics (e.g., the FBI's Popstats). Since the underlying mixture review methods have little or no scientific justification, and are known to give inaccurate or inconclusive results, they cannot solve DNA mixture problems [10]. Nor can validation demonstrate their accuracy.

Some PG software restricts data input, using only some of the STR data. There is PG software that relies on human judgement for data and parameter choices. Genotyping software results may be impaired by incomplete data or by human subjectivity.

Cybergenetics TrueAllele® software uses all the data, and eliminates subjective human decision making. The computed results reflect the STR data's identifying power. The system measures the identification information present in the DNA evidence.

What are any challenges associated with using these types of software?

Insufficient education or training can lead to software misuse and incorrect results. Developers and user laboratories usually have the knowledge and skill to properly operate their technology, and report accurate results. Less proficient users (e.g., students or scientists without adequate training) may get it wrong.

Session 4: Key Issues Affecting Usage of Forensic Algorithms

What key issues affect the use of these algorithms?

Biased decision making can make the software less effective. Subjective manual review has been used in forensic DNA interpretation for over twenty years. Human shortcuts can bypass reliable mathematics, yielding inaccurate or unreliable match statistics.

What aspect of the algorithm affects these issues?

The term "algorithm" does not properly characterize forensic genotyping software that measures DNA identification information. An information measurement (even if based on validated mathematics or statistics) can become less accurate when human choices affect the computational process.

Session 5: Policy Options for Forensic Algorithms

What policy options could address key issues to using forensic algorithms?

Data transparency

In many jurisdictions, outside scientists cannot check the DNA conclusions reached by government laboratories. With DNA evidence, the prosecution may refuse to provide the original instrumentation data needed for independent computer assessment. With DNA investigation, FBI policies often block an independent search of the CODIS database.

There is a twenty-year history of misinterpreting hundreds of thousands of DNA mixtures [11]. This DNA information failure was spearheaded by NIST, a federal agency that promoted ineffective methods over proven science. Bad policy has harmed criminal justice.

CODIS database match algorithms are based on ineffective DNA comparison methods. CODIS software limitations block genotype upload and comparison. More effective science overcomes these limitations. Better genotyping software can use DNA databases to free the innocent and find the guilty.

The law should promote data transparency. Government should share crime lab DNA data with outside experts, who can then independently re-examine the data using sophisticated interpretation software. The FBI should open CODIS to better science, with the goal of revealing important criminal justice information.

Software transparency

Computer software that assists one side in a criminal case should be made available to the other side. This software access provides a transparent check on method limitations and operator errors.

Protecting the confidentiality of proprietary source code incorporating the developer's trade secrets is unrelated to software transparency. Source code is not needed to run computer software (e.g., user labs do not have or need source code). Moreover, protecting trade secrets fosters commercial innovation that advances criminal justice.

Additional Information

Establishing scientific reliability through empirical testing

Empirical testing is essential. Such testing was not done for ineffective DNA mixture manual review methods. Instead, in 2010 the FBI and NIST mandated "stochastic thresholds," a human algorithm for discarding data that lacked empirical support. This centralized policy caused considerable forensic failure [11]. Empirical testing of these unproven methods could help reopen past cases of forensic injustice.

Sophisticated commercial forensic genotyping systems have undergone thorough empirical testing that establishes their scientific reliability. This empirically tested reliability has been published in peer-reviewed journals and other available documents. Such empirical validation testing is reiterated by every crime laboratory using the software.

Determining error rates for probabilistic forensic methods

Quantifying error rates is essential when reporting scientific results. Recent statistical improvements show that highly accurate error rates can be quickly and precisely calculated for probabilistic genotypes and statistical match results [12].

Current PG error methods can establish false positive rates for inclusionary match results, and false negative rates for exclusionary nonmatch results. They can be used on evidence items in casework, or on genotype collections in validation studies [12]. For better science, DNA match statistics should be reported along with their error rates.

However, older methods of calculating error rates have limited functionality. They are less accurate and far more expensive, resource intensive, time consuming, and wasteful. NIST champions these less-sophisticated methods, which unnecessarily inflate validation costs.

Defining "validation study"

A validation study determines the accuracy and efficacy of a forensic analysis method based on empirical assessment using actual data. The DNA samples can come from casework [2, 3, 7, 8, 9], or be made in a laboratory [4, 5, 6]. Advanced validation methods can handle either situation [12]. The "correct" answer is not needed to elicit genotype information.

Typical validation axes include sensitivity (e.g., false negative rates), specificity (e.g., false positive rates), and reproducibility (particularly when using randomized algorithms). Other axes that are relevant to DNA mixture interpretation include the impact of contributor number, and showing predictable information response.

The most thorough validation studies are done by commercial developers, since they have the resources and motivation for conducting extensive testing. Crime laboratory users also perform thorough studies, since they need to understand their systems in order to present them in court. Less skilled or motivated software operators may not be as effective.

The GAO has proposed a definition of validation as: "A formal, empirical process in which an authority/independent third party determines and certifies the performance characteristics of a given method." Such "authority/independent third party" groups do not have a role in normative science. Inexperienced users may lack the training and knowledge needed to conduct a valid study. Authorities and third parties can have undisclosed conflicting agendas.

In normative science, study authors, innovators, grant recipients, etc. are rarely "authorities" or "independent third parties." GAO's proposed definition would effectively invalidate most established science. It would erase the credentials of their own scientists and advisors, whose CV's would then omit most of their peer-reviewed empirical studies.

There are many ways to improve scientific diversity and software reliability. Replacing accountable scientists and developers with unaccountable central authorities and partisan third parties does not help. But more data transparency would.

Role of the Courts

Judges are experienced, effective and impartial gatekeepers regarding the admissibility of scientific evidence. Prosecutors and defenders currently have the right to present scientific evidence that could help their case. In my experience (with over 25 admissibility hearings), a well-prepared lawyer can present scientific testimony and validation studies that can make the case for admitting complex DNA evidence.

Replacing an impartial judiciary with a potentially partisan executive agency represents a radical change in criminal justice. For almost a century (Frye 1923; Daubert 1993), judges have served as gatekeepers, deciding on reliable science for criminal and civil justice.

In the proposed legislation, defendants could potentially lose their constitutional due process right to present their case. Federal agencies could enable government prosecutors to suppress vital DNA evidence needed for exoneration or acquittal. Even prosecutors may be unable to use reliable DNA evidence for securing convictions of violent criminals. Thankfully, unbiased courts eliminate this untenable proposed conflict of interest.

Role of NIST

NIST should not be centralizing validation studies or supplanting the judiciary in determining the reliability of scientific evidence. The agency lacks the requisite knowledge and expertise to understand, operate or report on these methods. NIST's DNA mixture group has long promoted favored companies, improperly operated PG software, not disclosed their errors, suppressed scientific results, advocated unvalidated methods, ignored better technology, championed wasteful approaches, suppressed scientific information, and misled policy makers.

NIST could better help society by expanding their Standard Reference Material (SRM) program with a few more DNA mixtures. Forensic biology laboratories could then incorporate these NIST-traceable DNA mixture SRMs into their own validation studies. The labs could generate STR data from these SRM mixtures, and interpret the data using their genotyping software to produce log(LR) match statistics. Documenting how much DNA information they found would improve scientific transparency for criminal justice.

References

[1] Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *J Forensic Sci*. 2001;46(6):1372-7.

[2] Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. *PLoS ONE*. 2009;4(12):e8327.

[3] Ballantyne J, Hanson EK, Perlin MW. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. *Science & Justice*. 2013;52(2):103-14.

[4] Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*. 2011;56(6):1430-1447.

[5] Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® Casework validation study. *Journal of Forensic Sciences*. 2013;58(6):1458-66.

[6] Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. TrueAllele® Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. *PLoS ONE*. 2014:9(3):e92837.

[7] Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*. 2015; 60(4):857-868.

[8] Greenspoon SA, Schiermeier-Wood L, Jenkins BC. Establishing the limits of TrueAllele® Casework: a validation study. *Journal of Forensic Sciences*. 2015;60(5):1263-1276.

[9] Bauer DW, Butt N, Hornyak JM, Perlin MW. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *Journal of Forensic Sciences.* 2020; 65(2):380-398.

[10] Perlin MW. Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information. *Journal of Pathology Informatics*, 6(1):59, 2015.

[11] Perlin MW, When DNA is not a gold standard: failing to interpret mixture evidence. *The Champion,* May, 2018; 42(4):50-56.

[12] Perlin MW. Efficient construction of match strength distributions for uncertain multi-locus genotypes. *Heliyon*, 4(10):e00824, 2018.