Investigative DNA Databases that Preserve Identification Information

Mark W. Perlin, PhD, MD, PhD[a,*]

[a]*Cybergenetics, Pittsburgh, PA, USA*

Abstract. At the heart of the science of genetics is the *genotype*, a genetic type comprised of allele pairs at a set of loci. Since the time of Mendel in the 19th century, genotypes have been understood to be uncertain quantities represented by probability. Forensic DNA has uprooted that scientific tradition, seeking genotype certainty where none exists in the evidence. The result is a tremendous loss of identification information through the application of misguided scientific models. That information loss extends to forensic DNA databases, where genotypes are incorrectly stored as allele pairs or lists, rather than preserving their full identification power through a standard probability representation. The consequences to society are severe, since the loss of DNA database investigative information leads to the needless victimization of innocent citizens by crimes that could have been prevented. The solution is to deploy investigative DNA databases that properly preserve identification information using probabilistic genotypes.

Keywords: Forensic DNA, DNA Database, Probabilistic Genotype, Likelihood Ratio, DNA Mixture, DNA Investigation

1. Introduction

A DNA database can link crime scenes to suspects, providing investigative leads. These DNA associations can solve cold cases, track terrorists, and stop criminals before they inflict further harm. However, current government databases do not fully preserve DNA identification information, and cannot maximize public safety.

DNA data is summarized in a genotype. The genotype can be stored on a database, and compared with other genotypes to form a likelihood ratio (LR) match statistic. Data uncertainty, present in most evidence, translates into genotype probability.

Highly informative interpretation uses all the quantitative DNA data, placing higher probability on more likely genotype values. Most evidence, though, is interpreted by qualitative human review, which diffuses probability across infeasible solutions. Since the LR is proportional to the true genotype probability, weaker interpretation methods lead to weaker (or nonexistent) DNA matches.

The weakest DNA interpretation method is random man not excluded (RMNE), which thresholds quantitative data into all-or-none qualitative "allele" events. The current DNA databases (including CODIS) use an RMNE allele representation that discards considerable genotype information, losing sensitivity and specificity.

The "probabilistic genotype" representation is part of the new ANSI/NIST-ITL data exchange standard. Unlike allele lists, this database representation can preserve all DNA identification information, and be quantified dynamically into LR match statistics. Every interpretation method has a corresponding genotype probability representation.

ISFG's 2006 mixture guidelines recommend the more informative LR over RMNE. Unfortunately, current databases transform hard won LR genotypes into less informative RMNE alleles. This paper shows how genotype probability can preserve identification information for DNA investigation.

2. Information Failure

Taxpayers fund crime laboratories so that DNA can help apprehend and convict criminals, hoping to prevent further victimization [1]. In that light, every incorrect DNA miss is a moral failure. Government DNA policies over the last decade have magnified these scientific failures into a public safety crisis [2].

Most biological evidence is not pristine, comprising mixed, low level or damaged DNA. Crime labs excel at generating superb DNA data from these specimens. But their approximate human review methods cannot fully extract the identification information from their data.

In science, informative inference is achieved by fully explaining the observed data. Computers can explain DNA signals by examining every conceivable quantitative genotype explanation [3]. Without computer assistance, people cannot conduct a thorough examination of their data.

Instead, human review reduces highly informative DNA data to qualitative all-or-none possibilities [4]. These "threshold" methods discard most of the DNA match strength [5, 6]. For example, on homicide mixture DNA data from a national laboratory, evidence was given in court to a 189 billion computer-inferred match statistic; using thresholds, the lab had only assigned 13 thousand [7-10].

[*] Corresponding author. *Tel*: 001-412-683-3004. *Fax*: 001-412-683-3005.
E-mail address: perlin@cybgen.com. Cybergenetics, 160 North Craig Street, Suite 210, Pittsburgh, PA 15213, USA.

With thresholds, the false negative rate (failure to identify) exceeds 100% on typical mixture data [11]. This error rate is unprecedented in science, and would be unacceptable in any other field affecting human lives (e.g., medicine, engineering). These errors bring into question the scientific rigor of human DNA mixture interpretation.

Threshold interpretation of DNA evidence lets a forensic analyst testify comfortably in court. But these weak methods often fail to identify criminals or prevent crime. Indeed, human review can misrepresent 70% of computer interpretable DNA mixture items as "inconclusive", providing no match score at all [12].

The same threshold methods that lose a million-fold factor of DNA information are also used for national DNA database evidence. By storing "alleles" instead of genotypes, these databases discard vast amounts of identification information. However, a probabilistic genotype DNA database [3] can preserve the evidence information, and thus solve far more cold cases.

3. Conclusions

Forensic DNA investigative databases should preserve the full amount of identification information present in criminal evidence. Currently they do not, since scientifically unfounded "thresholds" are used that instead deplete DNA data of its probative force. With the advent of computer systems that can properly model quantitative data and its uncertainty, there is no longer any need for these crude "threshold" approximations. Forensic DNA evidence interpretation should instead infer fully informative probabilistic genotypes, and store them on investigative databases in order to better protect society from harm.

4. Acknowledgements

Cybergenetics staff members William Allan, Meredith Clarke, Dr. Ria David and Matthew Legler assisted in the performance and presentation of this research.

5. Conflict of interest

Dr. Mark Perlin is a shareholder, officer and employee of Cybergenetics, an American company that develops the TrueAllele® Casework computer system that infers and matches probabilistic genotypes from forensic DNA evidence, including complex mixtures that contain up to six contributors, and provides a fully informative DNA database.

6. References

[1] Wickenheiser R. The business case for using forensic DNA technology to solve and prevent crime. *J Biolaw & Bus*. 2004;7(3).

[2] Gill P, Brenner CH, Buckleton JS, et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Sci Int*. 2006;160:90-101.

[3] Perlin MW, Legler MM, Spencer CE, et al. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*. 2011;56(November):in press.

[4] Perlin MW. Explaining the likelihood ratio in DNA mixture interpretation. *Promega's Twenty First International Symposium on Human Identification*; San Antonio, TX. 2010.

[5] Perlin MW. *The DNA Investigator™ Newsletter*: Validating DNA Mixture Interpretation Methods. Pittsburgh: Cybergenetics; 2010.

[6] Perlin MW, Duceman BW. Casework validation of genetic calculator mixture interpretation (A77). *AAFS 62nd Annual Scientific Meeting*, February 22-27; Seattle, WA. American Academy of Forensic Sciences; 2010. p. 62-3.

[7] Perlin MW. *The DNA Investigator™ Newsletter*: Same Data, More Information – Murder, Match and DNA. Pittsburgh: Cybergenetics; 2009.

[8] Perlin MW, Cotton RW. Three match statistics, one verdict (A78). *AAFS 62nd Annual Scientific Meeting*, February 22-27; Seattle, WA. American Academy of Forensic Sciences; 2010. p. 63.

[9] Perlin MW, Kadane JB, Cotton RW. Match likelihood ratio for uncertain genotypes. *Law, Probability and Risk*. 2009;8(3):289-302.

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

[11] Perlin MW. Reliable interpretation of stochastic DNA evidence. *Canadian Society of Forensic Sciences 57th Annual Meeting*; Toronto, ON. 2010.

[12] Perlin MW, Duceman BW. Profiles in productivity: Greater yield at lower cost with computer DNA interpretation (Abstract). *Twentieth International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society*, September; Sydney, Australia. 2010.

(Most of these articles are freely downloadable from the web site www.cybgen.com/information.)