Poster 374

Investigative DNA Databases that Preserve Identification Information

Dr. Mark W. Perlin, Cybergenetics, Pittsburgh, PA USA

ABSTRACT

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 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 poster shows how genotype probability can preserve identification information for DNA investigation.

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.

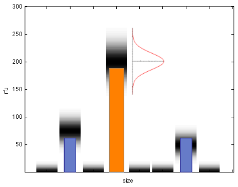
In science, informative inference is achieved by fully explaining the observed data. Computers can explain DNA signals by examining every conceivable quantitative genotype explanation (3). Without computer assistance, people cannot conduct a thorough examination of their data.

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 DNA data from a national laboratory, I testified to a 189 billion computer-inferred match statistic; using thresholds, the lab had only assigned 13 thousand (7-10).

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.

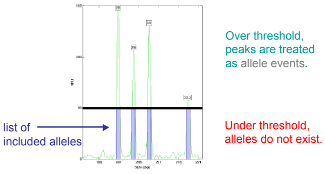
Threshold interpretation of DNA evidence lets a forensic analyst testify comfortably in court. But these weak methods often fail to identify criminals or prevent crime. Indeed, human review can misrepresent 70% of computer interpretable DNA mixture items as "inconclusive", providing no match score at 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 (such as TrueAllele) can preserve the evidence information, and thus solve far more cold cases.

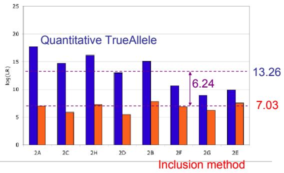


A. The quantitative modeling of DNA data accounts for peak height and uncertainty, producing allele pair possibilities. A **genotype** is a probability distribution over these allele pairs. Shown is a genotype possibility for a two contributor DNA mixture, with a minor (blue) and major (orange) allele pairs.

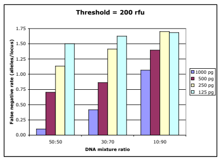
B. Applying a threshold discards the peak height information and its uncertainty, leaving only a guess at possible **alleles**. This allele list is currently stored on a DNA database, instead of more informative probabilistic genotypes.



C. Thresholds cause a **million-fold information loss** in DNA information, as measured by the likelihood ratio (LR) match statistic. Shown are the log(LR) values on the same DNA mixture data as inferred using probabilistic genotypes (blue) and RMNE thresholds (orange). See reference #3.



D. Thresholds introduce a **false-negative rate** that often exceeds 100% on DNA mixture data. These low information evidence alleles populate DNA databases, and fail to identify criminals. A more informative DNA database would instead use probabilistic genotypes.



Visit Cybergenetics website for papers and presentations:
http://www.cybgen.com/information

Visit Cybergenetics ISFG booth to see a live computer demonstration of TrueAllele® Casework.

Cybergenetics

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, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS. 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, Smith JL, Allan WP, Belrose JL, Duceman BW. 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.