Casework Validation
of Genetic Calculator
Mixture Interpretation

Mark W Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

Barry W Duceman, PhD
New York State Police, Albany, NY

American Academy of Forensic Sciences
February, 2010
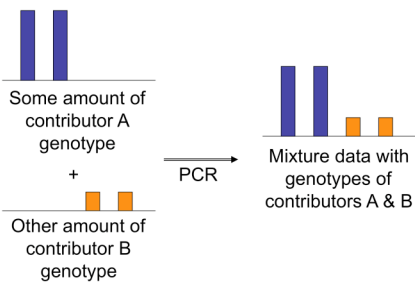
Cybergenetics

Information Gain (LR)

identification hypothesis:
the suspect contributed to the evidence

$$\text{information gain (likelihood ratio)} = \frac{\text{Odds}(\text{hypothesis} \mid \text{data})}{\text{Odds}(\text{hypothesis})} \quad \begin{array}{c} \text{after} \\ \uparrow \text{data} \\ \text{before} \end{array}$$

Additive information units: log(LR)
Order of magnitude, powers of ten

DNA Mixture Data

Some amount of
contributor A
genotype

+

Other amount of
contributor B
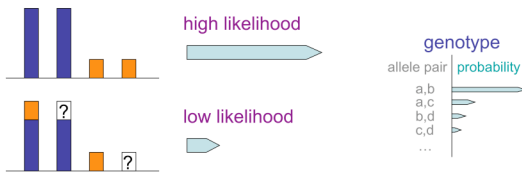genotype

PCR →

Mixture data with
genotypes of
contributors A & B

Quantitative Mixture Interpretation

Step 1: infer genotype

• consider every possible allele pair
• compare pattern with DNA data
• Rule: *better fit's more likely it*

high likelihood

low likelihood

genotype

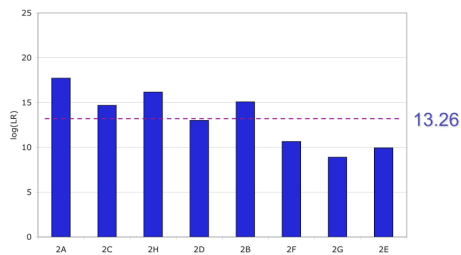| allele pair	probability
a,b	
a,c	
b,d	
c,d	
…	

Information Gain (LR)

Step 2: match genotypes

At the suspect's genotype allele pair,
what is the locus *information gain*?

$$\text{information gain (likelihood ratio)} = \frac{\text{Prob(allele pair | data)}}{\text{Prob(allele pair)}}$$
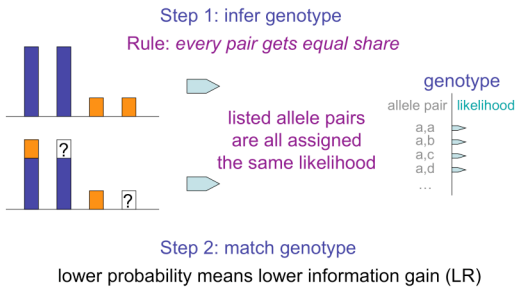
after
data
before
(population)

Computer objectivity:
(Step 1) infer evidence genotype from data
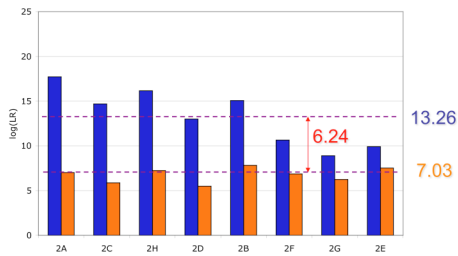(Step 2) compare genotype with suspect
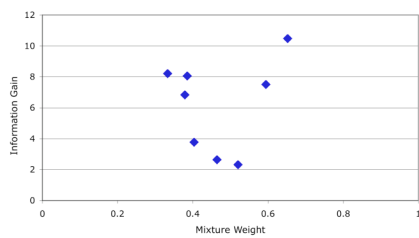

Efficacy (2 unknown)

13.26

(bar chart: log(LR) vs samples 2A, 2C, 2H, 2D, 2B, 2F, 2G, 2E)

Qualitative Manual Review

Step 1: infer genotype

Rule: *every pair gets equal share*

genotype

listed allele pairs are all assigned the same likelihood

allele pair	likelihood
a,a	
a,b	
a,c	
a,d	
…	

Step 2: match genotype

lower probability means lower information gain (LR)

Improvement

13.26

6.24

7.03

log(LR)

2A 2C 2H 2D 2B 2F 2G 2E

Improvement: Mixture Weight

Information Gain
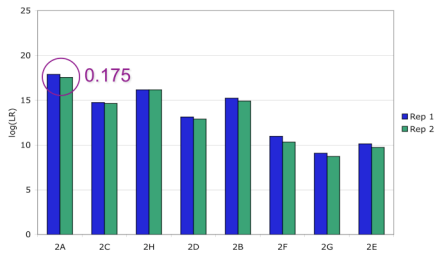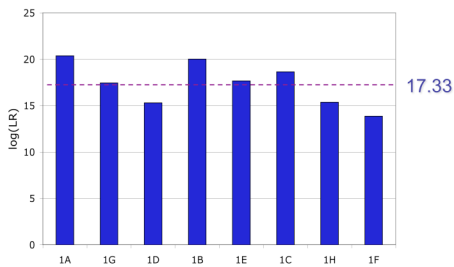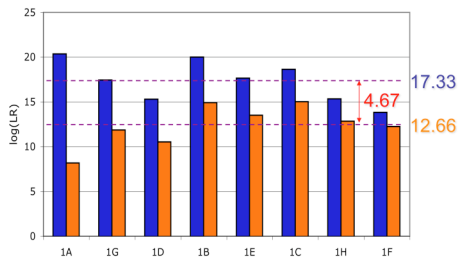
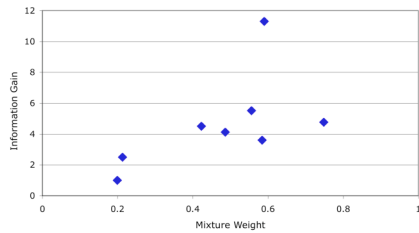Mixture Weight

Reproducibility

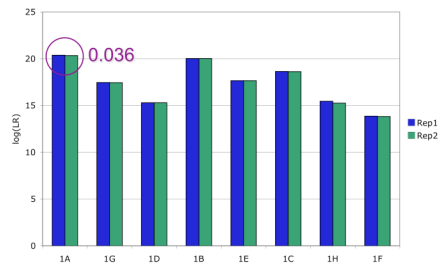

Efficacy (1 unknown)

Improvement

Improvement: Mixture Weight

Reproducibility

Validation Summary

interpretation method	two unknown (without victim)	one unknown (with victim)
quantitative computer	13.26 (0.175) (ten trillion)	17.33 (0.036) (hundred quadrillion)
qualitative human	7.03 (ten million)	12.66 (fifty trillion)
improvement	6.24 (one million)	4.67 (fifty thousand)

Conclusions

- information gain (LR) is a universal DNA metric
- *efficacy*: computer extracts useful information
- *improvement*: computer mixture interpretation is more informative than human review
  - with victim 50,000x - without victim 1,000,000x
- *reproducibility*: tenths of a log(LR) unit
- *objectivity*: "parallel unmasking", infer then match
- *productivity*: lab gives statistic for 1 of 3 items
- *utility*: science, investigation and evidence

Acknowledgements

Cybergenetics
William Allan
Meredith Clarke
Matthew Legler
Cara Spencer
Jessica Staab

Northeast Regional Forensics Institute
Jamie Belrose

New York State Police
Melissa Lee
Shannon Morris
Elizabeth Staude

Points of view in this presentation are those of the authors, and do not necessarily represent the official positions of Cybergenetics, the New York State Police, the Northeast Regional Forensic Institute or the U.S. Department of Justice. New York State Executive Law 995(c) permits the disclosure of certain DNA records for the purposes of creating or maintaining a population or statistics database, and for identification research and protocol development.

perlin@cybgen.com