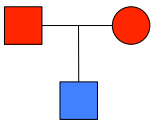
How TrueAllele® Works (Part 1)

Cybergenetics Webinar
October, 2014

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

Cybergenetics

Biological evidence



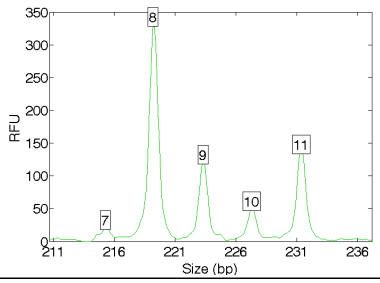
Q S

Forensic question

Did suspect S contribute his DNA
to biological evidence Q?

DNA laboratory

extract, amplify, separate, detect



TrueAllele® Casework: Infer

ViewStation User Client

Database Server

Interpret/Match Expansion

Visual User Interface
VUIer™ Software

Parallel Processing Computers
8 to 80 questions simultaneously

Bayesian probability model

Rev Bayes, 1765

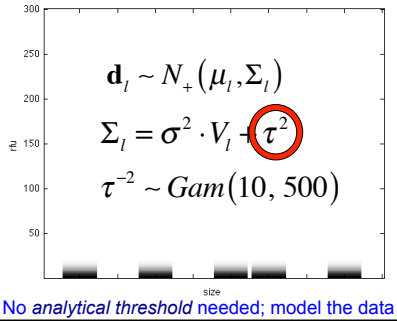
$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{data}}$$

Background noise
PCR variation
Mixture weight
PCR stutter
Relative amplification
Genotype (separated)
Differential degradation

No calibration needed; learn from the data

Perlin, M.W. and Sinelnikov, A. An information gap in DNA evidence interpretation. *PLoS ONE*, 4(12):e8327, 2009.

Background noise



$$\mathbf{d}_l \sim N_+(\mu_l, \Sigma_l)$$

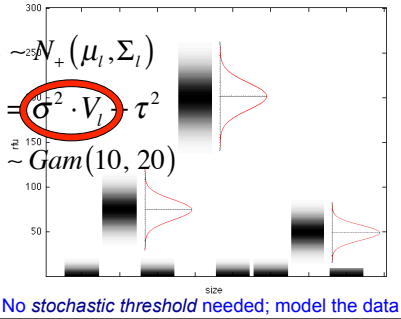
$$\Sigma_l = \sigma^2 \cdot V_l + \tau^2$$

$$\tau^{-2} \sim Gam(10, 500)$$

No *analytical threshold* needed; model the data

Perlin, M.W., Legler, M.M., Spencer, C.E., Smith, J.L., Allan, W.P., Belrose, J.L., and Duceman, B.W. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*, 56(6):1430-47, 2011.

PCR variation



$$\mathbf{d}_l \sim N_+(\mu_l, \Sigma_l)$$

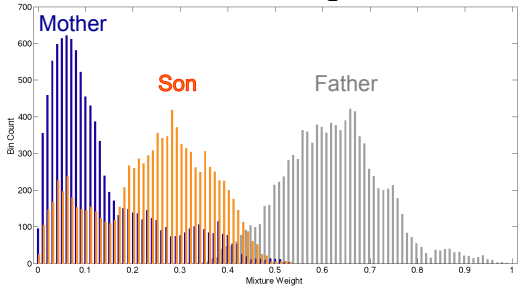
$$\Sigma_l = \sigma^2 \cdot V_l + \tau^2$$

$$\sigma^{-2} \sim Gam(10, 20)$$

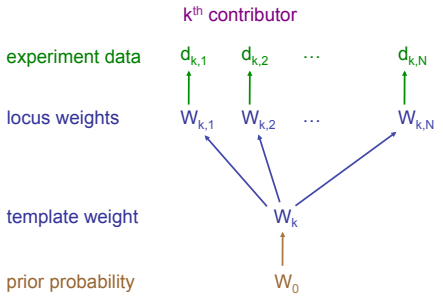
No *stochastic threshold* needed; model the data

Perlin, M.W. and Szabady, B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, 46(6), 1372-77, 2001.

Mixture weight

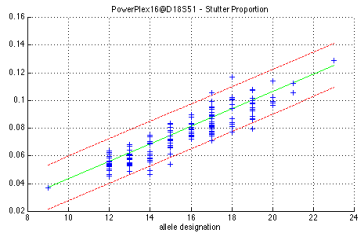


Hierarchical variables



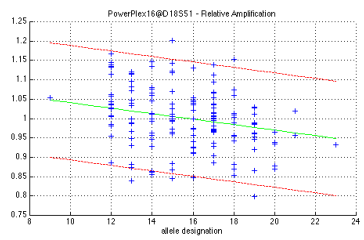
M.W. Perlin, "Expert systems for automated STR analysis", *Scientific Working Group on DNA Analysis Methods (SWGDAM)*, Quantico, VA, 14-Jan-2003.

PCR stutter



M.W. Perlin, "Expert systems for automated STR analysis", *Scientific Working Group on DNA Analysis Methods (SWGDAM)*, Quantico, VA, 14-Jan-2003.

Relative amplification



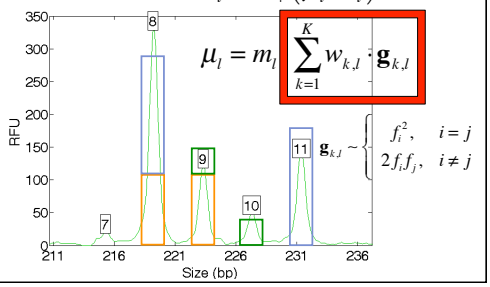
Perlin, M.W. and Szabady, B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, 46(6), 1372-77, 2001.

Genotype pattern

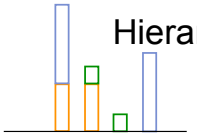
$$\mathbf{d}_l \sim N_+\left(\mu_l, \Sigma_l\right)$$

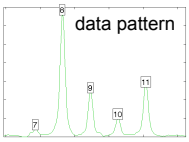
$$\mu_l = m_l \sum_{k=1}^{K} w_{k,l} \cdot \mathbf{g}_{k,l}$$

$$\mathbf{g}_{k,l} \sim \begin{cases} f_i^2, & i = j \\ 2 f_i f_j, & i \neq j \end{cases}$$



Hierarchical pattern



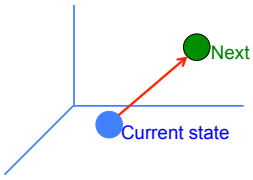


- DNA quantity
- degradation
- mixture weight
- relative amplification
- PCR stutter
- capillary injection

process transformation
pattern variation

Markov chain Monte Carlo

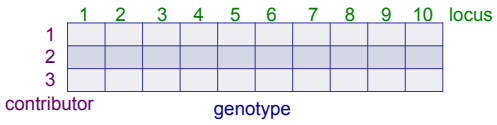
Local choices, global solution
Every cycle, visit all 100 variables



Transition probability = $\frac{\text{Pr(Next state)}}{\text{Pr(Current state)}}$

Joint probability distribution

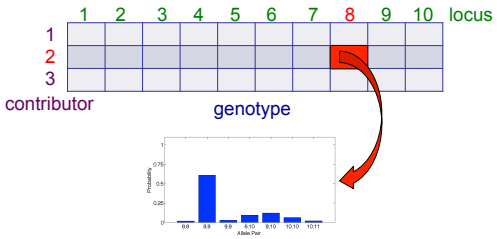
all the variables together in a high-dimensional space



genotype(contributor, locus)
mixture-weight(contributor, locus)
PCR-stutter(locus)
variance-parameters(locus)
hierarchical variables

Marginalize to separate

reduce to one variable, sum over all the other variables



objectively inferred genotype at locus 8 for contributor 2

TrueAllele® Casework: Match

ViewStation User Client

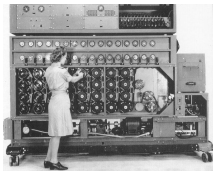
Database Server

Interpret/Match Expansion

Visual User Interface
VUIer™ Software

Parallel Processing Computers
2 hours for each question

The likelihood ratio



WWII Bombe computer cracks Enigma code

1940's, Bletchley Park, UK: Alan Turing measures how data changes belief

LR is a ratio of probabilities, information is log(LR), unit is the "ban"

exclusion | inconclusive | inclusion

-12 -6 0 +6 +12 ban

Hypothesis

Suspect S contributed his DNA to biological evidence Q

Known: the suspect's genotype is the allele pair: s = [a,b]

Odds form

Original Turing & Good formulation (1950)

$$LR = \frac{O(H \mid d)}{O(H)}$$

Bayes factor, or likelihood ratio (LR)

Perlin, M.W. Explaining the likelihood ratio in DNA mixture interpretation, in *Proceedings of Promega's Twenty First ISHI*. San Antonio, TX, 2010.

Likelihood form

Hypotheses correspond to H_P and H_D

$$\frac{\Pr(d \mid H)}{\Pr(d \mid \overline{H})}$$

Genotype expansion

TrueAllele can separate genotypes

$$\frac{\sum_{x \in G} \overbrace{\Pr(d \mid X = x)}^{\text{likelihood}} \overbrace{\Pr(X = x \mid H)}^{\text{prior}}}{\Pr(d)}$$

Apply hypothesis

Reduce summation to just suspect's term

$$\frac{\overbrace{\Pr(d \mid X = s)}^{\text{likelihood}}}{\Pr(d)}$$

Bayes theorem

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{data}}$$

Reorganize terms

$$\frac{\text{posterior}}{\text{prior}} = \frac{\text{likelihood}}{\text{data}}$$

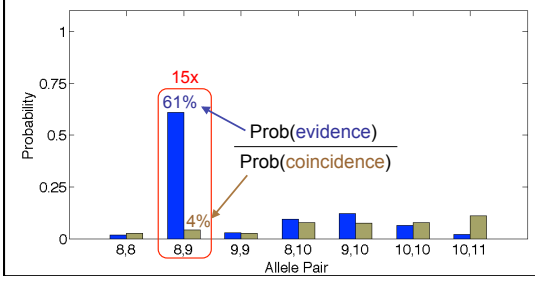
Genotype form

Posterior over prior

$$\frac{\overset{\text{posterior}}{\Pr\left(X = s \mid d_Q\right)}}{\underset{\text{prior}}{\Pr\left(X = s\right)}}$$

Visualize the LR

Ratio of genotype probabilities



Match form

Genotype match over coincidence

Q posterior S posterior

$$\frac{\Pr\left(X = s \middle| d_Q\right)\Pr\left(Y = s \middle| d_S\right)}{\Pr\left(X = s\right)}$$

prior

Match statement

Ratio of match probabilities

A match between the shotgun shell
and the son is
6 trillion times more probable
than coincidence

- accurate mathematics
- understandable language
- no conditionals to transpose
- compares separated genotypes

Evidence to evidence

Sum of genotype match over coincidence

Q posterior S posterior

$$\sum_{x \in G} \frac{\Pr\left(X = x \middle| d_Q\right)\Pr\left(Y = x \middle| d_S\right)}{\Pr\left(X = x\right)}$$

prior

standard form of LR in probability theory

Single contributor LR

Separated genotypes for each contributor
- TrueAllele does the heavy lifting
- separates out the genotypes
- single-source simplicity

Easy to understand, report and explain
Straightforward direct and cross examination

Relevant: focus is on one person
Compares genotype with genotype

The unseparated LR

H_P: X=x, Y=y, Z=z
H_D: X=?, Y=?, Z=z

$$LR_{unseparated} = \frac{\Pr("data"|H_P)}{\Pr("data"|H_D)}$$

- "data" is not quantitative data (thresholds)
- lacks Bayesian foundation (wrong math)
- subjective genotype values (just guessing)
- uses defendant genotype (not objective)
- implicates multiple people (not relevant)
- doesn't separate genotypes (incomplete)
- ignores most of the data (inaccurate)
- considers few possibilities (not thorough)
- many hypothesis formulations (subjective)
- difficult to state or explain (unworkable)

Part 2 continues with

Drop out, degraded DNA, kinship, DNA databases

http://www.cybgen.com/information/presentations/page.shtml

M.W. Perlin, "TrueAllele® interpretation of DNA mixture evidence", Keynote talk, *9th International Conference on Forensic Inference and Statistics*, Leiden University, The Netherlands, 20-Aug-2014.

M.W. Perlin, "TrueAllele® Casework", *Almost Everything You Wanted to Know About Probabilistic Software (But Were Afraid to Ask)*, *Promega's Twenty Fifth International Symposium on Human Identification*, Phoenix, AZ, 29-Sep-2014.

Cybergenetics

perlin@cybgen.com