How TrueAllele® Works (Part 1)

Cybergenetics Webinar
October, 2014

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

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
Visual User Interface VUIer™ Software
Database Server
Parallel Processing Computers
Interpret/Match Expansion
8 to 80 questions simultaneously

Bayesian probability model
posterior = \frac{\text{likelihood} \times \text{prior}}{\text{data}}

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

No calibration needed; learn from the data
Background noise

\[ d_i \sim N_i (\mu_i, \Sigma_i) \]

\[ \Sigma_i = \sigma^2 \cdot V_i \]

\[ \tau^2 \sim \text{Gam}(10, 500) \]

No analytical threshold needed; model the data

PCR variation

\[ d_i \sim N_i (\mu_i, \Sigma_i) \]

\[ \Sigma_i = \sigma^2 \cdot V_i \cdot \tau^2 \]

\[ \sigma^2 \sim \text{Gam}(10, 20) \]

No stochastic threshold needed; model the data

Mixture weight

Hierarchical variables

$k^{th}$ contributor

experiment data $d_{k,1}, d_{k,2}, \ldots, d_{k,N}$

locus weights $W_{k,1}, W_{k,2}, \ldots, W_{k,N}$

template weight $W_k$

prior probability $W_0$

PCR stutter

Relative amplification

Genotype pattern

\[ d_i \sim N_i(\mu_i, \Sigma_i) \]

\[ \mu_i = m_i \sum_{k=1}^{K} w_{i,k} g_{i,k} \]

\[ f_i = \begin{cases} f_i^2 & i = j \\ \frac{1}{2}f_i f_j & i \neq j \end{cases} \]


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 = \( \frac{Pr(\text{Next state})}{Pr(\text{Current state})} \)
Joint probability distribution

all the variables together in a high-dimensional space

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

contributor

genotype

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

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

objectively inferred genotype at locus 8 for contributor 2

TrueAllele® Casework: Match

ViewStation User Client
Visual User Interface VUler™ Software

Database Server
Parallel Processing Computers

Interpret/Match Expansion

2 hours for each question
The likelihood ratio

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"

<table>
<thead>
<tr>
<th>exclusion</th>
<th>inconclusive</th>
<th>inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>-12</td>
<td>-6</td>
<td>0</td>
</tr>
<tr>
<td>+6</td>
<td>+12</td>
<td></td>
</tr>
</tbody>
</table>

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)
Likelihood form

Hypotheses correspond to $H_p$ and $H_D$

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

Genotype expansion

TrueAllele can separate genotypes

$$\sum_{x \in G} \frac{Pr(d \mid X = x)Pr(X = x \mid H)}{Pr(d)}$$

Apply hypothesis

Reduce summation to just suspect's term

$$\frac{Pr(d \mid X = s)}{Pr(d)}$$
Bayes theorem

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{data}}
\]

Reorganize terms

\[
\frac{\text{posterior}}{\text{prior}} = \frac{\text{likelihood}}{\text{data}}
\]

Genotype form

Posterior over prior

\[
\frac{\text{Pr}(X = s \mid d_Q)}{\text{Pr}(X = s)}
\]

Visualize the LR

Ratio of genotype probabilities

![Graph showing probability distributions with ratios and evidence probability.](image)
Match form

Genotype match over coincidence

\[
\frac{\Pr(X = s|d_q) \Pr(Y = s|d_s)}{\Pr(X = s|\text{prior})}
\]

Q posterior \hspace{1cm} S posterior

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

\[
\sum_{x \in G} \frac{\Pr(X = x|d_q) \Pr(Y = x|d_s)}{\Pr(X = x|\text{prior})}
\]

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

\[ \begin{align*}
H_0 & : X=x, Y=y, Z=z \\
H_1 & : X=\?, Y=\?, Z=z
\end{align*} \]

\[ LR_{\text{unseparated}} = \frac{Pr(\text{“data”}|H_0)}{Pr(\text{“data”}|H_1)} \]

• "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® 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