DNA Identification: Mixture Weight & Inference

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

TrueAllele® Lectures
Fall, 2010

Cybergenetics

Mixture Weight: Uncertain Quantity

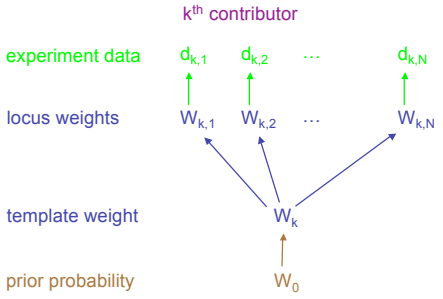
Infer mixture weight from STR experiments:

- quantitative peak data
- contributor genotypes

$Pr(W=w \mid data, G_1=g_1, G_2=g_2, \ldots)$
hierarchical Bayesian model

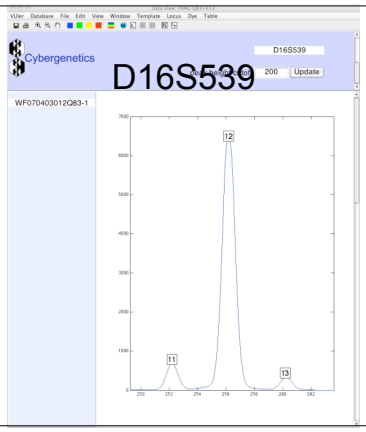
Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. Journal of Forensic Sciences. 2011;56(November):in press.

Mixture Weight Model



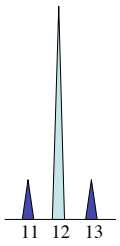
Experiment Estimate

$$w_k = \frac{\text{sum of peak heights from } k^{th} \text{ contributor}}{\text{sum of peak heights from all contributors}}$$



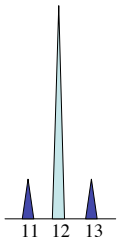
D16S539

Three Alleles



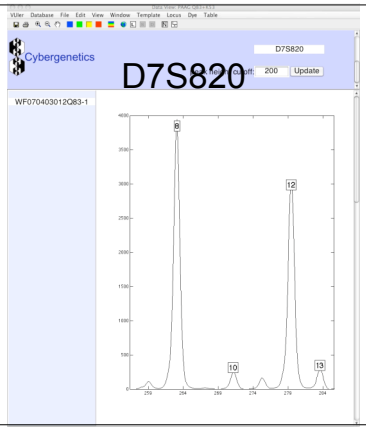
Allele	Quantity
11	500
12	6,750
13	250

Experiment Estimate



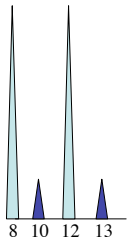
Allele	Quantity
11	500
12	6,750
13	250

$$\frac{500 + 250}{500 + 6{,}750 + 250} = \frac{750}{7{,}500} = 10\%$$



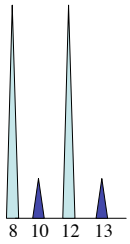
D7S820

Four Alleles



Allele	Quantity
8	4,000
10	250
12	3,000
13	250

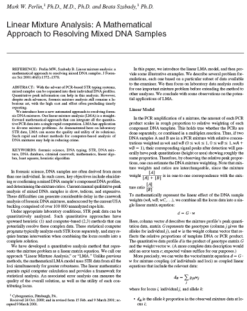
Experiment Estimate



Allele	Quantity
8	4,000
10	250
12	3,000
13	250

$$\frac{250 + 250}{4{,}000 + 250 + 3{,}000 + 250} = \frac{500}{7{,}500} = 6.7\%$$

Overlapping Alleles



Mark W. Perlin,[1] Ph.D., M.D., Ph.D. and Beata Szabady,[1] Ph.D.

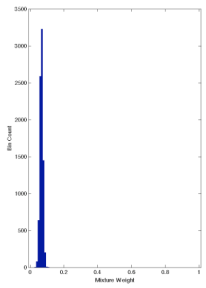
Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples

Template Average

mean $$\mu_k = \frac{1}{N}\sum_{n=1}^{N} w_{k,n}$$

variance $$\sigma_w^2 = \frac{1}{N}\sum_{n=1}^{N}\left(w_{k,n} - \mu_k\right)^2$$

Template Mixture Weight Probability Distribution



mean = 6.7%

standard deviation = 0.9%

Central Limit Theorem

- more data experiments for a template provide greater mixture weight precision
- double the precision by doing four times the number of experiments
- combine evidence from multiple experiments to obtain a more informative result

Probability Solution

interacting random variables

$w \mid d, g_1, g_2, \ldots$

$g_1 \mid d, g_2, w, \ldots$

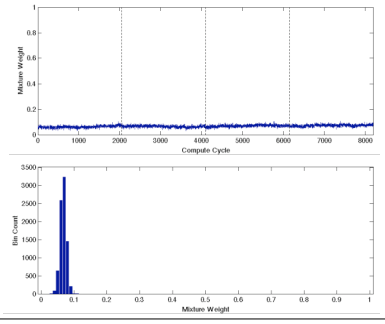
$g_2 \mid d, g_1, w, \ldots$

$z_i \mid d, g_1, g_2, w, \ldots$

find probability distributions by iterative sampling

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. J. American Statist. Assoc., 85:398-409.

Markov Chain Monte Carlo



Prior Probability

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

mixture weight $\mathbf{w} \sim Dir(\mathbf{1})$

DNA quantity $m_l \sim N_+\left(5000, 5000^2\right)$

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

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

$\psi^{-2} \sim Gam(1/2, 1/200)$

Joint Likelihood Function

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

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

$\mathbf{w}_l \sim N_{[0,1]^{K-1}}\left(\mathbf{w}, \psi^2 \cdot I\right)$

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

Posterior Probability

$$\Pr\{Q = x \mid d_{1,1}, d_{1,2}, \ldots, d_{1,z}, \ldots\} \propto \Pr\{Q = x\} \cdot \prod_{i=1}^{I} \Pr\{d_{i,j} \mid Q = x, \ldots\}$$

genotype

$$\Pr\{W = w \mid d_1, d_2 \ldots, d_J, \ldots\} \propto \Pr\{W = w\} \cdot \prod_{j=1}^{J} \Pr\{d_j \mid W = w, \ldots\}$$

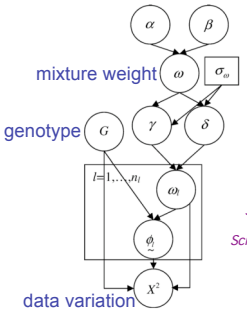
mixture weight

$$\Pr\{\sigma^2 = s^2 \mid d_1, d_2, \ldots, d_j, \ldots\} \propto \Pr\{\sigma^2 = s^2\} \cdot \prod_{j=1}^{J} \Pr\{d_j \mid \sigma^2 = s^2, \ldots\}$$

data variation

$$\Pr\{\tau^2 = t^2 \mid d_1, d_2, \ldots, d_j, \ldots\} \propto \Pr\{\tau^2 = t^2\} \cdot \prod_{j=1}^{J} \Pr\{d_j \mid \tau^2 = t^2, \ldots\}$$

Generally Accepted Method



James Curran. A MCMC method for resolving two person mixtures. *Science & Justice*. 2008;48(4):168-77.

Hierarchical Bayesian Model with MCMC Solution

- standard approach in modern science
- describes uncertainty using probability
- the "new calculus"
- replaces hard calculus with easy computing
- can solve virtually any problem
- well-suited to interpreting DNA evidence