Using Exact Likelihood Ratio (LR) Distributions for Probabilistic Genotyping Software Validation

Jennifer M. Bracamontes, MS, Erin E. Estus, MS, and Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

Abstract

After attending this presentation, attendees will understand a fast and easy approach to constructing exact LR distributions that help calculate accurate sensitivity and specificity error rates when validating probabilistic genotyping (PG) software.

This presentation will impact the forensic science community by showing a method for calculating precise LR distributions for sensitivity and specificity error rates. These distributions consider every possible reference genotype. They are helpful for PG software validation, and for establishing scientific reliability in the courtroom.

Testing DNA PG interpretation software is important in forensic science. Empirical testing ensures a method works as expected. Validation studies test the PG method on representative data sets, reporting likelihood ratio (LR) match statistics. These studies typically include sensitivity and specificity error rates. Sensitivity evaluates the inclusionary strength of true contributors to DNA. Specificity examines the ability of DNA evidence to statistically exclude non-contributors. The log(LR) number is used to measure sensitivity and specificity information. From LR distributions, false exclusion and false inclusion error rates can be immediately calculated.

To use PG software for DNA interpretation, applicable validation standards require sensitivity and specificity studies, as well as error rate determination. Legal admissibility standards encourage PG software validation and error rate calculation – both of which are Daubert prongs.

LR distributions for examining system sensitivity and specificity can be developed either by limited sampling or by exact convolution. Both calculation methods produce distributions of log(LR) statistics. The sampling method approximates exact log(LR) distributions by comparing a set of evidence genotypes with a set of randomly sampled reference genotypes. Sampling is incomplete, only testing a thousand (10^3) or so references, which is a miniscule fraction of possible genotypes. Calculating by sampling is tedious in validation; comparing a thousand (10^3) evidence genotypes with a thousand references entails a million (10^6) match statistic calculations.

The exact method accurately calculates log(LR) distributions for evidence genotypes. The requisite convolution can have any preset numerical resolution. The convolution approach is complete, with one distribution accounting for all (e.g., 10^{24}) possible reference genotypes [1]. The calculation is fast; a hundred genotype distributions can be constructed in one second. Many evidence genotype distributions can be averaged to represent a set of genotypes in one composite distribution. This composite feature is highly useful for validation studies.

We assessed both the sampling and convolution methods on the same DNA laboratory's mixture validation data set. We constructed contributor (posterior evidence probability weighted) and non-contributor (prior population probability weighted) genotype log(LR) distributions. We calculated error rates from these distributions to measure sensitivity and specificity. The data came from single source and DNA mixture samples.

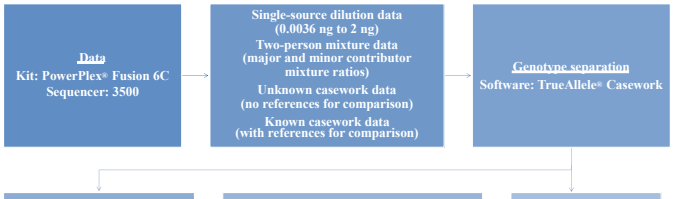
Sampled contributor distributions were limited to the provided matching references, which severely under sampled reference genotypes, and gave limited false exclusion rates. But the exact distributions spanned the entire range of expected log(LR) match values, and provided accurate false exclusion probability for the tested data sets.

Non-contributor distributions were calculated by limited sampling and exact convolution. The distributions from both methods appeared qualitatively similar. But more random reference sampling – and time – was needed to better approximate the true distribution. Building exact convolved distributions was far faster than using sampling.

Using exact convolution, rapid calculation of sensitivity and specificity from the log(LR) distributions on multiple datasets sped up the PG validation, relative to sampling methods. Human operator time was significantly reduced. User interfaces for noncontributor, contributor, and composite distributions simplified PG validation.

Calculating exact composite log(LR) distributions by convolution – and determining associated error rates on genotype subsets – improves on LR sampling methods. Convolution construction is easy, fast, complete, and accurate. The method lets forensic scientists readily determine error rates for PG methods of interpreting complex DNA evidence. Moreover, the exact convolution LR distribution construction approach has applicability to other forensic subdisciplines, providing accurate error rate determination for reliable scientific validation and reporting.

Materials and Methods



Results

Single-source exact composite distributions

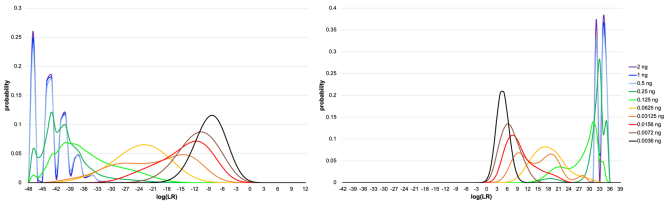


Figure 1. The line graphs show non-contributor (left) and contributor (right) exact composite distributions for a single-source data dilution set. Each template amount is shown in its own color.

Two-person exact mixture composite distributions

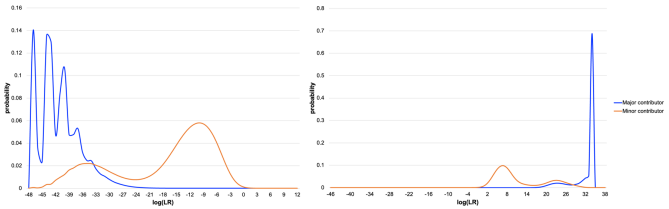


Figure 2. The line graphs show non-contributor (left) and contributor (right) exact composite distributions for the two-person mixture data. Major (blue) and minor (orange) contributor LR results are shown.

Casework exact mixture composite distributions

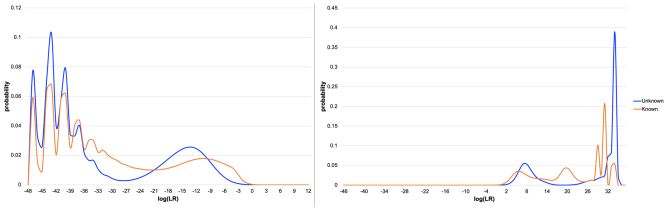


Figure 3. The line graphs show non-contributor (left) and contributor (right) exact composite distributions for the casework mixture data. Unknown (blue) and known (orange) LR results are shown.

Sensitivity comparison

	Two-person mixtures				Casework mixtures			
	Sampling		Exact		Sampling		Exact	
	Major	Minor	Major	Minor	Unknown	Known	Unknown	Known
false exclusions	0	5.5782E-01	1.5335E-18	8.0343E-04			2.6392E-05	2.7339E-04
minimum	13.9930	-37.6844	-44.0160	-45.0990	7.6587	1.2293	-45.1220	-46.0000
mean	31.0881	1.6376	32.8611	12.2650	26.4841	20.1525	25.5282	22.0047
maximum	33.3251	24.2907	37.3340	55.5590	35.9918	33.2207	48.1220	50.7810
std deviation	4.6321	12.3219	3.9505	7.8874	12.2794	10.1663	11.7613	10.0571

Table 1. Two-person mixture and known casework sensitivity statistics for both exact and sampled methods. These statistics include false exclusion probability, and log(LR) minimum, maximum, and standard deviation values. The table shows these sampling statistics as KL values for unknown casework data. Exact probability distributions encompass the entire range of minimum to maximum values for expected LR results.

Specificity comparison of exact and sampled methods

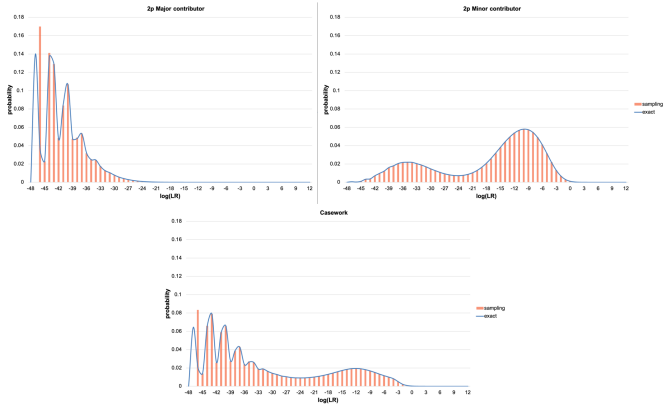


Figure 4. The exact composite non-contributor distribution (blue line) compared with the sampled non-contributor match distribution (orange histogram) for three different mixture data sets. The co-ancestry coefficient was set to 0%.

	Two-person mixtures					
	Major		Minor		Casework mixtures	
	Sampling	Exact	Sampling	Exact	Sampling	Exact
false inclusions	0	4.5863E-11	1.5565E-03	1.5776E-03	3.5522E-04	3.5429E-04
minimum	-46.0000	-46.0000	-46.0000	-46.0000	-46.0000	-46.0000
mean	-40.4706	-40.5076	-17.1624	-17.2023	-30.8114	-30.8326
maximum	-12.5590	37.3330	4.6503	55.4890	4.4640	50.7010
std dev	4.3980	4.4074	11.3306	11.3372	13.0654	13.0571

Table 2. Specificity statistics include false inclusion probability, as well as log(LR) minimum, mean, maximum, and standard deviation. These statistics are shown for both the exact and sampling methods. Exact log(LR) distributions encompass the entire range of minimum to maximum values.

Conclusions

- The sensitivity and specificity summary statistics of the sampled distributions converge to the exact distributions [1]. LR distribution similarity cross-validates the independently derived exact and sampled approaches. Statistics from sampled distributions have been published in earlier peer-reviewed probabilistic genotyping validation studies [2].
- Exact LR distributions are more accurate than sampled ones. Sampled distributions are only approximate, due to their limited number of genotype comparisons. Thus, exact method sensitivity and specificity error rates are more accurate.
- With a positive co-ancestry coefficient, exact LR distribution calculation is very fast; exact distribution calculation takes seconds. With sampling under positive theta, computing and collating many genotype comparisons can take days.
- As DNA amount for a contributor decreases, so too does genotype information [3]. The mean of the sensitivity and specificity log(LR) distributions will then tend to zero.
- Exact LR distributions can be rapidly calculated to accurately assess probabilistic genotypes produced by PG computer DNA software. The exact approach provides a speed, accuracy, and workflow improvement over sampled methods.

References

[1] Perlin, M.W. Efficient construction of match strength distributions for uncertain multi-locus genotypes. *Heliyon*, 4(10):e00824, 2018.

[2] Perlin, M.W., Hornyak, J.M., Sugimoto, G., and Miller, K.W.P. TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*, 60(4):857-868, 2015.

[3] Bauer, D.W., Butt, N., Hornyak, J.M., and Perlin, M.W. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *Journal of Forensic Sciences*, 65(2):380-398, 2020.

Cybergenetics

Cybergenetics © 2023