

Article 451
Articles Title

Mathematics in Forensic Genetics

Keywords – Forensic mathematics, DNA identification, product rule, population structure, Bayes' Theorem, likelihood ratio

Author – Brenner, Charles H.
Consulting in forensic mathematics
<http://dna-view.com>

Institution – (Visiting) School of Public Health, University of California, Berkeley

Country – USA

Mathematical model of tribunal decision- making

Genetic evidence – DNA evidence – is preeminent, among evidence that arises in court, in offering scope for mathematical treatment. The objective character of DNA typing and the developed theory for the principles of distribution of genetic traits in populations (“population genetics”) offer opportunity to attach meaningful numbers to explain the strength of DNA evidence.

The underlying issue is to quantify evidence in a way that gives a rational basis to decide how probabilities change as evidence is assimilated. (“Evidence” here means information. For “evidence” in the police sense – physical evidence from which information may be available – the term “evidence sample” is unambiguous.)

Likelihood ratio

Evidence gives support to one hypothesis rather than another if it is a more likely consequence of the one hypothesis than of the other. The degree of support is quantified by the *likelihood ratio*, defined as the ratio of two probabilities of the same event (i.e. the appearance of the evidence) under different hypotheses (i.e. prosecution theory vs. defense theory).

From this formulation it is easy to see that the likelihood ratio by which the combination of several independent pieces of evidence favors one hypotheses over another, is the product of the likelihood ratios of the pieces of evidence individually. The reason is that likelihood ratios inherit from probabilities the multiplication rule for combining independent events.

Likelihood ratios can stand alone to quantify degrees of support, but their utility is most easily understood by noting how they arise in Bayes' Theorem. For this purpose it is most convenient to make a formulation in terms of *odds*.

Odds means the ratio of the probabilities of two mutually exclusive events, such as the prosecution theory H_1 and the defense theory H_0 of a crime. H_0 could be simply $\sim H_1$ (\sim means “not”), but

sometimes we take the slightly more general point of view that $\mathbf{H}_0 \subseteq \sim \mathbf{H}_1$.

Consider a criminal dispute with prosecution theory \mathbf{H}_1 that the defendant committed the crime. Various bits of scientific evidence $\mathbf{E}_1, \mathbf{E}_2, \dots$ will be introduced, but before considering them assume that the jurors have some private notion of $P(\mathbf{H}_1)$, the “prior probability” that the defendant is guilty, and consequently of the prior odds $\mathcal{O}(\mathbf{H}_1)$, defined as $P(\mathbf{H}_1)/P(\sim \mathbf{H}_1) = P(\mathbf{H}_1)/(1-P(\mathbf{H}_1))$. The posterior odds $\mathcal{O}(\mathbf{H}_1 | \mathbf{E}_1, \mathbf{E}_2, \dots)$ are then given by

$$\mathcal{O}_{\text{posterior}} = \mathcal{O}_{\text{prior}} \cdot \mathbf{L}_1 \cdot \mathbf{L}_2 \cdot \dots, \quad (1)$$

where $\mathbf{L}_i = P(\mathbf{E}_i | \mathbf{H}_1) / P(\mathbf{E}_i | \sim \mathbf{H}_1)$, by an application of Bayes’ Theorem.

This means that an initial belief $\mathcal{O}_{\text{prior}}$ in a certain proposition, upon presentation of new independent pieces of evidence with likelihood ratios $\mathbf{L}_1, \mathbf{L}_2$, etc. logically must be modified to a new or posterior assessment as given by the formula.

Obviously “prior” and “posterior” are relative terms. Any product of initial factors on the right side of (1) can be regarded as the odds prior to consideration of yet further factors. The only exception is the empty product; there seems to be no way to reduce the right hand side of (1) purely to a product of likelihood ratios. Prior odds seem to have an inevitable element of subjectivity.

The remaining factors may be more or less subjective. If a criminal defendant’s DNA type is found at the crime scene, then it is a fairly objective matter to compute the likelihood ratio comparing the chance for such a finding to occur under the prosecution’s claim that it’s the suspect’s DNA, versus the chance of a coincidental DNA match between the suspect and the real contributor.

On the other hand, testimony by a witness who claims to have seen the defendant at the crime scene is much harder and more controversial to quantify. In principle the judge or jury can try to estimate a likelihood ratio for witness testimony. But thinking about likelihood ratios does not change the inevitable fact that assessing testimony is a subjective matter, from which the juror can get no particular help from mathematics.

Sometimes the lay juror is impatient with a likelihood ratio for (say) DNA evidence, and would prefer to hear the *probability* that the suspect left the stain, analogous to fingerprint claims of identity or the claim of an eyewitness that his identification is “99% certain.” That would be backward. Instead, the eyewitness’ posterior probability assertion must be taken with a grain of salt, and re-cast into a likelihood ratio: divide (a) the probability that the witness would make such a claim if it were true, by (b) the probability that the witness would make the claim assuming hypothetically it to be false. The result depends on various subjective factors such as the demeanor of the witness including the words chosen, but clearly bears no particular numeric relationship to the number 99%. The fingerprint claim is based on the approximation – perhaps practical but logically tenuous – that the likelihood ratio for a fingerprint match is infinite.

“Presumed innocent until proven guilty”

If this cliché of American jurisprudence (or equally the German: “Bis zum Beweis der Schuld gilt die Unschuldsvermutung” – Until guilt is proven, the presumption of innocence is valid.) means the prior odds $\mathcal{O}_{\text{prior}}$ should be taken to be exactly 0, then from (1) obviously no amount of evidence \mathbf{L}_i will ever increase the odds above 0. A more malleable interpretation is “that the mere fact of an indictment

... is not evidence and may not be considered supportive of guilt” (Faigman *et al*, p 759). (In Japanese: “Utagawashiki wa bassezu” – The suspect should not be punished for being suspected.) Thus, it seems fair that even before hearing any evidence, the jury can view the defendant with the same minuscule but non-zero suspicion with which they would regard any anonymous person.

“Beyond a reasonable doubt”

If the final assessment $\mathcal{O}_{\text{posterior}}$ is sufficiently large then the jury is entitled to rule that the matter before it has been proven, even though mathematically there is always some amount of doubt. At any rate, the posterior odds threshold for proof must be finite, just as the prior odds must be non-zero, in order to have any scope for a mathematical treatment. The actual threshold value is no doubt vague, but must depend on the statement in the law (i.e. “preponderance of the evidence”, “beyond a reasonable doubt”, etc.) and on the conscience and understanding of each juror. Nonetheless we can assume that in concept a threshold value

$$\text{DNA odds evidential threshold} = \text{posterior odds threshold} / \text{prior odds}$$

exists, by solving (1), elusive though its value may be.

DNA matching

A typical DNA case involves the comparison of two samples — an unknown or evidence sample, such as semen from a rape, and a known or reference sample, such as a blood sample from a suspect.

If the DNA profile obtained from the two samples are indistinguishable (they "match"), that of course is evidence for the court that the samples have a common source — in this case, that the suspect contributed the semen.

How strong is the evidence? If the DNA profile consists of a combination of traits that figure to be extremely rare, the evidence is very strong that the suspect is the contributor. To the extent that the DNA profile is not so rare, it is easier to imagine that the suspect might be unrelated to the crime and that he matches only by chance.

For purposes of formulating a likelihood ratio to measure the evidence against the suspect, the simplest point of view is to regard the crime stain DNA profile S as fixed, and the evidence, \mathbf{E} , as the fact that the suspect also has S . Then if the prosecution theory, \mathbf{H}_1 , is that the suspect is the donor; $P(\mathbf{E}|\mathbf{H}_1)=1$. If the defense assumes \mathbf{H}_0 , that some person unrelated to the suspect is the donor, then $P(\mathbf{E}|\mathbf{H}_0)=m$ =probability that a randomly selected person would have profile S . Then the likelihood ratio favoring \mathbf{H}_1 over \mathbf{H}_0 is $1/m$.

DNA profile probability

Therefore it is essential to have some idea as to the probability m that a match would occur by chance. Classically m is estimated as the expected population frequency of the stain profile. It is easiest to illustrate how the frequency is computed for the typical situation of a battery of co-dominant autosomal genetic loci by an example:

DNA Profile		Allele frequency from population survey			P(E H ₀)≈ estimated genotype frequency for locus		
Locus	Alleles	times allele observed	size of survey (2N, N=people)	frequency		formula	number
CSF1PO	10	109	432	p=	0.25	2pq	0.16
	11	134		q=	0.31		
TPOX	8	229	432	p=	0.53	p ²	0.28
	8						
THO1	6	102	428	p=	0.24	2pq	0.07
	7	64		q=	0.15		
vWA	16	91	428	p=	0.21	p ²	0.05
	16						
expected profile frequency=							0.00014

The allele 10 at the locus CSF1PO was observed 109 times in a population sample of 432 chromosomes (216 people). Therefore it is reasonable to estimate that there is a chance $p=0.25$ that any particular CSF1PO allele, selected at random, would be a 10. Similarly, the chance is about $q=0.31$ for a random CSP1PO allele to be 11. Prior to typing the suspect, if we assume that he is not the donor of the crime stain then the chance that he will match its CSF1PO genotype is equal to the chance he received a CSF1PO 10 allele and an 11 allele from his parents. The chance to receive 10 from his mother and 11 from his father is pq , and to receive 11 from mother and 10 from father is another pq , so the probability to be 10,11 by chance is $2pq$. Hence about 16% of people have the 10,11 genotype at the CSF1PO locus.

At the TPOX locus, since both alleles are the same there is only one term — pp or p^2 , which represents the combined probability of inheriting the allele 8 from each parent. Hence about 28% of people have the same TPOX genotype as does the evidence. It is to be expected that the proportion of TPOX 8,8 people is still 28% even if attention is restricted only to people who have a particular CSF1PO genotype such as 10,11. Therefore the chance for a person to have the combined genotype in the two loci is 28% of 16% — about 4%.

The calculations for the THO1 and vWA loci are similar, and taking them into account whittles the overall chance for a random person to have the combined genotype from 4% down to about 1/7000. Equivalently, the likelihood ratio or *matching odds* is 7000.

product rule

In summary, the *matching chance* for a particular multiple-locus genotype is obtained by multiplication — by multiplying together the frequencies of the per-locus genotypes, which is to say, by multiplying together the frequencies of all the individual alleles and including in addition a factor of 2 for each heterozygous locus. This method is called the product rule.

verbal explanation

In the example case, the expected profile frequency is 0.00014 or about 1/7000. Therefore, a summary of the evidence is that

Either the suspect contributed the evidence, or an unlikely coincidence happened — the once-in-7000 coincidence that an unrelated person would by chance have the same DNA profile as that obtained from the evidence.

A shorter summary is "common source, or unlikely coincidence."

Fallacies

"Prosecutor's fallacy"

A dangerous paraphrase of the above is the statement

There is a 1/7000 chance that someone other than the suspect could have left the observed evidence sample.

If "someone" means "one particular random untyped person" and "could have left" means merely "has the appropriate genetic profile," it is correct. But the words are ambiguous. Most journalists – and putatively some prosecutors – misinterpret "someone" to mean "the collection of all people other than the suspect", and end up saying "There is only a 1/7000 chance than anyone other than the suspect left the crime scene DNA," which is unreasonably presumptuous. Other evidence in the case (even if the "suspect" is a dead woman, or even if the suspect was filmed in the act) would be irrelevant. In effect, the fallacious prosecutor imputes prior odds of 1:1.

As (1) shows, DNA evidence alone cannot be a proof. Some additional information – some prior odds – are necessary. However, the amount of additional information that is necessary might be a very small amount. For example, add to the DNA matching evidence (of 7000 to one) the mere knowledge that the suspect was arrested before his DNA type was known; that may be enough.

"Defense attorney's fallacy"

Sometime the defense tries to minimize the impact of 7000 to one matching odds by saying, "Since that means that there are hundreds of men in this city with the same profile, there is only one chance in several hundred that my client is the donor of the semen." That might be good logic if the other evidence suggests that every man in the city had the same access to the crime scene as did the suspect; not otherwise. In terms of (1), the defense attorney implicitly assumes that the prior odds are equal for all men in the city.

Laboratory error

Besides "common source," and "unlikely coincidence," a third possible explanation for a match between suspect and evidence is error. The chance of an error that would cause a spurious match — mishandling the evidence, PCR contamination — although unquantifiable, is probably very small. Nonetheless, it seems a plausible guess that the chance of error is often much larger than the extremely small random match chances (such as 1 in 10^{15}) that occur, so it may be more realistic and

more fair in such cases to say "same source, or (unlikely) error" rather than to say "same source, or unlikely coincidence."

Mathematically speaking the point is to refine the ideal point of view

The evidence is **E**, that the crime stain has DNA profile **S** and the suspect matches it, to the more precise statement

The evidence is **E'**, that the laboratory claims **E**.

Such statistical evidence as is available gives little help in quantifying the distinction between these two, that is, in quantifying the chance of making the "false match" error of reporting that the DNA profiles for the crime and reference samples are similar when they are actually different. The arguments are mainly in the specific realms of biological technology, laboratory procedure, and legal procedure. However, there is one mathematical point to make. Suppose there is a record of 1000 relevant comparable assays that the laboratory performed correctly. Then the defense may argue that even this excellent record gives no reason to believe that the long-run error rate is less than 3/1000, because statistically the observed good historical performance is within reason (95% confidence) even if the true rate is 3/1000. That is true. Statistics alone can prove little; normally a persuasive argument requires in addition a model – i.e. perhaps the testing protocol makes false-match errors inherently unlikely.

Limitations

The method of calculation described above makes several idealized assumptions. Since those assumptions are always more or less false it is important to be aware of them. For a more thorough discussion of these issues see Crow 1996, Morton 1992, Balding & Nichols 1994.

relatives

The analysis above assumes that if suspect is not the donor, he is unrelated to the donor. But common sense shows immediately that if the suspect can make a case that a relative of his, especially his brother, is the donor, then that goes a long way towards explaining away the coincidental similarity between the suspect and the evidence. The effect of augmenting the defense hypothesis from **H**₀ to include also **B**, "the culprit is a brother to the suspect," is to change the likelihood ratio from $1/P(\mathbf{E}|\mathbf{H}_0)$ to $1/P(\mathbf{E}|\mathbf{H}_0 \cup \mathbf{B}) = 1/(aP(\mathbf{E}|\mathbf{H}_0) + bP(\mathbf{E}|\mathbf{B}))$, where a and b are the prior probabilities for **H**₀ and **B**, normalized so that $a+b=1$. The defense tactic in offering the "brother defense" thus involves a trade-off: a reduced burden in explaining away the coincidence of DNA similarity since $P(\mathbf{E}|\mathbf{B}) \gg P(\mathbf{E}|\mathbf{H}_0)$, but a new burden of suggesting possible guilt by a brother, of arguing $b > 0$. Of course the same argument is available for possible relatives other than brother.

For an autosomal locus with homozygosity= h , the average matching chance between two unrelated people is about $2h^2$ (Brenner & Morris 1989). The average matching chance between brothers is $(1+h)^2/4$. Therefore assuming a battery of 13 loci, each with $h \approx 1/5$ (typical for forensic practice as of this writing), the likelihood ratio for the value of the DNA match exceeds 10^{13} if the jury believes $b=0$, but drops linearly with b to under 10^6 when $b=1$. Thus even a small value of b is possibly significant, although the benefit to the defense probably comes more from gaining moral stature by

impugning excessive claims by the prosecution, than from any innate exculpatory virtue of an adverse likelihood ratio of 10^6 .

independence

The application of the product rule assumes independent events, i.e. presumes the relevant loci and population to be in Hardy-Weinberg equilibrium and linkage equilibrium, which is to say that the alleles are randomly distributed among the population as they would be if mating were random.

These population genetic concepts have been found to hold to a reasonable degree of accuracy for major populations and typical forensically-used loci. However, there are two sorts of objections that arise against the independence assumption.

Population substructure

For mixed populations and inbred populations the product rule is not as accurate. Therefore it is usual to consider the various races or ethnicities separately. That covers the obvious cases, but an argument can be made that it may not be sufficient.

Suppose, for example, that contrary to common wisdom and unnoticed for centuries, U.S. Caucasians consist of two equal non-interbreeding sub-populations – perhaps Republicans and Democrats. The vWA 16 allele, with a frequency of $p=0.21$ among all Caucasians, thanks to genetic drift would have different frequencies, $r=p+\delta$ and $d=p-\delta$, within the two subpopulations. Consequently, the fraction of U.S. Caucasians who are homozygous for the 16 allele is $(r^2+d^2)/2=p^2+\delta^2$ so the product rule estimate of p^2 is too low. To the extent that the product rule is inaccurate, the error on average, as here, works against the suspect, unfairly exaggerating the strength of the evidence. Such is the effect of *Hardy-Weinberg disequilibrium*.

Rather than rely on a hand-waving argument that such errors are insignificant, a more forceful procedure for forensic debate is to compensate for them. If θ (the “inbreeding coefficient”) is the mean probability of identity by descent for two alleles selected randomly from the population, the probability of a homozygous type is $p^2(1-\theta)+p\theta$ (Wright 1930). Since θ is small – $\theta < 1\%$ for major populations (Morton 1992) – the effect of revising the product rule is to surrender a mere factor of about $1+\theta/p$ in the suspect’s favor (typically 1.05 – very rarely¹ as much as 2 – for each homozygous locus in the DNA profile).

Balding and Nichols (1994) take the argument further. The population may not only be substructured, but the suspect and the true contributor, if different people, are particularly likely to belong to the same subpopulation. They supply formulas in terms of θ that compensate for the expected increased chance for a random individual to have a genotype, given that it has already been observed in his subpopulation.

Linkage

The possibility of population substructure suggests not only that allele probabilities at a locus may

¹ not because rare types are rare – collectively rare types may be common. But homozygosity for a rare type is a rare event.

not be independent, but also that allele probabilities at different loci may be statistically linked. *Linkage association* (or *gametic disequilibrium*) is a general term meaning that the product rule between loci fails, for whatever reason. Various computations including statistical tests of population data show that at worst the extent of association cannot be very much. However, unlike Hardy-Weinberg disequilibrium, there are presently no formulas and there is no common practice to compensate for linkage association when computing profile probabilities.

Linkage can also arise from a biological mechanism. If two loci are close neighbors on the same chromosome, they will be in *linkage disequilibrium*, meaning that they tend to be inherited as a unit, with a consequent tendency to linkage association. Through the passage of generations, recombination dilutes the association. In practice the impact of linkage disequilibrium on the accuracy of the product rule for computing evidential significance is negligible even for rather close pairs of loci. It is usual to be cautious and avoid such pairs altogether.

Even alleles at loci on different chromosomes hypothetically could be associated if there is a peculiar combination of selective pressures that favors a particular allele A at one locus preferentially in the presence of some allele X at the other locus. Forensic loci are presumed to be non-functional and hence immune to selection. However, they are sometimes within introns of genes so there is a theoretical possibility that they may be indirectly selected by hitchhiking. The effect on profile probabilities of a hitchhiking link figures to be weak though (Schneider 1997), and it's hard to find an instance of the necessary kind of epistasis.

Additional topics

Some further topics in DNA identification that provide interesting scope for mathematical analysis include analysis of an evidence sample that consists of a mixture from several people (Weir *et al* 1997, but see Gill *et al* 1998), how to analyze when the suspect is found through a database search, and how to analyze relationship cases including paternity and missing bodies (Brenner 1997, Krawczak 2001 in this encyclopedia).

References

1. Balding DJ, Nichols RA (1994),
[DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands](#)
Forensic Science International 64:125-140
2. Brenner CH (1997),
[Symbolic Kinship Program](#)
Genetics 145:535-542
3. Brenner CH, Morris J (1990),
[Paternity Index Calculations in Single Locus Hypervariable DNA Probes: Validation and Other Studies](#)
<http://dna-view.com/promeg89.htm> in Data Acquisition and Statistical Analysis for DNA Typing Laboratories, Proceedings for The International Symposium on Human Identification 1989, pp21-54, Promega Corporation, Madison, Wisconsin

4. Crow J, National Research Council (1996),
The evaluation of forensic DNA evidence
National Academy Press
5. Ewens WJ (1972),
The sampling theory of selectively neutral alleles
Theoretical population biology **3**:87-112
6. Faigman D, Kaye DH, Saks M, Sanders J eds. (1997),
Modern Scientific Evidence: The Law and Science of Expert Testimony
vol. 1, sec. 19-1.5, pp. 759-60
7. Gill, P; Sparkes, R., Pinchin, R; Clayton, T; Whitaker J., Buckleton J (1998),
Interpreting simple STR mixtures using allele peak areas
Forensic Science International **91**:41-53
8. Morton NE (1992),
Genetic structure of forensic populations
Proceedings of the National Academy of Science USA **89**:2556-2560
9. Schneider, PM (1997),
Basic issues in DNA typing
Forensic Science International **88**:17-22
10. Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton J (1997),
Interpreting DNA mixtures
Journal of Forensic Science **42**(2)213-222
11. Wright S (1930),
Evolution in Mendelian populations
Genetics **16**:97-159

Further reading

12. Balding David J, Donnelly Peter (1995),
Inference in Forensic Identification
Journal of the Royal Statistical Society A **158** Part 1, 21-53.
13. Balding DJ (1999),
When can a DNA profile be regarded as unique?
Science & Justice, **39**(4)257-260.
14. Brookfield JFY (1995),
Statistical issues in DNA evidence
Electrophoresis **16**:1665-1669

15. Crow JF (1989),
[Twenty-five Years Ago in Genetics: The Infinite Allele Model](#)
Genetics **121**:631-634
16. Dawid AP, Mortera J (1996),
[Coherent Analysis of Forensic Identification Evidence](#)
Journal of the Royal Statistical Society B **58** (2) 425-443
17. Evett IW, Foreman LA, Jackson G, Lambert JA (2000),
[DNA profiling: A Discussion of Issues Relating to the Reporting of Very Small Match Probabilities](#)
Criminal Law Review, May 2000, 341-355
18. Gale, JS (1990),
[Theoretical Population Genetics](#)
Unwin Hyman Ltd, London
19. Lewontin RC, Hartl DL (1991),
[Population genetics in forensic DNA typing](#)
Science 20 Dec:1745-1750
20. Morton NE (1995),
[Alternative approaches to population structure](#)
Genetica 96:139-144
21. Palmirotta R, Verginelli R, Cama A, Mariani-Costantini R, Frati L, Battista P (1998),
[Origin and gender determination of dried blood on a statue of the Virgin Mary](#)
Journal of Forensic Science **42**(2):431-434
22. Weir BS (1992),
[Population genetics in the forensic DNA debate \(review\)](#)
Proceedings of the National Academy of Science USA Vol 89 11654-11659
23. Weir BS, Cockerham CC (1984),
[Estimating F-statistics for the analysis of populations structure](#)
Evolution 38(6):1358-1370