Monte Carlo Bayesian Interpretation of STR mixtures

Donald I. Promish
68 Richardson Street
Burlington, Vermont
05401-5026
U.S.A.

DonaldPromish@cs.com

Abstract: This is an essay on the hypothesis that a suspect profile has been contributed to an STR mixture. One type of result is, for example, that the maximum random "match" likelihood for a suspect profile having 13 loci, of which 3 are homozygous, is $(0.38^3 \times 0.19^{10} =) \ 3.4\times10^{-9}$. Consequently, the minimum likelihood ratio regarding the hypothesis that the suspect contributed to the mixture is 3.0×10^{8}.

Monte Carlo Interpretation of STR mixtures

Donald I. Promish

This article is an attempt to deal with a problem which appears to underly STR mixture interpretation. The approach has its origins in detection theory, particularly as it relates to the detection of targets in the presence of extraneous signals. (See, for example, [1][1].)

At issue is the hypothesis that a suspect profile has been contributed to an STR mixture.

The evidence related to the hypothesis is that all of the alleles in the profile, as defined above, appear in the mixture. The profile has N complete loci, and n of those loci are homozygous. A "complete" locus possesses a pair of alleles. In the context of this essay, a locus possessing an unpaired allele is uninformative, because information can be gotten only from the difference (which may be zero) between members of a pair.

Let the expression (uv) denote an allele pair at any locus of a profile. Let u denote the length of the shorter allele; let v denote the length of the longer allele, if the locus is heterozygous. Otherwise, u = v. Then the difference between the alleles, which is denoted $|\Delta uv|$, is found by subtracting u from v; that is, $|\Delta uv| = v - u$. For example, suppose STR locus D18S51 has alleles 14 and 17. Then $|\Delta uv|$ for this locus is (17 - 14 =) 3.

Also, let the expression p(uv) denote the elementary probability that the allele pair (uv) will occur, at random, at the locus. Further, let the expression (u and v) denote the appearance of the alleles u and v in a forensic STR mixture at the same locus and let P(u and v) denote the probability that (u and v) occurs at random in the mixture.

The elementary probability p(uv) can be obtained with the Monte Carlo Bayesian (MCB) identification method as applied to STR profiles, by positing a single-locus "profile", (uv), and setting the prior probability of identity to 0**.**5 . [The MCB method is explained in the **Appendix**. The related software is freely available from the author.] As mentioned above, the MCB method shows that the random occurrence probability of any STR allele pair (uv) depends only on the absolute difference, $|\Delta(uv)|$, between its members, and not their actual sizes. Thus, random occurrence probabilities p(uv) can be tabulated for as many values of $|\Delta(uv)|$ as is necessary. Table 1 gives the values of p(uv) for the $|\Delta(uv)|$ values {0, 1, 2, ..., 10}.

$\lvert\Delta(uv)\rvert$	p(uv)
0	$1.1(3) \times 10^{-1}$*
1	$1.1(0) \times 10^{-1}$*
2	1.9×10^{-2}
3	4.1×10^{-3}
4	1.6×10^{-3}
5	8.3×10^{-4}
6	4.3×10^{-4}
7	2.8×10^{-4}
8	2.2×10^{-4}
9	1.7×10^{-4}
10	1.5×10^{-4}

* Third figures are shown in parentheses merely to demonstrate monotonicity.

Table 1. Probability of random occurrence of STR allele pair (uv), p(uv), as a function of the absolute difference, $\lvert\Delta(uv)\rvert$, between the alleles.

Now suppose that alleles (a, b, c, d, ...) appear at a locus of an STR mixture. Suppose, also, that a known suspect has the allele pair (ab) at that locus, as well as similarly "matching" pairs at all other loci. Let the probability that (a and b) occurs at random in the mixture if the suspect pair (ab) is heterozygous (i.e., $a \neq b$) be denoted by P(a and b | $a \neq b$). Then

$$P(a \text{ and } b \mid a \neq b) = p(ab) + [p(aa) + p(ac) + p(ad) + ...] \times [p(bb) + p(bc) + p(bd) + ...], \quad (1)$$

where the elementary probabilities on the right-hand side of the equation cover all possible ways that allele a and allele b can enter the mixture. In addition to (ab) alone, for example, allele a and allele b could randomly enter the mixture by way of contributions (ac) and (bd), with probability p(ac)×p(bd).

If, however, the suspect pair (ab) is homozygous (i.e., a = b), then the alleles a and b are identical. Thus, the probability, P(a and b|a = b), that (a and b) occurs at random in the mixture, reduces to:

$$P(a \text{ and } b \mid a = b) \equiv P(a) = [p(aa) + p(ac) + p(ad) + ...] . \quad (2)$$

It is possible, using Table 1, and equations (1) and (2), to calculate the probability of a random "match" between a specific STR profile and an STR mixture. However, it may be useful merely to estimate the <u>maximum</u> likelihood of such an event. For a single homozygous locus, the maximum value is 0.38 . For a single heterozygous locus, the maximum value is 0.19 .

Let N be the number of loci in the profile and let n be the number of homozygous loci among them ($0 \leq n \leq N$). Then the <u>maximum</u> random "match" likelihood is:

$$P_{max} \text{ (random "match"; N loci, n homozygous)} = 0.38^{n} \times 0.19^{(N-n)} . \quad (3)$$

As an example, the maximum random "match" likelihood for an individual profile having 13 loci, of which 3 are homozygous, is ($0.38^3 \times 0.19^{10}$ =) 3.4×10^{-9}. Consequently, the minimum likelihood ratio regarding the hypothesis that an individual contributed to the mixture is 3.0×10^{8}. Bayes's theorem for this likelihood ratio is plotted in Figure 1, below. It gives, for any prior ("input") value, the minimum ("output") probability that an individual contributed that profile to the mixture.

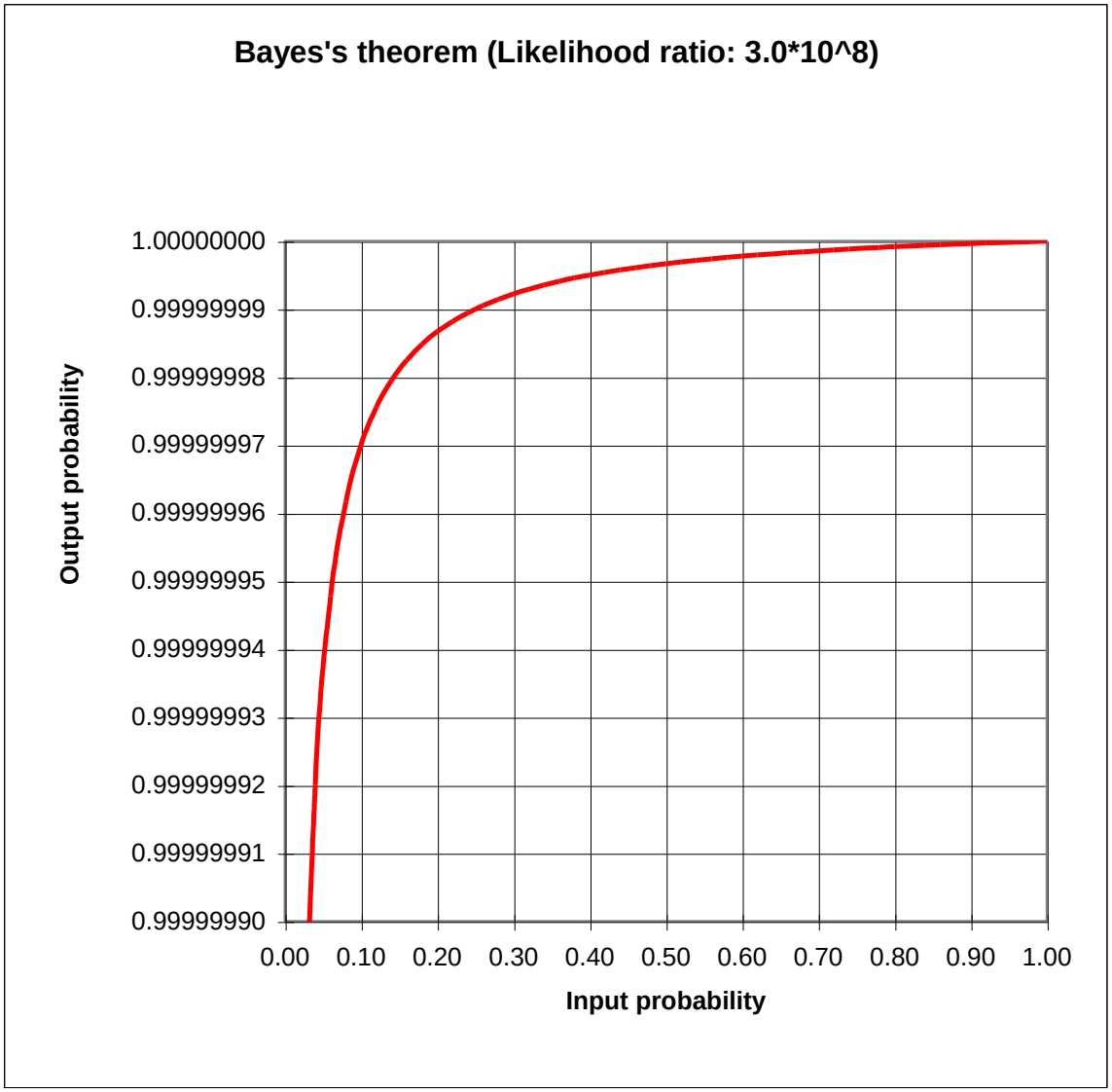

Figure 1. Bayes's theorem for a minimum likelihood ratio of 3.0*10^8.

Appendix: Calculating the value of the elementary probability p(uv)

The probability, p(uv), that an allele pair (uv) will occur at random at a locus of a forensic mixture, can be obtained with the Monte Carlo Bayesian (MCB) method as applied to STR profiles. This is true because the pair (uv) can be regarded as a single-locus "profile", which the MCB method can analyse as well as it does, say, a 13-locus profile. That being the case, the question, "What is the value of p(uv)?", is identical to that which may be asked in court regarding the random occurrence of a potentially incriminating simple STR profile.

In a criminal trial, the question may arise, "What is the likelihood of a match, given that the culprit is not the suspect?" In other words, "What is the likelihood of a random match?" The random match likelihood is easily obtained with the MCB method because of its eqivalence with Bayes's theorem. However, the random match likelihood, by itself, is not enough to make an identification; it must first be compared with the likelihood of a match between suspect and culprit by means of the likelihood ratio in order to estimate the culprit-suspect identity probability; and then the likelihood ratio must be combined with a prior probability derived from non-DNA sources to complete the Bayesian calculation.

The equivalence of the "standard" Bayesian calculation and the MCB method is illustrated in Figure A1, below. The Input axis shows the culprit-suspect identity probability prior to obtaining the STR profile; the Output axis shows the identity probability as affected by the profile.

Budowle - Moretti ID #: C003
Markers: D3 (14,18); vWA (17,17); FGA (18,24); D8 (13,13)
STRMCBAIN random match probability for those markers: $4.0*10^{-10}$
STRMCBAIN likelihood ratio (LR) for those markers: $2.5*10^{9}$

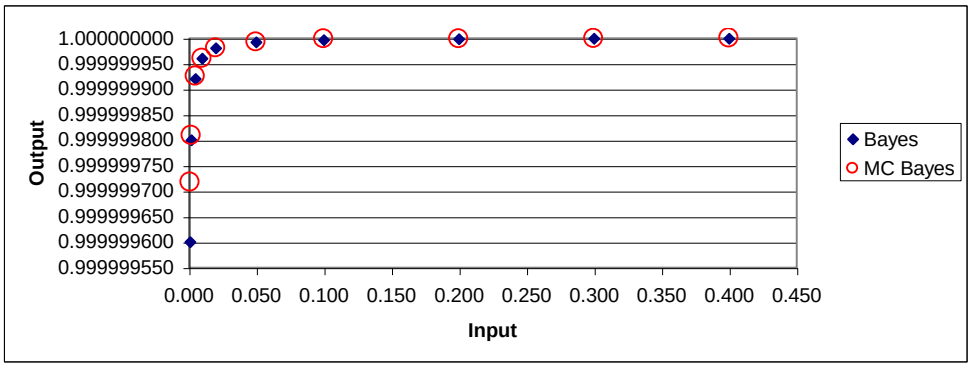

Figure A1. Comparison of Bayes's theorem (using STRMCBAIN LR) with STRMCBAIN (using Budowle - Moretti ID # C003 markers) [Reference 2]. [2]

Bayes's theorem, stated in terms of odds instead of probabilities, tells us that if the prior odds on a hypothesis are even ("50-50", 1/1, "same chance either way"), then the posterior odds are numerically equal to the likelihood ratio. Therefore, setting the prior probability of identity to 0.500... results in the following relationship between the posterior probability, $P_{post,\,0.5}$, and the random match likelihood, L(match|non-suspect):

$$L(\text{match}|\text{non-suspect}) = (1 - P_{post,\,0.5})/\,P_{post,\,0.5}\,. \qquad \text{(A1)}$$

For example, the MCB method assigns a maximum random match likelihood of 8.05×10^{-10} to a 13-locus STR profile. An 8-locus profile has a maximum random match likelihood of 9.58×10^{-7} ; and a 5-locus profile has a maximum random match likelihood of 1.48×10^{-4}. The trend, not surprisingly, is toward increased random occurrence likelihood as profile size decreases.

The MCB method consists of iterative Bayesian analysis of stratified random sample arrays. A verbal description is given here, followed by the mathematics of the method.

Typically, the method, in the form of a computer program, is applied to a criminal case in which there are an unknown culprit and a known suspect whose DNA profiles are identical. The investigator might first ask: "How inbred is the group which produced the culprit?" The answer to this question is useful in the search for other possible suspects. Ultimately, the investigator has to decide whether culprit and suspect are actually the same person, so she/he asks, "Could any group, whether or not it produced the culprit, have produced someone else with exactly the same profile?"

In this article, in order to answer the first of these questions, the MCB program uses a mathematical array consisting of 10 discrete, equal-sized homozygosity ranges, called "demes"*. The demes divide the entire homozygosity range from 0 to 1 into 10 equal-sized parts. So the first question becomes: "What is the chance that the culprit is a member of this or that deme?"

The program applies Bayes' theorem to each of several Monte Carlo samples taken from the deme array, in order to get the culprit's deme membership probabilities. Each deme contributes one randomly-generated group to a sample array. Each group differs from all the others in the sample array in terms of its profile allele frequencies.

Each sample array, when processed according to Bayes' theorem, yields a set of 10 probabilities with respect to culprit membership. Each probability is assigned to a different randomly-generated group in the sample array.

In essence, the program has evaluated each group for its ability to produce the profile.

By making repeated samples, the program develops a collection of probability sets on sample arrays of groups. By taking the average of this collection, deme by deme, the MCB program calculates the set of probabilities, with respect to culprit membership, on the array of demes. This set answers the investigator's question: "What is the chance that the culprit is a member of this or that deme?"

Next, in order to answer the question, "Could any group, whether or not it produced the culprit, have produced someone else with exactly the same profile?", the investigator instructs the MCB program to add an eleventh, distinct element to the array of 10 demes.

The eleventh element is unusual because it contains only one group, in contrast to the essentially infinite number of possible groups in any deme. Further, that single group is unusual because it contains only one member, that is, it contains only the suspect. Also, the suspect's DNA profile exactly matches the culprit's, and therefore the likelihood of getting the profile from this group, given that the culprit is indeed the suspect, is always exactly one. There is no need to calculate the suspect's profile likelihood from allele frequencies.

This 11th array element is called "the singular group". Adding it to the deme array is as simple as making its prior probability non-zero. For example, if the investigator gets a "cold hit" profile match from a DNA data base, she/he may say, conservatively, that the smallest chance of getting such a match at random is one out of the estimated world population in the year 2050. She/he would then input 0.0000000001, as the prior probability for the singular group, into the MCB program. In order to obtain a random match likelihood, the investigator would input a prior probability of 0.5 .

(When a non-zero prior is assigned to the singular group, the program automatically adjusts the deme priors so that they all add up to 1.000... .)

The program now applies Bayes' theorem to expanded sample arrays that comprise, not only groups from the 10 demes, but also the singular group. It then collects the results and takes the average of the collection, as before. The investigator thus finds the chances that the culprit is either the suspect or someone other than the suspect. (Again, an input prior probability of 0.5 will also result in a random match likelihood.)

Each MCB computation for this essay comprised 500 iterations on a MicroSoft Excel spreadsheet, and took less than 20 seconds.

* ["Deme" fits Sewall Wright's concept of the "neighborhood" of the singular group. He writes, "A term is needed to designate the local population of which the parents may be representative. ... An essential property of the population in question is that the individuals are neighbors in the sense that their gametes may come together." (Isolation by distance under diverse systems of mating, Genetics, January 1946, Volume 31, pp. 39 - 59.)]

The following version of Bayes' theorem is the core of the method described in this work. Let

(1) $P_0(g_k \mid h)$ be the prior probability, on the basis of knowledge, h, from sources other than the culprit's STR data, that the culprit is a member of the group g_k. The index k runs from 1 through 11; when $1 \le k \le 10$, g_k is a deme; when $k = 11$, g_k is the singular group.

(2) $P(g_k \mid h, d)$ be the posterior probability that the culprit is a member of the group g_k, given the culprit's STR data, d, in addition to h.

(3) $L(d \mid g_k)$ be the likelihood of the STR data, d, if the culprit were, in fact, a member of the group g_k.

Then Bayes' theorem appears as

$$P(g_k \mid h, d) = \frac{P_0(g_k \mid h) \times L(d \mid g_k)}{\sum_{j=1}^{11} P_0(g_j \mid h) \times L(d \mid g_j)} \quad . \qquad \text{(A2)}$$

By setting $P_0(g_{11}|h)$, the prior probability for the singular group, equal to zero, one can obtain the posterior probability that the culprit is a member of each of the 10 demes. This is a useful result, because each deme represents a different one of the 10 homozygosity intervals **{**(0.0-0.1),(0.1-0.2),(0.2-0.3), ... ,(0.7-0.8),(0.8-0.9),(0.9-1.0)**}**.

Calculating the likelihood $L(d \mid g_k)$, for $1 \le k \le 10$, is the main computational task of the method presented here. In particular, the likelihood of a single locus involves the product of the frequencies of the two alleles, at that locus, that are part of the culprit's STR profile.

Because only the homozygosity interval of a deme is given, the method samples, at each locus and within each interval, the space of all possible allele frequency products, as follows.

The allele frequency distribution at each STR locus is modelled by a Gaussian density function f(x | μ , σ) as is illustrated in Figure A2. The figure shows two alleles. The midpoint between the alleles is defined as the origin of the length variable, x, i.e., (x ≡ 0). The smaller allele of the two is at (x = -a) and the larger is at (x = +a). The difference in their lengths is thus 2×a. It is this difference, not the lengths themselves, that affects the product of the alleles' frequencies.

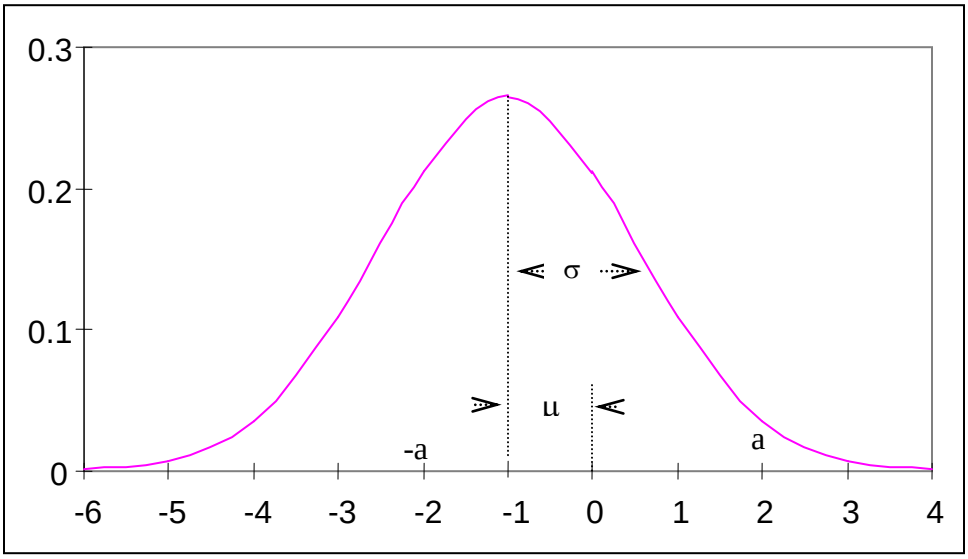

Figure A2. Relation of allele-pair having length difference 2a = 4 with Gaussian density function having mean μ = -1 (i.e. offset with respect to midpoint between allelele lengths) and standard deviation σ.

(NOTE: Number of STR repeats is the measure of length, x. All frequency calculations are thus referred to the same scale, regardless of locus. That is, they are independent of the physical lengths of repeats.)

The homozygosity of each deme is determined by the Gaussian function's standard deviation, σ. Because demes belong to homozygosity intervals, the median homozygosity of each interval is chosen to represent any homozygosity within the interval. For example, a deme belonging to the interval (0.0 - 0.1) is represented by the median value 0.05, for which the Gaussian standard deviation, σ, is 5.60. By way of contrast, a deme belonging to the interval (0.9 - 1.0) is represented by the median value 0.95, for which $\sigma = 0.41$.

The mean, μ , of the Gaussian function is defined as a random variable whose value is zero at the midpoint between the two STR profile alleles at a locus. The value of μ for each locus is chosen independently of that of any other locus, including those belonging to other demes.

Thus, for a 13-locus profile, for example, one iteration of the sampling process comprises a random, i.e."Monte Carlo", selection of (13 loci × 10 demes =) 130 values of μ. The likelihood, $L(d \mid g_k)$, of the STR data with respect to each deme can then be calculated as the product of the likelihoods of its 13 loci, each locus having an independently and randomly chosen value of μ. Because the product is independent of the order of the locus likelihoods, the entire profile can be encrypted by shuffling, or interchanging, the loci.

In the present context of a single-locus "profile" (uv), only 10 randomly-generated values of μ are needed for each iteration.

Taking into account the prior probability and likelihood of the singular group, a single iteration's calculation then generates a posterior probability for each deme and also for the singular group.

References

[1] [1] Naval Operations Analysis, Naval Institute Press, 1972, ISBN 0-87021-439-X.

[2] [2] B. Budowle and T.R. Moretti, Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci ... , Forensic Science Communications, July 1999, Volume 1, Number 2; http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm, dnaloci.txt .