MONTE CARLO BAYESIAN IDENTIFICATION USING STR PROFILES

Donald I. Promish, MS

68 Richardson St.

Burlington, Vermont

05401-5026

U.S.A.

VOX: 802 860 9441

E: DonaldPromish@cs.com

9 January 2008

Abstract

The method described here bears out the premise that a complete CODIS 13 STR profile contains all the information necessary to establish the probability that two individuals who both have that profile are really the same person. In order to do so, this paper combines three concepts: the concept of the culprit as a member of a group; the concept of the suspect as a one-person group; and the concept of groups, other than the suspect, which are distinguishable only by their homozygosity. It shows how to perform a Bayesian analysis of an STR profile with respect to the suspect group and a random sample of non-suspect groups. It demonstrates the robustness of the method with respect to a varied selection of real profiles and with respect to evidence other than the profiles.

Keywords

group membership, singular group, STR profile, substructure, Bayes, Monte Carlo, stratified random sample

Introduction

The Monte Carlo Bayesian (MCB) method described here has several features worth noting.

(a) The method is case-specific. Both evaluation of and adjustment for substructure are automatic, and they depend only on the STR profile at issue.

(b) The method accommodates variation in prior probabilities according to the investigator's judgment regarding non-profile data.

(c) The method produces probabilities as well as likelihood ratios.

(d) The method does not rely on reference group allele frequency data. The investigator can use the method when she/he lacks either knowledge of, or immediate access to, suitable frequency data.

Here is a brief survey of the differences between the MCB method and the conventional method that is based on Identity By Descent (IBD). Features of the latter are taken from Reference [1][1].

The MCB method does not depend on empirical data on ethnic/racial populations. It thus operates as well in a highly-diverse, enclave-mottled metropolitan environment as in a more ethnically/racially uniform one. (See, for example, Reference [2][2].) The IBD method, in contrast, not only requires empirical data, it also requires that choices be made from among the available populations. Failing that, IBD analysis may adopt the conservative assumption that "any racial group consists of isolated groups of second cousins or first cousins once removed." This assumption not only eliminates differences in substructure between populations, it produces uniform substructures within populations. Under the MCB method, this assumption is not necessary.

The MCB method centres on the culprit who, although unidentified, is known to have produced the crime scene profile. The IBD method centres on the defendant/suspect. As a consequence, MCB can always provide information about the culprit even in the absence of a defendant/suspect, whereas IBD may not be able to do so. In particular, MCB, using only the culprit's profile, produces a profile-specific substructure estimate. However, IBD, unless it has empirical data on the defendant/suspect's family and also assumes that the culprit is in that family, cannot make such an estimate.

The MCB method not only allows but actually requires the inclusion of non-profile evidence. The non-profile evidence appears as the input prior probabilities in the MCB calculations, which produce culprit-suspect identity probabilities. Conversely, the IBD method entirely ignores non-profile evidence and produces only profile likelihoods.

For example, at some point in a criminal case, the investigator may feel that the non-profile evidence alone has raised the probability above 50% that the culprit is a particular suspect. This 50% level might be considered "minimum probable cause" for arresting that suspect. If, after arrest, the culprit's profile is found to match that of the suspect, MCB will use the minimum probable cause probability as its prior probability in order to calculate the posterior probability that the culprit is the suspect. See Reference [3][3].

It is thus conceivable that, in a criminal trial involving a DNA profile, the task of the prosecution would be to argue for an MCB (non-profile) prior probability which is as high as possible; while the task of the defence would be to argue for a prior probability which is as low as possible.

Method

The core of the MCB method consists of iterative Bayesian analysis of stratified random sample arrays.

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. See Reference [4][4].

(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.

Each MCB computation for this article comprised 500 iterations on a MicroSoft Excel spreadsheet, and took less than 20 seconds. The software is available from the author on request by post.

* ["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 mathematics of the method are given in the Appendix.

Data

The CODIS profile data used here were taken from Reference [5][5]. They represent individuals who contributed to a database of U.S. Caucasians. Specific profiles are included in the **Results**; they are identified by the codes employed in Reference [5].

This article employs a uniform ("flat") distribution for the deme prior probabilities. The flat deme prior distribution is a prudent choice because, as Reference [2] shows, it can be impossible to take an accurate and stable account of the varied degrees of inbreeding in a population, especially a cosmopolitan one: "New York City's demography is not static, but [is constantly undergoing] a dynamic process defined by the ebb and flow of people. [I]n just 30 years, what was primarily a [city with a] European population has now become a place with no dominant race/ethnic or nationality group. Indeed, New York epitomizes the world city." The flat deme prior is also a safe choice because, in the MCB method, the evidence of the CODIS loci employed here overwhelms any reasonable prior. Each of the 13 loci in a person's profile makes an independent contribution to the description of their shared provenance.

In this article, two prior probabilities are applied to the culprit's membership in the singular group.

The first prior, as has been mentioned above, is the "cold hit" prior, 1 in 10 billion. It is derived from the rough projection of the world population in 2050 that appears in Reference [4].

The second prior is the "minimum probable cause" prior. It is based on Reference [3], which defines probable cause as corresponding to the verbal expression, "more likely than not", i.e. probability greater than 50%. The "minimum probable cause" prior corresponds to the more modest expression "as likely as not". It is exactly equal to 50%. It is related to the investigator's valuation of evidence other than the culprit's STR profile.

Results

Most of this article's results are part of a series of case reports. The rationale for developing the series is two-fold. First, all the profiles come from prison inmates; and, second, some of these inmates are likely to be repeat offenders. Thus, one can expect that a case report profile will occasionally appear at a new crime scene. When this happens, the report's results will add weight to the evidence of the profile alone.

The contextual uniformity of the reports facilitates quantitative comparisons among them.

The first three cases can easily be analysed for substructure by using the profile's within-locus allele differences (which are provided in all cases except Case 9). They thus illustrate the fact that MCB gives results that are consistent with common sense.

Cases 1 and 2 are similar with respect to homozygous loci, but they are dissimilar with respect to heterozygous loci. The profile differences clearly lead to substructure differences.

Case 3 illustrates the importance of having independent loci that are linked only by their common provenance. The evidence of even one locus can weigh heavily in the substructure estimate.

Cases 4 - 8 use profiles that all have three homozygous loci. This type of profile appears in roughly one quarter of the population. These cases demonstrate substructure diversity in a fairly common subpopulation in which everyone's profile has three homozygous loci. As mentioned above, within-locus allele differences are provided for these cases.

Cases 4 - 8 also demonstrate, although more subtly than in Cases 1 and 2, the effect of within-locus allele differences.

A survey of these 8 cases highlights the fact that MCB relies only on within-locus allele differences, not absolute allele sizes. This fact can be put to use in encrypting profiles, as Case 8A shows.

Finally, because MCB relies only on within-locus allele differences, it is a simple matter to analyse the exceptional case, Case 9, of an individual who is entirely homozygous with respect to the CODIS 13 loci. All that is needed is to give each allele the same size; 10 was used here; it could as well have been 117.

Case 1.

Subject: C091

Profile:

Locus	(Alleles)
D3S1358	(15, 15)
VWA	(16, 17)
FGA	(21, 22)
D8S1179	(11, 12)
D21S11	(30, 31)
D18S51	(14, 17)
D5S818	(11, 12)
D13S317	(8, 11)
D7S820	(10, 11)
CSF1PO	(12, 14)
TPOX	(8, 9)
TH01	(7, 8)
D16S539	(9, 12)

Inferences:

Local substructure probabilities for subject C091, using uninformative ("flat") prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.000
0.1 - 0.2	0.007
0.2 - 0.3	0.131
0.3 - 0.4	0.209
0.4 - 0.5	0.207
0.5 - 0.6	0.176
0.6 - 0.7	0.110
0.7 - 0.8	0.064
0.8 - 0.9	0.060
0.9 - 1.0	0.035

Comment: Subject C091 has a small chance (~ 14%) of coming from a normally inbred group (homozygosity range 0.0 - 0.3), a good chance (~42%) of coming from a moderately inbred group (homozygosity range 0.3 - 0.5), a fair chance (~29%) of coming from a highly inbred group (homozygosity range 0.5 - 0.7) and a small chance (~16%) of coming from an extremely inbred group (homozygosity range 0.7 - 1.0). This quantitative inference is consistent with the following qualitative observations. Although there is very little "identity" in the profile, there is a great deal of "similarity". The only homozygous locus is D3S1358; however, the alleles at seven other loci differ merely by 1, and the rest differ by no more than 3. The within-locus allele differences, in order of size, are: [0,1,1,1,1,1,1,1,1,2,3,3,3]. Hence, by inspection alone, one might suppose that Subject C091 is more inbred than normal, although not extremely inbred.

“Cold hit” probability, given a match between Subject C091 and a known individual, that the subject is the known individual: 1**.**000..., to 8 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion).

“Minimum probable cause” probability, given a match between Subject C091 and a known individual, that the subject is the known individual: 1**.**000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0**.**500... .

Case 2.

Subject: C039

Profile:

Locus	(Alleles)
D3S1358	(14, 18)
VWA	(14, 16)
FGA	(20, 21)
D8S1179	(8, 10)
D21S11	(27, 33.2)
D18S51	(14, 15)
D5S818	(11, 14)
D13S317	(11, 13)
D7S820	(10, 11)
CSF1PO	(12, 12)
TPOX	(9, 10)
TH01	(6, 9.3)
D16S539	(9, 13)

Inferences:

Local substructure probabilities for Subject C039, using uninformative ("flat") prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.007
0.1 - 0.2	0.378
0.2 - 0.3	0.470
0.3 - 0.4	0.141
0.4 - 0.5	0.004
0.5 - 0.6	0.000
0.6 - 0.7	0.000
0.7 - 0.8	0.000
0.8 - 0.9	0.000
0.9 - 1.0	0.000

Comment: Subject C039 very likely (~ 85%) comes from a normally inbred group (homozygosity range 0.0 - 0.3), and has a small chance (~ 15%) of coming from a moderately inbred group (homozygosity range 0.3 - 0.5). This quantitative inference is consistent with the following qualitative observations. As in Case 1, there is very little "identity" in this profile. However, in this case, there is a great deal of "dis-similarity", not "similarity" as in Case 1. The within-locus allele differences, in order of size, are: [0,1,1,1,1,2,2,2,3,3.3,4,4,6.2]. Hence, by inspection alone, one might expect to find that, as the MCB method shows, Subject C039 is fairly sure to be normally inbred.

“Cold hit” probability, given a match between Subject C039 and a known individual, that the subject is the known individual: 1**.**000..., to 14 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion). Note that “cold hit” probability for C039 is a little higher than the probability for the more inbred C091 in Case 1.

“Minimum probable cause” probability, given a match between Subject C039 and a known individual, that the subject is the known individual: 1**.**000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0**.**500... .

Case 3.

Subject: C089

Profile:

Locus	(Alleles)
D3S1358	(15, 16)
VWA	(16, 16)
FGA	(18, 23)
D8S1179	(10, 13)
D21S11	(28, 30)
D18S51	(12, 22)
D5S818	(9, 11)
D13S317	(12, 12)
D7S820	(7, 10)
CSF1PO	(11, 11)
TPOX	(8, 8)
TH01	(6, 6)
D16S539	(9, 13)

Inferences:

Local substructure probabilities for Subject C089, using uninformative (“flat”) prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.050
0.1 - 0.2	0.615
0.2 - 0.3	0.323
0.3 - 0.4	0.012
0.4 - 0.5	0.000
0.5 - 0.6	0.000
0.6 - 0.7	0.000
0.7 - 0.8	0.000
0.8 - 0.9	0.000
0.9 - 1.0	0.000

Comment: Subject C089 is practically certain (~99%) to have come from a normally inbred group (homozygosity range 0.0 - 0.3). This quantitative inference is consistent with the following qualitative observations. Although the subject’s profile has 5 homozygous loci and also has 5 loci whose alleles differ by no more than 3, the alleles at locus FGA differ by 5 and, more importantly, those at D18S51 differ by 10. These two loci thus argue very strongly against the subject’s coming from even a moderately inbred group. As unlikely as this profile is in a normal group, it is even more unlikely in an inbred one. One explanation for the profile is that Subject C089 is the offspring of an incestuous mating within an otherwise normal group. The within-locus allele differences, in order of size, are: [0,0,0,0,0,1,2,2,3,3,4,5,10].

“Cold hit” probability, given a match between Subject C089 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion). The large number of decimal places for zeroes is due to the very low probability that the subject is more inbred than normal.

“Minimum probable cause” probability, given a match between Subject C089 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 4.

Subject: C096

Profile:

Locus	(Alleles)
D3S1358	(16, 18)
VWA	(17, 17)
FGA	(21, 22)
D8S1179	(13, 15)
D21S11	(30, 31)
D18S51	(15, 18)
D5S818	(12, 12)
D13S317	(12, 13)
D7S820	(11, 12)
CSF1PO	(11, 12)
TPOX	(8, 11)
TH01	(8, 8)
D16S539	(12, 13)

Inferences:

Local substructure probabilities for Subject C096, using uninformative ("flat") prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.000
0.1 - 0.2	0.003
0.2 - 0.3	0.068
0.3 - 0.4	0.133
0.4 - 0.5	0.179
0.5 - 0.6	0.180
0.6 - 0.7	0.171
0.7 - 0.8	0.099
0.8 - 0.9	0.094
0.9 - 1.0	0.071

Comment: Subject C096 has a slight chance (~ 7%) of being normally inbred (homozygosity range 0.0 - 0.3), a good chance (~ 31%) of being moderately inbred (homozygosity range 0.3 - 0.5), a slightly better chance (~ 35%) of being highly inbred (homozygosity range 0.5 - 0.7) and a fair chance (~ 26%) of being extremely inbred (homozygosity range 0.7 - 1.0). The within-locus allele differences, in order of size, are: [0,0,0,1,1,1,1,1,1,2,2,3,3].

“Cold hit” probability, given a match between Subject C096 and a known individual, that the subject is the known individual: 1.000..., to 7 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion).

“Minimum probable cause” probability, given a match between Subject C096 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 5.

Subject: C083

Profile:

Locus	(Alleles)
D3S1358	(14, 18)
VWA	(14, 15)
FGA	(21, 22)
D8S1179	(12, 14)
D21S11	(32.2, 33.2)
D18S51	(13, 16)
D5S818	(12, 12)
D13S317	(11, 11)
D7S820	(9, 10)
CSF1PO	(10, 12)
TPOX	(8, 11)
TH01	(6, 7)
D16S539	(9, 9)

Inferences:

Local substructure probabilities for Subject C083, using uninformative ("flat") prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.000
0.1 - 0.2	0.040
0.2 - 0.3	0.192
0.3 - 0.4	0.258
0.4 - 0.5	0.246
0.5 - 0.6	0.137
0.6 - 0.7	0.077
0.7 - 0.8	0.037
0.8 - 0.9	0.006
0.9 - 1.0	0.006

Comment: Subject C083 has a fair chance (~ 23%) of being normally inbred (homozygosity range 0.0 - 0.3), a very good chance (~ 50%) of being moderately inbred (homozygosity range 0.3 - 0.5), a fair chance (~ 21%) of being highly inbred (homozygosity range 0.5 - 0.7) and a slight chance (~ 5%) of being extremely inbred (homozygosity range 0.7 - 1.0). The within-locus allele differences, in order of size, are: [0,0,0,1,1,1,1,1,2,2,3,3,4].

“Cold hit” probability, given a match between Subject C083 and a known individual, that the subject is the known individual: 1.000..., to 9 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion).

“Minimum probable cause” probability, given a match between Subject C083 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 6.

Subject: C058

Profile:

Locus	(Alleles)
D3S1358	(14, 16)
VWA	(15, 20)
FGA	(15, 20)
D8S1179	(14, 14)
D21S11	(28, 31)
D18S51	(18, 18)
D5S818	(11, 12)
D13S317	(11, 14)
D7S820	(10, 10)
CSF1PO	(11, 12)
TPOX	(9, 10)
TH01	(7, 8)
D16S539	(9, 11)

Inferences:

Local substructure probabilities for Subject C058, using uninformative ("flat") prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.001
0.1 - 0.2	0.222
0.2 - 0.3	0.429
0.3 - 0.4	0.266
0.4 - 0.5	0.071
0.5 - 0.6	0.011
0.6 - 0.7	0.000
0.7 - 0.8	0.000
0.8 - 0.9	0.000
0.9 - 1.0	0.000

Comment: Subject C058 has a very good chance (~ 65%) of being normally inbred (homozygosity range 0.0 - 0.3), a fairly good chance (~36%) of being moderately inbred (homozygosity range 0.3 - 0.5), a very slight chance (~ 1%) of being highly inbred (homozygosity range 0.5 - 0.7) and no chance (~ 0%) of being extremely inbred (homozygosity range 0.7 - 1.0). The within-locus allele differences, in order of size, are: [0,0,0,1,1,1,1,2,2,3,3,5,5].

“Cold hit” probability, given a match between Subject C058 and a known individual, that the subject is the known individual: 1.000..., to 13 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion).

“Minimum probable cause” probability, given a match between Subject C058 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 7.

Subject: C037

Profile:

Locus	(Alleles)
D3S1358	(16, 17)
VWA	(15, 17)
FGA	(19, 25)
D8S1179	(11, 14)
D21S11	(30, 32.2)
D18S51	(13, 14)
D5S818	(11, 12)
D13S317	(11, 12)
D7S820	(10, 12)
CSF1PO	(12, 12)
TPOX	(8, 8)
TH01	(6, 7)
D16S539	(13, 13)

Inferences:

Local substructure probabilities for Subject C037, using uninformative (“flat”) prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.000
0.1 - 0.2	0.105
0.2 - 0.3	0.353
0.3 - 0.4	0.325
0.4 - 0.5	0.161
0.5 - 0.6	0.046
0.6 - 0.7	0.009
0.7 - 0.8	0.000
0.8 - 0.9	0.000
0.9 - 1.0	0.000

Comment: Subject C037 has a good chance (~ 46%) of being normally inbred (homozygosity range 0.0 - 0.3), a somewhat better chance (~ 49%) of being moderately inbred (homozygosity range 0.3 - 0.5), a slight chance (~ 6%) of being highly inbred (homozygosity range 0.5 - 0.7) and practically no chance of being extremely inbred (homozygosity range 0.7 - 1.0). The within-locus allele differences, in order of size, are: [0,0,0,1,1,1,1,1,2,2,2.2,3,6].

“Cold hit” probability, given a match between Subject C037 and a known individual, that the subject is the known individual: 1.000..., to 11 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion).

“Minimum probable cause” probability, given a match between Subject C037 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 8.

Subject: C056

Profile:

Locus	(Alleles)
D3S1358	(14, 18)
VWA	(14, 17)
FGA	(19, 26)
D8S1179	(11, 13)
D21S11	(30, 32.2)
D18S51	(15, 18)
D5S818	(11, 11)
D13S317	(8, 12)
D7S820	(10, 10)
CSF1PO	(10, 11)
TPOX	(8, 11)
TH01	(9, 9)
D16S539	(10, 11)

Inferences:

Local substructure probabilities for Subject C056, using uninformative (“flat”) prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.007
0.1 - 0.2	0.462
0.2 - 0.3	0.438
0.3 - 0.4	0.092
0.4 - 0.5	0.001
0.5 - 0.6	0.000
0.6 - 0.7	0.000
0.7 - 0.8	0.000
0.8 - 0.9	0.000
0.9 - 1.0	0.000

Comment: Subject C056 has a very high chance (~ 91%) of being normally inbred (homozygosity range 0.0 - 0.3), a slight chance (~ 9%) of being moderately inbred (homozygosity range 0.3 - 0.5) and practically no chance of being either highly or extremely inbred (homozygosity range 0.5 - 1.0). The within-locus allele differences, in order of size, are: [0,0,0,1,1,2,2.2,3,3,3,4,4,7].

“Cold hit” probability, given a match between Subject C056 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion). As with Subject C089, the large number of decimal places for zeroes is due to the very low probability that the subject is more inbred than normal.

“Minimum probable cause” probability, given a match between Subject C056 and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 8A.

Subject: C056 (encrypted)

Profile:

Locus	(Alleles)
D3S1358	(10, 14)
VWA	(10, 13)
FGA	(10, 17)
D8S1179	(10, 12)
D21S11	(10, 12.2)
D18S51	(10, 13)
D5S818	(10, 10)
D13S317	(10, 14)
D7S820	(10, 10)
CSF1PO	(10, 11)
TPOX	(10, 13)
TH01	(10, 10)
D16S539	(10, 11)

Inferences:

Local substructure probabilities for Subject C056 (encrypted), using uninformative (“flat) prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.007
0.1 - 0.2	0.419
0.2 - 0.3	0.450
0.3 - 0.4	0.121
0.4 - 0.5	0.004
0.5 - 0.6	0.000
0.6 - 0.7	0.000
0.7 - 0.8	0.000
0.8 - 0.9	0.000
0.9 - 1.0	0.000

Comment: Subject C056 (encrypted) has a very high chance (~ 88%) of being normally inbred (homozygosity range 0.0 - 0.3), a slight chance (~ 13%) of being moderately inbred (homozygosity range 0.3 - 0.5) and practically no chance of being either highly or extremely inbred (homozygosity range 0.5 - 1.0). The within-locus allele differences, in order of size, are: [0,0,0,1,1,2,2.2,3,3,3,4,4,7].

“Cold hit” probability, given a match between Subject C056 (encrypted) and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion). As with Subject C089, the large number of decimal places for zeroes is due to the very low probability that the subject is more inbred than normal.

“Minimum probable cause” probability, given a match between Subject C056 (encrypted) and a known individual, that the subject is the known individual: 1.000..., to at least 30 decimal places. This is based on non-profile evidence that sets the prior probability at 0.500... .

Case 9.

Subject: Max13 (CODIS 13 fully homozygous)

Profile:

Locus	(Alleles)
D3S1358	(10, 10)
VWA	(10, 10)
FGA	(10, 10)
D8S1179	(10, 10)
D21S11	(10, 10)
D18S51	(10, 10)
D5S818	(10, 10)
D13S317	(10, 10)
D7S820	(10, 10)
CSF1PO	(10, 10)
TPOX	(10, 10)
TH01	(10, 10)
D16S539	(10, 10)

Inferences:

Local substructure probabilities for Subject Max13, using uninformative (“flat”) prior:

Homozygosity interval	Posterior probability
0.0 - 0.1	0.000
0.1 - 0.2	0.000
0.2 - 0.3	0.000
0.3 - 0.4	0.000
0.4 - 0.5	0.000
0.5 - 0.6	0.043
0.6 - 0.7	0.103
0.7 - 0.8	0.175
0.8 - 0.9	0.260
0.9 - 1.0	0.407

Comment: Subject Max13 has a very small chance (~ 1%) of being either normally or moderately inbred (homozygosity range 0.0 - 0.5), a slight chance (~15%) of being highly inbred (homozygosity range 0.5 - 0.7) and a very high chance (~ 84%) of being extremely inbred (homozygosity range 0.7 - 1.0).

“Cold hit” probability, given a match between Subject Max13 and a known individual, that the subject is the known individual: ~ 94%. The probability is ~ 6% that the subject is not the known individual, but comes from a highly or extremely inbred group. These results are based solely on the estimated world population in 2050 that sets the prior probability at 1/(10 billion).

“Minimum probable cause” probability, given a match between Subject Max13 and a known individual, that the subject is the known individual: 1**.**000..., to 9 decimal places. This is based on non-profile evidence that sets the prior probability at 0**.**500... .

Conclusions

In addition to the features mentioned in the Introduction, the case analysis results shown here lead, through a series of subordinate conclusions, to one major one.

Here are the subordinate conclusions.

(a) The results of a Monte Carlo Bayesian (MCB) analysis appear to be consistent with what an experienced investigator might infer by inspection.

(b) Because MCB relies only on within-locus allele differences, it can analyse encrypted (real) profiles. It thus has potential for use where personal privacy is of concern.

(c) Consequently, MCB can also analyse synthetic profiles, such as the extreme case, Case 9, of an individual who is homozygous at all of the CODIS 13 loci. The results of Case 9 lead to the following major conclusion.

It seems safe to say that any CODIS 13 “cold hit”, regardless of substructure, is at least a very good lead; and that a CODIS 13 profile match coupled with a “minimum probable cause” prior amounts to an investigative, if not a juridical, certainty. Table 1 shows how the number of loci in the culprit’s profile affects the minimum probability that, given a match, the culprit is the suspect. It is followed by a discussion of random match likelihood and likelihood ratios.

Number of loci in culprit’s profile	“Cold hit” prior: 10^{-10}	“Minimum probable cause” prior: 0.50
13	0.92	1.000... to 8 decimal places
12	0.81	1.000... to 7 decimal places
11	0.64	1.000... to 7 decimal places
10	0.45	1.000... to 7 decimal places
9	0.23	1.000... to 6 decimal places
8	0.08	1.000... to 5 decimal places
7	0.01	1.000... to 4 decimal places
6	---	1.000... to 4 decimal places
5	---	1.000
4	---	0.999
3	---	0.996
2	---	0.982
1	---	0.901
0	10^{-10}	0.500...

Table 1. Minimum probability, given a match, that the culprit is the suspect. This minimum probability depends only on the number of loci in the culprit’s profile, and not on which loci they are.

In a courtroom, the question may arise, "What is the likelihood of a match, given that the culprit is not the defendant?" In other words, "What is the likelihood of a random match?" Bearing in mind that, by definition, the likelihood of a match between defendant and culprit is exactly 1, the random match likelihood is easily obtained from Bayes's theorem. The 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-defendant):

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

For example, the MCB method assigns a <u>maximum</u> 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 corresponding <u>minimum</u> likelihood ratios are 1.24×10^{9}, 1.04×10^{6}, and 6.76×10^{3}, respectively.

Plans

Because this article tacitly assumes that crime scenes yield complete CODIS 13 profiles, a new article will explore the sensitivity of MCB to reduction of the profile, particularly by elimination of the more informative loci.

On the grounds that someone who is inbred with regard to the CODIS 13 is even more inbred with regard to the genome, it seems that MCB might be usefully employed to find “leads” to groups at risk for genetic disease. An exploration along these lines seems worthwhile.

APPENDIX: Mathematics of the Method

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, in this context, appears as

$$\mathrm{P}(\mathrm{g}_{\mathrm{k}} \mid \mathrm{h}, \mathrm{d}) = \frac{\mathrm{P}_0(\mathrm{g}_{\mathrm{k}} \mid \mathrm{h}) \times \mathrm{L}(\mathrm{d} \mid \mathrm{g}_{\mathrm{k}})}{\sum_{j=1}^{j=11} \mathrm{P}_0(\mathrm{g}_{\mathrm{j}} \mid \mathrm{h}) \times \mathrm{L}(\mathrm{d} \mid \mathrm{g}_{\mathrm{j}})} \qquad .$$

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 $1 \leq k \leq 10$, 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 $f(x \mid \mu , \sigma)$ as is illustrated in Figure 1. The figure shows two alleles. The midpoint between the alleles is defined as the origin of the length variable, x, i.e., $(x \equiv 0)$. The smaller allele of the two is at $(x = -a)$ and the larger is at $(x = +a)$. The difference in their lengths is thus $2\times a$, and it is this difference, not the lengths themselves, that affects the product of the alleles' frequencies. Thus any locus can be encrypted, with no effect on results, by translation, that is, by changing the lengths of both of its alleles by the same amount. The **Results** section makes further use of the allele-length difference $2\times a$.

(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, in this work, 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, 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.

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.

Acknowledgments

Humble gratitude to Professor Judith E. Stone, of the University of Vermont, for relentless guidance in matters of written English.

Deepest thanks to Karen L. Ayres, Ph.D., of the University of Reading, for generous counsel and encouragement.

Highest admiration to several anonymous reviewers (and to one anonymous editor) for their patience and support over several years.

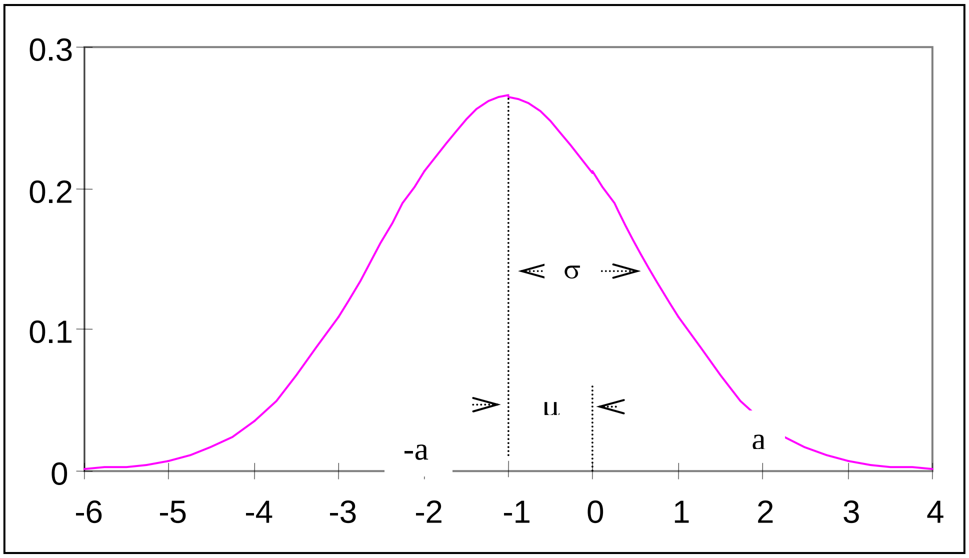

Figure 1. Relation of allele-pair having length difference 2a = 4 with Gaussian density function having mean μ = -1 (i.e. offset with respect to midpoint between allelele lengths) and standard deviation σ.

References

[1] [1.] B.S. Weir, The coancestry coefficient in forensic science, Proceedings of the Eighth International Symposium on Human Identification, 17-20 September, 1997, pp 87-91.

[2] [2.] Population Division, New York Department of City Planning, The newest New Yorkers, 2000: immigrant New York in the new millennium; http://www.nyc.gov/html/dcp/html/census/nny.html; January 2005.

[3] [3.] FBI Special Agent Coleen Rowley, Memorandum of 21 May 2002 to FBI Director Robert Mueller; http://www.com/time/covers/1101020603/memo.html .

[4] [4.] Columbia Encyclopedia, 5th edition, Columbia University Press, 1993, p 2193.

[5] [5.] 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 .