MONTE CARLO BAYESIAN IDENTIFICATION USING SNP PROFILES

Donald I. Promish, MS

68 Richardson Street

Burlington, Vermont

05401-5026

U.S.A.

VOX: 802 860 9441

E: DonaldPromish@cs.com

18 September 2008

Abstract. The method demonstrated here enables an investigator to analyse individual culprits' dimorphic SNP profiles quickly without depending on "reference groups", to analyse composites of discrete SNP profiles, and to analyse combinations of STR and SNP profiles.

Keywords: group membership; singular group; SNP profile; Bayes; Monte Carlo; stratified random sample

1. Introduction

The Monte Carlo Bayesian (MCB) method has several operational features worth noting. (a) The method is case-specific. Both evaluation of and adjustment for substructure are automatic, and they depend only on the profile at issue. (b) The method accommodates variation in prior probabilities according to the investigator's judgment regarding non-profile evidence. (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.

2. Method

The MCB method, in the form of a computer program, iteratively applies Bayes' theorem to stratified random sample arrays comprising specimens taken, in part, from 10 discrete, equal-sized allele frequency ranges, called "demes". The demes are modelled after Wright's definition, which is as follows: "Most species contain many small, random breeding local populations (demes) that are sufficiently isolated, if only by distance, to permit differentiation of their sets of gene frequencies, but that are not so isolated as to prevent the gradual spreading of favorable gene complexes throughout the species from their centers of origin. This differentiation need not be associated with conspicuous phenotypic differences." [Reference 1[i]]

Within the demes, as modelled here, the allele frequencies are the sampling random variables. In addition to the 10 demes, an 11th array element, a "singular group" representing a suspect with a matching profile may be included. The computer program, during each iteration, evaluates each deme (and also the suspect's singular group) for its ability to produce the profile. That is, the program computes the profile likelihood for each array element.

The contribution of the homozygous portion of the profile to the likelihood is obtained by using the Binomial Theorem to produce an expected value. The contribution of the heterozygous portion of the profile is the product of the individual heterozygous SNP pair frequencies. For details, see the Appendix.

The MCB program develops a collection of probability sets on the sample arrays. By taking the average of this collection, the program calculates the set of probabilities that the culprit is either the suspect (a member of the singular group) or somebody else (a member of a non-suspect deme).

Each MCB computation for this article consisted of 1000 iterations on a Pentium-4-equipped machine running a MicroSoft Excel spreadsheet, and took less than 40 seconds. The software is available from the author on request by post.

3. Results

The following two tables offer a sample of SNP identification results. The computation for each table entry starts with the investigator's judgment, on evidence other than the SNP profile. This judgment may produce a small subset, perhaps only one, of uncountably many possible values of prior probabilities ("priors") that span the probability range.

The three prior values used in these tables are their column headers. The lowest value, 0.0000000001, is based on the estimated world population, 10 billion, in 2050. It can be regarded as a very conservative "cold hit" prior probability that the culprit is the suspect. The highest prior value, 0.5, interprets the non-profile evidence as saying that the culprit is as likely as not to be the suspect; it is the "mimimum probable cause" prior.

Two parameters for the MCB computation are the number of all loci in the profile, and the number of homozygous loci in the profile. Table 1 corresponds to a 50-SNP profile; Table 2 corresponds to a 20-SNP profile. In each table, the number of homozygous loci in the profile is a row header.

4. Discussion

Even though a very large proportion of homozygous loci, say, 40 or more out of 50, may hamper its usefulness, it appears that the 50-SNP profile is greatly superior to the 20-SNP profile, especially under low (e.g., "cold hit") prior probability conditions. As will be shown in the demonstration to follow, however, this disparity in effectiveness is no reason to discard a 20-SNP database, if it exists, in favor of a 50-SNP one.

For this demonstration, the context shifts from a "culprit-suspect" scenario to one that involves unknown remains and a known missing person.

In this case, the investigator receives three distinct sets of DNA evidence. The first set consists of only four CODIS STR loci: D3S1358, with allele pair (17,17); VWA, with allele pair (19,19); FGA, with allele pair (22,23); and D8S1179, with allele pair (12,14). The second set of DNA evidence is a 20-SNP profile, of which 18 SNPs are homozygous. The third set of evidence is a profile comprising 50 SNPS, all different from those in the 20-SNP profile; of these 50 SNPs, 47 are homozygous. Although all three sets match the corresponding loci of an individual entry in a comprehensive database, the investigator chooses to start the analysis of the data with the very conservative "cold hit" prior probability, 0.0000000001, that the unknown source and the known entry are the same person.

The first evidence to be analysed is that of the partial CODIS profile. Using the "STR" version of the MCB method that is described in Reference 2[ii], the investigator obtains a Bayesian posterior probability of 0.00005 that the unknown and known persons are the same. This posterior can now be used as the prior for the analysis of the two SNP profiles.

However, instead of analysing the two SNP profiles serially, the investigator realises that, because they comprise entirely different SNPs, they can be combined into one profile. Thus, the 20-SNP profile, of which 18 SNPs are homozygous, combined with the 50-SNP profile, of which 47 SNPs are homozygous, yields a 70-SNP profile, of which 65 SNPs are homozygous.

A "cold hit" prior, when applied to this combined profile, results in a posterior probability of only 0.18. When, however, the results of the STR analysis, 0.00005, are used as the prior, the final result becomes the probability 0.99998 that the unknown source and the known individual are the same person.

The following part of the discussion returns to the context of a "culprit-suspect" scenario.

In a courtroom, the question may arise, "What is the likelihood of a SNP profile 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 a 50-SNP profile, the MCB method gives, as examples, the following random match likelihoods. If all 50 SNPs are homozygous, the random match likelihood is ~ 0.02; if 49 SNPs are homozygous, the random match likelihood is ~ 0.0004; and if 45 SNPs are homozygous, the random match likelihood is ~ 0.000000007. The corresponding likelihood ratios are ~ 50, ~ 2500, and ~ 1.43×10^8, respectively.

5. Conclusion

This article has demonstrated, through more than 110 computations taking a total of about 1 ¼ hours, that the Monte Carlo Bayesian approach enables the investigator to analyse SNP profiles quickly without depending on “reference groups”, to analyse composites of discrete SNP profiles, and to analyse combinations of STR and SNP profiles.

Appendix: The likelihood of a SNP profile

In the derivation that follows, it must be kept in mind that the investigator knows only how many loci the SNP profile contains, how many of the loci are heterozygous and how many are homozygous. It is, thus, impossible to know which of the two possible alleles at a locus is the A allele (with frequency *f*) and which is the B allele (with frequency $1 - f$).

The homozygous loci in the profile could be all AA, or all BB, or some mixture of the two. Therefore all possible mixtures are considered, and a weighted average, or expectation, of these mixtures' likelihoods is taken as the factor F_{HOM} that contributes to the profile likelihood. For this purpose, the reduced sample space that contains only homozygous loci, with correspondingly modified locus frequencies, is analysed by means of the Binomial Theorem, summing over a running index, *k*, that scans the field of possible mixtures from all-AA to all-BB.

Let there be a single nucleotide polymorphism (SNP) profile consisting of *N* dimorphic loci.

Let any dimorphic locus i contain either one or both of 2 alleles, denoted by A_i and B_i . That is, the locus can be described by (A_i,A_i) or (B_i,B_i) or (A_i,B_i) .

Let the population frequency of allele A_i be denoted by *f*; then the population frequency of B_i is $(1-f)$. The frequency *f* is assumed to be the same for all loci in the profile; $0 \leq f \leq 1$.

Let the number of homozygous loci in the profile be denoted by *n*. These *n* homozygous loci contribute the factor F_{HOM} to the profile's likelihood product.

The number of heterozygous loci in the profile in the profile is thus (*N*-*n*) . These (*N*-*n*) heterozygous loci contribute the factor F_{HET} to the profile's likelihood product.

The profile's likelihood, given the allele frequency *f*, is therefore

$P(\text{profile} \mid f) = F_{HET} * F_{HOM}$.

Let the number of (A,A) homozygous loci in the profile be denoted by *k*. Then the number of (B,B) homozygous loci is (*n*-*k*) .

Trivially, the factor F_{HET} can then be calculated as

$$F_{HET}(N,n,f) = [2f\,(1-f)]^{(N-n)} \quad .$$

In order to compute F_{HOM}, we define a reduced locus sample space that contains only homozygous loci; that is, it contains only those loci in the profile whose form is either (A,A) or (B,B) .

Now, in the complete sample space, the frequency of (A,A) loci is f^2; and, similarly, the frequency of (B,B) loci is $(1 - f)^2$.

Therefore, in the reduced space, the frequency, *p*, of (A,A) is

$$p = \frac{f^2}{f^2 + (1-f)^2} \quad .$$

Similarly, in the reduced space, the frequency of (B,B) is

$$(1-p) = \frac{(1-f)^2}{f^2 + (1-f)^2} \quad .$$

By the Binomial Theorem, the probability of having k loci of type (A,A), with single-trial probability *p*, in *n* homozygous loci, is ($0 \le k \le n$)

$$P[k; n, p] = \binom{n}{k} p^k (1-p)^{n-k} \quad .$$

Complementarily, the probability of having (*n*-*k*) loci of type (B,B), with single-trial probability (1-*p*), in *n* homozygous loci, is

$$P[(n-k); n, (1-p)] = \binom{n}{n-k} (1-p)^{n-k} \, p^k \quad .$$

The factor $F_{HOM}(n,f)$ can then be computed as the expectation, over k, of $f^{2k} * [(1-f)^2]^{(n-k)}$. That is,

$$F_{HOM}(n,f) = \sum_{k=0}^{k=n} \binom{n}{k} p^k (1-p)^{(n-k)} \left[f^{2k} (1-f)^{2(n-k)} \right] \quad , \text{ where}$$

$$p \equiv \frac{f^2}{f^2 + (1-f)^2} \quad .$$

	Prior probabilities		
	0.0000000001	0.00001	0.5
Number of homozygous SNPs in the profile	Posterior probabilities		
50	0.0000007	0.04	0.981
49	0.0000034	0.13	0.99962
48	0.000029	0.48	0.999985
47	0.000241	0.923	0.99999912
46	0.0026	0.993	0.999999931
45	0.024	0.9993	0.999999993
44	0.15	0.99992	0.9999999992
43	0.50	0.999989	0.9999999999
42	0.85	0.9999982	1.0000000000
41	0.968	0.99999967	. . .
40	0.993	0.99999993	. . .
30	0.999999924	1.000000000	. . .
20	0.9999999994	. . .	. . .
10	0.9999999968	. . .	. . .
0	0.9999989	. . .	. . .

Table 1. The probability, given a match between 50-SNP profiles, that the culprit is the suspect.

	Prior probabilities		
	0.0000000001	0.00001	0.5
Number of homozygous SNPs in the profile	Posterior probabilities		
20	0.000000005	0.00046	0.955
19	0.00000005	0.00051	0.99771
18	0.00000052	0.047	0.999782
17	0.0000035	0.26	0.999968
16	0.000018	0.62	0.999994
15	0.000068	0.865	0.99999843
14	0.00021	0.954	0.99999952
13	0.00058	0.9826	0.999999823
12	0.00136	0.9924	0.999999924
11	0.00267	0.9961	0.999999961
10	0.00444	0.99768	0.9999999768
9	0.00623	0.99838	0.9999999838
8	0.00757	0.99867	0.9999999867
7	0.00793	0.99874	0.9999999874
6	0.00716	0.99859	0.9999999860
5	0.00561	0.99822	0.9999999822
4	0.00398	0.99749	0.9999999749
3	0.00259	0.99613	0.9999999615
2	0.00162	0.99378	0.9999999375
1	0.00096	0.9895	0.999999894
0	0.00056	0.9820	0.999999814

Table 2. The probability, given a match between 20-SNP profiles, that the culprit is the suspect.

References

[i] [1.] S. Wright, Random drift and the shifting balance theory of evolution (Paper No. 1203 from the Department of Genetics, University of Wisconsin), in: Ken-Ichi Kojima (Ed.) Mathematical Topics in Population Genetics, Springer-Verlag, New York, Heidelberg, Berlin, 1970; pp. 1 - 30.

[ii] [2.] D. I. Promish, Monte Carlo Bayesian identification using STR profiles, Progress in Forensic Genetics 11, International Congress Series, vol. 1288, pp. 471 - 473, Elsevier B.V., 2006.