Genetica **96**: 107–117, 1995.

Applications of the Dirichlet distribution to forensic match probabilities

Kenneth Lange
Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, USA

Received 7 February 1994 Accepted 26 July 1994

Key words: genetic equilibrium, racial admixture, Monte Carlo, upper bounds

Abstract

The Dirichlet distribution provides a convenient conjugate prior for Bayesian analyses involving multinomial proportions. In particular, allele frequency estimation can be carried out with a Dirichlet prior. If data from several distinct populations are available, then the parameters characterizing the Dirichlet prior can be estimated by maximum likelihood and then used for allele frequency estimation in each of the separate populations. This empirical Bayes procedure tends to moderate extreme multinomial estimates based on sample proportions. The Dirichlet distribution can also be employed to model the contributions from different ancestral populations in computing forensic match probabilities. If the ancestral populations are in genetic equilibrium, then the product rule for computing match probabilities is valid conditional on the ancestral contributions to a typical person of the reference population. This fact facilitates computation of match probabilities and tight upper bounds to match probabilities.

Introduction

The introduction of DNA profile evidence in criminal cases has sparked one of the most spirited and, at times, acrimonious debates in recent scientific history (Chakraborty & Kidd, 1991; Devlin, Risch & Roeder, 1992; Jeffreys, Wilson & Thien, 1985; Lander, 1989; Lewin, 1989; Lewontin & Hartl, 1991). At the heart of this debate are disagreements about techniques for computing match probabilities. Defense experts contend that match probabilities as currently computed are meaningless because of the failure of the product rule in contemporary American populations. Application of the product rule requires that the reference population for the evidentiary DNA be in genetic equilibrium. Because of the racial heterogeneity in the United States, no one can honestly claim that the population as a whole is in genetic equilibrium. However, this fact should not signal the end of the debate. There is a conditional form of the product rule that holds even in the presence of genetic heterogeneity (Lange, 1993). This conditional form provides a wedge for valid calculation of match probabilities. Of course, to make this proposal workable, we must introduce some approximations about the nature of the genetic contributions from the various subpopulations. This is where the Dirichlet distribution proves useful.

The Dirichlet distribution provides a flexible way of parameterizing the contributions of the subpopulations to a proposed reference population. By the reference population for a particular crime, we simply mean the collection of people who could have conceivably contributed the evidentiary DNA. This vague hypothetical construct may vary from locale to locale and from case to case. The particular ancestry of the alleged contributor of the evidentiary DNA, be he perpetrator or victim, is largely irrelevant to the choice of the reference population. Of course, the postulated racial composition of the reference population should be broad enough to reflect the ancestry of the alleged contributor.

The mathematical attractiveness of the Dirichlet distribution stems from the fact that it makes necessary expectations trivial to compute. The following sections document the mathematical manipulations involved in computing match probabilities via the Dirichlet distribution. Readers primarily interested in general conclusions can skip these details and turn directly to the discussion at the end of the paper. Some limitations of the Dirichlet model are mentioned there.

The Dirichlet distribution is also relevant to the related problem of allele frequency estimation. If one adopts a Bayesian framework for estimation, then a Dirichlet prior for the multinomial distribution of allele counts leads to a Dirichlet posterior. Classical statisticians will object that the choice of any particular prior from the Dirichlet family is bound to be arbitrary. When data from several populations are available, one can choose the prior empirically from the data. This empirical Bayes procedure incorporates some of the best features of classical and Bayesian statistics. The resulting Bayesian estimates of allele frequencies tend to moderate the extremes seen in classical estimates based on sample proportions. We develop this perspective in a preliminary digression that may be of independent interest to many readers.

Empirical Bayes estimation of allele frequencies

Consider a locus with k codominant alleles. To estimate the frequencies $p_1, \ldots, p_k$ of these alleles in some population, suppose one takes a random sample from the population and observes n_i genes of type i. Then $n_i / n_.$ is the maximum likelihood estimate of p_i, where the abbreviation $n_. = \sum_{i=1}^{k} n_i$ relies on the usual summation convention. This frequentist estimate based on the multinomial distribution can be contrasted to a Bayesian estimate using a Dirichlet prior for the allele frequencies (Good, 1965). The Dirichlet distribution (Kingman, 1993) with parameters $\gamma_1, \ldots, \gamma_k > 0$ has density

$$\frac{\Gamma(\gamma_.)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \prod_{i=1}^{k} p_i^{\gamma_i - 1} \qquad (1)$$

on the simplex

$$\Delta_k = \left\{ (p_1, \ldots, p_k) \colon p_1 > 0, \ldots, p_k > 0, \ \sum_{i=1}^{k} p_i = 1 \right\}$$

endowed with the uniform measure. One of the virtues of the Dirichlet distribution is the elegant moment formula

$$E\left(\prod_{i=1}^{k} p^{t_i} \right) = \frac{\Gamma(\gamma_.)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \int_{\Delta_k} \prod_{i=1}^{k} p_i^{t_i + \gamma_i - 1} dp$$

$$= \frac{\Gamma(\gamma_.)}{\Gamma(t_. + \gamma_.)} \prod_{i=1}^{k} \frac{\Gamma(t_i + \gamma_i)}{\Gamma(\gamma_i)}. \qquad (2)$$

The Dirichlet prior is a conjugate prior for the multinomial distribution (Lee, 1989). In the current context, this means that if the allele frequency vector $p = (p_1, \ldots, p_k)$ has a Dirichlet prior with parameters $\gamma_1, \ldots, \gamma_k$, then based on the sample, p has a Dirichlet posterior with revised parameters $n_1 + \gamma_1, \ldots, n_k + \gamma_k$. This fact follows from an application of the moment formula (2) in the conditional density computation

$$\frac{\frac{\Gamma(\gamma_.)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \binom{n}{n_1 \ldots n_k} \prod_{i=1}^{k} p_i^{n_i + \gamma_i - 1}}{\frac{\Gamma(\gamma_.)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \binom{n}{n_1 \ldots n_k} \int_{\Delta_k} \prod_{i=1}^{k} q_i^{n_i + \gamma_i - 1} dq}$$

$$= \frac{\Gamma(n_. + \gamma_.)}{\prod_{i=1}^{k} \Gamma(n_i + \gamma_i)} \prod_{i=1}^{k} p_i^{n_i + \gamma_i - 1}.$$

A second application of (2) now implies that $(n_i + \gamma_i)/(n_. + \gamma_.)$ is the posterior mean of p_i. The posterior mean is a strongly consistent, asymptotically unbiased estimator of the true p_i.

The primary disadvantage of taking a Bayesian stance on allele frequency estimation is that there is no obvious way of selecting a reasonable prior. However, if data from several distinct populations are available, then one can select an appropriate prior empirically. Consider the marginal distribution of the allele counts $(N_1, \ldots, N_k)$ in a sample of genes from a single population. Integrating out the prior on the allele frequency vector $p = (p_1, \ldots, p_k)$ yields the predictive distribution (Mosimann, 1962)

$$\Pr(N_1 = n_1, \ldots, N_k = n_k)$$

$$= \binom{n}{n_1 \ldots n_k} \frac{\Gamma(\gamma_.)}{\Gamma(n_. + \gamma_.)} \prod_{i=1}^{k} \frac{\Gamma(n_i + \gamma_i)}{\Gamma(\gamma_i)}. \qquad (3)$$

This distribution is known as the Dirichlet-multinomial distribution. Its parameters are the γ's rather than the p's

With independent data from several distinct populations, one can estimate the parameter vector $\gamma = (\gamma_1, \ldots, \gamma_k)$ of the Dirichlet-multinomial distribution by maximum likelihood. Newton's method offers by far the fastest means of finding the maximum likelihood estimate. To implement Newton's method, one needs the loglikelihood $L(\gamma)$, the score vector $dL(\gamma)$, and the observed information matrix $-d^2L(\gamma)$ for each population. Elementary calculus based on the likelihood (3) shows that the score has entries.

$$\begin{aligned} \frac{\partial}{\partial \gamma_i} L(\gamma) &= D(\gamma.) - D(n. + \gamma.) \\ &\quad + D(n_i + \gamma_i) - D(\gamma_i), \end{aligned} \tag{4}$$

where $D(s) = d/ds \ln \Gamma(s)$ is the digamma function (Hille, 1959). The observed information has entries

$$\begin{aligned} -\frac{\partial^2}{\partial \gamma_i \partial \gamma_j} L(\gamma) &= -T(\gamma.) + T(n. + \gamma.) - \chi_{\{i=j\}} \\ &\quad \times [T(n_i + \gamma_i) - T(\gamma_i)], \end{aligned} \tag{5}$$

where $\chi_{\{i=j\}}$ is the indicator function of the event $\{i = j\}$, and where $T(s)$ is the trigamma function $d^2/ds^2 \ln \Gamma(s)$ (Hille, 1959). The digamma and trigamma functions appearing in the expressions (4) and (5) should not be viewed as a major barrier to computation since good software for evaluating these transcendental functions exists (Bernardo, 1976; Schneider, 1978).

Extending the above notation for the loglikelihood, score, and observed information from a single population to the entire random sample from several populations, Newton's method updates the current parameter iterate γ^m by

$$\gamma^{m+1} = \gamma^m - d^2L(\gamma^m)^{-1} dL(\gamma^m). \tag{6}$$

For Newton's method to move in an uphill direction, the observed information matrix $-d^2L(\gamma^m)$ should be positive definite. This may not always be the case. The obvious remedy is to replace the observed information $-d^2L(\gamma^m)$ by an approximating matrix that is positive definite.

Equation (5) for a single population evidently can be summarized in matrix form by

$$-d^2L(\gamma) = D - c\mathbf{1}\mathbf{1}^t, \tag{7}$$

where D is a diagonal matrix with ith diagonal entry $d_i = T(\gamma_i) - T(n_i + \gamma_i)$, c is the constant $T(\gamma.) - T(n. + \gamma.)$, and $\mathbf{1}$ is a column vector of all 1's. Because the trigamma function is decreasing (Hille, 1959), $d_i > 0$ when $n_i > 0$; the constant $c > 0$ always. Since the representation (7) is preserved under finite sums, it holds, in fact, for the entire sample.

The idea now is to approximate $-d^2L(\gamma)$ by the right side of (7) with a decreased value of the constant c if necessary. If the right side of (7) is to be positive definite, then

$$\mathbf{1}^t D^{-1}(D - c\mathbf{1}\mathbf{1}^t)D^{-1}\mathbf{1} = \mathbf{1}^t D^{-1}\mathbf{1}(1 - c\mathbf{1}^t D^{-1}\mathbf{1})$$

must be positive. This implies

$$1 - c\mathbf{1}^t D^{-1}\mathbf{1} = 1 - c\sum_{i=1}^{k} \frac{1}{d_i} > 0. \tag{8}$$

Conversely, inequality (8) is sufficient for the right side of (7) to be positive definite. This fact can be most easily demonstrated by noting the Sherman-Morrison formula (Miller, 1987)

$$(D - c\mathbf{1}\mathbf{1}^t)^{-1} = D^{-1} + \frac{c}{1 - c\mathbf{1}^t D^{-1}\mathbf{1}} D^{-1}\mathbf{1}\mathbf{1}^t D^{-1} \tag{9}$$

Formula (9) proves that $(D - c\mathbf{1}\mathbf{1}^t)^{-1}$ exists and is positive definite under assumption (8). Since the inverse of a positive definite matrix is positive definite, it follows that $D - c\mathbf{1}\mathbf{1}^t$ is positive definite.

These results suggest that c be replaced by $\min\left\{c, (1-\epsilon)/\left(\sum_{i=1}^{k} d_i^{-1}\right)\right\}$, where ϵ is a small positive constant. With this substitution and with occasional backtracking to avoid overshooting the maximum of $L(\gamma)$ along the current Newton direction, Newton's method can proceed safely. Near the maximum likelihood point, $-d^2L(\gamma)$ will be positive definite, and no adjustment of it is necessary. Throughout the iterations the Sherman-Morrison formula can be used to invert $-d^2L(\gamma)$ or its substitute.

Example of the empirical Bayes procedure

Edwards *et al.* (1992) gathered population data in Houston, Texas on the eight alleles of the HUMTH01 locus on chromosome 11. This is a tandem repeat locus whose allele names refer to numbers of repeat units. From the four separate subpopulations of Caucasians, blacks, Chicanos, and Asians, the eight γ's are estimated by maximum likelihood to be .11, 4.64, 7.33, 2.97, 5.32, 5.26, .27, and .10. Using these estimated Dirichlet parameters, Table 1 compares the maximum likelihood estimates (top row) and posterior mean estimates (bottom row) of the allele frequencies within each subpopulation. It is noteworthy that all posterior means are within one standard error of the maximum likelihood estimates. (These standard errors are given in Table 2 of Edwards *et al.*, 1992.) Nonetheless, the empirical Bayes procedure does tend to moderate the

Table 1. Classical and Bayesian allele frequency estimates.

Estimator	Allele	Caucasian	Black	Chicano	Asian
Classical	5	.0054	.0000	.0000	.0000
Bayesian	5	.0053	.0003	.0003	.0006
Classical	6	.2258	.1351	.2083	.1039
Bayesian	6	.2227	.1380	.2064	.1147
Classical	7	.1586	.3703	.3333	.2597
Bayesian	7	.1667	.3645	.3301	.2630
Classical	8	.1102	.2108	.0677	.0519
Bayesian	8	.1105	.2045	.0707	.0609
Classical	9	.1425	.1459	.1432	.4416
Bayesian	9	.1465	.1498	.1471	.4073
Classical	10	.3522	.1378	.2474	.0909
Bayesian	10	.3424	.1421	.2445	.1070
Classical	11	.0054	.0000	.0000	.0455
Bayesian	11	.0057	.0007	.0007	.0404
Classical	12	.0000	.0000	.0000	.0065
Bayesian	12	.0002	.0002	.0002	.0061
Sample size *n*		372	370	384	154

extremes in estimated allele frequencies seen in the different subpopulations. In particular, all posterior mean estimates are positive. The maximum likelihood estimates suggest that those alleles failing to appear in a sample are nonexistent in the corresponding subpopulation. The posterior mean estimates suggest more reasonably that such alleles are simply rare in the subpopulation.

Match probabilities for a population at equilibrium

For a single population in Hardy-Weinberg equilibrium, forensic match probabilities can be computed from either a frequentist or a Bayesian perspective. If the allele frequencies are known without error, then the match probabilites for homozygous *i*/*i* genotypes and heterozygous genotypes *i*/*j* are p_i^2 and $2p_ip_j$, respectively. In practice, the frequencies p_i can only be estimated. Assuming codominant alleles and the maximum likelihood estimates $\hat{p}_i = n_i/n_{.}$, the frequentist genotype estimates have the following sampling properties (Chakraborty, Srinivasan & Daiger, 1993):

$$\begin{aligned} E(\hat{p}_i^2) &= p_i^2 + \frac{p_i(1-p_i)}{n_{.}} \\ Var(\hat{p}_i^2) &= \frac{4p_i^3(1-p_i)}{n_{.}} + O\left(\frac{1}{n_{.}^2}\right) \\ E(2\hat{p}_i\hat{p}_j) &= 2p_ip_j - \frac{2p_ip_j}{n_{.}} \\ Var(2\hat{p}_i\hat{p}_j) &= \frac{4p_ip_j}{n_{.}}(p_i + p_j - 4p_ip_j) + O\left(\frac{1}{n_{.}^2}\right). \end{aligned}$$

From the Bayesian perspective, the genotype probabilities are random variables whose distributions depend on the posterior distribution of the allele frequencies. Based on the moment formula (2), it is straightforward to compute that

$$\begin{aligned} &E(p_i^2|N_1 = n_1, \ldots, N_k = n_k) \\ &= \frac{(n_i+\gamma_i)^{\bar{2}}}{(n_{.}+\gamma_{.})^{\bar{2}}} \\ &Var(p_i^2|N_1 = n_1, \ldots, N_k = n_k) \\ &= \frac{(n_i+\gamma_i)^{\bar{4}}}{(n_{.}+\gamma_{.})^{\bar{4}}} - \left[\frac{(n_i+\gamma_i)^{\bar{2}}}{(n_{.}+\gamma_{.})^{\bar{2}}}\right]^2 \\ &E(2p_ip_j|N_1 = n_1, \ldots, N_k = n_k) \\ &= 2\frac{(n_i+\gamma_i)(n_j+\gamma_j)}{(n_{.}+\gamma_{.})^{\bar{2}}} \\ &Var(2p_ip_j|N_1 = n_1, \ldots, N_k = n_k) \\ &= \frac{4(n_i+\gamma_i)^{\bar{2}}(n_j+\gamma_j)^{\bar{2}}}{(n_{.}+\gamma_{.})^{\bar{4}}} \end{aligned}$$

$$- \left[\frac{2(n_i+\gamma_i)(n_j+\gamma_j)}{(n_.+\gamma_.)^2}\right]^2,$$

where $x^{\overline{r}} = x(x+1)\cdots(x+r-1)$ denotes a rising power. It is interesting that the above mean expressions entail

$$\begin{aligned} E(p_i^2|N_1 = n_1, \ldots, N_k = n_k) &> \tilde{p}_i^2 \\ E(2p_ip_j|N_1 = n_1, \ldots, N_k = n_k) &< 2\tilde{p}_i\tilde{p}_j, \end{aligned}$$

where $\tilde{p}_i$ and $\tilde{p}_j$ are the posterior means of p_i and p_j.

As a numerical example, consider the 5/5 and 5/6 genotypes for Caucasians at the HUMTH01 locus of Table 1. Using the maximum likelihood estimates of p_5 and p_6, these genotypes have predicted frequencies of .0000289 and .00243, respectively. These values can be compared to the Bayesian values $E(p_5^2) = .0000412$ and $E(2p_5p_6) = .00235$ and their standard deviations $\sqrt{Var(p_5^2)} = .0000608$ and $\sqrt{Var(2p_5p_6)} = .00162$. Let us emphasize here that the reference population for the two matches is Caucasian and that conditional expectations and variances are abbreviated for the sake of convenience as ordinary expectations and variances.

Chakraborty, Srinivasan and Daiger (1993) point out that it is probably more relevant to consider the sampling distribution of $\ln p_i^2$ and $\ln 2p_ip_j$, since the log transformation focuses attention on the order of magnitude of a match probability and makes the central limit theorem applicable when a match probability is computed over several independent loci. In the Bayesian context, one can compute the moments of the random vector $(\ln p_1, \ldots, \ln p_k)$ through its multivariate moment generating function

$$\begin{aligned} &E\left(e^{\sum_{i=1}^k t_i \ln p_i}|N_1 = n_1, \ldots, N_k = n_k\right) \\ &= E\left(\prod_{i=1}^k p_i^{t_i}|N_1 = n_1, \ldots, N_k = n_k\right), \end{aligned}$$

which is given explicitly by equation (2) with $n_i + \gamma_i$ replacing γ_i. Entirely straightforward, but slightly tedious calculations show that

$$\begin{aligned} &E(\ln p_i^2|N_1 = n_1, \ldots, N_k = n_k) \\ &= 2[D(n_i+\gamma_i) - D(n_.+\gamma_.)] \\ &Var(\ln p_i^2|N_1 = n_1, \ldots, N_k = n_k) \\ &= 4[T(n_i+\gamma_i) - T(n_.+\gamma_.)] \\ &E(\ln 2p_ip_j|N_1 = n_1, \ldots, N_k = n_k) \\ &= \ln 2 + D(n_i+\gamma_i) + D(n_j+\gamma_j) - 2D(n_.+\gamma_.) \\ &Var(\ln 2p_ip_j|N_1 = n_1, \ldots, N_k = n_k) \\ &= T(n_i+\gamma_i) + T(n_j+\gamma_j) - 4T(n_.+\gamma_.). \end{aligned}$$

Consider again the 5/5 and 5/6 genotypes at the HUMTH01 locus among Caucasians. Using the maximum likelihood estimates of p_5 and p_6, these genotypes have predicted log frequencies of −10.45 and −6.02 to the base e, respectively. These classical values can be compared to the Bayesian values $E(\ln p_5^2) = -10.99$ and $E(\ln 2p_5p_6) = -6.31$ and their standard deviations $\sqrt{Var(\ln p_5^2)} = 1.55$ and $\sqrt{Var(\ln 2p_5p_6)} = .78$.

Now contemplate m codominant loci in Hardy-Weinberg and linkage equilibrium. Let G_l be the probability of an observed genotype at locus l. Thus, each G_l corresponds to either an expression p_i^2 for a homozygote or to an expression $2p_ip_j$ for a heterozygote. Given a random sample of genotypes at locus l, the moments of the random variables G_l and $\ln G_l$ can be calculated as indicated above based on a Dirichlet posterior distribution of allele frequencies. Since the random variables G_l are independent under the assumption of linkage equilibrium, a multilocus match probability $\prod_{l=1}^m G_l$ has posterior mean and variance

$$\begin{aligned} E\left(\prod_{l=1}^m G_l\right) &= \prod_{l=1}^m E(G_l) \\ Var\left(\prod_{l=1}^m G_l\right) &= \prod_{l=1}^m E(G_l^2) - \prod_{l=1}^m E(G_l)^2 \\ &= \prod_{l=1}^m [Var(G_l) + E(G_l)^2] \\ &\quad - \prod_{l=1}^m E(G_l)^2. \end{aligned}$$

Similarly, $\ln \prod_{l=1}^m G_l$ has posterior mean and variance

$$\begin{aligned} E\left(\sum_{l=1}^m \ln G_l\right) &= \sum_{l=1}^m E(\ln G_l) \\ Var\left(\sum_{l=1}^m \ln G_l\right) &= \sum_{l=1}^m Var(\ln G_l). \end{aligned}$$

Of course, all of these means and variances are conditional on the sampled data.

Match probabilities in admixed populations

Once the assumptions of Hardy-Weinberg and linkage equilibrium fail, calculation of match probabilities become problematic. In particular, the product rule for combining match probabilities across separate loci no longer holds. One device for rescuing the product rule is to condition on the ancestry of a typical person from the reference population (Lange, 1993; Mickey *et al.*, 1983). This ancestry should consist of contributions from a finite number of specified ancestral populations that individually are assumed to be at equilibrium even when the reference population is not. A convenient way of parameterizing these contributions is to postulate that a proportion x_i of the maternal genes and y_i of the paternal genes of the typical person originate from ancestral population *i*. The mother and father of the typical person are assumed to be unrelated.

Assuming that there are n ancestral populations, it is again convenient to assign Dirichlet distributions to the random vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. If the common Dirichlet parameter associated with the components x_i and y_i is α_i, then ancestral population *i* contributes a proportion $\alpha_i/\alpha_{.}$ of the genes observed in the reference population. Of course, the exact ancestral contributions will vary from person to person in the reference population. If the total $\alpha_{.}$ of the α_i is close to 0, then most people will exhibit maternal and paternal vectors *x* and *y* having a single component close to 1 and the remaining components close to 0. This is consistent with little mixing of the races in the reference population. Large values of $\alpha_{.}$ suggest a thorough mixing of the races.

Two extremes relating *x* and *y* are apt to be important in practice. On one hand, the maternal contributions *x* might be independent of the paternal contributions *y*. On the other hand, total endogamy might prevail, in which event *x* and *y* coincide. In between these two extremes is partial endogamy, where *x* and *y* are independent with probability β and coincide with probability $1 - \beta$.

To compute a multilocus match probability, we adopt the allele frequency notation p_{ijk} with three subscripts *i*, *j*, and *k*, indicating population, locus, and allele, respectively. The posterior Dirichlet parameter that corresponds to p_{ijk} we denote by θ_{ijk}, condensing into a single symbol the sum of the prior Dirichlet parameter and the number of genes sampled for this combination of population, locus, and allele. Because sampling to estimate allele frequencies in ancestral populations has little to do with determining the racial composition of a hypothetical reference population, it is sensible to assume that the racial admixture proportions are independent of the posterior allele frequencies. It is also reasonable to assume that posterior allele frequencies are independent from one ancestral population to the next and in view of linkage equilibrium from one locus to the next within an ancestral population.

Now consider a multilocus genotype defined by genotype k_j/l_j at locus *j*, where *j* ranges over some prescribed set of m loci situated on m different chromosomes. The conditional probability of observing this multilocus genotype is

$$\begin{aligned} & w\prod_j \left[\left(\sum_u x_u p_{ujk_j}\right)\left(\sum_v y_v p_{vjl_j}\right)\right. \\ & \left. +\chi_{\{k_j\neq l_j\}}\left(\sum_u x_u p_{ujl_j}\right)\left(\sum_v y_v p_{vjk_j}\right)\right] \\ & +(1-w)\prod_j \left[(1+\chi_{\{k_j\neq l_j\}})\left(\sum_u x_u p_{ujk_j}\right)\right. \\ & \left. \times\left(\sum_v x_v p_{vjl_j}\right)\right], \end{aligned} \tag{10}$$

where the random variable *w* is independent of all other random variables in sight and indicates whether the ancestral proportions *x* and *y* are independent $(w = 1)$ or agree $(w = 0)$. In similar fashion, the function $\chi_{\{k_j\neq l_j\}}$ indicates whether the observed genotype at locus *j* is heterozygous. The expectation of (10) is by definition the multilocus match probability.

The best methods of evaluating the match probability all hinge on first computing the conditional expectation of (10) with respect to the ancestral proportions *x* and *y*. Owing to the various independence assumptions and to the fact that $E(w) = \beta$, this conditional expectation reduces to

$$\begin{aligned} & \beta\prod_j \left[\sum_u\sum_v x_u y_v E(p_{ujk_j}p_{vjl_j})\right. \\ & \left. +\chi_{\{k_j\neq l_j\}}\sum_u\sum_v x_u y_v E(p_{ujl_j}p_{vjk_j})\right] \\ & +(1-\beta)2^h\prod_j\left[\sum_u\sum_v x_u x_v E(p_{ujk_j}p_{vjl_j})\right] \end{aligned} \tag{11}$$

if *h* heterozygous genotypes are observed among the m loci. The expression (11) evidently reflects the fact that conditional on ancestry, match probabilities obey the product rule. Motivation and explanation for this conditional product rule is given in Lange (1993) and will not be repeated here.

The ordinary expectations appearing in (11) fortunately yield to the moment formula (2). Indeed,

$$E(p_{ujk_j}p_{vjl_j}) = \begin{cases} \dfrac{\theta_{ujk_j}}{\theta_{uj.}}\dfrac{\theta_{vjl_j}}{\theta_{vj.}} & u \neq v \\ \dfrac{\theta_{ujk_j}\theta_{ujl_j}}{\theta_{uj.}^{\bar{2}}} & u = v \; k_j \neq l_j . \\ \dfrac{\theta_{ujk_j}^{\bar{2}}}{\theta_{uj.}^{\bar{2}}} & u = v \; k_j = l_j \end{cases}$$

As before, the dot subscript indicates summation over an omitted index.

If we now naively take the expectation of (11) and use the distributive rule, we are faced with evaluating n^{2m} terms of type

$$E\left(\prod_j x_{u_j}x_{v_j}\right)\prod_j E(p_{u_jjk_j}p_{v_jjl_j}) \qquad (12)$$

and $2^h n^{2m}$ terms of type

$$E\left(\prod_j x_{u_j}\right)E\left(\prod_j y_{v_j}\right)\prod_j E(q_{u_jj}r_{v_jj}), \quad (13)$$

where $q_{u_jj} = p_{u_jjk_j}$ and $r_{v_jjl_j} = p_{v_jjl_j}$, or $q_{u_jj} = p_{u_jjl_j}$ and $r_{v_jj} = p_{v_jjk_j}$. To evaluate $E\left(\prod_j x_{u_j}\right)$, suppose the variable m_k counts the number of $u_j = k$. Then

$$E\left(\prod_j x_{u_j}\right) = \frac{\prod_k \alpha_k^{\overline{m_k}}}{\alpha_{.}^{\overline{m}}}.$$

Similarly, if m_k counts the number of $u_j = k$ plus the number of $v_j = k$, then

$$E\left(\prod_j x_{u_j}x_{v_j}\right) = \frac{\prod_k \alpha_k^{\overline{m_k}}}{\alpha_{.}^{\overline{2m}}}.$$

Thus, the principal barrier to computation is not evaluation of the various expectations, but rather the sheer number of terms that must be summed.

If we are satisfied with an upper bound on the match probability, we can consider the x_u and y_v appearing in (11) to be parameters rather than random variables and then find the maximum of (11) with respect to x and y. This is the point of view taken in Lange (1993). The resulting bound is valid regardless of the joint distribution assigned to x and y. A better bound is available under our current Dirichlet assumptions about x and y. This improved bound follows from the simple observation that the function $s \rightarrow (a+s)/(b+s)$ is increasing provided $0 < a \leq b$. Given this fact, the expectation $E(p_{u_jjk_j}p_{v_jjl_j})$ satisfies the inequality

$$\begin{aligned} E(p_{u_jjk_j}p_{v_jjl_j}) &\leq c_{u_jjk_j}d_{v_jjl_j} \\ &\leq d_{u_jjk_j}d_{v_jjl_j}, \end{aligned} \qquad (14)$$

where $c_{u_jjk_j} = \theta_{u_jjk_j}/\theta_{u_jj.}$ and $d_{v_jjl_j} = (\theta_{v_jjl_j} + 1)/(\theta_{v_jj.} + 1)$. Both upper bounds in (14) will be close to $c_{u_jjk_j}c_{u_jjl_j}$ if the random gene sample from population v_j at locus j is reasonably large.

When the first upper bound in (14) is substituted in (12) and (13), it follows that the match probability is bounded above by

$$\begin{aligned} &\beta E\left[\prod_j \left\{\left(\sum_u x_u c_{ujk_j}\right)\left(\sum_v y_v d_{vjl_j}\right)\right.\right. \\ &\left.\left. +\chi_{\{k_j \neq l_j\}}\left(\sum_u x_u c_{ujl_j}\right)\left(\sum_v y_v d_{vjk_j}\right)\right\}\right] \\ &+(1-\beta)2^h E\left[\prod_j \left(\sum_u x_u c_{ujk_j}\right)\right. \\ &\left.\times\left(\sum_v x_v d_{vjl_j}\right)\right]. \end{aligned} \qquad (15)$$

In the absence of endogamy, this bound is vastly simpler to evaluate than the original match probability. Indeed, the expectation (15) now splits into 2^h terms of type

$$\begin{aligned} &E\left[\prod_j \left(\sum_u x_u e_{uj}\right)\left(\sum_v y_v f_{vj}\right)\right] \\ &= E\left[\prod_j \left(\sum_u x_u e_{uj}\right)\right] E\left[\prod_j \left(\sum_v y_v f_{vj}\right)\right], \end{aligned}$$

where $e_{uj} = c_{ujk_j}$ and $f_{vj} = d_{vjl_j}$, or $e_{uj} = c_{ujl_j}$ and $f_{vj} = d_{vjk_j}$. Evaluation of

$$E\left[\prod_j \left(\sum_u x_u e_{uj}\right)\right]$$

$$= \sum_{u_1} \cdots \sum_{u_m} E\left(\prod_j x_{u_j}\right) \prod_j e_{u_j j}$$

involves summing only n^m terms rather than n^{2m} terms. It is easy to imagine evaluating match probabilities involving $n = 10$ or 20 populations at $m = 5$ or 6 loci.

As an alternative to exact computation of an upper bound, one can estimate the match probability by Monte Carlo simulation based on expression (11). To simulate the ancestral proportions x and y, it is helpful to note that if $Z_1, \ldots, Z_n$ are independent random variables with Z_i having gamma density $z_i^{\alpha_i - 1} e^{-z_i}/\Gamma(\alpha_i)$, then the random proportions $W_1, \ldots, W_n$ defined by

$$W_i = \frac{Z_i}{\sum_{j=1}^{n} Z_j}$$

follow a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_n$ (Kingman, 1993). Gamma distributed random variables can be simply and economically simulated by an acceptance-rejection method for $\alpha_i \leq 1$ (Ahrens & Dieter, 1974) or by a ratio method for $\alpha_i > 1$ (Cheng & Feast, 1979).

Sample match probability calculations

As an example of the above calculations, consider the four loci HUMHPRTB, HUMTH01, HUMRENA, HUMFABP featured in Edwards *et al.* (1992). These tandem repeat loci occur on chromosomes *X*, 11, 1, and 4, respectively. For the purposes of this example, we will pretend that locus HUMHPRTB is autosomal and seek the match probability of the typical multilocus genotype 6/12, 7/9, 8/8, and 10/11. As already described for the HUMTH01 locus, it is possible to estimate the parameters of the Dirichlet-multinomial distribution for each locus from the data of Table 2 of Edwards *et al.* (1992). Once this is done, the posterior allele frequency parameters θ_{ijk} are immediately available. For the ancestral parameters, we select the approximate American proportions

$$\frac{\alpha_{\text{Caucasian}}}{\alpha_.} = .7$$
$$\frac{\alpha_{\text{black}}}{\alpha_.} = .15$$
$$\frac{\alpha_{\text{Chicano}}}{\alpha_.} = .1$$
$$\frac{\alpha_{\text{Asian}}}{\alpha_.} = .05.$$

The total $\alpha_.$ of the ancestral parameters is more difficult to determine. As mentioned earlier, large values of $\alpha_.$ are consistent with widespread racial admixture and small values of $\alpha_.$ with little admixture. The parameter β controls the degree of endogamy between the parents of a typical person of the reference population. Because of doubt about the exact values of $\alpha_.$ and β, it is crucial to test the sensitivity of the match probability to a range of values of these two parameters.

Table 2 presents the results of our calculations. The Monte Carlo estimates are based on 10,000 samples each of the ancestral proportions *x* and *y*. As noted above, the Dirichlet upper bound is relevant only when endogamy is absent ($\beta = 1$). By way of comparison, the upper bound to the match probability previously suggested in Lange (1993) is 6.3×10^{-6}, about double the Dirichlet upper bound. If one naively ignores the racial stratification of the population, then one arrives at a match probability of 3.5×10^{-6} or 1.9×10^{-6} (Chakraborty & Kidd, 1991). The first of these depends on allele frequencies estimates derived by pooling the sampled genes from the four separate subpopulations. The second depends on allele frequencies that are weighted averages of the maximum likelihood allele frequencies calculated for the separate subpopulations, with weight $\alpha_i/\alpha_.$ given to subpopulation *i*.

It is noteworthy how close the naive match probabilities are to the match probabilities based on the Dirichlet model. The parameters $\alpha_.$ and β appear to have little effect on computed values under the Dirichlet model.

Discussion

The empirical Bayes approach to allele frequency estimation has much to offer. Because populations are never totally isolated, allele information from one population is bound to be relevant to other populations. Empirical Bayes procedures permit propagation of partial knowledge from the whole to its parts. At the same time, Bayesian estimates conform to the dictates of data as more data are gathered. The practical effect of these adaptive features is a moderation of the extreme allele estimates produced by the sample proportions in each population. In particular, small allele frequency estimates tend to increase under the Bayesian procedures. Since match probabilities are sensitive to small allele frequencies, Bayesian multilocus match probabilities

Table 2. Calculations for a multilocus match probability.

Ancestral Total $\alpha_.$	Monte Carlo $\beta = 1$	Monte Carlo $\beta = .5$	Monte Carlo $\beta = 0$	Dirichlet Bound
.1	2.7×10^{-6}	2.5×10^{-6}	2.4×10^{-6}	2.8×10^{-6}
1	2.9×10^{-6}	2.8×10^{-6}	2.7×10^{-6}	3.0×10^{-6}
10	3.1×10^{-6}	3.1×10^{-6}	3.1×10^{-6}	3.3×10^{-6}
100	3.2×10^{-6}	3.2×10^{-6}	3.2×10^{-6}	3.3×10^{-6}

are less influenced by rare genotypes than classical match probabilities.

In our example dealing with just four populations, the empirical Bayes estimates show these desirable properties. Doubtless, sampling a larger number of populations would enhance the empirical choice of a prior. The empirical Bayes procedure is apt to be most advantageous when the number of populations is large and the sample size per population is relatively small. Economic constraints on data collection suggest that this situation is likely to occur as more populations are sampled.

Stochastic variation enters at two levels in computing match probabilities. First, match probabilities are usually computed under the assumption that allele frequencies are known with infinite precision. When limited samples are available to estimate allele frequencies, it is helpful to include this uncertainty in match probability estimation (Chakraborty, Srinivasan & Daiger, 1993). We have shown how to accomplish this in the Bayesian context for a population at genetic equilibrium. Our explicit expressions for posterior means and variances of genotype probabilities and their logarithms are straightforward to evaluate.

The second source of stochastic variation arises from our uncertainty about the degree of racial admixture in contemporary populations. Even within groups that society lumps into a single ethnic category, there can be considerable genetic heterogeneity. Hispanics are a good case in point. Because genetic equilibrium is required for application of the product rule, it is necessary to revert to a weaker form of the product rule. As argued in Lange (1993), the product rule is valid conditional on the ancestry of a typical person in the reference population. In other words, if we imagine tracing back the pedigree of the typical person to his ancestors in defined populations at equilibrium, then the presence of equilibrium in the ancestral populations implies independence of his genotypes at loci occurring on different chromosomes. In practice, we replace this hypothetical pedigree by the proportion x_i of his maternal genes and y_i of his paternal genes contributed by the *i*th ancestral population.

If we make the further assumption that these proportions follow Dirichlet distributions, then we can write explicit expressions for multilocus match probabilities. Even with a moderate number of loci, these expressions involve large numbers of terms. Nonetheless, it is possible to evaluate them accurately by simulation and to provide simple upper bounds amenable to exact evaluation. The example in Table 2 illustrates the close agreement between simulated values and upper bounds that can be achieved in practice. Although this particular example suggests that naive use of the product rule is acceptable, it is premature to make this generalization. As more allele frequency data accumulate from different ancestral populations, further comparisons of the Dirichlet match probabilities and those computed by the product rule should be undertaken.

Although the Dirichlet model does provide a more sophisticated basis for calculation, it still entails approximations and intentional simplifications. For instance, the model omits laboratory errors, the possibility of confused DNA samples, and fraud within the criminal justice system. The later two issues are hard to quantify, but juries need to bear them in mind. Incorporation of data from the currently used VNTR loci necessarily involves an analysis of laboratory measurement errors (Devlin, Risch & Roeder, 1992). It is debatable whether the additional information content of these highly polymorphic loci is worth their price in phenotypic ambiguity. One can argue that substituting more loci with less information content per locus is preferable (Lange, 1991). This substitution would help, for example, in distinguishing the culprit or victim from his close relatives such as siblings, an important issue not directly addressed by the Dirichlet model. Evett (1992) suggests computational remedies for the sibling problem.

One can also criticize the Dirichlet model as insufficiently flexible in parameterizing the contributions of the various ancestral populations to a typical person of the reference population. While this may be true, application of the Dirichlet model surely is better than incorrectly assuming the validity of the product rule when genetic equilibrium fails. Note that numerically bounding a match probability by adjusting ancestral gene proportions entails no distributional assumptions at all (Lange, 1993). Unfortunately, there are no obvious alternatives to the Dirichlet model, either mechanistic or phenomenological, that permit exact calculation of match probabilities in the absence of genetic equilibrium. The fact that the Dirichlet model involves substantial computation should not be a deterrent to its use. These calculations can easily be done on a personal computer. Whether a judge or jury can understand the nature of the calculations is another matter. However, this objection is irrelevant if a scientific consensus develops affirming the usefulness of the model and the feasibility of its attendant calculations. Finally, the Dirichlet model fails to be fully Bayesian. A match probability is not, after all, a posterior probability that the suspect or victim contributed the evidentiary DNA.

Despite these reservations, the Dirichlet model advances our approximate understanding of how to compute match probabilities. It is not ideal, but no scientific theory or technique ever is. Doubtless the Dirichlet model can be refined and improved. At some point, however, the legal and scientific professions need to reach a consensus about how to compute match probabilities; otherwise, DNA profiling can serve only to exonerate the innocent and never to convict the guilty. Our legal system constantly contends with approximations to the truth. This attitude, embodied in the phrase 'beyond a reasonable doubt', is as much a part of our scientific heritage as it is of our legal heritage. A rigid insistence on infallible procedures is antithetical to both professions.

Acknowledgments

I thank Ethan Lange for computing assistance and Thomas Belin and David Gjertson for reading a preliminary version of this manuscript and suggesting clarifications. This work was supported in part by the University of California, Los Angeles, and USPHS Grant CA 16042.

References

Ahrens, J.H. & U. Dieter, 1974. Computer methods for sampling from gamma, beta, Poisson and binomial distributions. Computing 12:223–246.

Bernardo, J.M., 1976. Algorithm AS 103: psi (digamma) function. Appl. Stat. 25:315–317.

Chakraborty, R. & K.K. Kidd, 1991. The utility of DNA typing in forensic work. Science 254:1735–1739.

Chakraborty, R., M.R. Srinivasan & S.P. Daiger, 1993. Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their applications in DNA forensics. Am. J. Hum. Genet. 52:60–70.

Cheng, R.C.H. & G.M. Feast, 1979. Some simple gamma variate generators. Appl. Stat. 28:290–295.

Devlin, B., N. Risch & K. Roeder, 1992. Forensic inference from DNA fingerprints. J. Am. Stat. Assoc. 87:337–350.

Edwards, A., H.A. Hammond, L. Jin, C.T. Caskey & R. Chakraborty, 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241–253.

Evett, I.W., 1992. Evaluating DNA profiles in a case where the defence is 'It was my brother'. J. Forensic Sci. Soc. 32:5–14.

Good, I.J., 1965. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press, Cambridge, MA.

Hille, E., 1959. Analytic Function Theory Vol. 1. Blaisdell Ginn, New York.

Jeffreys, A.J., V. Wilson & S.L. Thein, 1985. Individual-specific 'fingerprints' of human DNA. Nature 316:76–79.

Kingman, J.F.C., 1993. Poisson Processes. Oxford University Press, Oxford.

Lander, E., 1989. DNA fingerprinting on trial. Nature 339:501–505.

Lange, K., 1991. Comment on 'Inferences using DNA profiling in forensic identification and paternity cases' by D.A. Berry. Stat. Science 6:190–192.

Lange, K., 1993. Match probabilities in racially admixed populations. Am. J. Hum. Genet. 52:305–311.

Lee, P.M., 1989. Bayesian Statistics: An Introduction. Edward Arnold, London.

Lewin, R., 1989. DNA typing on the witness stand. Science 244:1033–1035.

Lewontin, R.C. & D.L. Hartl, 1991. Population genetics in forensic DNA typing. Science 254:1745–1750.

Mickey, M.R., J. Tiwari, J. Bond, D. Gjertson & P.I. Tersaki, 1983. Paternity probability calculations for mixed races. pp. 325–347 in Inclusion Probabilities in Parentage Testing, edited by R.H. Walker, Amer. Assoc. Blood Banks, Arlington, VA.

Miller, K.S., 1987. Some Eclectic Matrix Theory. Robert E. Krieger Publishing, Malabar, FL.

Mosimann, J.E., 1962. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 49:65–82.

Schneider, B.E., 1978. Algorithm AS 121: trigamma function. Appl. Stat. 27:97–99.

Editor's comments

The author continues the formal Bayesian analysis introduced by Gjertson & Morris in this volume. He invokes Dirichlet distributions, and so brings rigor to the discussion of the effects of population structure on match probabilities. The increased computational burden this approach entails should not be regarded as a hindrance.