Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2004/09

Inference from mitochondrial data in forensic identification

Paola Berchialla

Università degli Studi
di Firenze

Statistics

Inference from mitochondrial DNA data in forensic identification

Paola Berchialla
Department of Statistics "G. Parenti", University of Florence, Viale Morgagni 59, I–50134, Florence, Italy
berchial@ds.unifi.it

Summary

The strength of the evidence against an incriminated individual in cases of forensic identification is presented in the form of a likelihood ratio or its reciprocal (profile match probability). This paper is concerned with methods for calculating profile match probability, for personal identification, when analysis is based on mitochondrial DNA. The current method for estimating profile match probability is based on the frequency of sequences within databases. Such estimate is not statistically rigorous since complete database has not been yet compiled. The aim of this paper is to develop a method for analysing data, that allow for the effect of the mutation process which affect mitochondrial DNA molecule evolution, and computing profile match probability in the framework of a fully likelihood based approach.

1. Introduction

The interest in maternally inherited mitochondrial DNA (mtDNA) for forensic identification is partially due to the possibility of typing sequences from very small or degradated biological samples.

While most population genetic models for analysing DNA sequence polymorphisms were developed under the infinite-sites model, which assumes that every mutation occurs at a different site in the sequence, for mtDNA this assumption is violated. A more realistic framework is the finite-sites model which allows for multiple substitutions at a single locus.

A general problem in forensic identification arises when a suspect is observed to have a genetic profile also known to be possessed by the offender whose mtDNA is recovered from a biological sample left at the scene of a crime. The problem consists in quantifying the evidential strenght, for the suspect's guilt, of such observation.

Consider an individual X_1 accused to be the offender. We take a locus as a single position in the mitochondrial DNA. Given that X_1 has an allele A_i at a locus, our aim is to compute the probability that another individual X_2, who is not related to X_1, shares the X_1's allele A_i. We refer to this probability as *conditional match probability.* This is a natural measure of the weight of evidence in support of the event that the suspect is the offender since it indicates how likely it is another individual shares the suspect's genetic profile.

In section 2, two results are given for computing the conditional match probability described above. Since we operate in the framework of the finite-sites model, a mutated locus is a single position in the mtDNA that has been hit by at least one mutation. Furthermore it is assumed that no recombination occurs within the locus.

In Proposition 1 we consider a one-locus model with two alleles A_1 and A_2. We assume that A_1 mutates to A_2 and A_2 to A_1 at a same rate μ, and an allele can mutate at most once per generation.

In Proposition 2 we consider a one-locus model with four alleles that correspond to the four DNA bases $\{A, G, T, C\}$. In order to allow for multiple subsitutions, a Markov process model is used for base mutation at a single locus.

Finally, section 3 contains some examples of how to compute the conditional match probability, under different scenarios, using results of section 2.

2. Results

Proposition 1. *Consider a one locus model with two alleles A_1 and A_2, and let π_1 and $\pi_2 = 1 - \pi_1$ be the frequencies of A_1 and A_2 respectively. Moreover let θ be the limit* $\lim_{N\to\infty} 2N\mu$ *where μ is the mutation* *rate per locus per generation and N is the population size of a haploid population $\mathcal{P}$. Finally let t be the coalescence time of two individuals chosen at random in the population $\mathcal{P}$, this is, two individuals who are not related. Then given that an individual X_1 has the allele A_i, the probability that another individual X_2, chosen at random in the population $\mathcal{P}$, who has a common ancestor with X_1 tN generations ago, share the same allele is*

$$P(X_2 = A_i | X_1 = A_i, \theta, t) = \left(1 - \frac{\theta t}{2}\right)\left(\pi + (1-\pi)\frac{\theta t}{2}\right). \tag{2.1}$$

Proof. In the finite-sites model hypothesis, the same locus can be hit several times by mutations or not. In order to compute probability (2.1) we have to consider two cases: whether the common ancestor shares the same allele with X_1 and X_2 or not. If the common ancestor shares the X_1 and X_2's allele A_i, we have to compute the probability that the X_2's locus has been hit by an even number of mutations, in case none. If the common ancestor doesn't share the X_1 and X_2's allele, we have to compute the probability that X_2's locus has been hit by an odd number of mutations. Then

$$P(X_2 = A_i | X_1 = A_i, \mu, N, t) = \pi \sum_{j=0}^{\lfloor tN/2 \rfloor} (1-\mu)^{\lfloor tN \rfloor - 2j} \mu^{2j} + (1-\pi) \sum_{j=1}^{\lfloor tN/2 \rfloor} (1-\mu)^{\lfloor tN \rfloor - 2j + 1} \mu^{2j-1} \tag{2.2}$$

where $\lfloor tN \rfloor$ is the largest integer less than or equal to tN.

From the first summation of equation (2.2) we have

$$\sum_{j=0}^{\lfloor tN/2 \rfloor} (1-\mu)^{\lfloor tN \rfloor - 2j} \mu^{2j} = (1-\mu)^{\lfloor tN \rfloor} \sum_{j=0}^{\lfloor tN/2 \rfloor} \left[\left(\frac{\mu}{1-\mu}\right)^2\right]^j = \frac{1 - \left(\frac{\mu}{1-\mu}\right)^{\lfloor tN \rfloor + 2}}{1 - \left(\frac{\mu}{1-\mu}\right)^2} \cdot (1-\mu)^{\lfloor tN \rfloor} =$$

$$= \frac{(1-\mu)^{\lfloor tN \rfloor + 2} - \mu^{\lfloor tN \rfloor + 2}}{1 - 2\mu} \tag{2.3}$$

where $\sum_{j=0}^{\lfloor tN/2 \rfloor} (1-\mu)^{-2j} \mu^{2j}$ is the partial sum of a geometric series with ratio $\mu^2 \cdot (1-\mu)^{-2}$. This series converges if and only if $\mu < 1/2$. Such condition is satisfied since the mutation rate μ is tipically quite small, namely of the order of 10^{-5} or 10^{-6}. Moreover we can use the following approximation:

$$(1-\mu)^{\lfloor tN \rfloor + 2} - \mu^{\lfloor tN \rfloor + 2} \simeq 1 - (\lfloor tN \rfloor + 2)\mu. \tag{2.4}$$

In order to apply the diffusion limit for large but finite populations, we assume the limit $\lim_{N \to \infty} 2N\mu$ exists and we indicate it with θ. Then,

$$\lim_{N \to \infty} \frac{1 - (\lfloor tN \rfloor + 2)\mu}{1 - 2\mu} = \lim_{N \to \infty} 1 - \frac{\lfloor tN \rfloor \mu}{1 - 2\mu} \simeq 1 - \frac{\theta t}{2}. \tag{2.5}$$

Even the second summation of equation (2.2) is a geometric series which converges if and only if $\mu < 1/2$. Then

$$\sum_{j=1}^{\lfloor tN/2 \rfloor} (1-\mu)^{\lfloor tN \rfloor - 2j + 1} \mu^{2j-1} = (1-\mu)^{\lfloor tN \rfloor} \sum_{j=1}^{\lfloor tN/2 \rfloor} \frac{\mu^{2j-1}}{(1-\mu)^{2j-1}} = (1-\mu)^{\lfloor tN \rfloor} \left[\frac{1}{1 - \frac{\mu}{1-\mu}} - \sum_{j=0}^{\lfloor tN/2 \rfloor} \frac{\mu^{2j}}{(1-\mu)^{2j}}\right]$$

$$\simeq (1-\mu)^{\lfloor tN \rfloor} \left[1 - \sum_{j=0}^{\lfloor tN/2 \rfloor} \frac{\mu^{2j}}{(1-\mu)^{2j}}\right] \tag{2.6}$$

Hence for the same as before, we obtain the following result:

$$\lim_{N \to \infty} (1-\mu)^{\lfloor tN \rfloor} \left[1 - \sum_{j=0}^{\lfloor tN/2 \rfloor} \frac{\mu^{2j}}{(1-\mu)^{2j}}\right] = (1 - \frac{\theta t}{2}) \cdot \left[1 - \left(1 - \frac{\theta t}{2}\right)\right] = \left(1 - \frac{\theta t}{2}\right) \frac{\theta t}{2} \tag{2.7}$$

So thesis derives from (2.5) and (2.7). □

Proposition 2. *Consider a one locus model with four alleles A_i, $i = 1, 2, 3, 4$, that correspond to the four DNA-bases $\{A, G, T, C\}$ respectively, and two type of mutations. We indicate with μ and ν the transition ($T \leftrightarrow C$, $A \leftrightarrow G$) and transvertion ($T, C \leftrightarrow A, G$) mutation rate respectively. Let π_i, $i = 1, 2, 3, 4$, $\sum_{i=1}^{4} \pi_i = 1$, be the allele frequencies. Moreover let θ_μ and θ_ν be the limits $\lim_{N\to\infty} 2N\mu$ and $\lim_{N\to\infty} 2N\nu$ respectively, where N is the size of a haploid population $\mathcal{P}$. Given that an individual X_1 has an allele A_i, the probability that an individual X_2 share the same allele is:*

$$P(X_2 = A_i|X_1 = A_i, \theta, t) = \pi_i \left((1 - \frac{\theta_\mu t}{2}) - \frac{1}{2}\theta_\nu t \right) + (1 - \pi_i) \left(\frac{\theta_\mu t}{2} + \frac{\theta_\nu t}{2} \right). \tag{2.8}$$

Proof. We model the mutation process as a Markov process with transition probabilities

$$P = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 1-(\mu+\nu) & \nu/2 & \mu & \nu/2 \\ \nu/2 & 1-(\mu+\nu) & \nu/2 & \mu \\ \mu & \nu/2 & 1-(\mu+\nu) & \nu/2 \\ \nu/2 & \mu & \nu/2 & 1-(\mu+\nu) \end{pmatrix} \end{array} \tag{2.9}$$

The nth power of P is the matrix of n-step transition probabilities. The transition matrix P can be written on the form VDV^{-1} where D is a diagonal matrix with the eigenvalues of P as its entries and V is an invertible matrix consisting of the eigenvectors corresponding to the eigenvalues in D:

$$D = \mathrm{diag}\{1 - 2\mu - 2\nu, 1 - 2\mu - 2\nu, 1 - 2\nu, 1\}, \qquad V = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix} \tag{2.10}$$

Then

$$P^n = VD^nV^{-1} = \begin{pmatrix} a_n & b_n & c_n & b_n \\ b_n & a_n & b_n & c_n \\ c_n & b_n & a_n & b_n \\ b_n & c_n & b_n & a_n \end{pmatrix} \tag{2.11}$$

with

$$a_n = \frac{1}{2}(1 - 2\mu - \nu)^n + \frac{1}{4}(1 - 2\nu)^n + \frac{1}{4} \tag{2.12a}$$

$$b_n = -\frac{1}{4}(1 - 2\nu)^n + \frac{1}{4} \tag{2.12b}$$

$$c_n = -\frac{1}{2}(1 - 2\mu - \nu)^n + \frac{1}{4}(1 - 2\nu)^n + \frac{1}{4} \tag{2.12c}$$

The probability that X_2 shares the X_1's allele is the probability to observe no differerences if X_1 and X_2's common ancestor $\lfloor tN \rfloor$ generations has the same allele too; or to observe one difference if their common ancestor doesn't share the same allele. Hence

$$P(A_i|A_i, \lfloor tN \rfloor \text{ steps}) = \pi_i \cdot a_{\lfloor tN \rfloor} + (1 - \pi_i) \cdot (1 - a_{\lfloor tN \rfloor}) \tag*{2.13)}$$

Applying the diffusion limit for large but finite populations, we obtain

$$a_{\lfloor tN \rfloor} = \frac{1}{2}(1 - 2\mu - \nu)^{\lfloor tN \rfloor} + \frac{1}{4}(1 - 2\nu)^{\lfloor tN \rfloor} + \frac{1}{4} =$$

$$\frac{1}{2}(1 - (2\mu + \nu)\lfloor tN \rfloor) + \frac{1}{4}(1 - 2\nu\lfloor tN \rfloor) + \frac{1}{4} \tag{2.14}$$

and when $N \to \infty$

$$a_{\lfloor tN \rfloor} \to (1 - \frac{\theta_\mu t}{2}) - \frac{1}{2}\theta_\nu t \tag{2.15}$$

from which the thesis derives. □

3. Examples.

Suppose that an mtDNA sequence recovered from a biological sample left at a crime scene is found to match the mtDNA sequence obtained from an individual X_1. This observation supports the hypothesis that X_1 is the source of the recovered biological sample. In order to assess the evidential strenght of this support, we need to evaluate the probability that another individual X_2 would also match the recovered biological sample left at crime scene.

Suppose a reference sample $\mathcal{D}$ is available consisting of the mtDNA sequences of n unrelated individuals drawn from the same racial group as X_1. Then the conditional match probability another individual X_2, who is not maternally-related to X_1, shares X_1's mitochondrial haplotype is

$$P(X_1 = A_i | X_2 = A_i, \bar{H}, \mathcal{D}) = \int_{\theta,\pi,t} P(X_1 = A_i | X_2 = A_i, \bar{H}, \theta, \pi, t, \mathcal{D}) P(\theta, \pi, t | \bar{H}, \mathcal{D}) d\theta d\pi dt =$$

$$\int_{\theta,\pi,t} P(X_1 = A_i | X_2 = A_i, \bar{H}, \theta, \pi, t) P(\theta, \pi | \mathcal{D}) P(t | \bar{H}, \mathcal{D}) d\theta d\pi dt. \tag{3.1}$$

In this equation $\bar{H}$ indicates the hypothesis X_1 and X_2 doesn't pertain to the same maternal lineage and so, for instance, they aren't brothers—this hypothesis ,which is weaker than supposing X_1 and X_2 are the same person, is due to the fact we analyse maternally inherited DNA.

The probability $P(X_1 = A_i | X_2 = A_i, \bar{H}, \theta, \pi, t)$ is given by result 1 or result 2 in preceding section. Instead $P(\theta, \pi | \mathcal{D})$ denotes the posterior distribution of parameters θ and π which can be obtained by a mutational model. Finally, $P(t | \bar{H}, \mathcal{D})$ is the posterior coalescence times distribution of two people who are not related but present the same mitochondrial haplotype.

Often in databases between different individuals within a population, a single position exibits only two variants—this is, for instance, the case of genetic data namely single-nucleotide polymorphisms (SNPs) for which the more polymorphic such a locus is, the larger the relative frequency of the less common variant. When analysing data of that kind, in defining a mutational model we have information about only two alleles, hence we can consider result 1 in calculating the match conditional probability. We refer to this case to present some examples.

Let us consider a databse $\mathcal{D}$ consisting of $n = 49$ individuals who have 28 polymorphic loci. Each position is labeled with 0 or 1 whether the allele presents the common variant or the rare one respectively. Since the polymorphic loci are distanced each other, we assume they evolve independently, so that

$$\int_{\theta,\pi,t} P(X_1 = A_1 \dots A_{28} | X_2 = A_1 \dots A_{28}, \bar{H}, \theta, \pi, t) P(\theta, \pi | \mathcal{D}) P(t | \bar{H}, \mathcal{D}) d\theta d\pi dt =$$

$$\prod_{i=1}^{28} \int_{\theta_i,\pi_i,t} P(X_1 = A_i | X_2 = A_i, \bar{H}, \theta_i, \pi_i, t, \mathcal{D}) P(\theta_i, \pi_i, t | \bar{H}, \mathcal{D}) d\theta_i d\pi_i dt \tag{3.2}$$

Example 1. In this first example we consider a plug-in estimate approach in calculating probability (3.2). Under this way of dealing with estimates, π is fixed to be its empirical estimate $\hat{\pi}$, so that the population frequency of each allele is simply estimated by its corresponding frequency as observed in the data set $\mathcal{D}$. In addition let us assume no randomness about mutation and demographic parameters, for which $\mu = 10^{-9}$ and $N = 5000$ (). For the posterior of t coalescence time between the two individuals X_1 and X_2, we observe that the probability of not observing differences between two sequences given separation time t

is $P(\text{no differences}|t) = (1 - \theta t/2) \simeq \exp(-\theta t/2)$ while the natural prior of coalescence time between two individual who are not related is an exponential with mean equals to 2. So,

$$P(t|\text{no differences}) = P(\text{no differences}|t) \cdot P(t) = \frac{1}{2}\exp(-(\theta+1)t/2) \tag{3.3}$$

which is approximately an exponential with mean 2 since θ is very small. With this assumption, for two individuals who has the same sequence with the rare variant on positions 11 and 25, then $P(X_1 = A_1 \ldots A_{28}|X_2 = A_1 \ldots A_{28}, \bar{H}, \mathcal{D}) = 0.000713$ We observe that if two individual share the same sequence with just the rare allele on position 28, which is the most polimorphic one, we obtain $P(X_1 = A_1 \ldots A_{28}|X_2 = A_1 \ldots A_{28}, \bar{H}, \mathcal{D}) = 0.0326$.

Example 2. In this example we introduce randomness for mutation process. We modelize data as a Binomial(n, p) where $p = \exp(-\theta_j t)(1/2 - \pi_j) + 1/2$ is the rare variant proportion in the population. Than we assume $N = 5000$ constant and used a gamma distribution with shape equal to 2 and scale parameter equal to $10^{(-9)}$. There are no compelling reasons for the particular choices of a gamma distribution. This distribution has desiderable properties of being smooth, unimodal and excluding negative values. Having adopted this functional form, the parameter values were chosen to give desired mean and variance. (Tavaré et al. 1997). Finally we complete the specification of the model by taking a symmetric beta*(a,a)* distribution on π with $a = 0.1$ (Nicholson et al. 2002) As in example 1 the coalescence time between two individuals who are not related has an exponential distribution with mean 2. With this assumption, if two individuals have the same sequence with the rare variant on positions 11 and 25, the conditional match probability is $P(X_1 = A_1 \ldots A_{28}|X_2 = A_1 \ldots A_{28}, \bar{H}, \mathcal{D}) = 0.00074$. We observe that, for instance, if two individuals share the same sequence with just the rare allele on position 28, which is the most polimorphic one, we obtain $P(X_1 = A_1 \ldots A_{28}|X_2 = A_1 \ldots A_{28}, \bar{H}, \mathcal{D}) = 0.031$.

Results show that our method is robust under the different scenarios assumed in the examples. Furthermore it is interesting to compare our results whith those based on mtDNA sequences, that is analyses based on the frequency of a sequence within a database. For instance in the FBI database, which contains 2087 sequences, there is not the mithocondrial aplotype analysed in previous examples with rare variant on positions 11 and 25. So in the augmented database we have a probability of order 10^{-3} which is more conservative than ours. This finding would suggest that our method captures some information that, although contained in the considered data, is not recognized by a pure descriptive method.

References

Balding, D. J. & Donnelly, P. (1995) Inference in forensic identification. *J. R. Statist. Soc. A,* **158**, 21–53.

Foreman, L. A., Smith, A. F. M. & Evett, I. W. (1997) Bayesian analysis of DNA profiling data in forensic identification. *J. R. Statistic. Soc. A*, **160**, 429–469.

Neuhauser, C. (2001). Mathematical models in population genetics, pp. 153–177 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Nicholson G., Smith, A. V., Jonssón, F. Gústafsson, O., Stefánsson, K. & Donnelly, P. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Statist. Soc. A,* **64**, 695–715.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequences data. *Genetics*, **145,** 505–518.

Weir, B. S. (2001) Forensics, pp. 721–739 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Wilson, I. J., Weale, M. E & Balding, D. J. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Statist. Soc. A*, **166**, 1–33.