Statistical Genetics and Genetical Statistics: a Forensic Perspective*

THORE EGELAND

National University Hospital, Oslo

PETTER F. MOSTAD

Chalmers Technical University, Gothenburg

ABSTRACT. This review paper focuses on forensic aspects of the relation between statistics and genetics. Some of the scientific achievements of the Swedish psychiatrist and geneticist Erik Essen-Möller may be viewed in the above interdisciplinary context. In a number of situations the correct familial relation between a group of individuals is required. We discuss recent work done to relax some assumptions involved in the classical calculations in Essen-Möller (1938). Moreover, we extend the discussion to identification problems. In a given case there may be a large number of possible family constellations or pedigrees. A prior probability distribution is established. The posterior model accounts for the combinatorial complexities of the pedigrees, mutations, kinship, and uncertainty in allele frequencies. Examples are based on the shareware program FAMILIAS, see http://www.nr.no/familias. A main message of the present paper is that the Bayesian approach is a convenient framework to down-weight unreasonable (e.g. incestuous) pedigrees that may always appear likely if only DNA-measurements are used.

Key words: Bayes, DNA, forensic statistics, genetics, identification, kinship, likelihood ratio, mutation

1. Introduction

The remarkable development of genetics from Mendel's discoveries in the middle of 19th century to the micro-array techniques introduced in the last decade of the 20th century (Dudoit *et al.*, 2000), has obviously influenced everyday life and science in many ways. Modern DNA technology was applied for identification purposes the first time in 1985, 85 years after a previous breakthrough in identification: the Galton–Henry system for fingerprinting, see Evett & Williams (1996). The paper by Jeffreys *et al.* (1985) describes an immigration case. A boy from Ghana was initially not granted residence in the United Kingdom, as the authorities did not believe the declared family relations to be true. Conventional genetic markers at that time, e.g. AB0 and RH, could not determine whether a certain woman was the mother or aunt of the boy. The case was resolved by means of what was then called DNA fingerprints and the boy was granted residence. The impact of modern genetics and molecular biology on forensics, i.e. science relevant to the court, has been particularly obvious to the public. In fact, the O. J. Simpson case (*People vs Simpson*, see Weir, 1995), might well be the single issue that has done most to advance the knowledge of DNA among ordinary people. Major newspapers all over the world carried material explaining the structure of DNA in surprising detail. The same story repeats over and over again when major criminal cases hit the headlines. These criminal cases involving biological evidence communicate statistical issues to a wide audience. Scientific discoveries, like the discovery of genes relating to major diseases, tend not to get publicity of the same order.

*This paper was presented by Thore Egeland as an Invited Lecture at the 18th Nordic Conference on Mathematical Statistics, Grimstad, Norway, June 2000.

There is an intrinsic link between DNA evidence and statistical questions like: What is the probability that the defendant's DNA-profile match the evidence by chance? What is the relevant reference sample upon which statistical estimates should be based? This paper focuses on the interplay of genetics and statistics in forensic genetics. In particular, we are concerned with how DNA can serve to identify people or families. To indicate the wide range of applications, it suffices to mention some diverse papers. The identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster was described in Olaisen *et al.* (1997). Bowers *et al.* (1999) deals with the parentage of Chardonnay, Gamay, and other wine grapes of Northeastern France, while the title of Foster *et al.* (1998) is "Jefferson fathered slave's last child".

Section 2 reviews paternity cases from a historical perspective. It is followed by a discussion of how the classical assumptions may be relaxed. Moreover, additional challenging identification problems requiring a computer for their solution, are addressed. Various aspects are discussed in the closing section including alternative approaches and problems for future research. Small bits and pieces of genetics and molecular biology required to understand the statistics are presented when needed.

2. Essen-Möller's index

The discussion below is brief, a number of papers and books including Evett & Weir (1998) provide more details. Resolving disputed paternity remains a cornerstone of forensic genetics. In Norway (pop. 4.5 million) the number of paternity cases exceeds 1000 annually. A typical case involves a child and a mother (not disputed) and a putative father. The case could be raised because the man does not accept paternity or because the mother claims he is not the father. Every year some less standard cases occur. Matters are complicated if incest is suspected (in which case it is a criminal case) or if DNA profiles are available only for relatives of the persons involved. This section will review some of the statistical aspects of standard cases, delaying the discussion of more complicated cases. The Swede Erik Essen-Möller (1901–92) made lasting contributions to various scientific fields including forensic statistics and genetics (Hummel & Gerschow, 1981). Essen-Möller (1938) continues to be cited.

Figure 1 shows a simple pedigree. There are parents, a female F_1 and a male M_1, and a son M_2. Females are conventionally depicted by circles, men by squares. Consider first data available for one locus (also called system or genetic marker), i.e. one specific location of the human genome. As shown in the figure, F_1 is homozygous $\{A, A\}$ (i.e. she has the allele or genetic variant A in both her chromosomes with this locus), M_1 homozygous $\{B, B\}$ while M_2 is heterozygous $\{A, B\}$. The precise definition of an allele will depend on the type of DNA involved. We choose the vWA locus to exemplify. Table 1 displays a Norwegian database of 300 persons (600 alleles).
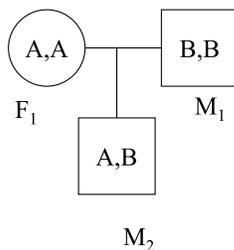

Fig. 1. A pedigree showing a mother F_1 (alleles A and A), father M_1 and son M_2.

For instance, A is called '14' implying that a specific sequence of bases, 'TCTA' for this locus, is repeated 14 times. The length of the allele is said to be 14. More extensive information on forensic loci are provided on the net, e.g. http://www.mclink.it/personal/MD1696/data/freqask.htm.

M_2 has inherited one allele, A, from his mother and the other, B, from his father. It is not always possible to distinguish the maternal and the paternal allele. Assume now that the paternity of M_2 is disputed. Certainly, the data presents evidence in favour of the hypothesis of Fig. 1, i.e. H_1: M_1 *is the father of* M_2. The strength of this evidence depends on the probabilities (or population frequencies) of the alleles, denoted p_A and p_B. To assess the DNA evidence, we calculate the likelihood of the data as

$$\Pr(F_1 = \{A, A\},\ M_1 = \{B, B\},\ M_2 = \{A, B\}|H_1) = p_A^2 p_B^2. \tag{1}$$

The assumptions underlying these types of calculations are: (i) F_1 and M_1 are unrelated; (ii) Hardy–Weinberg equilibrium; (iii) mutations do not occur and (iv) Mendelian inheritance. Considering next the hypothesis H_2: *An unknown man unrelated to both* F_1 *and* M_1 *is the father of* M_2, we find a likelihood of $p_B p_A^2 p_B^2$. The paternity index is the ratio of these likelihoods. In modern parlance the term likelihood ratio (LR) is preferred and so for this locus

$$LR = p_A^2 p_B^2 / (p_B p_A^2 p_B^2) = 1/p_B. \tag{2}$$

To obtain stronger evidence, the analysis is based on several loci or genes and combined statistically using Bayes theorem on odds form:

$$\frac{\Pr(H_1|E)}{\Pr(H_2|E)} = \frac{\Pr(E|H_1)}{\Pr(E|H_2)} \times \frac{\Pr(H_1)}{\Pr(H_2)}. \tag{3}$$

E typically denotes genetical evidence. The conventional usage of the term odds implies that $\Pr(H_1|\cdot) = 1 - \Pr(H_2|\cdot)$, where the dot indicates some general conditioning. However, the theorem is true also if this last relation does not apply. In such cases we will use the term posterior probability ratio (PPR) rather than posterior odds. The analogue of the usual informal version of the theorem then reads

Posterior probability ratio = Likelihood ratio × prior probability ratio

and so the PPR coincides with the LR whenever the prior probabilities are equal. The LR's of the various loci may be multiplied to obtain an overall likelihood ratio provided the genes involved are independent. The independence is usually genetically reasonable provided the loci are from different chromosomes or at least sufficiently far apart on the same chromosome. The question of independence between loci and also independence within loci, i.e. Hardy–Weinberg equilibrium, has been extensively debated, see Weir (2000) for a review of the exchange.

Essen-Möller (1938) did not refer specifically to Bayes, but he wrote (p. 12) "Wir wollen nun die weitere Annahme machen, daß wahre und falsche Väter gleich häufig Begutachtung kommen". In other words, he *a priori* assumed $\Pr(H_1) = 1 - \Pr(H_2) = 0.5$. The posterior

Table 1. *Data for the locus vWA*

Allele label	A	B	C	D	E	F	G	H
Repeat number	14	15	16	17	18	19	20	21
Count	44	49	127	175	133	58	12	2
Proportion	0.073	0.082	0.212	0.292	0.222	0.097	0.02	0.003

probability of paternity $\Pr(H_1|E)$, denoted W by Essen-Möller to allude to Wahr-scheinlichkeit, is then $W = LR/(LR + 1)$ as a direct consequence of (3). Evett & Weir write (1988 p. 164) "We do not advocate the use of this probability of paternity because of the implicit assumption of a prior probability of 0.5, irrespective of the non genetic evidence". Hummel (1994) discusses and compares paternity indices and Essen-Möller's W.

The practical treatment of paternity cases varies from country to country or even to some extent from laboratory to laboratory. Generally, a new sample should be obtained from a man who is excluded based on the first analysis. This is a useful precaution to avoid the sample coming from the wrong person: the man could have sent someone else. Furthermore, an exclusion could be the result of switching of samples whereas a match is less likely to be the result of such an error.

There appears to be two principally different traditions when it comes to reporting the results. One is to compute the LR or W based on a set of loci. The authorities making the decision will then base their conclusion on numbers. The other, is to conclude according to a validated decision rule, like for instance the following: five loci are investigated. If a man has DNA profiles consistent with all five systems, this will be considered very strong evidence in favour of paternity, particularly if close relatives of the putative father are excluded *a priori*. On the other hand, if none or one loci fit, he is excluded initially. Numbers like LRs are not routinely reported, but provided if requested. The laboratory would typically provide general statements like: "If a man fits for all loci, the LR exceeds 10 000". In the remaining cases, i.e. two, three or four loci match, further investigations are undertaken. In some of these cases a relative of the putative father may be the father, see section 3.3.

3. Beyond standard paternity cases

The objective of this section is to extend the previous in several directions. In particular, we discuss how the assumptions following (1) may be relaxed. Moreover, whereas the classical approach normally restricts attention to two hypotheses or pedigrees, we would like to consider a large number simultaneously in a Bayesian framework. Finally, confidence statements regarding LRs, reflecting various types of uncertainty, will often be requested.

Population stratification and relatedness may invalidate calculations like the one leading to (1). For instance, Hardy–Weinberg may not apply in paternity cases. The man and woman involved may be remotely related, although not knowing so, simply because they belong to the same subpopulation. This kinship has nothing to do with incest which is dealt with later in section 3.3. Balding & Nichols (1995) propose a practical way of handling these problems that has been endorsed in the recommendations of the National Research Council (1996). The approach may be given a genetic or evolutionary argument as well as a more statistical one. For practical calculational purposes it suffices to consider the Dirichlet distribution (Lange, 1995) presented next.

3.1. Kinship and the Dirichlet distribution

Assume the frequencies of a set of n alleles in a population is $p = (p_1, \ldots, p_n)$, with $\sum p_i = 1$. If $Y = (y_1, \ldots, y_n)$ are the number of observations of each allele in a database, with $\sum y_j = m$,

$$Y|p \sim \text{multinomial}(m, p_1, \ldots, p_n),$$

assuming Hardy–Weinberg equilibrium and a random sample of persons in the database. If we want to estimate p from Y, we may represent p with a random variable $\tilde{p}$, put a prior on it, and find its posterior given Y. We will use a Dirichlet prior. This is practical, as the Dirichlet distribution is a conjugate to the multinomial. With the prior

$$\tilde{p} \sim \text{Dir}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right),$$

we get

$$\tilde{p}|Y \sim \text{Dir}\left(y_1 + \frac{1}{n}, \ldots, y_n + \frac{1}{n}\right).$$

If we are using a database with zero observations of some allele that may exist, the above prior is also convenient since there is a positive probability of observing the allele in later data.

If we write $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_n) \sim \text{Dir}(kp_1, \ldots, kp_n)$, where $\sum p_i = 1$ as before, then

$$E(\tilde{p}_i) = p_i,$$
$$\text{Var}(\tilde{p}_i) = F_{ST} p_i (1 - p_i),$$
$$\text{cov}(\tilde{p}_i, \tilde{p}_j) = -F_{ST} p_i p_j,$$

where $F_{ST} = 1/(k + 1)$ (Roeder, 1994). The following relation (see Balding & Nichols, 1995) is particularly useful

$$E(\tilde{p}_A^{r+1} \tilde{p}_B^s \tilde{p}_C^t \tilde{p}_D^u) = E(\tilde{p}_A^r \tilde{p}_B^s \tilde{p}_C^t \tilde{p}_D^u) \frac{rF_{ST} + p_A(1 - F_{ST})}{1 + (r + s + t + u - 1)F_{ST}}. \tag{4}$$

Here A, B, C and D are arbitrary indices and r, s, t and u non-negative integers.

Example 1. In presence of kinship, the likelihood ratio in (2) becomes, using (4),

$$LR = \frac{1 + 3F_{ST}}{2F_{ST} + (1 - F_{ST})p_B}.$$

Thus, the likelihood ratio decreases as the kinship increases (as long as $p_B < 0.5$), which is reasonable. In other words, the evidence in favour of paternity is in this case weakened in the presence of kinship.

3.2. Mutation models

We next discuss the impact of various mutation models on *LR*-calculations. Typically, the numerical calculations require a computer program even in quite simple cases. One such shareware program, FAMILIAS, is described in Egeland *et al.* (2000) and is obtainable from http://www.nr.no/familias. However, some analytical calculations are possible in simple cases as discussed below. The mutation rates for forensic loci are relatively high, otherwise these genes might not have been so polymorphic, i.e. displayed some many alleles. Assume a laboratory handles 1000 paternity cases every year. For each case, something like five loci are considered. If the mutation rate is set to 0.005, which is not totally unrealistic, we would then expect $1000 \cdot 5 \cdot 2 \cdot 0.005 = 50$ mutations observed annually.

It may be prudent to compute with several mutation models testing the robustness of the conclusions, and even be able to operate with different mutation rates for paternal and maternal alleles. Three different models have been implemented. The first is called the uniform: We assume N is the number of alleles observed in the underlying database, and M ($M \geq N$) is the number of "possible" alleles. We then assume that if a mutation occurs, the mutated allele is one of the $M - 1$ other possible alleles, with equal probability.

The above model has not reached stationarity if the allele probabilities differ. Stationarity is not a natural biological condition, as allele frequencies do change over time. However,

non-stationarity has the somewhat unpleasant consequence that the exact *LR* will be changed by including extra irrelevant persons in the calculations, see Egeland *et al.* In fact, a person's allele frequencies will be different if they are derived directly from the database compared to if, with mutations, they are derived, from parents having the database allele frequencies.

A simple stationary model, referred to as the proportional, is one where the probability of a mutation ending up at an allele is proportional to the frequency of that allele, and independent of the allele from which the mutation occurred. If we describe the mutation model in terms of a transition matrix $[m_{ij}]$, where m_{ij} denotes the probability that allele i is inherited as allele j $(i, j = 1, \ldots, n)$, the model is described by $m_{ij} = kp_j$ for $i \neq j$, and $m_{ii} = 1 - k + kp_i$, where k is a constant. This model clearly satisfies the stationarity condition $\sum_{i=1}^{n} p_i m_{ij} = p_j$. The overall mutation rate becomes $R = k \sum_{i=1}^{n} p_i(1 - p_i)$, so with R fixed, we must set $k = R / \sum_{i=1}^{n} p_i(1 - p_i)$.

In general, many stationary models are possible. In particular, for some loci a model where the probability of mutating from allele a to allele b decreases as a function of the difference in length between the alleles seems reasonable, and has support in experimental data; a model is presented in Dawid *et al.* (2000). However, a database may not contain all possible lengths, and so we have considered a non-stationary model in this case which we denote the decreasing model for simplicity. Assume the list of alleles is expanded to include all "possible" alleles, and that they are listed by increasing lengths. Then we simply set $m_{ii} = 1 - R$, where R is the overall mutation rate, and for $i \neq j$, $m_{ij} = k_i r^{|i-j|}$, where r, $0 < r < 1$, is a constant, and k_i is chosen so that $\sum_{j=1}^{n} m_{ij} = 1$. A simple calculation gives $k_i = R(1 - r)/(r(2 - r^{i-1} - r^{n-i}))$.

Example 2. Assume the alleged father (AF) has alleles A and B and the child (CH) C and D for the locus shown in Table 1. (This case is discussed by Stockmarr (2000), but our models and objectives differ.) The likelihood of the data assuming "H_2: The two are unrelated", equals $4p_A p_B p_C p_D$. Denote by m_{AC} the probability that an allele mutates from A to C, and similarly for other transitions. The likelihood assuming AF is the father, H_1, is $p_A p_B (p_C(m_{AD} + m_{BD}) + p_D(m_{AC} + m_{BC}))$ leading to

$$LR = \frac{\Pr(E|H_1)}{\Pr(E|H_2)} = \frac{p_C(m_{AD} + m_{BD}) + p_D(m_{AC} + m_{BC})}{4p_C p_D}. \tag{5}$$

The remaining part of this example compares the mutation models and we specify a mutation rate of 0.005 for all three models. The first mutation model, the "uniform", with $M = 10$ leads to $m_{AC} = m_{AD} = m_{BC} = m_{BD} = 0.005/9$, and $LR = 0.0023$. In the second model, the "proportional", $k = 0.0063$ and also $LR = 0.0063$. For the third model, the "decreasing", observe that A, B, C, and D appear as the four shortest alleles in the database, in that order, and we then get $LR = 0.0047$ using $r = 0.5$. As a preliminary conclusion, we observe that all models lead to very small *LR*s as expected. However, the relative differences are considerable and the choice of model might well influence the overall *LR* considerably. A similar observation is made by Dawid *et al.* (2000). The statistical properties of *LR*s for the various mutation models may be studied most simply by simulation.

We may of course combine mutation models with kinship. Applying (4) we arrive at

$$LR_K = \frac{\Pr(E|H_1)}{\Pr(E|H_2)} = \frac{(1 + 2F_{ST})}{(1 - F_{ST})} LR \tag{6}$$

where the *LR* is given in (5). When computing the expectation under the model for $\tilde{p}$, we have here ignored the dependence the different *m*-values might have on $\tilde{p}$.

3.3. Identification cases

In traditional paternity cases, only two alternative hypotheses are considered. However, a more careful consideration may in some cases reveal that one should rather divide into a larger set of hypotheses. For example, there might in a case be evidence indicating that the brother of a putative father might be the real father. (It may not always be appropriate for the laboratory to direct the authorities to alternative hypotheses. This problem we disregard throughout.) Traditional computations may give a high *LR* for the putative father, just because he shares some rare alleles with his brother and the brother's offspring. We recommend analysing such cases in a Bayesian framework: The relevant mutually exclusive hypotheses are assembled, and a prior distribution assigned. Then, the posterior probabilities are computed by introducing the evidence from the DNA measurements. As the specification of a prior is often difficult, it is advantageous to separate it from calculations based on DNA measurements. In the case of two hypotheses, this is easy: because of (3), one may simply report the *LR*. With more than two hypotheses, one might report all the likelihoods of the hypotheses, but this information is difficult to understand and use in a sensible way without a prior.

In other uses of DNA fingerprinting, the need for multiple hypotheses is even more apparent. When identifying the victims of a mass disaster, there is often a large number of possible pedigrees from which one must find the most probable. This is also the case when determining the identities of bodies in, for example, a family grave. Clearly, the number of possible pedigrees linking a set of persons grows extremely fast with the number of persons. In some cases, the prior set of possible pedigrees may be so large that only simulation approaches, for example MCMC, may be able to span the pedigree set. We have considered the more modest cases, where one may generate a limited set (say, a few thousands) of pedigrees, which is then believed to contain the true pedigree. A prior is assigned to this set, using some simple algorithms, and a posterior is computed using the DNA measurements. The idea is that the data should be able to pick out the true pedigree fairly clearly, thus making the details of how the prior is constructed less important.

In our method, one starts with generating the set of "possible" pedigrees involving the relevant persons. Family relations are described through parent–child relations. To include other family relations, it is necessary to introduce extra persons, i.e. persons for which no data is available. An algorithm then generates all possible pedigrees linking all these persons. The algorithm may be constrained by information like known age differences between persons, that some are children, and so on. A prior probability $\Pr(g)$ is assigned to each pedigree g, and the following choice has proven useful:

$$\Pr(g) = c M_I^{b_I(g)} M_P^{b_P(g)} M_G^{b_G(g)}, \tag{7}$$

where c is the normalization constant, and M_I, M_P, and M_G are non-negative parameters provided by the user of the program. The corresponding integer exponents $b_I(g)$, $b_P(g)$, and $b_G(g)$ explained next are included to accommodate the possibility of weighing pedigrees according to certain specific characteristics. $b_I(g)$ is the number of children where both parents are present in the pedigree and where the parents have a common ancestor in the pedigree and so accounts for inbreeding. The number of pairs of children having precisely one parent in common is calculated and denoted $b_P(g)$. Finally, the longest chain of generations starting with a named person and ending in an adult, is calculated and assigned the value $b_G(g)$. The user may assign 0 prior probability to all incestuous pedigrees by letting $M_I = 0$, and similarly for the other parameters. A flat prior corresponds to setting the M-parameters to unity.

3.4. Posterior model

The DNA data available for each locus as well as a mutation rate for each system is used to compute the likelihood of all pedigrees considered in the prior model. These likelihoods are multiplied with the priors to obtain the posterior probability. All details related to the computation of the likelihoods are provided in Egeland *et al.* (1996).

Example 3. This example discusses the Romanov case documented in Gill *et al.* (1994) and subsequent papers including Ivanov *et al.* (1996). The Romanovs, i.e. Tsar Nicholai II, the Tsarina, three of their five children as well as three servants and a doctor were executed in 1918. A grave was found in Ekaterinburg in 1991 and various DNA analyses were performed to verify that the Royal Family had been found. There is a total of 4536 possible family relations between the one male and the four females of Table 2 according to the algorithm implemented in FAMILIAS (recall that only parent–child relations are considered). The parenthesized information of the first column of Table 2, reflects the accepted identification.

The allele frequencies correspond to the ones used in Egeland *et al.* (2000). Equal prior probabilities were assigned to all these alternatives and it was shown that the previously published and accepted solution emerged as the most probable. The strength of the conclusion in terms of likelihood ratios or posterior probabilities depend on the prior assumptions; the accepted solution always came out as the one with highest posterior probability. Table 3 summarizes the findings in the case where mutations and kinship are disregarded.

Alternative I in Table 3 corresponds to the hypotheses H_1: "4 and 7 are the parents of 3, 5, and 6 in Table 2" and H_2: "An unknown man (unrelated to both 4 and 7) and 7 are the parents of 3, 5, and 6 in Table 2". Turning to alternative II, we consider all families where 3, 5 and 6 have no children of their own. Consequently, the children may have none, one, or both parents among 4 and 7 resulting in four possibilities. Furthermore, there are three possible relations among the two adults and so there are $3 * 4 * 4 * 4 = 192$ alternatives to consider. The PPR in Table 3 is the ratio $Pr(H_1|data)$ to $Pr(H|data)$ where H is the next to most likely alternative or one of them if there are several equiprobable. All incestuous pedigrees have been

Table 2. *STR genotypes for the nine skeletons Gill* et al. (1994). *The numbers in parentheses are allele frequencies*

Skeleton	HUMVWA/31	HUMTH01	HUMF13A1	HUMFES
3 (Child1)	15(0.087),16	8(0.088),10	5(0.184),7	12(0.197),13
4 (Tsar)	15,16(0.22)	7,10(0.332)	7(0.354),7	12,13(0.0228)
5 (Child2)	15,16	7(0.228),8	5,7	12,13
6 (Child3)	15,16	8,10	3(0.077),7	12,13
7 (Tsarina)	15,16	8,8	3,5	12,13

Table 3. *Five alternative models for the Romanov case are shown*

Alternative models	Number of pedigrees	Posterior probability ratio
I	2	2692.4
II	192	63.1
III	2020	63.1
IV ($M_I = 0.1$)	4536	20
V	4536	1.2

removed for alternative III. For the next alternative, the incestuous families are not removed but rather down-weighted by entering a parameter value $M_I = 0.1$ in (7). This implies that *a priori* a non-incestuous pedigree is ten times likelier than an incestuous one. The last alternative corresponds to a flat prior on all generated pedigrees. We have incorporated various models for mutations and as expected these have very little impact on the resulting figures. Accounting for kinship by varying F_{ST} between 0.01 and 0.15 tend to give somewhat larger *LR*s.

The analyses is next supplemented by a simulation exercise to get some idea of the performance of the approach. Alleles were sampled with replacement from loci with ten alleles of uniform frequency for persons 4 and 7 while the children were assigned alleles according to Mendel's laws from the parents. The data is shown in Table 4.

To mimic the real case, only the first four loci were considered initially. Assigning equal prior probabilities to all 4536 possible pedigrees, the correct pedigree emerged as the one with highest posterior probability, namely 0.3363. This figure was four times larger than the second most likely alternative, i.e. the PPR was 4. Increasing the number of loci to 8, the Posterior Probability and PPR became 0.8100 and 32.0. Stronger results were obtained when all incestuous pedigrees were removed *a priori* and the remainder were assigned equal prior probabilities. The Posterior Probability and PPR was 0.9561 (0.9999) and 312.5 (97656.3) where the parenthesized numbers correspond to 8 loci.

4. Discussion

There are a number of programs and algorithms dedicated to paternity and identification cases. For instance, Brenner (1997) presents a symbolic kinship program. Dawid *et al.* (2000) demonstrate that many problems may be approached using probabilistic network systems (Cowell *et al.*, 1999). Once the problem is suitably reformulated, general software such as HUGIN (http://www.hugin.dk) performs the numerical calculations.

In many criminal cases (e.g. the O. J. Simpson case, Weir, 1995) the biological sample reveals that several people have contributed. The DNA profile of each contributor may not be extracted with certainty. Cases of this sort are referred to as mixture problems. Some of the challenges are sorted out, see e.g. Evett & Weir (1998), but there are remaining research problems.

The methods of the present paper require extensions in some cases to solve practical problems. The magnitude of the problem, i.e. the number of different *a priori* pedigrees, needs to be limited to, say, a few thousand as mentioned previously. Thus an identification problem involving hundreds of persons would require other methods to establish the correct pedigree with some certainty. In addition, further studies are required to give advice on the amount of

Table 4. *Simulated data inspired by the Romanovs*

Locus	(Tsar)	(Tsarina)	(Child1)	(Child2)	(Child3)
s1	4 10	8 10	8 10	4 10	8 10
s2	4 9	6 8	4 6	4 8	4 8
s3	6 10	2 10	2 6	2 10	6 10
s4	7 8	2 6	6 8	6 7	2 7
s5	1 9	2 6	2 9	1 2	1 6
s6	2 4	3 7	4 7	3 4	2 3
s7	2 9	2 5	2 9	2 2	2 9
s8	3 9	1 2	1 9	2 9	1 9

information required to resolve various identification problems. The latter problem could be approached initially by simulation, i.e. extending the approach of example 3. In this context it would be useful to be able to supplement the posterior probabilities with credibility intervals.

We have restricted attention to Mendelian inheritance. However, mitochondrial DNA, abbreviated mtDNA, (inherited essentially maternally, i.e. the mtDNA of a woman is transferred to all her children) has become important in forensic case work, see Bar *et al.* (2000). One reason is that mtDNA may be obtained in many cases where a biological stain is too degraded to obtain ordinary (nuclear) DNA. Recently, also *Y*-chromosome genes (transferred from a father to his sons) have been used in identification cases (Foster *et al.*, 1998). Observe that (3) may be used to combine various types of DNA or diverse evidence for that matter.

It seems fair but perhaps trivial to conclude that DNA-measurements may be used to identify family relations. The Bayesian approach is a convenient framework to down-weight unreasonable (e.g. incestuous) pedigrees that may always appear likely if only DNA measurements are used.

References

Balding, D. J. & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Bar, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, J., Mayr, W., Morling, N., Olaisen, B., Schneider, P. M., Tully, G. & Wilson, M. (2000). Guidelines for mitochondrial DNA typing. *Vox Sang* **79**, 121–125.

Bowers, J., Boursiquot, J., This, P., Chu, K., Johansson, H. & Meredith, C. (2000). Historical genetics: the parentage of Chardonnay, Gamay, and other wine grapes of North-Eastern France. *Science* **285**, 162–164.

Brenner, C. H. (1997). Symbolic kinship program. *Genetics* **145**, 535–542.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer-Verlag, New York.

Dawid, A. P., Mortera, J., Pascali, V. & van Boxel, D. (2000). Probabilistic expert systems from forensic inference from genetic markers. Submitted to *Scand. J. Statist*.

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report, **578**. http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html.

Egeland, T., Mostad, P. F. & Olaisen, B. (1996). Computerized probability assessments of family relations. *Sci. Justice* **37**, 269–275.

Egeland, T., Mostad, P. F., Mevåg, B. & Stenersen, M. (2000). Beyond traditional paternity and identification cases: selecting the most probable pedigree. *Forensic Sci. Internet* **110**, 47–59.

Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitt. Anthropol. Ges.* **68**, 9–53.

Evett, I. W. & Williams, R. L. (1996). A review of the sixteen points fingerprint standard in England and Wales. *J. Forensic Ident*. 48–72.

Evett, I. W. & Weir, B. S. (1998). *Interpreting DNA evidence*. Sinauer, Sunderland MA.

Foster, E. A., Jobling, M. A., Taylor, P. G., Donnelly, P., de Knijff, P., Miermet, R. & Tyler-Smith, C. (1998). Jefferson fathered slave's last child. *Nature* **396**, 27–28.

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I. W., Hagelberg, E. & Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.* **6**, 130–135.

Hummel, K. (1994). On the theory and practice of Essen-Möller's *W*-value and Gurtler's paternity index (PI). *Forensic Sci. Internet* **25**, 1–17.

Hummel, K. & Gerschow, J. (eds) (1981). *Biomedical evidence of paternity: Festschrift for Essen-Möller*. Springer-Verlag, Berlin.

Ivanov, P. L., Waldham, M. J., Roby, R. K., Holland, M. M., Weedn, V. M. & Parsons, T. J. (1996). Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat. Genet.* **12**, 417–420.

Jeffreys, A. J., Brookfield, J. F. Y. & Semeonoff, R. (1985). Positive identification of an immigration test-case using DNA fingerprints. *Nature* **317**, 818–819.

Lange, K. (1995). Applications of Dirichlet distributions to forensic match probabilities. *Genetica* **96**, 107–117.

National Research Council (1996). *The evaluation of forensic DNA evidence*. National Academy Press, Washington, DC.

Olaisen, B., Stenersen, M. & Mevåg, B. (1997). Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nat. Genet.* **15**, 402–405.

Roeder, K. (1994). DNA fingerprinting: a review of the controversy. *Statist. Sci.* **9**, 222–278.

Stockmarr, A. (2000). The choice of hypotheses in the evaluation of DNA evidence. In *Statistical science in the courtroom* (ed. J. L. Gastwirth), ch. 8, 143–160. Springer-Verlag, New York.

Weir, B. S. (1995). DNA statistics in the Simpson matter. *Nat. Genet.* **11**, 366–368.

Weir, B. S. (2000). The consequences of defending DNA statistics. In *Statistical science in the courtroom* (ed. J. L. Gastwirth), ch. 5, 87–98. Springer-Verlag, New York.

Thore Egeland, Center for Epidemiology, National University Hospital, N-0027 Oslo, Norway.
E-mail: thore.egeland@basalmed.uio.no