Probability assessments of family relations

T. Egeland* P. Mostad* B. Olaisen†

1 Introduction

Developments during the last few years have made it increasingly feasible to use measurements of DNA to calculate probabilities or likelihood ratios in identification cases. Roeder (1994) reviews the issue from a statistical point of view. Even in simple cases, like the one shown in Figure 2, considerable time may be required to work out the exact formulae, particularly for non–experts. In sufficiently complex cases, analytical calculations may become prohibitive. We have developed and implemented and algorithm called `pater` (Mostad & Egeland (1995)) addressing the following two related questions:

- Given a pedigree, what is the probability of making certain DNA measurements for persons in it?
- Given two alternative pedigrees, what is the odds ratio between them, given certain DNA measurements? More specifically, `pater` calculates a *likelihood ratio*

$$\frac{P(\text{data} \mid \text{one pedigree})}{P(\text{data} \mid \text{another pedigree})}$$

for each allele system involved and multiplies the odds to obtain the overall odds.

An early discussion of the closely related *paternity index* is provided in Essen-Möller (1938). Lindley (1977) and Evett (1991) discuss the likelihood ratio in a Bayesian context.

Section 2 describes the basic algorithm. A simple extension to allow for mutations is discussed in Section 2.1. Section 3 includes examples demonstrating the performance of the algorithm.

The discussion concluding the paper addresses mainly improvements of the mutation model and kinship, cf. Balding & Nichols (1995).

*Norwegian Computing Center, P.O.Box 114, Blindern, N–0314 OSLO, Norway.
†Institute of Forensic Medicine, Rikshospitalet, 0027 Oslo, Norway. Paper available from: http://www.nr.no/home/SAND/thore/

2 The 'pater' algorithm

Consider a pedigree consisting of persons $X_1, \ldots, X_n$, and focus on a single system. Let $t_1, \ldots, t_k$ be a list of the alleles in this system appearing as observations from the pedigree, and let t_0 denote all alleles different from these. Specifying all alleles for all persons in the pedigree, we obtain what we will refer to as a constellation. Disregarding for a moment family relations, we realize that there is a total of $(k+1)^{2n}$ constellations: Each of the $2n$ alleles $a_{11}, a_{12}, \ldots, a_{n1}, a_{n2}$ (a_{i1} denotes the paternal allele of person i while a_{i2} is the maternal) may be of either of $k+1$ types $t_0, t_1, \ldots, t_k$. Assuming knowledge of allele frequencies of t_i for $i = 1, \ldots, k$, the frequency of the rest allele becomes $P(t_0) = 1 - P(t_1) - \ldots - P(t_k)$. Including family relations but disregarding mutations (the modification required to take care of mutations is attended to in Section 3.2), we may compute the probability $P(a_{11}, a_{12}, a_{21}, a_{22}, \ldots, a_{n1}, a_{n2})$ of a constellation by writing

$$\begin{aligned} & P(a_{11}, a_{12}, a_{21}, a_{22}, \ldots, a_{n1}, a_{n2}) \\ = \; & P(a_{11}) \cdot P(a_{12}) \\ & \cdot P(a_{21} \mid a_{11}, a_{12}) \cdot P(a_{22} \mid a_{11}, a_{12}) \\ & \vdots \\ & \cdot P(a_{n1} \mid a_{11}, a_{12}, \ldots, a_{n-1,1}, a_{n-1,2}) \cdot P(a_{n2} \mid a_{11}, a_{12}, \ldots, a_{n-1,1}, a_{n-1,2}). \end{aligned} \tag{1}$$

If we order $X_1, \ldots, X_n$ chronologically implying that parents of X_i have indices smaller than i, the conditional probabilities are easy to compute. The conditional probability of an a_{ij} whose relevant parent is not among $X_1, \ldots, X_n$ coincides with the general allele population frequency $P(a_{ij})$. If, however, X_k is the appropriate parent (then we know that $k < i$))

$$P(a_{ij} \mid a_{11}, a_{12}, \ldots, a_{i-1,1}, a_{i-1,2}) = \begin{cases} 1 & \text{if } a_{ij} = a_{k1} \text{ and } a_{ij} = a_{k2}, \\ \frac{1}{2} & \text{if } a_{ij} = a_{k1} \text{ and } a_{ij} \neq a_{k2}, \\ \frac{1}{2} & \text{if } a_{ij} \neq a_{k1} \text{ and } a_{ij} = a_{k2}, \\ 0 & \text{if } a_{ij} \neq a_{k1} \text{ and } a_{ij} \neq a_{k2}. \end{cases}$$

We may now compute the probability of the data given the pedigree by summing all constellations compatible with data:

$$P(\text{data} \mid \text{pedigree}) = \sum_{\substack{\text{constellations} \\ \text{compatible with data}}} P(a_{11}, a_{12}, a_{21}, a_{22}, \ldots, a_{n1}, a_{n2}).$$

The outlined approach may demand large computer resources and so more efficient algorithms are called for. In particular, we would like to detect and discard zero probability constellations as early as possible. Intuitively, it is also reasonable to attend to persons assigned observations as early as possible. Moreover, as will be exemplified in in Section 3.1, it is advantageous to let parents precede children. We may do so by rearranging the factors appearing in (1). Define functions

$$p(a_{11}, a_{12}, \ldots, a_{i-1,1}, a_{i-1,2}, a_{i1}, a_{i2}) = f \cdot b_1 \cdot b_2 \cdot \ldots \cdot b_s, \tag{2}$$

where

$$f = \begin{cases} 1 & \text{if } X_i \text{ has both parents among } X_1, \ldots, X_n, \\ P(a_{i1}) & \text{if only mother is among } X_1, \ldots, X_n, \\ P(a_{i2}) & \text{if only father is among } X_1, \ldots, X_n, \\ P(a_{i1})P(a_{i2}) & \text{otherwise.} \end{cases} \tag{3}$$

There is one factor b_j for each parent/child pair appearing in the pedigree, where either the parent or the child is X_i. Furthermore,

$$b_j = \begin{cases} 1 & \text{if both parent alleles are identical} \\ & \text{to the allele the child has inherited from this parent,} \\ \frac{1}{2} & \text{if one parent allele coincides with the child's,} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Observe that b_j may be interpreted as the conditional probability of the paternal (maternal) allele given the father's (mother's) alleles. This interpretation prevails in Section 3.2 where mutations are considered. The definitions of f and b_j in (3) and (4) are illustrated in Figure 1.

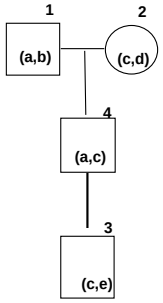

Figure 1: The figure illustrates Equations (2), (3) and (4). Specifically, $p((a,c),(a,b),(c,d),(c,e)) = fb_1b_2b_3 = 1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$. Assuming, on the other hand, no data is available for person 2: $p((a,c),(a,b),(c,e)) = P(c) \cdot \frac{1}{2} \cdot \frac{1}{2}$.

The equation below contains exactly the same factors as the product in (1), and so

$$\begin{aligned} & P(a_{11}, a_{12}, a_{21}, a_{22}, \ldots, a_{n1}, a_{n2}) \\ = \; & p(a_{11}, a_{12}) \\ & p(a_{11}, a_{12}, a_{21}, a_{22}) \\ & \vdots \\ & p(a_{11}, a_{12}, \ldots, a_{n1}, a_{n2}). \end{aligned} \tag{5}$$

This may be realized in several ways. A *formal* proof by induction follows (perhaps overdoing it a bit):

- Equation (5) holds for $n = 1$.
- Assume (5) to be true for $n = k$. The validity for $n = k + 1$ requires that

  $P(a_{k+1,1}, a_{k+1,2} \mid a_{11}, a_{12}, \ldots, a_{k,1}, a_{k,2}) = p(a_{11}, a_{12}, \ldots, a_{k+1,1}, a_{k+1,2})$.

  This concludes the proof since the persons are ordered chronologically implying the last equation.

Sorting $X_1, \ldots, X_n$ to achieve that persons with measured alleles are attended to first, we get

$$P(\text{data} \mid \text{pedigree}) \quad = \quad \sum_{\substack{\text{constellations} \\ \text{complying with data}}} P(a_{11}, a_{12}, \ldots, a_{n1}, a_{n2})$$

$$= \sum_{\substack{\text{constellations} \\ \text{complying with data}}} p(a_{11}, a_{12}) \cdot \ldots \cdot p(\ldots, a_{n1}, a_{n2})$$

$$= \sum_{\substack{(a_{11}, a_{12}) \\ \text{complying with data}}} p(a_{11}, a_{12})$$

$$\cdot \left[\sum_{\substack{(a_{21}, a_{22}) \\ \text{complying with data}}} p(a_{11}, a_{12}, a_{21}, a_{22}) \right.$$

$$\left. \cdot \left[\sum_{\substack{(a_{31}, a_{32}) \\ \text{complying with data}}} p(a_{11}, a_{12}, a_{21}, a_{22}, a_{31}, a_{32}) \cdot [\ldots] \right] \right] .$$

Rephrased as above, the algorithm is well suited for *recursive* implementation. At each level, one computes allele combinations (a_{i1}, a_{i2}) complying with data, and such that $p(a_{11}, a_{12}, \ldots a_{i1}, a_{i2}) > 0$. Pairs (a_{j1}, a_{j2}) with $j > i$ are only checked for such cases.

2.1 Modifications to account for mutations

The **pater** algorithm may be modified to include quite general mutation models by changing the definition of b_j in Equation (4). Currently, only a simple mutation model has been implemented. A mutation probability M is specified for each system as well as the total number n of alleles in the system. An allele mutates to any of the other $n-1$ equally probably. The required modification amounts to replacing (4) by $b_j = \frac{1}{2}g(a_1, b) + \frac{1}{2}g(a_2, b)$ where a_1 and a_2 are the alleles of the parent and b is the relevant allele of the child, k the number of specified alleles, and

$$g(a,b) = \begin{cases} 1-M & \text{if } a \neq t_0, b = a, \\ \frac{M}{n-1} & \text{if } a \neq t_0, b \neq a, \\ 1 - M\frac{k}{n-1} & \text{if } a = t_0, b = a, \\ \frac{M}{n-1}(n-k) & \text{if } a = t_0, b \neq a. \end{cases}$$

A simple example is provided in Section 3.2. As demonstrated in Example 3.3, the speed of the algorithm is seriously affected by introducing mutations since all constellation will then have non–zero probabilities.

3 Examples

The first two examples, Section 3.1 and 3.2 of this section, are simple in the sense that they can be worked out fairly quickly by a trained person. They are

included basically to check the performance of the algorithm. Certain transformations may be performed on pedigrees leaving the odds unchanged thus allowing checking `pater` without knowing the exact answer. One such example appears in Section 3.3.

3.1 The case of the missing father

A corpse is found, and one wants to determine whether this is in fact the missing father of two brothers. The pedigree is shown in Figure 2 and complete data is provided in Table 1. The odds in favor of F being the father of E and S becomes

$$\frac{(P(a1)+P(b1))\,(1+P(a1)+P(b1))}{4P(a1)P(b1)(1+P(a1)+P(b1)+2P(a1)P(b1))} = 2.9275, \quad (HLADQA1)$$

$$\frac{1+P(a2)+P(b2)}{4P(a2)P(b2)[(1+P(a2))]} = 4.905, \quad (HUMFES)$$

$$\frac{1+P(c3)}{4P(a3)P(b3)[1+2P(c3)]} = 371.85. \quad (HUMACTBP2)$$

The results agree with those obtained from `pater`. The efficiency of the algorithm in the example may be summarized as follows if F is the father and *HUMACTBP2* is the system. In this case $n = 4$ persons define the pedigree and $k = 3$. Consequently, there is a total of $(k+1)^{2n} = 4^8 = 65536$ *HUMACTBP2* constellations. **pater** orders F prior to E and S which in turn precede persons without observations; in this case the mother. The number of constellations considered by *pater* equals $2 \cdot 1 \cdot 1 \cdot 4 \cdot 4 = 32$. Several of the 16 terms corresponding to the mother vanish, but they are inspected and should thus be counted. Note that 64 combinations would be required if the sons were ordered prior to the father.

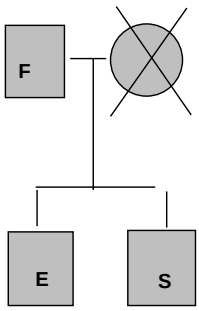

Figure 2: The pedigree shows that data are available for two brothers and their possible father.

3.2 Mutations

Assume there is an alleged father with alleles A and B in some system, and a son with alleles C and C, see Figure 3.

Obviously, with a mutation rate of 0, the odds that the alleged father is the real father is 0. With a positive mutation rate M however, complicates the

	HLADQA1			*F*	*E*	*S*
	a1	b1		(a1,b1)	(a1,b1)	(a1,b1)
frequency	0.125	0.237				
	HUMFES					
	a2	b2		(a2,a2)	(a2,b2)	(a2,b2)
frequency	0.298	0.197				
	HUMACTBP2					
	a3	b3	c3	(a3,c3)	(b3,c3)	(a3,b3)
frequency	0.063	0.01	0.072			

Table 1: The case of the missing father. Data for systems *HLADQA1*, *HUMFES* and *HUMACTBP2* for persons *F*, *E*, *S*, see Figure 2.

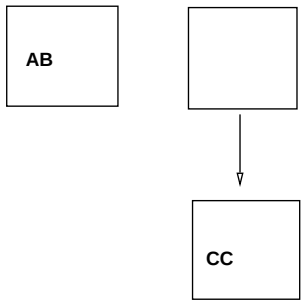

Figure 3: Illustration of model for mutation.

computation. If the alleged father *is* the real father, then we get a probability of observing the given data of

$$P(A)P(B)P(C)M\frac{1}{n-1},$$

where n is the total number of possible alleles in the system. If he is *not*, then `pater` computes the probability of the data by considering whether the allele inherited by the child from the real father was C or non-C. The probability of the data equals

$$P(A)P(B)P(C)\left[P(C)(1-M)+(1-P(C))M\frac{1}{n-1}\right].$$

while the odds becomes

$$\frac{M\frac{1}{n-1}}{P(C)(1-M)+(1-P(C))M\frac{1}{n-1}}=\frac{M}{P(C)(1-M)(n-1)+(1-P(C))M}.$$

Assuming $P(A)=0.1$, $P(B)=0.2$, $P(C)=0.3$, $M=0.02$ and $n=50$, leads to an odds of 0.00138696 as confirmed by `pater`.

3.3 Example based on similar pedigrees

Example 3.1 Consider the pedigree shown in Figure 4. V.1 is (a,b); the corpse or body, assumed to be II.2, is (a,c). Based on results in Mostad (1995) (Cock-

erham (1971) is also relevant) the body found is equally likely to be II.2 as I.1 provided mutations are disregarded. This is confirmed by the `pater` algorithm.

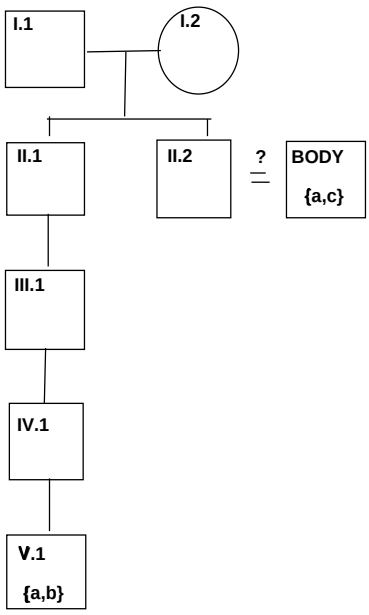

Figure 4: A body is found, suspected to be II.2. Only data from V.1 is available.

Performance. In the above case there is a total of $4^{14} = 268435456$ constellations provided the body is II.2. An odds calculation like the above takes 6 CPU–seconds [1] on a Sparc-solaris station. If mutations are included all constellations contribute and the execution time increases to 322 seconds. As a rough estimate, `pater` only considers 1.9% of the total number of constellations. If data is available also for persons IV.1, III.1, IV.1, say (a,d), (a,e) and (a,f), respectively, the total number of constellations equals $7^{14} = 0.7 \cdot 10^{12}$. However, `pater` calculates the odds instantaneously in this case since more constellations may be avoided.

Example 3.2 This example expands on the previous. Data is available for IV.2, IV.3 and V.1 as shown in Figure 5. The odds in favor of the body being the remaining persons of the pedigree are provided in Figure 5 assuming a to have frequency 0.01 and c 0.05. The CPU time is less than 135 seconds. In this case brute force considering all constellations seems prohibitive. After 12 hours, `pater` had not worked through all constellations allowing for mutations and we discontinued the run. If the body is, say II.2, the total number of constellations is $9.5 \cdot 10^{13}$.

4 Discussion

We would like to draw attention to

[1] CPU time may be a confusing concept. In the present case, system time is negligible and CPU time coincides with user time. Throughout times reported in seconds refer to CPU time.

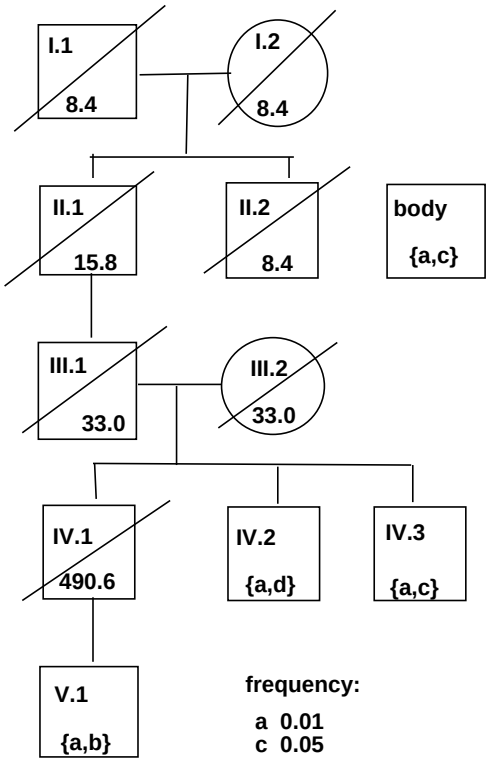

Figure 5: Data is available for IV.2, IV.3 and V.1 as shown. The odds in favor of the body being various persons is indicated assuming a to have frequency 0.01 and c 0.05. In this case brute force considering all constellations seems prohibitive. `pater` delivers an odds in less than 135 seconds on a Sparc–solaris.

- **Models for mutation**. The one we have presented may be criticized essentially because allele frequencies depend on the specification of the pedigree. Specifically the answer may differ slightly if irrelevant persons are added to the pedigree. This inconsistency is believed to be of no practical significance. It may however be alleviated by modifying the probability of mutating from A to B to $MP(B)$.

- **Kinship**. Consider n persons with no family relations. `pater` bases its calculations on

$$P(a_{11}, a_{12}, a_{21}, a_{22}, \ldots, a_{n1}, a_{n2}) = p_{11}p_{12}\cdots p_{n1}p_{n2} \tag{6}$$

  Following Balding & Nichols (1995) the right hand side of should be replaced by

$$E\left(p_{11}p_{12}\cdots p_{n1}, p_{n2}\right) \tag{7}$$

  where the vector of allele frequencies

$$(p_{11}, p_{12}, p_{21}, p_{22}, \ldots, p_{n1}p_{n2})$$

  is Dirichlet distributed. Closed formulae are then available to calculate 7. It is our intention to extend the **pater** algorithm to include kinship. However, based on our *Norwegian* experience, kinship is not likely to be important in the majority of identification cases.

- **Continuous models**, i.e., using allele measurements on an continuous scale as discussed in Devlin, Risch & Roeder (1992), require methods and algorithms beyond from what we have presented.

- **Other applications**. The *pater* algorithm may be applied to determine the most likely family relations. In practical cases, there should be no need for sophisticated optimization; *tour de force* methods, i.e., running through a large number of pedigrees is expected to be sufficient.

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*.

Cockerham, C. C. (1971), 'Higher order probability functions of identity of alleles by descent', *Genetics* **69**, 235–246.

Devlin, B., Risch, N. & Roeder, K. (1992), 'Forensic Inference From DNA Fingerprints', *JASA* pp. 337–349.

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

Evett, I. (1991), Interpretation: a personal odyssey, *in* C. G. G. Aitken & D. A. Stoney, eds, 'The Use of Statistics in Forensic Science', Ellis Horwood Ltd., Chichester, pp. 9–22.

Lindley, D. V. (1977), 'Statistics and the Law', *The Statistician* **26**(3), 203–220.

Mostad, P. (1995), 'Symmetries in pedigrees'. Manuscript.

Mostad, P. & Egeland, T. (1995), Probability assessments of family relations using the program 'pater', NR-note SAND/6/95, Norwegian Computing Center, P.O.Box 114 Blindern, N-0314 Oslo, Norway.

Roeder, K. (1994), 'DNA Fingerprinting: A review of the Controversy', *Statistical Science* pp. 222–278.